BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA
FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

Metodología para la identificación de emociones en un ambiente educativo con aprendizaje computacional

Una disertación presentada en cumplimiento de los requisitos para obtener el grado de
Doctora en Ingeniería del Lenguaje y del Conocimiento

Presenta
Yesenia Nohemí González Meneses

Directores
Dra. Josefina Guerrero García
Dr. Carlos Alberto Reyes García
Comité Evaluador
Dr. Iván Olmos Pineda
Dr. Ramón Zatarain Cabada
Dr. Juan Manuel González Calleros

Agosto 2021.

Resumen

En este trabajo de investigación se planteó la identificación de emociones en ambientes educativos, utilizando algoritmos de aprendizaje computacional y tecnologías de adquisición de señales fisiológicas y de comportamiento. Se consideraron dos emociones centradas en el aprendizaje: interés y aburrimiento. Con la finalidad de acotar el alcance de los objetivos, se seleccionaron las que se presentan más comúnmente en el proceso de enseñanza aprendizaje y que representan estados emocionales opuestos. Se probaron cuatro propuestas para la identificación de emociones utilizando algoritmos tradicionales de aprendizaje automático. Para la captura de datos se utilizaron dos tecnologías para la adquisición de señales fisiológicas (sensor de ritmo cardíaco y cámara térmica) y una de comportamiento (cámara de video); con ello se creó una base de datos con el propósito de identificar las emociones de una manera más precisa. El desarrollo de la base de datos es una tarea fundamental para el reconocimiento automático de emociones, por lo que se puso énfasis en el diseño de un protocolo formal para la captura de los datos; este, guio la ejecución de experimentos en un ambiente de aprendizaje real, con la participación de alumnos de nivel superior en un ambiente de aprendizaje en línea. Para la clasificación de las emociones, se entrenaron diferentes algoritmos de aprendizaje computacional y, se seleccionaron los que alcanzaron mejores resultados para la etapa de pruebas. Los algoritmos elegidos fueron k-vecinos más cercanos, ensamble de árboles y redes neuronales artificiales. Los resultados de la clasificación de las emociones interesado y aburrido, con redes neuronales, son satisfactorios y permiten una buena comparación con trabajos similares del estado del arte.

Abstract

The objective of this research work is to identify emotions in educational environments using machine learning algorithms and physiological and behavioral signal acquisition technologies. Two learning-centered emotions are considered: interest and bored. In order to limit the scope of the objectives, those that are most commonly presented in the teaching-learning process and that represent opposite emotional states were selected. Four proposals for the identification of emotions were tested using traditional machine learning algorithms. Moreover, two technologies for the acquisition of physiological signals (heart rate sensor and thermal camera) and one for behavioral signals (video camera) are used, with the aim of creating a data base that helps us identify emotions with better accuracy. The development of an appropriate database for the automatic recognition of emotions is a fundamental task, thus, the formal design of a protocol for data capture is put into practice. This guided the execution of experiments in a wild learning environment with the participation of college students in an online learning environment. For the classification of emotions, different machine learning algorithms were trained and those that achieved the best results for the testing stage were selected. The chosen algorithms were K- nearest neighbors, tree assembly and artificial neural networks. The results of the classification of the interested and bored emotions with neural networks are satisfactory and allow a good comparison with similar works identified in the state of the art.

Contenido

Capítulo 1. Introducción

La interacción entre los seres humanos y la computadora será cada vez más natural, si, estas últimas, son capaces de percibir y responder a la comunicación no verbal humana como las emociones. En el área de interacción humano-computadora, esto es conocido como clasificación de emociones, detección o reconocimiento automático de emociones de acuerdo a lo mencionado en Zatarain-Cabada et al. (2017a) y Bosch et al. (2015).

En este sentido, dos enfoques han sido usados para reconocer las emociones humanas: el objetivo y el subjetivo (Fuentes et al., 2016). En el enfoque objetivo, se usan sensores o captura de imágenes; como en el reconocimiento de la expresión facial, la voz, el ritmo cardíaco, el lenguaje corporal, la actividad térmica del cuerpo, la actividad muscular y las ondas cerebrales, entre otros. En el subjetivo, se proponen técnicas tales como el análisis contextual a través de observación directa, encuestas o interrogatorios a los propios individuos.

El reconocimiento automático de emociones, en la interacción humano computadora, se puede plantear a partir del uso de sensores fisiológicos, que permiten la adquisición de datos en forma de señales fisiológicas; y a través del análisis de señales de voz, de imágenes de video del rostro, ojos, cabeza o movimientos corporales de las personas (Fuentes et al., 2016). Dentro de las interfaces humano-computadora, para capturar señales fisiológicas que pueden ayudar al reconocimiento de emociones Zatarain et al.(2014) menciona las diademas de ondas cerebrales, que miden la actividad cerebral y mandan la información a un dispositivo electrónico en forma de electroencefalograma (EEG). También están las pulseras cardiovasculares, que miden el ritmo cardíaco y proporcionan información en forma de electrocardiograma (ECG). Además, se encuentran los sensores de actividad electrodermal que miden el nivel de conductividad de la piel a través del sudor en las manos, mientras que las cámaras térmicas permiten medir el cambio de temperatura del cuerpo humano asociada a los diferentes estados emocionales. También existen dispositivos para medir la respuesta muscular o la actividad eléctrica derivada de una estimulación nerviosa de un músculo, en forma de electromiografía (EMG). Por lo que respecta a dispositivos relacionados a la identificación del comportamiento de las personas -como posturas del cuerpo y gesticulaciones- están las cámaras de video tradicionales, cámaras web o cámaras de realidad aumentada que permiten grabar expresiones faciales y movimientos corporales, así como la mirada de ojos, importante para el reconocimiento de emociones. En este tipo de dispositivos también

están las grabadoras de voz, otro medio utilizado para la identificación de emociones (Cowie et al., 2001).

Los datos obtenidos de diferentes dispositivos deben ser procesados y clasificados de acuerdo con los objetivos buscados. En este trabajo serán utilizados para el reconocimiento de emociones, centradas en el aprendizaje, capturados en ambientes educativos reales, cuando los estudiantes estén ejecutando alguna actividad de aprendizaje. Dichas actividades pueden ser realizadas por medio de algún dispositivo electrónico a través del uso de un sistema tutorial inteligente o MOOC (*Massive Open Online Courses*) o de algún video juego educativo, o usando de forma tradicional una computadora, ya sea viendo algún video, estudiando, investigando o leyendo.

Desde el enfoque computacional, la detección y reconocimiento de emociones es un problema relevante dentro del área de investigación de cómputo afectivo, pues es el punto de partida en el estudio y desarrollo de sistemas de interacción humano–computadora, sensibles a las emociones. La complejidad de los problemas con los que trata el cómputo afectivo radica en el hecho de ser un área interdisciplinaria que abarca las ciencias computacionales, la psicología y las ciencias cognitivas (Picard, 1997).

Así que, para desarrollar modelos de reconocimiento de emociones, que generen resultados satisfactorios con un grado de precisión aceptable, será necesario el estudio profundo de estas tres áreas. Particularmente, en el área computacional, el reto es seleccionar y probar algoritmos de aprendizaje computacional que puedan ser integrados en un modelo completo de reconocimiento de emociones centradas en el aprendizaje (ECA).

1.1 Planteamiento del Problema

El problema del reconocimiento automático de emociones ha sido un área de investigación altamente activa en los últimos años. Aun así, se está lejos de una solución clara y que a su vez esté al alcance de la mayoría de las personas. Diversos inconvenientes han influido en la construcción de una solución apropiada desde el punto de vista computacional.

Por un lado, un factor que afecta el desempeño de los reconocedores de emociones, en contextos reales, es la dificultad de generar bases de datos con emociones espontáneas. Generalmente realizan trabajos con bases de datos formadas a partir de actuaciones de personas, las cuales proporcionan *retratos de emociones* representando expresiones prototípicas e intensas

que facilitan la búsqueda de correlaciones y la subsecuente clasificación automática. Este tipo de bases de datos suelen grabarse en un ambiente controlado, lo cual disminuye problemas en el procesamiento de la información (por ejemplo, ruido). Además, se puede garantizar una cantidad balanceada de muestras por clase. Como consecuencia, no se han tenido buenos resultados al trasladar el conocimiento extraído de estas bases de datos a contextos reales (Steidl, 2009).

En contraparte, los repositorios con datos de emociones espontáneas muestran información con contenido emocional no perteneciente a una sola clase, sino que son una mezcla de emociones. En otros casos, existen muestras con una carga emocional muy ligera, cercana a un estado emocional neutro. Además, las bases de datos con emociones espontáneas suelen grabarse en ambientes ruidosos como salones de clase, cuartos de estudio, áreas de entretenimiento, oficinas, fábricas o en conversaciones telefónicas, lo que conlleva la inclusión de ruido. Finalmente, por la naturaleza misma del problema, no es posible asegurar una cantidad balanceada de ejemplos por clase.

Otro reto por resolver es la extracción y selección de un conjunto de características que permitan reconocer emociones en los datos capturados de manera espontánea. A pesar de que los avances en el área han sido importantes, es evidente que en contextos realistas aún falta bastante por hacer. Por lo tanto, es necesario proponer y explorar otros enfoques que permitan llegar a un buen desempeño del reconocimiento de emociones en aplicaciones del mundo real. Un aspecto evidente a considerar es el hecho de que el área de aplicación influye de manera importante en la exactitud del reconocimiento de emociones (Cowie et al., 2001), ya que ésta puede variar por el contexto, la edad de los individuos participantes, las horas del día en que se trabaja, la comodidad, el grado de intrusión de las herramientas utilizadas y el tipo específico de actividad que se está realizando, entre otros.

Para abordar el problema, los investigadores en el área computacional han trabajado en dos aspectos principalmente. El primero es desarrollando técnicas de procesamiento, análisis y caracterización de los datos capturados (señales e imágenes). Para esta etapa de selección de características, las técnicas más utilizadas en las investigaciones referidas en el estado del arte son: el algoritmo RELIEF-F (para análisis del rostro y cabeza), redes neuronales convolucionales, patrón local binario, redes neuronales, algoritmos genéticos, análisis de regresión múltiple, análisis de componentes principales, análisis de discriminante lineal, entre otros. Las técnicas empleadas varían

de acuerdo con el tipo de datos que se procesan y pueden ser tan específicas como los datos lo requieran por lo que es importante realizar un análisis previo de ellos, lo que dará la pauta en su elección. En segundo lugar, se ha trabajado con diferentes técnicas de reconocimiento de patrones que modelen las propiedades de la información extraída en el procesamiento de los datos. Las técnicas más utilizadas para el proceso de clasificación son: redes neuronales artificiales, máquina de vectores de soporte (SVM, *Support Vector Machine*), algoritmo del KNN (K-*Nearest Neighbors*), clusterización, clasificador Bayesiano, regresión logística, Naïve Bayes y redes de aprendizaje profundo como redes neuronales convolucionales, entre otros (Martín de Serrano et al., 2006).

En esta propuesta se abordan ambos aspectos. Se trabajó en el análisis de las imágenes para encontrar las características más relevantes en el reconocimiento de emociones. También, se definió el uso de un método de clasificación apropiado para enfrentar la complejidad de las características que describen los datos procesados y para aprovechar la información provista por cada tipo de característica buscando una mejor precisión en la clasificación de emociones.

Por lo anterior, se identificó la carencia de una metodología adecuada para el reconocimiento de emociones centradas en el aprendizaje. Así, esta propuesta plantea el reconocimiento de dos emociones principales centradas en el aprendizaje: interés y aburrimiento. La metodología integra características diversificadas, obtenidas incluso de la fusión de los datos provenientes del uso de diferentes tecnologías de adquisición de señales fisiológicas y de comportamiento para la creación de la base de datos, además de emplear un modelo de emociones que permita apegarse más a la realidad y sobre todo al proceso cognitivo del aprendizaje.

La intención de este trabajo de investigación es el avance hacia el reconocimiento de emociones centradas en el aprendizaje producidas de manera espontánea y capturadas en contextos educativos utilizando técnicas de aprendizaje computacional mencionadas anteriormente. Por lo tanto, se espera contribuir con relación a las técnicas de adquisición de señales fisiológicas desde su captura, preprocesamiento, selección de características y finalmente con la identificación de las emociones centradas en el aprendizaje.

Se probaron algoritmos de aprendizaje computacional para procesar los datos provenientes de tres tecnologías de adquisición de datos (sensor de ritmo cardíaco, cámara térmica y cámara web) y capturados en ambientes educativos, con la finalidad de mejorar la precisión del reconocimiento y obtener un modelo de identificación de emociones con un nivel de confianza

aceptable más apegado a la realidad. Después de un preprocesamiento de los datos capturados, se contó con información que nos permitió identificar los algoritmos de selección de características y de reconocimiento de patrones que representan y clasifican de mejor manera los datos. Se inició con algoritmos tradicionales de aprendizaje computacional y finalmente se probaron técnicas de aprendizaje profundo.

El objetivo de identificar las emociones durante actividades educativas como parte de sistemas educativos inteligentes fue corroborar la relación entre emociones centradas en el aprendizaje y el nivel de aprendizaje obtenido por los estudiantes. Esta relación es la base para el planteamiento de estrategias educativas que ayuden a mejorar los niveles de aprendizaje y por ende el nivel educativo (D'Mello y Graesser, 2012). De tal forma que las emociones detectadas formen parte del modelado del estudiante dentro de la implementación de tutores inteligentes o software educativo.

Con esta investigación se espera contribuir en la identificación de las emociones (interesado y aburrido) que permitan un mejor desarrollo del proceso cognitivo de aprendizaje, aportando información para entender la relación dinámica entre emoción-aprendizaje (D'Mello y Graesser, 2012). En este sentido las aportaciones pueden ser muy valiosas para cualquier ambiente educativo en el que se monitoreen las actividades de aprendizaje con el fin de mejorar la motivación de los estudiantes sobre todo en tópicos difíciles de aprender. Lo anterior puede dar pauta a propuestas de estrategias específicas para mejorar los procesos de aprendizaje. Por ejemplo, ayudar en el diseño de ambientes de aprendizaje, de tal manera que respondan adecuadamente a las emociones de los estudiantes y que así promuevan el aprendizaje significativo. La relación entre emociones y aprendizaje significativo son la pauta para el diseño de juegos serios y el centro del diseño de software educativo que idealmente sean capaces de cambiar el trabajo a juego para minimizar el aburrimiento, optimizar el compromiso, presentar retos de diversas complejidades, prevenir la tendencia a la frustración, y generar estímulos que propicien el proceso de aprendizaje (D'Mello y Graesser, 2012).

1.2 Preguntas de Investigación

¿Qué algoritmos de aprendizaje computacional se identifican para la selección de características y para la clasificación de emociones?

¿Qué etapas es necesario definir en la metodología para el reconocimiento automático de emociones centradas en el aprendizaje haciendo uso de tecnologías de adquisición de datos y de algoritmos de clasificación?

¿Qué métricas permiten validar los resultados del reconocimiento y compararlos con los de trabajos relacionados?

1.3 Hipótesis

La metodología propuesta permite identificar emociones en contextos educativos con algoritmos de aprendizaje automático supervisado con una mejor precisión que lo encontrado en la literatura.

1.4 Objetivos

De acuerdo con lo planteado, el objetivo principal es reconocer automáticamente las emociones, centradas en el aprendizaje, de interés y aburrimiento, las cuales serán capturadas en ambientes educativos, utilizando tecnologías de adquisición y procesamiento de datos fisiológicos (sensor de ritmo cardíaco y cámara de temperatura) y de comportamiento (cámara web) para identificar relaciones emoción-aprendizaje. A continuación, se listan los objetivos específicos:

1. Identificar algoritmos de aprendizaje computacional para la selección de características y para la clasificación de emociones.
2. Diseñar e implementar una metodología para el reconocimiento automático de emociones centradas en el aprendizaje a partir del uso de las tecnologías de adquisición de datos propuestas y de la evaluación de algoritmos de selección de características y de clasificación.
3. Probar y validar la exactitud de reconocimiento con métricas que permitan comparar los resultados con los de trabajos relacionados.

1.5 Organización de la Tesis

Este documento consta de 8 capítulos. En el capítulo 2, se presenta el marco teórico que da sustento a la investigación. La primera parte de este capítulo aborda conceptos sobre la teoría de las emociones y cómputo afectivo. En la segunda parte describimos conceptos relacionados con el

reconocimiento automático de emociones, algoritmos de aprendizaje computacional y técnicas para el reconocimiento de expresiones faciales.

En el capítulo 3, se hace un análisis del estado del arte relacionado al reconocimiento automático de emociones centradas en el aprendizaje. El capítulo está organizado en cuatro secciones. En la primera se presentan y analizan trabajos enfocados en el problema computacional del reconocimiento automático de emociones. La segunda sección trata exclusivamente trabajos sobre reconocimiento automático de emociones enfocados en el análisis de la relación emoción-aprendizaje. En la tercera sección se hace una revisión sobre las bases de datos utilizadas para en el reconocimiento de emociones centradas en el aprendizaje. En la cuarta sección se analizan técnicas para el reconocimiento de expresiones faciales en imágenes visibles. Al final del capítulo se muestra un análisis estadístico de los trabajos revisados.

En el capítulo 4, se plantea la metodología para abordar el problema del reconocimiento automático de emociones centradas en el aprendizaje. En éste se describen las cuatro etapas que forman parte de la metodología propuesta, así como su relación con el proceso KDD (*Knowledge Discovery in Databases*).

En el capítulo 5, se describen las etapas 1 y 2 de la metodología. Primero se mencionan las actividades de investigación, análisis y selección de elementos claves para la captura de datos de señales fisiológicas y de comportamiento. En la segunda parte se explica el proceso para la creación de la base de datos y cómo está conformada.

El capítulo 6 corresponde a la etapa 3 de la metodología, reconocimiento de emociones. Aquí identificamos los algoritmos de clasificación que se utilizan para el entrenamiento, se presentan los resultados de éste y con los mejores modelos se hace la clasificación de las ECA.

La última etapa de la metodología se describe en el capítulo 7, aquí se presentan los resultados de las pruebas de clasificación realizadas y se hace una evaluación de ellas. Las conclusiones se describen en el capítulo 8, en el que se hace la interpretación, discusión de los resultados y el planteamiento de trabajos futuros.

Capítulo 2. Marco Teórico

Como parte del estudio de la teoría relacionada al problema planteado, en este capítulo se hace una reseña de los conceptos más importantes alrededor del reconocimiento automático de emociones. Iniciamos con una introducción al cómputo afectivo, el cual incluye, como una de sus áreas, el reconocimiento de emociones. Posteriormente abordamos los conceptos relacionados a la teoría de las emociones. En la segunda parte se describen conceptos sobre aprendizaje computacional y las principales técnicas para clasificación y reconocimiento de expresiones faciales usadas para el reconocimiento de emociones.

2.1 Cómputo Afectivo y Reconocimiento de Emociones

El cómputo afectivo es el estudio y desarrollo de sistemas y dispositivos que pueden reconocer, interpretar, procesar y simular el afecto humano. Se relaciona directamente con las emociones, cómo y cuándo son producidas y lo que ellas generan. El cómputo afectivo incluye interpretar emociones, y por lo tanto puede ayudar a desarrollar y probar nuevas y viejas teorías. También incluye muchas otras cosas, como dar a la computadora la habilidad de reconocer y expresar emociones, desarrollando su habilidad para responder inteligentemente a la emoción del humano, y hacer posible regularlas y utilizarlas (Picard, 2003).

Uno de los retos del cómputo inteligente es la habilidad para automáticamente reconocer emociones, inferir un estado emocional desde la observación de expresiones y razonar sobre la situación que genera la emoción. El reconocimiento puede requerir habilidades de visión y oído para recolectar expresiones faciales, gestos, y entonación de la voz. Adicionalmente, la computadora puede usar otras entradas de datos como las lecturas de temperatura infrarroja, medidas de respuestas electrodermales, etc. Una vez que las expresiones emocionales son censadas y reconocidas, el sistema puede usar su conocimiento sobre la situación y sobre la generación de la emoción para inferir el estado emocional

subyacente el cual muy probablemente da origen a las expresiones. Dar a una computadora estas habilidades perceptivas e interpretativas puede potencialmente darle la posibilidad de reconocer emociones tal como lo hacen las personas. Para saber cuándo una computadora ha logrado esta habilidad, se debe de evaluar la precisión del reconocimiento comparándola con el reconocimiento que un humano hace después de observar las expresiones de la cara, los gestos o voz de una persona (Picard, 2003).

Como lo menciona Rosalind Picard en su libro Picard (1995), el reconocimiento de emociones es un problema de reconocimiento de patrones, que incluso las mismas personas no pueden identificar de forma totalmente correcta, por lo tanto la expresión "*reconocimiento de emociones"* debe ser interpretada como inferir un estado emocional de la observación de expresiones y comportamiento y del razonamiento de la situación que genera la emoción. A partir de esto propone un conjunto de criterios para que una computadora pueda reconocer emociones (Picard, 1995):

- Entrada. Recibe una variedad de señales de entrada, por ejemplo: cara, voz, gestos de manos, postura y modo de caminar, respiración, respuesta electrodermal, temperatura, electrocardiograma, presión de la sangre, pulso, electromiograma, etc.
- Reconocimiento de patrones. Ejecuta extracción de características y clasificación de esas señales. Por ejemplo: análisis de las características de movimiento de video para discriminar la acción de fruncir el ceño o de una sonrisa.
- Razonamiento. Predice las emociones subyacentes basado en el conocimiento sobre cómo las emociones son generadas y expresadas. Últimamente esta habilidad requiere de percepción y razonamiento sobre el contexto, situación, objetivos personales y preferencias, reglas sociales y otro conocimiento asociado con la generación y expresión de emociones.
- Aprendizaje. Mientras la computadora "trata de conocer" a alguien, ésta aprende cuáles de los factores son más importantes para el individuo, y se vuelve más rápida y mejor al reconocer sus emociones.

- *Bias*. El estado emocional de la computadora, si ésta tiene emociones, influye en el reconocimiento de emociones ambiguas.
- Salida: La computadora nombra o describe las expresiones reconocidas, y las emociones posibles a ser presentadas.

Incluidos en estos criterios están numerosos requerimientos técnicos, por ejemplo, para capturar la entrada de datos, se requiere de tecnología precisa para digitalizar señales fisiológicas, de audio o video, así como también investigar cuáles señales son más importantes. En el reconocimiento de patrones, las características informativas de las señales necesitan ser identificadas –estadística, estructura, no linealidad, etc.- junto con variables que influyen en el significado de estas características.

2.2 Modelos de Emociones

En Cowie et al. (2001) se define a la emoción como un episodio de cambios sincronizados e interrelacionados en los estados de todos o la mayoría de los cinco subsistemas del organismo en respuesta a la evaluación de un evento de estímulo externo o interno relevante para las principales preocupaciones del organismo.

Hill (2014) las define como sucesos espontáneos que se desarrollan dentro de nosotros. Ayudan a movilizar al cuerpo para salir de una situación de urgencia. Tienen un inicio, una cúspide y por lo general se atenúan en pocos segundos. Son sumamente propensas a detonar estímulos que conducen a la acción. Son mucho más intensas que los estados de ánimo.

Al examinar las opiniones expuestas por diversos expertos en la materia, surge un consenso psicofisiológico (mente/cuerpo). Existen tres cualidades universales que caracterizan las emociones (Cornelius, 1996):

- Un componente de sentimiento: sensaciones físicas, incluyendo cambios químicos en el cerebro.
- Un componente de pensamiento: apreciaciones “racionales” conscientes o intuitivas.

- Un componente de acción: reacciones expresivas (sonrisas o fruncir el ceño), al igual que conductas de afrontamiento (pelea o huida).

A veces existe opcionalmente:

- Un componente sensorial: vista, sonido, etcétera, que se inmiscuye y sirve como detonador de la respuesta emocional.

El reconocimiento de emociones en los seres humanos tiene sus inicios con el análisis gestual haciendo observaciones detalladas, ya que las expresiones faciales son uniformes y universales. Charles Darwin fue el primero en descubrir la sorprendente verdad acerca de la naturaleza innata y preprogramada de las expresiones faciales. Por desgracia, no fue sino hasta mediados de 1960 que el Doctor Paul Ekman, profesor de la Universidad de California San Francisco y su colega Wally Friesen crearon el Sistema de Codificación de Acción Facial (FACS, *Facial Action Coding System*), haciendo posible cuantificar las emociones y definir siete emociones básicas: sorpresa (neutra), temor, enojo, tristeza, repugnancia, desprecio (negativas), y felicidad (positiva) (Paul Ekman, 2003).

2.2.1 Sistema de Codificación de Acción Facial (FACS)

El FACS es un sistema basado en cambios de músculos faciales, y puede caracterizar acciones de la cara para expresar emociones humanas individuales. El FACS codifica los movimientos de músculos del rostro específicos llamados unidades de acción (*AUs*, *Action Units*), las cuales reflejan distintos cambios momentáneos en la apariencia facial (Hjortsj et al., 2019). Las unidades de acción facial (*AUs*) codifican las acciones fundamentales (hay 46 *AUs* básicas) de un músculo o de un grupo de músculos típicamente vistos cuando la expresión facial de una emoción en particular es producida. Para reconocer la emoción facial, la *AU* individual es detectada y el sistema de clasificación facial la categoriza de acuerdo con la combinación de *AUs*. La Tabla 1 muestra las principales unidades de acción facial de la lista de Ekman. Así, por ejemplo, para la emoción de sorpresa, las *AUs* presentes son: *AU1, AU2, AU5, AU25 y AU26* como se observa en la Figura 1.

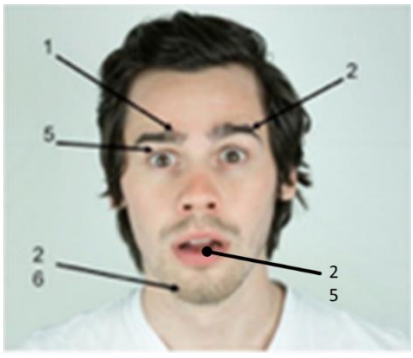
Tabla 1

Unidades de Acción Facial (AUs)

# *AU*	Nombre de la acción	# *AU*	Nombre de la acción
0	Rostro Neutral	24	Presionando los labios
1	Levantar las cejas internas	25	Separar los labios
2	Levantar las cejas externas (unilateral, lado derecho)	26	Dejar caer la mandíbula
4	Bajar las cejas	27	Estiramiento de la boca (abriendo la boca)
5	Levantar parpado superior	28	Chuparse los labios
6	Levantar mejillas	29	Sacar mandíbula
7	Apretar parpado(s)	30	Mandíbula hacia los lados
8	Labios encimados uno de otro	31	Apretar mandíbula
9	Arrugar la nariz	32	Mordida labial
10	Levantamiento del labio superior	33	Succión de mejillas (soplar)
11	Profundidad naso labial	34	Inflar mejillas
12	Tiramiento labial esquinal	35	Chupar
13	Tiramiento labial frontal	36	Protuberancia de lengua (bulto)
14	Hoyuelos	37	Repasar los labios con la lengua
15	Bajar las esquinas de los labios	38	Dilatar fosas nasales
16	Bajar el labio inferior	39	Comprimir fosas nasales
17	Levantar la barbilla o mentón	41	Dejar caer el parpado superior (sin cerrar totalmente)
18	Arrugar los labios	42	Entre cerrar los ojos
19	Mostrar la lengua	43	Ojos cerrados
20	Ensanchar (estirando) los labios	44	Vistazo (haciendo pequeños los ojos como bizco)
21	Apretamiento del cuello	45	Parpadear
22	Poner en embudo los labios	46	Guiño
23	Tensar los labios (hacia enfrente en pico)		

Figura 1

Unidades de acción facial presentes en la emoción de sorpresa: AU1, AU2, AU5, AU26 y AU25 (Hill, 2014)



A partir del modelo, de emociones básicas, es posible ampliar el conjunto de emociones para anticipar más situaciones y considerar emociones específicas para un contexto. Cada emoción secundaria es una combinación de dos emociones primarias, por ejemplo, indignación es una combinación de enojo y sorpresa. Cuando tomamos en cuenta todas las combinaciones de la lista de Ekman, se crean en total 30 emociones. Así las emociones secundarias se pueden tomar como referencia universal mensurable, contra las cuales se pueden cuantificar metas específicas en forma científica (Hill, 2014). Dentro de este grupo de emociones secundarias se encuentra la clasificación de las emociones centradas en el aprendizaje como interés, aburrimiento, confusión, frustración y sorpresa.

2.2.2 ¿Cómo Medir las Emociones?

Los psicólogos han usado dos métodos para obtener reportes personales de la experiencia emocional: (1) el enfoque de emociones discreto, y (2) el enfoque dimensional. El primero describe las expresiones como estados claramente separados. Su historia científica inicia junto con el estudio del comportamiento humano para analizar la experiencia emocional. Darwin ha hecho de este enfoque la base para las ciencias biológicas y sociales mostrando la evolución continua de las emociones básicas identificando indicios fisiológicos y de expresiones que las acompañan (Darwin, 1890). El enfoque de emociones discreto se basa en la categorización que es reflejada en la organización de los campos semánticos en los lenguajes naturales. La justificación de aceptar la estructura provista por el lenguaje es el hecho de que las categorías basadas en el lenguaje parecen corresponder a patrones de respuesta únicos, por ejemplo, los patrones específicos categoría-emoción de las expresiones faciales y de la boca, así como de respuestas fisiológicas. El método de evaluación que los investigadores adoptan con el enfoque de emociones discreto es el uso de escalas nominales, ordinales o intervalos de características, en una o varias listas de emociones que varían de acuerdo con la categorización o escalas que manejen. Aunque existen algunos instrumentos

estandarizados de este tipo, muchos investigadores prefieren crear categorías de emociones que son relevantes para un contexto de investigación específico (Scherer, 2005).

En el segundo método, el enfoque dimensional, se sugiere que los sentimientos subjetivos pueden ser descritos por su posición en un espacio tridimensional formado por la dimensión de valencia (positiva-negativa), incitación (calma-excitación), y tensión (tenso-relajado). Estas tres dimensiones se usan para describir el fenómeno mental de los sentimientos y varían con estados medibles del cuerpo tales como la incitación fisiológica. Dada la dificultad de identificar consistentemente una tercera dimensión -como incitación o tensión- muchos teóricos se limitan a la dimensión de valencia e incitación. Algunos otros sugieren estructuras circulares más adaptadas para mapear los sentimientos emocionales en el espacio de dos emociones (Scherer, 2005).

Por lo anterior, en este proyecto se utilizó el modelo de emociones secundarias, seleccionando, específicamente, las que presentan los alumnos con mayor frecuencia en actividades de aprendizaje. Se tomaron las emociones centradas en el aprendizaje sugeridas por Graesser y D'Mello (2012), que son las más utilizadas en los trabajos relacionados: confusión, frustración, aburrimiento, compromiso/interés, excitación y sorpresa. Para su clasificación se inicia con el enfoque discreto considerando las emociones de aburrido e interesado.

2.2.3 Tecnologías de Adquisición de Datos Fisiológicos y de Comportamiento

Las emociones pueden ser monitoreadas a través de varias técnicas (Lopatovska y Arapakis, 2011):

- **Usando sensores para medir las señales neuro-fisiológicas,** las modalidades pueden ser: actividad cerebral, rango del pulso, presión de la sangre, conductividad de la piel, actividad muscular, actividad térmica, etc. La ventaja de estos dispositivos es que detectan cambios a corto plazo que no es posible medirlos por otros medios; los cambios neuro-fisiológicos no pueden ser falsificados. Pero por otro lado sus desventajas son que no hay

suficiente confianza en ellos. Los sensores son invasivos cuando son colocados sobre alguna parte del cuerpo humano ya que pueden reducir la movilidad de las personas, causando distracción de las reacciones emocionales. También son propensos al ruido debido a cambios imprevistos en las características fisiológicas, además de ser incapaces de mapear datos a emociones específicas. Asimismo, requieren de experiencia y uso de equipo especial frecuentemente costoso.

- **Observando gestos, expresiones faciales y voz,** con técnicas para medir las emociones que no obstruyan el comportamiento de las personas. No se puede realizar una interpretación dependiente del contexto de los datos censados; son altamente dependientes de las condiciones ambientales (iluminación, ruido, etc.); algunas respuestas pueden ser falsificadas; reconocen la presencia de expresiones emocionales, no necesariamente emociones.
- **Preguntando a los usuarios sobre sus propias emociones,** a través de la modalidad de un diario, entrevistas o cuestionarios. Tiene las ventajas de alta correlación a la evidencia neuro-fisiológica; no obstruye, sencilla y simple –no requiere del uso de un equipo especial. Como desventajas es que se debe confiar en la suposición de que las personas son conscientes y están dispuestas a informar sus emociones; sujeto al prejuicio del encuestado; los resultados de diferentes estudios pueden no ser directamente comparables.

Para el procesamiento computacional lo más referido es combinar el uso de sensores para medir las señales neuro-fisiológicas, el análisis de imágenes de gestos y expresiones faciales y de señales de la voz. Algunos trabajos complementan sus datos con la aplicación de cuestionarios. Nosotros planteamos utilizar cámara térmica y pulsera de ritmo cardíaco, complementando las señales fisiológicas con análisis facial por medio de imágenes.

2.3 Aprendizaje Computacional

Desde el comienzo de las computadoras se cuestionó si eran capaces de aprender. El darles la capacidad de aprendizaje a las máquinas abre una amplia gama de nuevas aplicaciones. El aprendizaje humano en general es muy diverso e incluye entre otras cosas:

- Adquisición de conocimiento
- Desarrollo de habilidades a través de instrucción y práctica
- Organización de conocimiento
- Descubrimiento de hechos

De la misma forma el aprendizaje computacional (ML, *Machine Learning*) se encarga de estudiar y modelar los procesos de aprendizaje en sus diversas manifestaciones. Así, el aprendizaje computacional se refiere a las técnicas empleadas para crear programas que aprendan a realizar una tarea de manera eficiente como lo define Tom Mitchell (2009). Establece que un programa de computadora aprende de la experiencia *E*, con respecto a alguna tarea *T*, y a alguna medida de ejecución *P*, si su ejecución sobre *T* medida por *P* mejora con la experiencia *E*. Dos tipos principales de aprendizaje computacional son identificados:

1. Aprendizaje supervisado, la idea es enseñarle a la computadora cómo hacer algo.
2. Aprendizaje no supervisado, la idea es dejar que la computadora aprenda por sí misma, sin indicarle cómo hacer algo.

2.3.1 Aprendizaje Supervisado

Se refiere al hecho de que damos al algoritmo un conjunto de datos en el cual se encuentran las respuestas correctas. La idea es que, en cada ejemplo del conjunto de datos, decimos cuál es la respuesta correcta que el algoritmo debe predecir sobre ese ejemplo. Los problemas de aprendizaje supervisado son categorizados en (Mitchell, 2009):

- *Regresión:* Se trata de predecir resultados en una salida continua, lo que significa que se está entrenando para mapear variables de entrada a alguna función continua.

- *Clasificación:* Se trata de predecir resultados en una salida discreta, es decir mapear variables de entrada en categorías discretas.

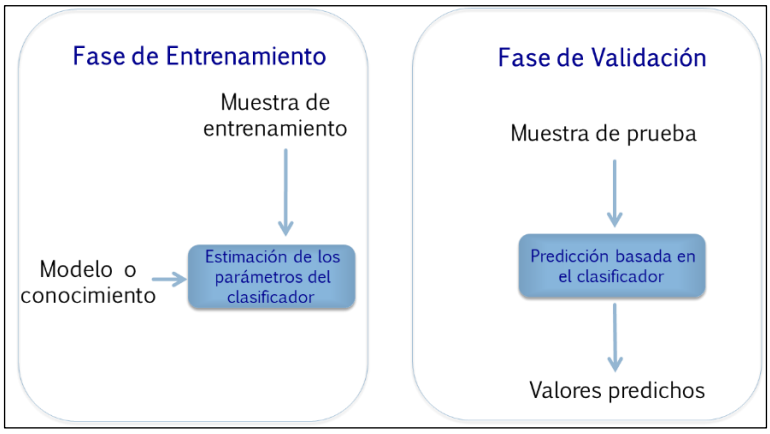
Por lo tanto, el problema de reconocimiento automático de emociones puede ser descrito como sigue: dada una muestra de objetos (caras, discursos, etc.), encontrar una función que asigne cada objeto a una de las emociones predefinidas, de modo que se minimice el error promedio de clasificación para futuras observaciones.

Existen dos etapas básicas en el diseño de un clasificador: la fase de entrenamiento y la fase de validación. En la primera fase se emplean los datos capturados, llamados muestra de entrenamiento. Pueden imponerse restricciones sobre el clasificador, generalmente relativas a hipótesis sobre la distribución de las observaciones. Una vez construido el modelo, dichas hipótesis han de ser evaluadas. Para construir el clasificador se emplea la muestra de entrenamiento, considerando las restricciones si las hubiera. Una vez que se dispone de una regla de clasificación que asigna los datos a las emociones, se pasa a la fase de validación. En ésta, el clasificador obtenido en la fase de entrenamiento es empleado para clasificar las observaciones pertenecientes a la muestra de prueba (Martín de Serrano et al., 2006).

El proceso de aprendizaje aparece resumido en la Figura 2. Tanto en la fase de entrenamiento como en la fase de validación, el clasificador asigna cada observación a una emoción. La suma del número de objetos de la muestra de entrenamiento que no son correctamente clasificados, es decir, aquellas en las que la emoción observada no coincide con la emoción predicha por el clasificador, recibe el nombre de error de entrenamiento. Así mismo, la suma de las observaciones de la muestra de prueba que no son correctamente clasificadas recibe el nombre de error de validación. La habilidad de un clasificador para clasificar correctamente observaciones que no pertenecen a la muestra de entrenamiento recibe el nombre de capacidad de generalización (Martín de Serrano et al., 2006).

Figura 2

Proceso de aprendizaje (Martín de Serrano et al., 2006)



Técnicas de clasificación supervisada más usadas en el reconocimiento automático de emociones:

1. *Redes Neuronales.* Este método fue desarrollado simultáneamente en los ámbitos del análisis estadístico y de la inteligencia artificial. La idea central consiste en extraer combinaciones lineales de los atributos presentes en los objetos obteniendo una serie de características y modelizar las clases como funciones no lineales de dichas características. Yacoub, et al. (2003), Karpouzis y Votsis (1999) y González-Hernández, et al. (2017) presentan algunos ejemplos de la aplicación de esta técnica de clasificación en el problema de reconocimiento de emociones. En el primero de estos artículos, las redes neuronales son adaptadas para identificar y agrupar los músculos de la cara que contribuyen a detectar las emociones. En el segundo, las emociones tratan de ser reconocidas a partir de señales obtenidas del discurso. En el tercer artículo implementan una red neuronal convolucional para el reconocimiento de emociones centradas en el aprendizaje. Las redes neuronales obtuvieron mejores resultados de clasificación que las técnicas alternativas con las que son comparadas: máquinas de vectores de soporte, árboles de decisión y k-vecinos más cercanos.

2. *Máquinas de Vectores de Soporte* (SVM, *Support Vector Machine*). Son procedimientos de clasificación y regresión basados en la teoría estadística del aprendizaje. Se define a una SVM como una clase específica de algoritmos preparados para el entrenamiento eficaz de una máquina de aprendizaje lineal en un espacio inducido por una función núcleo (*o kernel*), de acuerdo con unas reglas de generalización empleando técnicas de optimización. Las dos ideas fundamentales para la construcción de un clasificador SVM son la transformación del espacio de entrada en un espacio de alta dimensión y la localización en dicho espacio de un hiperplano separador óptimo. La transformación inicial se realiza mediante la elección de una función *kernel* adecuada. La ventaja de trabajar en un espacio de alta dimensión radica en que las clases consideradas serán linealmente separables con alta probabilidad, y, por tanto, encontrar un hiperplano separador óptimo será poco costoso desde el punto de vista computacional. Además, dicho hiperplano vendrá determinado por unas pocas observaciones, denominadas, vectores soporte por ser las únicas de las que depende la forma del hiperplano. Una de las principales dificultades en la aplicación de este método radica en la elección adecuada de la función *kernel*. Es decir, construir la función de transformación del espacio original a un espacio de alta dimensión es un punto crucial para el buen funcionamiento del clasificador (Martín de Serrano et al., 2006).
3. *Árboles de Decisión.* El propósito de este método de clasificación es crear una estructura de árbol que represente de forma eficiente las características más relevantes de los objetos considerados. El algoritmo de clasificación se representa como un árbol, donde las ramas son las condiciones establecidas sobre las características de los objetos y las hojas son las emociones consideradas. El proceso de clasificación comienza en la raíz y las ramificaciones del árbol se deciden a partir de las características del objeto más significativas. Dado un objeto, se desplaza a lo largo de las ramas del árbol hasta que finalmente se determina como emoción detectada aquella que define la última hoja. Existen varios algoritmos para construir un árbol de decisión (por ejemplo: el algoritmo

ID3, el C4.5, el CART, o el ID4). Sin embargo, la idea fundamental en todos ellos es la misma: los ejemplos en cada rama han de ser homogéneos, mientras que el tamaño del árbol ha de ser pequeño (es decir, la descripción ha de ser lo más simple posible). El algoritmo general de árbol de clasificación comienza considerando todas las posibles divisiones del conjunto inicial en subconjuntos. Se estima la calidad de cada una de las divisiones y se elige la mejor de todas ellas (la de menor entropía). El proceso se repite para cada una de las particiones realizadas, hasta que la entropía en cada uno de las hojas creadas es mejor que un valor predeterminado (Martín de Serrano et al., 2006).

4. *Redes Bayesianas.* Son clasificadores empleados para representar distribuciones conjuntas de modo que permitan calcular la probabilidad a posteriori de un conjunto de clases (emociones) dado un conjunto de características observadas en los objetos, y así clasificar los objetos en la clase más probable. Una red bayesiana se compone de un grafo dirigido en el cual cada nodo está asociado con una característica y con una distribución de probabilidad condicional. El grafo representa la estructura y las distribuciones de probabilidad los parámetros de la red. La idea general consiste en usar una estrategia que pueda buscar de modo eficiente en el espacio de posibles estructuras y extraer aquella que dé mejores resultados de clasificación (Ko et al., 2009).
5. *Algoritmos de votación Bagging, Boosting***.** El método llamado *Bagging (‘bootstrap aggregating’*) fue propuesto por Leo Breiman para clasificación y árboles de regresión (Speybroeck, 2012). Suponiendo que se dispone de un modelo ajustado a la muestra de objetos, tal que se obtiene un clasificador asociado a cada objeto. Se repite la estimación del clasificador, modificando en cada caso la muestra de entrenamiento. Cada una de estas muestras recibe el nombre de muestra *bootstrap*. El método *Bagging* consiste en obtener una media de las predicciones sobre un conjunto de muestras *bootstrap*. El *Adaboost* algoritmo del tipo ‘*boosting*’ donde el clasificador final se obtiene a partir de una ponderación de clasificadores ‘débiles’, esto es, con poca capacidad de generalización.

6. *Modelos Ocultos de Márkov* (HMM, *Hidden Markov Model*). Asumen que el sistema estudiado sigue un proceso de Márkov con parámetros desconocidos. La tarea fundamental consiste en determinar los parámetros ocultos a partir de los parámetros observados. La diferencia fundamental respecto a un modelo de Márkov habitual consiste en que los estados no son directamente visibles para el observador, pero sí lo son las variables influenciadas por el estado. Cada estado tiene una distribución de probabilidad asociada sobre el conjunto de posibles valores de salida. La secuencia de valores de salida generados a partir de un HMM nos dará cierta información sobre la secuencia de estados. Los tres problemas fundamentales a resolver en el diseño de un modelo HMM son: la evaluación de la probabilidad (o verosimilitud) de una secuencia de observaciones dado un modelo HMM específico; la determinación de la mejor secuencia de estados del modelo; y el ajuste de los parámetros del modelo que mejor se ajusten a los valores observados (Martín de Serrano et al., 2006).

2.3.2 Aprendizaje No Supervisado

En estos algoritmos se trabaja con un conjunto de datos no etiquetados. Dado este conjunto de datos el algoritmo no supervisado debe decidir qué datos pertenecen a determinados clústeres.

En la clasificación no supervisada diremos que hemos obtenido una buena clasificación cuando los grupos creados sean homogéneos respecto a los individuos que los forman y heterogéneos entre sí. Por ejemplo, Cao, et al. (2003) ha empleado uno de estos métodos no supervisados, el análisis de componentes independientes (ICA, *Independent Component Analysis*), para la detección de emociones en el discurso. Dada la naturaleza del problema de reconocimiento de emociones, este tipo de técnicas son muy útiles cuando las bases de datos disponibles se recogen en entornos no controlados y los individuos reflejen emociones no indicadas a priori por el investigador.

2.4 Reconocimiento de Expresiones Faciales

Las expresiones faciales son de los canales de comunicación no verbal más importante para expresar las emociones internas y las intenciones. Como se describió anteriormente Ekman y Rosenberg (2012) definen seis expresiones básicas (angustia, disgusto, miedo, felicidad, tristeza y sorpresa) que son universales entre los seres humanos. El reconocimiento de expresiones faciales (FER, *Facial Expression Recognition*) automatizado ha sido un tema de estudio por décadas. Aunque hay muchos avances en el desarrollo de sistemas de FER, muchos de ellos muestran comportamiento inadecuado en aplicaciones prácticas o carencia de generalización debido a las condiciones controladas en las cuales fueron desarrollados. El reconocimiento de expresiones es un proceso muy retador dividido en tres fases: inicio, ápice y culminación. El ápice describe la expresión en su máxima intensidad y la culminación describe la expresión desvaneciéndose. La mayoría de las veces la entrada del evento de la expresión facial desde el inicio hasta su culminación es muy rápido, lo cual hace el proceso de reconocimiento muy complejo.

Muchos métodos han sido propuestos para el FER. Los enfoques tradicionales principalmente consideran imágenes independientes e ignoran las relaciones temporales de los *frames* consecutivos en una secuencia lo cual es esencial para reconocer cambios sutiles en la apariencia de imágenes faciales especialmente en la transición de emociones entre *frames*. Ya sea considerando imágenes individuales o una secuencia de *frames*, en los enfoques tradicionales el primer paso consiste en la extracción de características bajo diferentes enfoques.

2.4.1 Modelos de Apariencia Activa

Los modelos de apariencia activa (AAM, *Active Appearance Models*) se han usado para describir imágenes generando un modelo de coordenadas ajustadas de acuerdo a la forma de la imagen (Lucey et al., 2010). Ajustar un AAM a una imagen consiste en minimizar el error entre la imagen de entrada y la instancia más cercana del modelo resolviendo un

problema de optimización no lineal. El enfoque común es usar un algoritmo iterativo para encontrar los parámetros por medio de incrementos con el objetivo de hacer coincidir el modelo generado con el modelo de la imagen de entrada, ver Figura 3.

Figura 3

Ajuste del modelo inicial de la cara a la imagen de entrada (Vargas, 2017)

a) Inicialización b) Ajuste

El modelo facial en un espacio 2D es triangulado a partir de los n puntos de referencia ($s = [x_1, y_1;\ x_2, y_2; \ldots .;\ x_n, y_n]$) ajustados a la imagen de entrada donde x y y son coordenadas en una imagen. Suponemos que dado una imagen de entrada $I(z)$ se quiere ajustar por un modelo AAM donde se conoce la forma óptima p y los parámetros de apariencia δ para el ajuste. Esto significa que la imagen $I(z)$ y la instancia del modelo $A(z)$ deben ser similares. El proceso de ajuste consiste en minimizar el error entre las coordenadas de los *frames* $I(z)$ y $A(z)$ (Lucey et al., 2010).

2.4.2 Transformación Afín (Affine Transformation)

Las transformaciones afines son ampliamente utilizadas para corregir las distorsiones geométricas en imágenes o en sus puntos de referencia pues preservan la forma. Es un método de mapeo lineal del plano euclidiano. Pueden ser usadas en un espacio $n - dimensional$, las transformaciones más usadas son: rotación, traslación y escala (Berger, 2006).

2.4.3 Análisis Procrustes (PA, Procrustes Analysis)

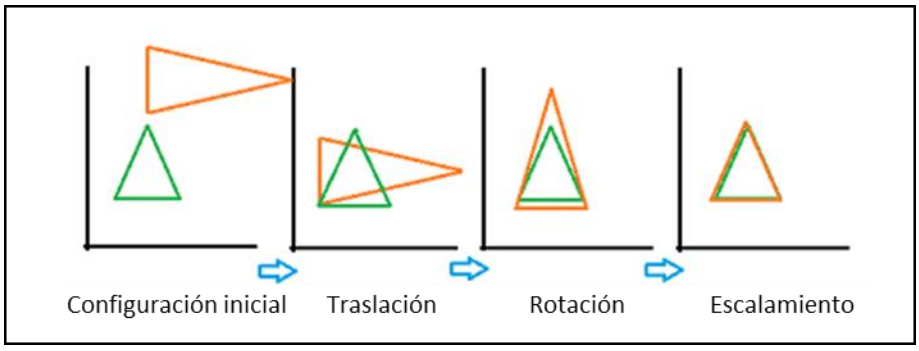
También es conocido como super imposición procrustes, ha sido usado para análisis de formas en diferentes aplicaciones. Superpone formas por medio de una óptima traslación, rotación y escalamiento uniforme de objetos, mitiga las distorsiones geométricas en imágenes o puntos de referencia usando transformaciones afines para minimizar la distancia entre dos formas usando una función de error. Se divide en tres pasos (Gower, 1975):

1) Traslación, pone el centroide de todas las formas analizadas en un punto convergente.
2) Cambio de escala de todas las formas dándoles un tamaño de centroide de 1.
3) Una rotación es ejecutada iterativamente de tal manera que la distancia entre todas las formas es minimizada.

En la Figura 4 los tres pasos de super imposición son mostrados.

Figura 4

Representación gráfica del proceso de super imposición Procrustes (Gower, 1975)



Por otro lado, recientemente, con la ayuda de las redes neuronales profundas (DNNs, *Deep Neural Networks*), los resultados reportados en el campo del reconocimiento de expresiones faciales han sido prometedores. Mientras en los enfoques tradicionales las características son usadas para entrenar clasificadores, las DNNs tienen la habilidad de

extraer más características discriminantes que pueden proporcionar una mejor interpretación de la textura de la cara humana en datos visuales (Li y Deng, 2018).

Capítulo 3. Estado del Arte

El análisis del estado del arte se ha ido acotando hacia el objetivo de la investigación. Se inicia con consultas generales sobre el reconocimiento automático de emociones como las mostradas en la

Figura 5, donde se observa el número de publicaciones por año de 2010 a junio de 2018. De este total de publicaciones aproximadamente el 60% de ellas corresponden a investigaciones en el campo de las ciencias computacionales y el 40% corresponden a investigaciones de otras áreas como educación, psicología, neurología, neuropsicología, psiquiatría, entre otras. Así que se considera el 60% para iniciar el análisis estado del arte.

En la primera etapa de revisión se analizaron trabajos sobre interacción humano computadora (HCI, *Human Computer Interaction*) con la idea de identificar las diferentes tecnologías de adquisición de datos fisiológicos y de comportamiento en seres humanos, así como para conocer los avances en la teoría del cómputo afectivo. Posteriormente, debido a que identificamos que gran parte de los avances en el reconocimiento automático de emociones ha sido utilizando cámaras de video y diademas para obtener lecturas de las ondas cerebrales, se hizo una revisión de estos trabajos. La mayor parte de estas investigaciones (aproximadamente el 80%) se enfocan en el reconocimiento de las emociones básicas: alegría, tristeza, enojo, miedo, asco, sorpresa y neutro; y sólo algunos pocos (aproximadamente el 20%) intentan reconocer emociones centradas en el aprendizaje: interés, aburrimiento, frustración, confusión, excitación y sorpresa; esto se representa gráficamente en la Figura 6.

Figura 5

Gráfica de la consulta sobre reconocimiento de emociones (fuente: http://wcs.webofknowledge.com/RA/analyze.do Consultada: 01 de diciembre de 2019)

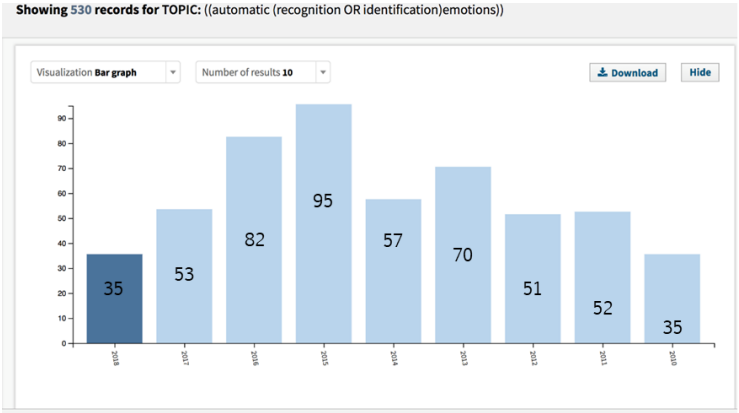
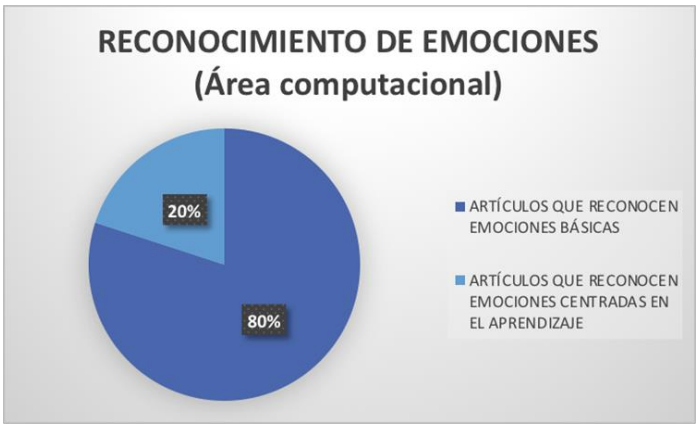


Figura 6

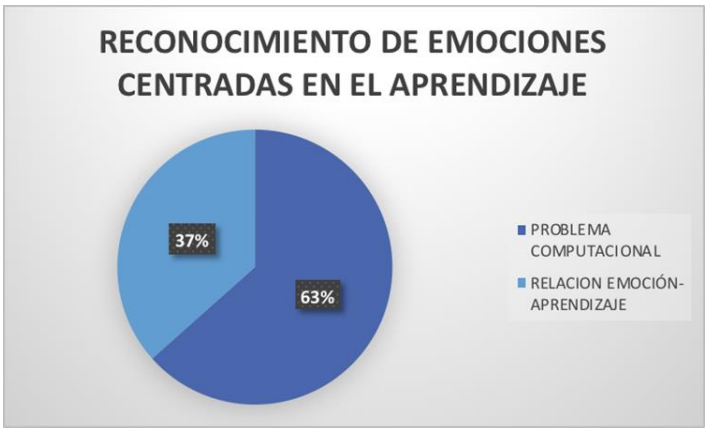
Gráfica de investigaciones del área computacional



Así finalmente, el análisis comparativo del estado del arte que se realizó consiste en una recopilación de artículos enfocados al reconocimiento de emociones en actividades de aprendizaje y a la identificación de la relación emoción-aprendizaje. Hay una clara diferencia en la literatura encontrada. Existen trabajos que, aunque realizan una identificación de emociones le dan mayor relevancia al análisis de la relación emoción-aprendizaje (aproximadamente el 37%, ver **¡Error! No se encuentra el origen de la referencia.**), mientras que otros hacen énfasis en los algoritmos para la identificación automática de emociones y la precisión de reconocimiento alcanzada dando menor importancia al análisis de la relación emoción-aprendizaje (aproximadamente el 63%, ver **¡Error! No se encuentra el origen de la referencia.**). En el capítulo se incluyen trabajos específicamente sobre tratamiento de emociones centradas en el aprendizaje correspondientes a las publicaciones más recientes. Los primeros artículos revisados se enfocan en el tratamiento computacional del problema. En el apartado dos se analizan trabajos enfocados en la relación emoción-aprendizaje. Posteriormente se revisan bases de datos de señales fisiológicas y de comportamiento para reconocimiento de emociones y por último trabajos sobre el reconocimiento de expresiones faciales en video.

Figura 7

Gráfica del reconocimiento de emociones (con dos enfoques diferentes)



3.1. Trabajos Enfocados en el Problema Computacional del Reconocimiento Automático de Emociones

El trabajo más reciente encontrado sobre reconocimiento de ECA es el de Nezami, et al. (2020). En este entrenan una red neuronal convolucional VGG-B con la base de datos de emociones básicas FER formada 35,887 muestras de imágenes de video. Este modelo es usado como una parte de su modelo para el reconocimiento de la emoción "*interesado*". Entrenan una CNN con una base de datos creada por ellos a partir de la grabación de video de 20 alumnos de secundaria mientras interactúan con el software OMOSA para aprendizaje de habilidades de investigación. La base de datos es etiquetada por seis anotadores estudiantes de psicología. La CNN para el reconocimiento de la emoción *"interesado"* es inicializada con los pesos obtenidos de la CNN VGG-B para emociones básicas. Con esta propuesta logran una exactitud para la emoción de interesado=72.38%. También entrenan una SVM con la que alcanzan una exactitud para la emoción de interesado=59.88%.

El siguiente trabajo analizado es el Gupta, et al. (2018), en el que utilizan cinco diferentes propuestas de aprendizaje profundo para probar su base de datos DAiSEE creada por ellos a partir de experimentos ejecutados haciendo uso de ambientes de aprendizaje en línea. La base de datos está formada por 9068 videos capturados de 112 usuarios para reconocer los estados afectivos de aburrido, confundido, interesado y frustrado capturados en ambientes reales y es etiquetada por cuatro estudiantes de psicología. Ejecutan pruebas con la red CNN *Inception* v3, con una CNN 3D y con una CNN *Long-Term* con *frames* y con video. Obtienen los siguientes valores promedio para *exactitud* por emoción: interesado=51.07%, aburrido=35.89%, confundido=57.45% y frustrado=73.09%

En el trabajo de González-Hernández, et al. (2017), utilizan una red neuronal convolucional para el reconocimiento de emociones centradas en el aprendizaje. Su arquitectura tiene tres capas convolucionales, cada una seguida de una capa de máxima

agrupación y finalmente tres capas de redes neuronales completamente conectadas. Ejecutan pruebas usando tres bases de datos: RaFD, base de datos de expresiones faciales posadas conteniendo imágenes de 8 emociones básicas y dos bases de datos espontáneas creadas por ellos mismos especialmente con contenido de emociones centradas en el aprendizaje. Las emociones que reconocen son: interesado, excitado (*excitement*), aburrido y relajado. Concluyen que en la literatura no hay ningún trabajo que aplique el enfoque de aprendizaje profundo para el reconocimiento de emociones en la educación. La precisión que alcanzan al utilizar la base de datos RaFD es de 95% y con las otras dos bases de datos alcanzan 88% y 74% respectivamente. Sus resultados son comparados contra algoritmos de clasificación como máquina de vectores de soporte, k-vecinos más cercanos y redes neuronales artificiales. Como trabajos futuros proponen hacer más grandes las bases de datos propias y probar con diferentes filtros y métodos de preprocesamiento.

En Mehmood y Lee (2017), proponen un método asistido por computadora para instructores de escuelas especiales, donde enseñan a alumnos con desórdenes mentales o problemas emocionales, con un sistema que trabaja tecnologías usables y reconocimiento inteligente de emociones. El módulo de reconocimiento de emociones inicia con la captura de la señal de diademas cerebrales. Los datos crudos son pasados a una etapa de procesamiento de datos donde las señales son filtradas en una frecuencia de 0.5 a 30 Hz. Estos filtros incluyen cuatro tipos de frecuencias: alfa, beta, theta y delta. Los datos filtrados son enviados al módulo de extracción de características. Generan tres tipos de características: 1) los datos de frecuencia filtrados, 2) los parámetros Hjort, con 14 señales cerebrales y 3) los parámetros Hjort con 6 señales cerebrales. Después de extraer estas características las procesan con dos clasificadores: máquina de vectores de soporte y k-vecinos cercanos con una validación cruzada de 10 iteraciones. El estado emocional identificado es enviado a un módulo de administrador de información para continuar con el módulo de tratamiento de la información de acuerdo con la condición mental detectada. Las emociones que reconocen son: felicidad, calma, tristeza y miedo. Un módulo de expresión muestra al instructor las

sugerencias de tratamiento de acuerdo con el estado emocional detectado. No presentan una evaluación de precisión de reconocimiento. Concluyen que el EEG proveniente de las diademas cerebrales es cada vez más usado para el reconocimiento de emociones por su bajo costo y fácil acceso. La contribución de su propuesta la centran en el reconocimiento de emociones y en las respuestas a las emociones de los estudiantes para mejorar sus capacidades de aprendizaje. Este trabajo identifica emociones básicas y aunque analiza el comportamiento de alumnos en un proceso de enseñanza no se enfoca a reconocer emociones centradas en el aprendizaje.

En Zatarain-Cabada, et al. (2017a)**,** implementan un patrón local binario para el reconocimiento de emociones centradas en el aprendizaje. El propósito de este trabajo fue construir una base de datos de expresiones faciales espontáneas correspondientes a estados afectivos en educación para ser usada en diferentes sistemas tutoriales inteligentes. Las tecnologías para captura de datos que utilizan son video y diademas de EEG (*Emotiv-EPOC*). Las emociones, centradas en el aprendizaje, que reconocen son: frustración, aburrimiento, compromiso y entusiasmo. Para construir la base de datos, toman fotografías de las expresiones faciales de los estudiantes y con el estado afectivo, detectado con las señales del EEG, etiquetan cada una de las imágenes. Obtienen una base de datos con 7019 fotogramas las cuales pasan a un proceso de filtrado quedándose con 730 fotogramas etiquetados. Inician el proceso de reconocimiento de emociones aplicando cinco diferentes filtros a las imágenes, para después aplicar el operador uniforme del patrón local binario (PLB), los histogramas de la imagen del PLB son usados como descriptores de características. Cada histograma es concatenado y normalizado en un vector. Un clasificador de máquina de vectores de soporte recibe el vector de característica para hacer la clasificación de emociones. Después de aplicar el reconocedor su precisión calculada fue de 80% con una desviación estándar de 2%.

Otra investigación es la de Zatarain-Cabada, et al. (2017b). En este trabajo se explica la construcción y validación de una base de datos de expresiones faciales que se recopila tomando fotografías con una cámara web cada cinco segundos mientras los alumnos

programan en código Java. Cada fotografía es etiquetada con las emociones de los usuarios obtenidas en ese momento desde el dispositivo *Emotiv-Epoc* Para el reconocimiento de expresiones faciales usan una técnica basada en geometría que mide las distancias entre el punto central de la cara y otros 68 puntos de referencia. Estas medidas (coordenadas de los puntos relevantes, distancias y ángulos relativos) son transformadas en características para entrenar una máquina de vectores de soporte. Miden la precisión del reconocimiento aplicando validación cruzada de 10 iteraciones. Obtienen una precisión por emoción con la siguiente distribución: aburrido de 64%, interesado de 64%, excitado de 83% y frustrado de 62%. El mismo modelo para el reconocimiento de emociones lo usan también en Zataraín, et al. (2017), como parte de un medio ambiente de aprendizaje afectivo basado en la Web 3.0 para aprender a programar en Java. El objetivo principal de este trabajo es proveer a los alumnos de instrucciones adaptadas e individualizadas de forma automática usando tecnologías de aprendizaje como sistemas recomendadores, sistemas de extracción de información, reconocedores de afecto, analizadores de sentimientos, y herramientas de auto tutoría. Sus evaluaciones están enfocadas en analizar el impacto de la herramienta de software sobre el comportamiento de los estudiantes y en evaluar el aprendizaje obtenido después de utilizar la herramienta por lo que está incluido en la clasificación de trabajos que le dan mayor importancia al análisis de la relación emoción-aprendizaje. La base de datos utilizada en los trabajos antes mencionados de Zatarain es implementada en Barrón-Estrada, et al. (2016). Explican ampliamente el proceso de construcción de la base de datos, la depuración de esta y su evaluación en una aplicación para el reconocimiento de emociones a partir de expresiones faciales usando la técnica de patrón local binario.

En el siguiente trabajo de Arana-Llanes, et al. (2017), proponen diferentes actividades recomendadas para inducir un determinado estado mental y la respuesta del EEG para cada uno de ellos. También presentan la definición del estado emocional ideal para aprender. Estas actividades están basadas en pruebas psicológicas que están dedicadas a medir el nivel de

atención, concentración y otras funciones. En el proceso de clasificación utilizan k-*means* y *clustering* con una tasa total de concentración de 96% de 3592 instancias.

En Botelho, et al. (2017), intentan mejorar la detección de afecto libre de sensores a través de "aprendizaje profundo" específicamente con Redes Neuronales Recurrentes (RNNs). Codificadores humanos observaban a los estudiantes mientras hacen uso de la plataforma de aprendizaje en línea ASSISTments y etiquetan el afecto de los estudiantes en intervalos de 20 segundos. Las emociones etiquetadas fueron aburrido, frustrado, confundido, concentrado e imposible de codificar. Sus datos los obtuvieron de 646 alumnos de 6 escuelas diferentes. Obtienen un conjunto de 51 características nivel-acción por cada intervalo de tiempo, de las cuales obtienen características estadísticas, para finalmente trabajar con 204 características por intervalo de tiempo. Usan las etiquetas y características en tres modelos de aprendizaje profundo: red neuronal recurrente tradicional (RNN), red neuronal de unidad recurrente cerrada (GRU) y red de memoria de término largo-corto (LSTM). Evalúan su modelo con tres estadísticas: área bajo la curva (AUC ROC/A'), Cohen's Kappa y Fleiss' Kappa con una validación cruzada de 5-folds. Los mejores resultados los obtienen para AUC= 0.78% con RNN, para Cohen's Kappa= 0.21% con LSTM y para Fleiss' Kappa= 0.27% con LSTM. En este trabajo lo que puede hacer imprecisos sus resultados es el etiquetado hecho por terceros, humanos que tienen la probabilidad de errar en el momento de juzgar el estado emotivo de los estudiantes y cuyas etiquetas son las que utilizan para validar la interacción del estudiante con el sistema.

En los artículos de Bosch, et al. (2016a) y Bosch, et al. (2016b) utilizaron visión por computadora, análisis del aprendizaje y aprendizaje computacional para detectar el afecto de los estudiantes en un entorno real en el laboratorio de computación de una escuela con al menos treinta estudiantes. Los estudiantes se movían, gesticulaban y hablaban entre ellos, lo que dificultó la tarea. A pesar de estos desafíos, pudieron detectar el aburrimiento, la confusión, el deleite, la frustración y el interés. Utilizan el algoritmo de selección de características estadísticas RELIEF-F, sobre un total de 78 características faciales y 3

características de los movimientos gruesos del cuerpo. Utilizan 14 clasificadores diferentes, incluidos clasificadores bayesianos, regresión logística, clasificación mediante *clustering* (con *k-means*), árboles C4.5, etc., utilizando implementaciones estándar de la herramienta de aprendizaje computacional WEKA (*Waikato Environment for Knowledge Analysis*). Presentan la mejor tasa de reconocimiento por clasificador: aburrimiento 64% (*k-means*), confusión 74% (Bayes net), placer 83% (Naïve Bayes), interés 64% (Bayes *net*) y frustración 62% (Bayes *net*). Finalmente hacen un análisis de las características detectadas para cada una de las emociones, análisis de generalización temporal (de día y de periodo) y de generalización demográfica (género y etnicidad). Identifican claramente que la distribución de los estados afectivos depende de la interfaz usada.

En Monkaresi, et al. (2016) usan técnicas de visión por computadora para detectar el compromiso (*engagement*) de 22 alumnos mientras realizan actividades de escritura. Para la captura de datos utilizan el sensor del Kinect de Microsoft y graban las señales del electrocardiograma utilizando el sistema BIOPAC MP150, usando tres electrodos colocados en la cintura y en el tobillo de los estudiantes. Extraen tres conjuntos de características de videos, ritmo cardíaco y de las unidades de animación del rastreador de rostros del Kinect de Microsoft. Usan técnicas psicológicas para obtener reportes de los propios alumnos que son generados de manera concurrente mientras realizan la actividad de escritura y posteriormente al ver los videos. Tres métodos de extracción de características son utilizados sobre los segmentos de los videos grabados. Del rastreador de rostros del Kinect de Microsoft, obtienen 6 características de la cara, 4 representan el movimiento de los labios y 2 el movimiento de las cejas. Del mismo rastreador de rostros del Kinect detectan los movimientos de la cabeza: cabeceo, guiño y balanceo. Obtienen 84 características, correspondientes a 7 medidas estadísticas, sobre 12 niveles de *frames*. Usan el patrón local binario en tres planos ortogonales para describir la apariencia y dinámica de los objetos faciales sobre las imágenes de video. Detectan ojos y boca y sobre cada secuencia de video extraen 2304 características. Finalmente, siete características estadísticas fueron extraídas de

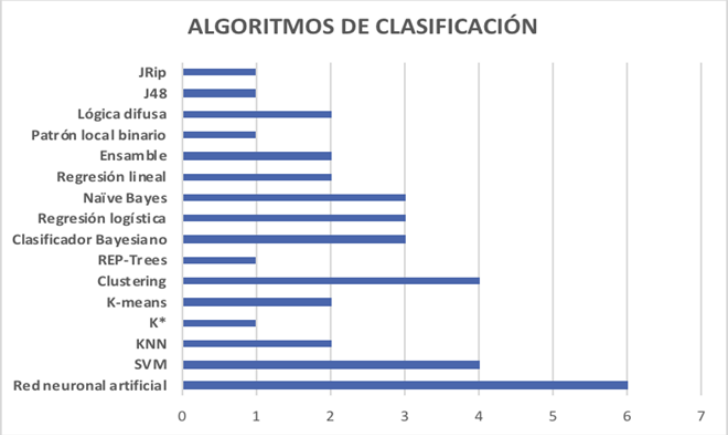
la estimación del ritmo cardíaco para cada segmento de video. Sobre estas características aplicaron técnicas de selección de características y algoritmos de aprendizaje computacional como Naïve Bayes, red Bayesiana, regresión logística, clusterización, bosque de rotación y ensamble de algoritmos de clasificación base diez. Obtienen con validación cruzada un área bajo la curva ROC de 0.758, para anotaciones concurrentes y 0.733 para anotaciones de retrospectiva. Este trabajo está enfocado en la detección de una única emoción, por lo que la clasificación se limita a dos clases (interesado y no interesado) y se identifica un rango de error en el cálculo del ritmo cardíaco a partir de video, que debe de ser tratado utilizando otras técnicas para su cálculo y adquisición.

En Arroyo, et al. (2009), el objetivo de su investigación es realizar una revisión sistemática de la relación estado afectivo y resultado deseado. También pretende evaluar la importancia y efectividad del seguimiento de los estados emocionales dentro del salón de clases. Usan dispositivos para captar datos fisiológicos (cámara, mouse, silla y pulsera). Generan su propia base de datos y solo mencionan que usan algoritmos de clasificación para el reconocimiento de emociones.

Como conclusión se observa que la mayoría de estos trabajos tratan de reconocer las principales emociones centradas en el aprendizaje: frustración, interés, aburrimiento, confusión, excitación y sorpresa. Los algoritmos más utilizados para este propósito son: máquina de vectores de soporte (SVM), regresión lineal y ensambles de algoritmos, como se muestra en la **¡Error! No se encuentra el origen de la referencia.**. El porcentaje de exactitud de reconocimiento varía desde el 62% al 88%, dependiendo de las métricas utilizadas para evaluar sus resultados.

Figura 8

Gráfica de algoritmos de clasificación más utilizados



La gran mayoría de trabajos generan sus propias bases de datos con un número de alumnos que van desde 20 hasta 646 (en una investigación libre de sensores), algunos otros utilizan bases de datos actuadas como RAFD, JAFFEE o Grimace, lo cual se observa gráficamente en la Figura 9. Las tecnologías de adquisición de datos que predominan son las cámaras web y las diademas *Emotiv*, algunos otros trabajos hacen uso del Kinect de Windows, aplicación de detección del ritmo cardíaco, sensores de silla y pulseras cardiovasculares, la popularidad de estas tecnologías es mostrada en la

Figura 10. Estos trabajos permiten plantear estudios similares utilizando ambientes de aprendizaje bajo configuraciones diferentes priorizando la naturalidad y comodidad de los estudiantes. Después de este análisis, es posible identificar áreas de oportunidad con el objetivo de lograr mejores precisiones de reconocimiento utilizando otros dispositivos de captura de datos fisiológicos como las cámaras térmicas, la captura del ritmo cardíaco con sensor de dedo en combinación con grabaciones de video.

En esta tesis se hizo uso de las imágenes de video, extrayendo características faciales, para ser procesadas por los algoritmos de clasificación más populares. Posteriormente se hicieron pruebas con algoritmos de aprendizaje profundo.

Figura 9

Gráfica de las bases de datos más utilizadas en las investigaciones para el reconocimiento de emociones

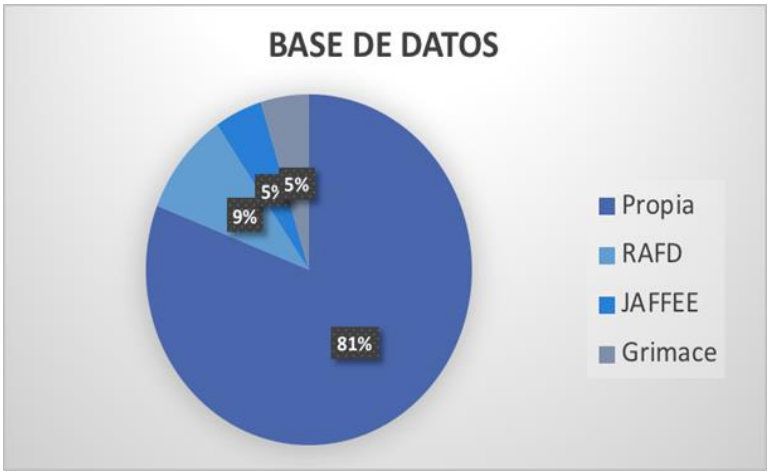
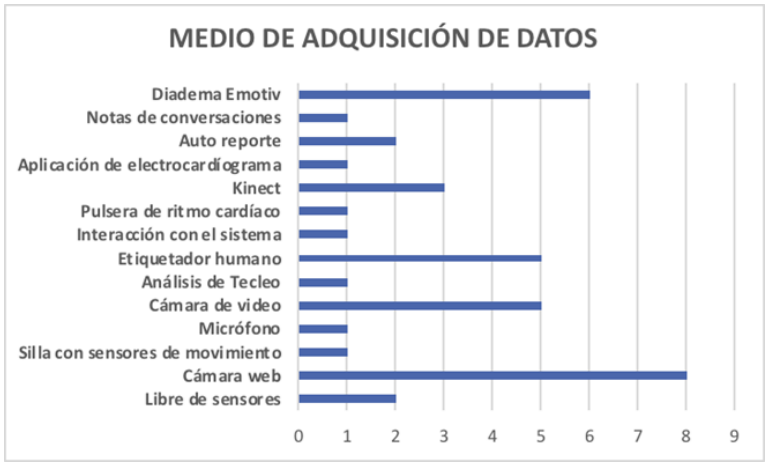


Figura 10

Gráfica de la popularidad de las tecnologías más utilizadas para el reconocimiento de emociones



3.2 Trabajos Sobre Reconocimiento Automático de Emociones Enfocados en el Análisis de la Relación Emoción-Aprendizaje

En Nye, et al. (2017) analizan cómo los sistemas tutores basados en escenarios influyen en el estado afectivo de los alumnos. Observan la relación de la emoción hacia otros componentes de la experiencia como corrección de respuestas, uso de pistas o de ayuda. Recolectan datos de 39 alumnos de una Universidad privada de California, a través de una cámara web, mientras interactúan con un sistema de ambiente inmersivo. Crean un corpus de grabaciones de video de 30 a 60 minutos por alumno. Usan un software de análisis de video comercial llamado C-CERT (*Computer Expression Recognition Toolbox*) desarrollado por el laboratorio de emociones de la Universidad de California, el cual ejecuta reconocimiento de expresiones en tiempo real a partir de análisis de video. Identifican 20 unidades de acción (AUs) y expresiones prototípicas tales como frustración, confusión, tristeza, alegría, angustia y miedo. Con las emociones identificadas hacen un análisis de correlación experiencias en línea – emoción. Concluyen identificando que los estudiantes muestreados están en un estado de interés (equilibrio) en la experiencia y con el contenido del sistema tutor basado en escenarios.

En Xiao, et al. (2017) analizan la dinámica de los estados afectivos de estudiantes (interés, aburrimiento, confusión, frustración, etc.) durante sesiones cortas de aprendizaje en MOOC´s (*Massive Open Online Courses*) basados en video. También muestran la factibilidad de predecir los estados afectivos “momento a momento” vía fotopletismiografía grabando las señales del ritmo cardíaco con la cámara de un celular utilizando una aplicación móvil. A través del estudio de 22 participantes, cuantifican la frecuencia de la ocurrencia y la transición dinámica de estados afectivos comunes durante el aprendizaje con MOOC´s. Coleccionan 1034 auto reportes de afecto de los estudiantes participantes y seleccionan 11 características de la señal de variación de ritmo cardíaco. Usan una máquina de vector de soporte con una función de base radial para construir tres clasificadores con los que detectan: 1) si el estudiante está interesado, aburrido o confundido (clasificación binaria); 2) si el estudiante

está en un estado negativo (baja valencia); y 3) la ocurrencia de eventos críticos marcados por emociones fuertes (alta excitación). Calculan el coeficiente Kappa y la precisión con los que evalúan sus resultados y concluyen identificando una clara relación entre los estados afectivos y la señal de la variación del ritmo cardíaco.

Este trabajo permite identificar como área de oportunidad el estudio de estados afectivos de estudiantes mientras interactúan con MOOC´s, pues como lo mencionan pocas investigaciones se han realizado a la fecha. En este trabajo sólo utilizan grabaciones de video tradicional y auto reportes del estado afectivo de los estudiantes, quedando abierta la posibilidad de integrar más dispositivos de captura de señales de comportamiento y fisiológicas como las propuestas en este trabajo. También, se identifica la posibilidad de aplicar otras técnicas computacionales para los procesos de clasificación y por otro lado corroborar el acierto de elegir las emociones centradas en el aprendizaje de interés y aburrimiento.

En Sawyer, et al. (2017) proponen un marco de modelado del estudiante afecto-mejorado que se basa en el seguimiento de expresiones faciales para aprendizaje basado en juego y en el análisis del video juego. Concluyen que el modelado basado en unidades de codificación de acción facial individuales es más efectivo que los modelos de emociones compuestas. Sugieren que incluir el seguimiento de las expresiones faciales puede mejorar la precisión para los modelos del estudiante, tanto para predecir la ganancia de aprendizaje como para predecir el interés de los estudiantes.

Su experimento inicia con la grabación videos de 33 alumnos interactuando en el medio ambiente de aprendizaje basado en el juego *Crystal Island* para microbiología en periodos de tiempo de 26 a 105 minutos. Antes de iniciar las grabaciones los estudiantes son evaluados con un *test* de 20 preguntas sobre microbiología y al finalizar su sesión nuevamente contestan el mismo test de conocimientos sobre el tema. También contestan un cuestionario de 30 preguntas para caracterizar su presencia dentro del medio ambiente virtual. Con la presencia miden el nivel de interés de los estudiantes. Las características de las expresiones

faciales fueron extraídas automáticamente del sistema de seguimiento de expresiones faciales de *iMotions* basado en video, extraen las características que corresponden al sistema de codificación acción facial (FACS, *facial action coding system*) con la que obtienen la probabilidad de la presencia de un estado afectivo en particular y 20 unidades de acción para las emociones de angustia, sorpresa, frustración, felicidad, confusión, miedo, disgusto y tristeza.

En Bosch y D'Mello (2015) hacen un estudio de los estados afectivos que se originan cuando los alumnos aprenden con tecnología, en este caso con un tutorial para aprender los fundamentos de la programación en Python. Dan seguimiento a los estados afectivos de 99 alumnos participantes de la Universidad de Midwestern de Estados Unidos, mediante la grabación de sus rostros y de las acciones que realizan sobre la pantalla de la computadora usando el tutorial. La sesión de aprendizaje dura 25 minutos y después realizan un ejercicio de evaluación de 10 minutos. Una vez terminadas ambas actividades le piden al alumno identificar las emociones que creen presentaron durante el uso del tutorial pidiéndoles visualicen las grabaciones de sus rostros y de la actividad que en ese momento realizaban en la pantalla. Los alumnos identifican sus emociones aproximadamente en 100 momentos (cada 15 segundos). Las posibles emociones que pueden identificar se les presenta en una lista de la que seleccionan las que para ellos son las más representativas de su rostro y de la actividad que en ese momento realizaban con el tutorial.

Las emociones que pueden identificar son: interés, confusión, frustración, aburrimiento, curiosidad (que fueron los estados afectivos más frecuentes), ansiedad, felicidad, angustia, sorpresa, disgusto, tristeza y temor (que fueron los más raros). Con el análisis de estos resultados identifican la relación emoción-aprendizaje indicando que aburrimiento (r=-.149) y frustración (r=-.218) fueron correlacionadas negativamente mientras que confusión (r=.087) e interés (r=.093) fueron correlacionadas positivamente con el aprendizaje. Por lo que respecta al análisis de transiciones entre confusión-frustración (r=.103), frustración-confusión (r=.105) y aburrimiento-interés (r=.282) fueron

correlacionadas positivamente con el aprendizaje. Concluyen que no hay una coocurrencia de los estados afectivos que se pueda generalizar; corroboran el modelo de dinámica afectiva de D'Mello y Graesser (2012) con las transiciones entre estados afectivos que identificaron y finalmente hacen el análisis de la relación entre afecto-aprendizaje desde cinco enfoques diferentes.

En Barrón, et al. (2014) presentan el desarrollo de un tutor inteligente con reconocimiento y manejo de emociones para las matemáticas. Por su parte, Zatarain-Cabada, et al. (2017c) desarrollan un sistema de aprendizaje afectivo para la lógica algorítmica aplicando gamificación. En ambos trabajos integran un módulo para el reconocimiento de emociones. Este proceso está basado en el análisis de rostros. En el segundo trabajo agregan la captura de señales de datos de EEG para etiquetar las imágenes faciales. Dividen el reconocimiento de emociones en dos etapas: una de entrenamiento de una red neuronal y otra de uso-producción de esta. En la primera etapa preparan una red neuronal para que sea capaz de reconocer el estado emocional de los estudiantes. En la segunda utilizan la red neuronal como parte del submódulo afectivo de un sistema tutor inteligente. Primero extraen la información del rostro del estudiante, y con la información y la ayuda de la red neuronal se clasifica (identifica) el estado emocional, y se envía al tutor inteligente de la red social. La red neuronal produce su propia salida (la emoción). Una vez que se reconoce la emoción, ésta es enviada al tutor inteligente de la red social *Fermat*. El sistema integra el estado emocional junto con los valores pedagógicos obtenidos en los ejercicios resueltos por el estudiante, para así enviar una retroalimentación afectiva por medio de un agente animado.

En ambos trabajos las conclusiones están en términos de la relación entre la emoción y el aprendizaje adquirido al usar el tutorial inteligente. Así como también sobre la relación "estrategias de enseñanza–emociones" como el hecho de unificar técnicas de gamificación con tecnologías de inteligencia artificial para la detección de emociones.

En la Tabla 2 se hace el análisis comparativo de los trabajos de reconocimiento de emociones que le dan mayor importancia al análisis de la relación emoción-aprendizaje.

Tabla 2

Resumen de trabajos que le dan mayor importancia al análisis de la relación emoción-aprendizaje

Título	Tecnología de adquisición de datos	Emociones reconocidas	Base de datos	Método de extracción de características	Algoritmo de clasificación	Métricas de evaluación
Affective Learning System for Algorithmic Logic Applying Gamification. (Zatarain-Cabada, Barrón-Estrada y Ríos-Félix., 2017)	Video y señales de EEG	**Centradas en el aprendizaje**: aburrido, frustrado, interesado, neutral	RafD (955 imágenes)	Vector de 10 características	Red neuronal artificial	-
Analyzing Learner Affect in a Scenario-Based Intelligent Tutoring System. (Nye et al., 2017)	Cámara web	**Centran el aprendizaje**: frustración, confusión, tristeza, alegría, angustia y miedo	Propia (39 alumnos, video de 30 a 60 min por alumno)	-	C-CERT (Computer Expression Recognition Toolbox)	Análisis de correlación: experiencias en línea– emoción.
Dynamics of Affective States During MOOC Learning. (Xiao et al., 2017)	Cámara integrada de un celular	**Centradas en el aprendizaje**: interés, aburrimiento, frustración, confusión, sorpresa, placer, curiosidad, felicidad y neutro	Propia (22 alumnos universitarios)	Vector de 11 características obtenidas de la señal de variación del ritmo cardíaco	SVM	Precisión y coeficiente Kappa, con los que identifican una clara relación entre los estados afectivos y la variación del ritmo cardíaco
Enhancing Student Models in Game-based Learning with Facial Expression Recognition. (Sawyer et al., 2017)	Video	angustia, sorpresa, frustración, felicidad, confusión, miedo, disgusto y tristeza	Propia (33 alumnos universitarios)	Extracción de características del sistema de seguimiento facial de iMotions	Clasificadores del sistema de reconocimiento facial de iMotions, AUs y emociones compuestas.	
An Affective and Web 3.0-Based Learning Environment for a Programming Language. (Zataraín et al., 2017)	Video (cámara C920 Logitech HD Pro-Web cam) y diadema de EEG (Emotiv-EPOC)	**Centradas en el aprendizaje**: frustración, aburrimiento, interés y excitación	Base de datos (propia) de expresiones faciales etiquetadas	Técnica basada en geometría (coordenadas de los puntos relevantes, distancias y ángulos relativos)	SVM	-
The Affective Experience of Novice Computer Programmers. (Bosch y D'Mello, 2015)	Video	compromiso, aburrimiento, frustración, confusión, ansiedad, felicidad, curiosidad, angustia, sorpresa, disgusto, tristeza y miedo.	Propia (99 alumnos)	Observación gesticular	Puntuación de la proporción obtenida de los reportes del estado afectivo de cada estudiante y pruebas no paramétricas.	Análisis de correlación emoción-aprendizaje
Intelligent Tutor with Emotion Recognition and Student Emotion Management for Math Performance. (Barrón et al., 2014)	Video	alegría, enojo, sorpresa, miedo, disgusto/desprecio e interés	RafD (8040 imágenes)	Vectores de características normalizados	Red neuronal artificial	-
Towards an Understanding of Affect and Knowledge from Student Interaction with an Intelligent Tutoring System. (Bixel y D´Mello, 2013)	Video y Auto-reporte	**Centradas en el aprendizaje**: frustrado, confundido, interesado, aburrido	Base de datos propia con datos de 29 estudiantes	-	-	Tasa de ocurrencia: interés 23%, confusión 22%, frustración 14%, aburrimiento 12%
Exploring the Relationships between Design, Students' Affective States, and Disengaged Behaviors within an ITS. (Bixel y D´Mello, 2013)	Auto-reporte	**Centradas en el aprendizaje**: frustrado, confundido, interesado, aburrido	Base de datos propia con datos de 58 estudiantes	-	K*, Naïve Bayes, JRip, REPtree	Interés 31%, aburrimiento 28%, confusión 40%, frustración 23%,

Towards an Understanding of Affect and Knowledge from Student Interaction with an Intelligent Tutoring System. (Bixel y D´Mello, 2013)	Dos codificado-res humanos de las acciones de los estudiantes	**Centradas en el aprendizaje**: frustrado, interesado, aburrido	Base de datos propia con datos de 229 estudiantes	-	Sistema ASSISTment. Algoritmos K*, Jrip y J48	Interés 0.678% con K*, aburri-miento 0.63% con Jrip y frustración 0.681% con J48

Se observa que en estas publicaciones no es relevante la tasa de precisión en el reconocimiento de emociones, aunque sí mencionan los métodos para extracción de características y los algoritmos de clasificación que utilizaron. Pero su principal aportación está enfocada en el análisis de la relación emoción-aprendizaje como la correlación entre experiencias en línea y emoción, relación del estado afectivo y la variación del ritmo cardíaco y la correlación emoción-aprendizaje. A pesar de esto, algunos trabajos incluyen el reconocimiento de emociones básicas y no de emociones centradas en el aprendizaje, al parecer con la finalidad de encontrar otras aportaciones.

El análisis de estos trabajos ayudó a identificar las áreas de oportunidad de esta investigación y a definir las estrategias en la metodología. Posteriormente también fueron considerados para discriminar y elegir los algoritmos de selección de características y clasificación con la finalidad de avanzar en el reconocimiento automático de emociones centradas en el aprendizaje y contribuir en esta área.

3.3 Bases de Datos de Señales Fisiológicas y de Comportamiento para Reconocimiento de Emociones.

Una base de datos de expresiones de emociones es una colección de imágenes, videos, voz y señales fisiológicas relacionadas con un amplio rango de emociones. Su contenido corresponde a expresiones de emociones relacionadas al contexto en donde fueron capturadas y en base al cual son etiquetadas, esto es esencial para el entrenamiento, prueba y validación de algoritmos para el desarrollo de sistemas de reconocimiento de expresiones. El etiquetado de emociones puede ser hecho en escala discreta o continúa. Muchas de las bases de datos usualmente se basan en la teoría de emociones que asume que existen en una escala discreta seis emociones básicas y una variedad aproximada de 22 de

emociones secundarias. Sin embargo, en algunas bases de datos las emociones son etiquetadas en la escala continua de excitación-valencia. Otras bases de datos incluyen las AUs del FACS (P. Ekman, et al., 2002).

Las bases de datos de expresiones de emociones en su mayoría están formadas solo por expresiones faciales y se clasifican en actuadas y espontáneas. En las bases de datos de expresiones actuadas, se les pide a los participantes que muestren diferentes expresiones emocionales, mientras que en las bases de datos espontáneas las expresiones son naturales. Las expresiones espontáneas difieren de las actuadas notablemente en términos de intensidad, configuración y duración. En la mayoría de los casos, las expresiones actuadas son exageradas, mientras que las espontáneas son sutiles y difieren en apariencia. Además de esto, la síntesis de algunas AUs apenas es alcanzable sin experimentar el estado emocional asociado por lo que no es posible capturar datos fisiológicos pues estos no pueden ser controlados por las personas y por lo tanto no corresponden a la emoción actuada.

A continuación, se muestra una recopilación de bases de datos de expresiones faciales, aunque el principal interés de esta investigación recae en bases de datos híbridas que contengan imágenes faciales y datos fisiológicos; hasta el momento solo se ha encontrado, disponible públicamente, una base de datos de este tipo; la base de datos DEAP de Soleymani, et al. (2012), que contiene grabaciones fisiológicas (de EEG) y video facial de un experimento donde 32 voluntarios vieron un subconjunto de 40 videos musicales. En este experimento también se les pide a los participantes calificar cada video de acuerdo con la emoción que provocaba en ellos.

El análisis de las bases de datos mostrado en la Tabla 3 incluye solamente bases de datos de expresiones faciales disponibles al público. Estas bases de datos corresponden a expresiones faciales de emociones básicas, que se han capturado mientras las personas realizan diversas actividades. La mayoría de las bases de datos se componen de espectadores viendo contenido de medios (es decir, anuncios, avances de películas, programas de televisión, gifs animados y campañas virales en línea) y de expresiones faciales actuadas. De

ellas la base de datos más robusta por la cantidad de datos que almacena es *Affectiva* (Kaliouby y Picard, 2019).

En la Tabla 4 se muestran los detalles de bases de datos de expresiones faciales espontáneas. Estas corresponden a emociones centradas en el aprendizaje que han sido capturadas mientras los estudiantes realizan alguna actividad de aprendizaje en un entorno natural. Algunas de ellas contienen datos fisiológicos, pero desafortunadamente no están disponibles al público, ni proporcionan información sobre características de los datos, ni mucho menos detalles del procesamiento que hacen con ellos.

De estas bases de datos, cuatro incluyen datos fisiológicos; de ellas la más completa es la mencionada en Arroyo et al. (2009), en la que participan 67 estudiantes y utilizan tres sensores de datos fisiológicos y una cámara de video. En general la mayoría de estas bases de datos han sido creadas con datos de muy pocos participantes.

Tabla 3

Bases de datos de expresiones faciales de emociones básicas

Base de datos	Expresión facial	Número de muestras	Número de imágenes / videos	Tipo
DEAP (Soleymani et al., 2012)	Análisis de emociones del modelo continuo.	32	Señales fisiológicas periféricas de EEG y 22 videos de la cara	Espontánea
RAVDESS (Livingstone y Russo, 2018)	Voz y canto: Calma, Felicidad, tristeza, miedo, enojo, sorpresa, disgusto y neutral.	24	7356 archivos de video y audio	Posada
F-M FACS 3.0 (EDU, PRO & XYZ versions) (Freitas-Magalhães, 2018)	Neutral, tristeza, sorpresa, felicidad, miedo, enojo, disgusto, desprecio.	10	4877 videos y secuencias de imágenes	Espontánea / Posada
CK ++ (Lucey et al., 2010)	Neutral, tristeza, sorpresa, felicidad, miedo, enojo, y disgusto.	123	593 secuencias de imágenes (327 con etiquetas de emociones secundarias)	Posada / sonrisa espontánea
JAFFE (Lyons, Gyoba y Kamachi, 1997)	Neutral, tristeza, sorpresa, felicidad, miedo, enojo, y disgusto.	10	213 imágenes estáticas	Posada
MMI (Valstar y Pantic, 2010)	Disgusto, Felicidad y tristeza	43	1280 videos y 250 imágenes	Espontánea / Posada

BELFAST (Sneddon, Mcrorie, Mckeown y Hanratty, 2012)	Conjunto 1 (disgusto, miedo, frustración, sorpresa, diversión)	114	570 video clips	Emociones naturales
	Conjunto 2 (disgusto, miedo, frustración, sorpresa, diversión, enojo y tristeza)	82	650 video clips	
DISFA (Mavadati et al., 2013)	Intensidad de las AUs	27	4845 frames de video	Espontánea
MUG (Aifanti, Papachristou y Delopoulos, n.d.)	Neutral, tristeza, sorpresa, felicidad, miedo, enojo y disgusto.	82	1462 secuencias	Posada
ISED (Happy, Patnaik, Routray y Guha, 2017)	Tristeza y felicidad	50	428 videos	Espontánea
RaFD (Langner et al., 2010)	Neutral, tristeza, sorpresa, felicidad, miedo, enojo, disgusto y desprecio.	67	3 direcciones diferentes de la mirada y 5 ángulos de la cámara (8,040 imágenes)	Actuada
Oulu-CASIA NIR-VIS (Zhao, n.d.)	Tristeza, sorpresa, felicidad, miedo y disgusto.	80	3 condiciones de iluminación diferentes: normal, clara y obscura (2880 videos)	Actuada
FERG (Aneja, Colburn, Faigin, Shapiro Y Mones, 2016)	Neutral, tristeza, sorpresa, felicidad, miedo, enojo, disgusto y alegría.	6	55,767 imágenes	Pose frontal
Affectnet (Mollahosseini, Hasani y Mahoor, 2017)	Neutral, tristeza, sorpresa, felicidad, miedo, enojo, disgusto y desprecio.	-	~450,000 anotaciones manuales ~500,000 anotaciones automáticas	Espontánea
IMPA-FACE3D (Mena-Chalco, Jesus Marcondes y Velho, 2008)	Neutral, tristeza, sorpresa, diversión, enojado, disgustado, abierto, cerrado y beso.	38	534 imágenes estáticas	Actuada
FEI (Thomaz, 2012)	Neutral y sonriente	200	2850 imágenes estáticas	Actuada
Aff-Wild (Zafeiriou, Kollias, Nicolaou, Papaiooannou y Kotsia, 2017)	-	200	~ 1,250,000 etiquetadas manualmente	Espontánea
Affectiva (Kaliouby y Picard, 2019)	frustración, sorpresa, diversión, enojo, tristeza, miedo, disgusto y desprecio.	-	7,860,463 caras analizadas	Espontánea

Tabla 4

Bases de datos de expresiones faciales correspondientes a ECA

Base de datos	Expresión facial	Número de muestras	Número de imágenes / videos	Tipo/Software
Engagement Recognition (ER), base	Interés.	20 estudiantes de secundaria	20 videos de una hora cada uno	Espontánea/Mundo virtual OMOSA para aprendizaje

de datos propia (Nezami et al., 2020)		(11 mujeres, 9 hombres)		de habilidades de investigación
DAiSEE, base de datos propia (Gupta et al., 2018)	Interés, aburrimiento, frustración, confusión.	112 estudiantes	9068 videos	Espontánea/Aprendizaje en línea
Base de datos propia (Nye et al., 2017)	frustración, confusión, diversión, tristeza, miedo, angustia.	39 estudiantes	30 a 60 minutos de video por estudiante	Espontánea/Sistema tutor inteligente
Base de datos propia (Xiao et al., 2017)	frustración, confusión, complacencia, interés, aburrimiento, sorpresa, curiosidad, felicidad y neutral.	22 estudiantes universitarios	imágenes	Espontánea/Sistema administrador de aprendizaje
Base de datos propia (Zataraín et al., 2017) (Barrón-Estrada et al., 2016) (Zatarain-Caba-da, Barron-Estrada, et al., 2017)	interés, aburri-miento, frustración y excitación	8 estudiantes	Video y señales de EGG (etiquetas de emociones discretas)	Espontánea/Sistema tutor inteligente
Base de datos propia (Bosch y D'Mello, 2015)	interés, aburrimiento, frustración y confusión	29 estudiantes	Video y auto reportes	Espontánea/Ambiente para programar en python
Base de datos propia (M. Harley, Bouchet, y Azebedo, 2013)	Emociones básicas y ECA (19 emociones y 1 un estado neutral)	67 estudiantes	Video y auto reportes	Espontánea/Sistema tutor inteligente multi agente para enseñar el sistema circulatorio humano
Base de datos propia (Graesser y D'Mello, 2012a)	interés, aburrimiento, frustración y confusión.	28–30 estudiantes	Video de movimiento de ojos, notas de diálogos con el tutor, expresiones faciales, posturas del cuerpo, sensor de presión de mouse y teclado y voz	Espontánea/Sistema tutor inteligente para reparar computadoras
Base de datos propia (Arana-Llanes et al., 2017)	Atención y concentración	23,072 instancias	Video y señales de EGG (Emotiv-EPOC). Extracción de ondas alfa y beta.	Espontánea/Test psicológico para inducir los estados mentales de atención y concentración.
Base de datos propia (Bosch, D'Mello, et al., 2016a) (Bosch, D'mello, et al., 2016b) (Bosch et al., 2015)	Complacencia, frustración, confusión, interés y aburrimiento.	137 estudiantes	Video y método de observación BROMP	Espontánea/Juego educativo de física *Playground*
Base de datos propia (Monkaresi et al., 2016)	Interés	22 estudiantes	Video de Kinect y Sistema de adquisición de señales BIOPAC MP150 de ECG	Espontánea/ Escritura de un resumen
Base de datos propia (Almoha-mmadi et al., 2017)	Interés	30 estudiantes	Video de Kinect V2	Espontánea/ curso de excel

Base de datos propia (Bixel y D´Mello, 2013)	Interés, aburrimiento, neutral	44 estudiantes	Análisis de tecleo y video	Espontánea/ Escritura de un ensayo
Base de datos propia (Arroyo et al., 2009)	Interés, excitación, confianza y frustración	2 grupos de 38 y 29 estudian-tes de nivel medio superior	Sensores fisiológicos (cámara de video, mouse, silla y brazalete)	Espontánea/Sistema tutor multimedia adaptivo de geometría

A partir de este análisis se identificó como área de oportunidad la creación de una base de datos fisiológicos y de comportamiento respaldada por un protocolo formal para la captura de los datos. La definición del protocolo nos permitió ejecutar un experimento controlado en un ambiente natural y que puede replicarse el número de veces necesarias para crear una base de datos robusta.

3.4 Reconocimiento de Expresiones Faciales en Imágenes Visibles.

Aunque el reconocimiento de expresiones faciales puede ser conducido usando múltiples sensores, uno de los enfoques más estudiados es a través del análisis de imágenes faciales visibles debido a que la expresión visual es uno de los principales canales de información en la comunicación interpersonal. Una cámara es el sensor más prometedor porque provee información clave para el FER y no es invasivo. El FER automático es una tecnología que usa marcadores biométricos para detectar emociones en caras humanas.

Los enfoques más utilizados para el FER automático se dividen en dos grupos basados en si las características son obtenidas manualmente o son generadas a través de la entrada de datos a una red neuronal profunda. En el enfoque convencional el FER está compuesto por tres pasos principales: 1) detección de la cara, 2) extracción de características, y 3) clasificación de la expresión. En el primer paso, la cara es segmentada desde una imagen de entrada, y los puntos de referencia y los componentes faciales (como ojos, nariz y boca) son detectados. En el segundo, varias características espaciales y temporales son extraídas de los componentes faciales. En el último paso, los clasificadores de expresiones faciales pre entrenados, tales como máquina de vectores de soporte (SVM, *Support Vector Machine),*

Adaboost o árboles de decisión, producen los resultados de reconocimiento usando las características extraídas.

En contraste al enfoque tradicional que usa características extraídas manualmente, surge el aprendizaje profundo como un enfoque general para el aprendizaje automático, produciendo resultados en el estado del arte en muchos estudios de visión por computadora con la disponibilidad de grandes datos. El FER basado en aprendizaje profundo reduce en gran medida la dependencia de los modelos basados en la cara física y otras técnicas de procesamiento al permitir que el aprendizaje ocurra "de extremo a extremo" directamente desde las imágenes de entrada (Walecki, et al., 2017).

Entre los diferentes modelos de aprendizaje profundo disponibles, la red neuronal convolucional (CNN, *convolutional neural network*), un tipo particular de aprendizaje profundo es el modelo de red más popular. En los enfoques basados en la CNN, la imagen de entrada es convolucionada a través de una colección de filtros en las capas convolucionales para producir un mapa de características. Luego, cada mapa de características es combinado con redes completamente conectadas y la expresión facial es reconocida como perteneciente a una clase particular basada en la salida de un algoritmo.

El FER también se divide en dos grupos de acuerdo a si usa *frames* o imágenes de video (Walecki et al., 2017). El primero, el FER estático (basado en *frames*), se basa únicamente en características faciales estáticas obtenidas por la extracción manual de *frames* seleccionados de una secuencia de imágenes. El segundo, FER dinámico (basado en video), utiliza características espaciotemporales para capturar la expresión dinámica en secuencias de expresiones faciales. Aunque es conocido que el FER dinámico tiene una tasa de reconocimiento más alta que el estático porque provee información temporal adicional, presenta algunos inconvenientes. Por ejemplo, las características extraídas dinámicamente tienen diferentes duraciones de transición y diferentes características de la expresión facial dependiendo de las caras particulares. Además, la normalización temporal usada para

obtener secuencias de expresiones con un número fijo de *frames* puede resultar en una pérdida de información de la escala temporal.

3.4.1 Enfoques Convencionales de FER

Para los sistemas de FER automáticos, varios tipos de enfoques convencionales han sido estudiados. Lo común en estos enfoques es detectar la región de la cara y extraer características geométricas, características de apariencia, o un híbrido de características geométricas y de apariencia sobre la cara etiquetada.

Para las características geométricas, la relación entre componentes faciales es usada para construir un vector de características para entrenamiento. Por ejemplo, en Ghimire y Lee (2013) usaron dos tipos de características geométricas basadas en la posición y ángulo de 52 puntos de referencia faciales. Primero, el ángulo y distancia Euclidiana entre cada par de puntos de referencia dentro de un *frame* son calculados, y después la distancia y ángulos son restados de la correspondiente distancia y ángulos del primer *frame* de la secuencia de video. Para la clasificación, dos métodos son presentados, uno usando *AdaBoost* multiclase y otro usando una máquina de vectores de soporte (SVM) sobre los vectores de características.

Las características de apariencia son usualmente extraídas de la región global de la cara o de diferentes regiones que contienen diferentes tipos de información. Ghimire, et al. (2017) extrae características de apariencia de una región específica dividiendo la región de la cara completa en regiones locales de dominio específico. La importancia de las regiones locales es determinada usando un enfoque de búsqueda incremental, el cual resulta en una reducción de la dimensión de características y en una mejor precisión de reconocimiento.

Para características híbridas, algunos enfoques han combinado características geométricas y de apariencia para complementar las debilidades de los dos enfoques y proveer mejores resultados en ciertos casos. Por ejemplo, en Benitez-Quiroz, et al. (2016) proponen un algoritmo que reconoce AUs y sus intensidades en grandes bases de datos de expresiones faciales. Así estas imágenes son automáticamente etiquetadas con AUs, intensidades de AUs

y categorías de emociones. Para la identificación de AUs utilizan estadísticas de segundo orden de los puntos de referencia de la cara (como distancias y ángulos entre puntos de referencia) y filtros de Gabor. Para categorizar las emociones utilizan un clasificador basado en subclases, el análisis discriminante de la subclase de *kernel* (KSDA, *Kernel Subclass Discriminant Analysis*). La contribución de su enfoque es que alcanzan una precisión de reconocimiento alta corriendo en tiempo real.

Otro trabajo híbrido es el presentado por Du, et al. (2014) en el que primero detectan los puntos de referencia facial para después hacer la identificación de AUs para 21 emociones compuestas determinando el grado de presencia de cada AU. La clasificación la llevan a cabo utilizando tres algoritmos diferentes: el clasificador *k-means*, KSDA y SVM, con los que obtienen precisiones muy similares.

Uno de los trabajos más recientes es el presentado por Rivera, et al. (2019). Este trabajo presenta el desarrollo de una herramienta para Matlab, la cual permite le análisis de la emoción facial en línea. El sistema está diseñado con cuatro módulos que permiten la adquisición de los datos, la extracción de características, la clasificación de la expresión y el reporte grafico del análisis. Sus resultados muestran que las seis clases de emociones básicas fueron reconocidas por su sistema, con precisión de 63.0% y 68.8% para los clasificadores de análisis discriminante lineal (LDA) y vecinos más cercanos (KNN).

3.4.2 Enfoques de FER Basados en Aprendizaje Profundo

En Kim, et al. (2019), proponen una representación de características espacio-temporal para el FER. Su método utiliza estados de expresiones representativas las cuales pueden ser especificadas en secuencias faciales a pesar de la intensidad de la expresión. Las características de las expresiones faciales son codificadas en dos partes. En la primera parte, las características de la imagen espacial de los *frames* representativos del estado de la expresión son aprendidos mediante una CNN. En la segunda parte, las características temporales de la representación de la característica espacial de la primera parte son

aprendidas a través de una memoria larga de corto plazo (LSTM*, Long Short-Term memory*) de la expresión facial. Sus experimentos fueron conducidos sobre conjuntos de datos de expresiones deliberadas (MMI) y un conjunto de datos de micro expresiones espontáneas (CASME II). Analizan seis emociones básicas.

En Breuer y Kimmel (2017), analizan ocho emociones y detectan 50 AUs. La extracción de características la hacen con una CNN y por inferencia identifican las emociones.

En Chu, et al. (2017) proponen una arquitectura de red híbrida. Específicamente representaciones espaciales son extraídas por una CNN. Para modelar las dependencias temporales las LSTMs son apiladas independientemente de la longitud de los videos. Las salidas de la CNN y de las LSTMs son integradas en una red para predecir por *frame* 12 AUs. Llevaron a cabo varios experimentos sobre dos grandes conjuntos de bases de datos espontáneas GFT y BP4D, con más de 400,000 *frames* codificados con 12 AUs. En ambos conjuntos de datos reportaron mejoras sobre la CNN estándar de multi-etiquetas y basadas en características.

Otro trabajo es el presentado en Hasani y Mahoor (2017). Ellos proponen el método de una CNN 3D para el FER en videos. Esta nueva arquitectura de red consiste en capas *Inception-ResNet* seguidas por una unidad de LSTM que juntas extraen las relaciones espaciales en imágenes faciales, así como también las relaciones temporales entre diferentes *frames* en el video. Los puntos de referencia facial también son usados como entrada a la red lo que enfatiza la importancia de los componentes faciales más que las regiones faciales que pueden no contribuir significativamente para la generación de expresiones faciales. Su método es evaluado usando cuatro bases de datos públicas.

A partir de la revisión del estado del arte sobre FER automático determinamos utilizar el video para iniciar con la extracción y selección de características necesarias para identificación de AUs. La presencia o ausencia de ellas nos ayudó a seleccionar las AUs asociadas a las emociones centradas en el aprendizaje que intentamos reconocer.

Capítulo 4. Metodología

En este capítulo se relacionan los diferentes tipos de investigación con las actividades realizadas durante el desarrollo de la tesis de acuerdo con la metodología propuesta. La metodología de investigación es de tipo mixta, pues se utilizó la investigación exploratoria en la adquisición de datos fisiológicos y de comportamiento. Experimental, en la selección y pruebas de algoritmos de aprendizaje computacional para el reconocimiento de emociones. Correlacional y descriptiva en el proceso de clasificación de las emociones. Por otro lado, los datos obtenidos de las señales fisiológicas y la precisión en el reconocimiento fueron evaluados cuantitativamente. A continuación, se presenta la metodología propuesta y se describen las etapas 1, 2 y 3.

4.1 Metodología

La metodología general de la investigación se muestra gráficamente en la Figura 11. Las etapas principales son:

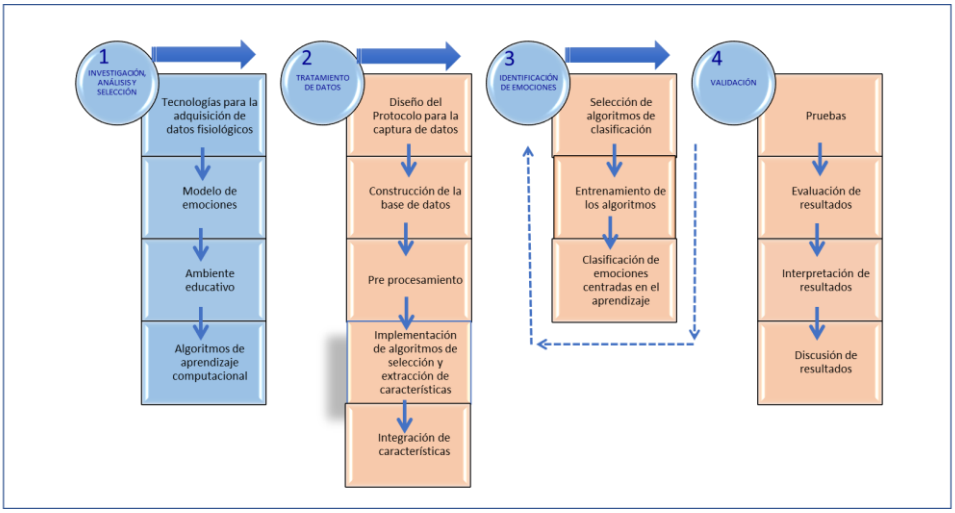
1. Investigación, análisis y selección de las tecnologías para adquisición de los datos fisiológicos, del modelo de emociones, del contexto de aplicación y de los algoritmos de aprendizaje computacional, que serán propuestos para el desarrollo de la metodología.
2. Tratamiento de datos, esta etapa abarca desde la grabación o adquisición de datos fisiológicos y de comportamiento para formar la base de datos, el preprocesamiento requerido para preparar los datos y la implementación de algoritmos de extracción, selección e integración de características relevantes provenientes de diferentes señales.
3. Identificación de emociones, en esta etapa se prueban y entrenan los algoritmos de aprendizaje computacional seleccionados para la clasificación de emociones centradas en el aprendizaje.
4. Validación, se realizan pruebas para evaluar la metodología para la identificación de emociones con métricas que miden la precisión y exactitud de reconocimiento.

Posteriormente se hizo la interpretación de los resultados y la discusión de los objetivos alcanzados.

El desarrollo e implementación de cada una de estas etapas será descrita en los siguientes capítulos.

Figura 11

Metodología general

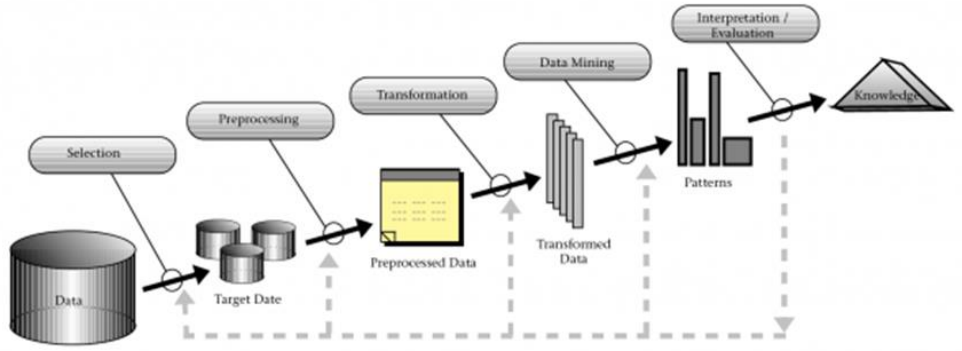


4.2 Proceso para el Descubrimiento de Bases de Datos (*KDD, Knowledge Discovery in Databases*)

Las etapas involucradas en la metodología propuesta corresponden a la implementación del proceso para el descubrimiento de información en bases de datos. En la Figura 12 se identifica cada uno de los pasos, los cuales de manera implícita forman parte del proceso computacional para el reconocimiento automático de emociones centradas en el aprendizaje que fueron propuestos en la metodología (aparecen en color naranja en la Figura 11). Enseguida se describe cada una de ellas:

Figura 12

Etapas del proceso KDD (Fayyad et al., 1996)



1. **Adquisición de datos** *(selection)***:** corresponde a la digitalización de los datos y su almacenamiento.
2. **Etiquetado de los datos** *(target data)***:** consiste en seleccionar e integrar el conjunto de los datos provenientes de fuentes múltiples y heterogéneas.

   En la primera etapa se etiqueta el video visible por observación:

   Etiquetamiento del video:

   a.- Etiquetamiento manual de *frames* de los videos por observadores humanos.

   b.- Etiquetamiento de *frames* de los videos de manera automática (considerando las etiquetas asignadas por el observador humano).

   Con los datos etiquetados de video se realizó el preprocesamiento de datos.
3. **Preprocesamiento** *(preprocessing)***:** si es necesario, eliminación de ruido y de datos aislados. Uso del conocimiento previo para eliminar inconsistencias y duplicados. Elección y uso de estrategias para manejar la información faltante de los conjuntos de datos.

   Sobre los *frames* etiquetados se realiza lo siguiente:

   a. Se mejora el contraste de la imagen, usando ecualización del histograma.

b. Se hace la detección y segmentación automática de la cara, eliminando el fondo como lo realizan en Morales-Vargas, et al. (2017):

1) Detección de puntos de referencia de la cara: consiste en identificar puntos de referencia de ojos, cejas, boca, nariz y contorno de la cara. Para la identificación de puntos de referencia faciales se utiliza el algoritmo del AAM (*Active Appearance Model*), que recuperan 68 pares de puntos faciales $st=\{\{x_i,y_i,x_2,y_2,\ldots x_{68},y_{68}\}, t\}$ los cuales describen la forma de la cara, cada par corresponde a un vértice de un descriptor de la cara definido para cada imagen ***I*** de la secuencia de imágenes $is=\{I,t\}$ de un video que corresponde a una etiqueta específica (aburrido o interesado). Donde *is*, es la secuencia de imágenes, *I* es un *frame* específico en un momento de tiempo *t.*

2) Alineación de puntos de referencia facial: como los puntos de referencia facial pueden ser afectados por la orientación, localización y escala (tamaño) y esto a su vez reflejarse en la eficiencia de los algoritmos de reconocimiento de expresiones, se debe utilizar un método para alinear los puntos de referencia. Un modelo heurístico basado en transformaciones afines es utilizado. Las transformaciones consisten en:

   - Rotación de puntos de referencia para alinear horizontalmente los cantos de los ojos *(s*=puntos de referencia alineados*)*.

   - Cálculo del tamaño máximo de todos los puntos de referencia del estado neutral *(sz).*

   - Normalización de los puntos de referencia *(s,sz)* entre [0,1], con base al tamaño máximo del estado neutral.

   - Superposición del estado neutral por translaciones afines: corrección, para mantener las deformaciones causadas por los movimientos faciales y translación, para alinear estado neutral con la expresión facial.

El *frame* #1 de cada secuencia de video etiquetado será considerado como el estado neutral.

4. **Transformación de los datos** *(transformation)***:** aquí se lleva a cabo la selección de características útiles para representar los datos, la reducción de dimensionalidad o métodos de transformación y fusión de los datos.

   En video se realiza lo siguiente Morales-Vargas, et al. (2019)**:**

   a. Extracción de características faciales:

   1) Cálculo de la magnitud y orientación de los puntos de referencia facial (68) del *frame* inicial *(o)* y del *frame* final *(f).*

   $$m_i = \sqrt{(dx_i)^2 + (dy_i)^2} \quad \text{y} \qquad o_i = \tan^{-1}\left(\frac{dx_i}{dy_i}\right)$$

   2) Concatenación de las tuplas (magnitud y orientación) por punto de referencia facial en un vector de la forma: $mo_i = [m_1, o_1; m_2, o_2; \ldots; m_n, o_n]$, de tamaño 136.

   3) Obtención de una forma triangular (107 triángulos) a partir de los puntos de referencia facial:

   $ts = [a_1, b_1, c_1\,; a_2, b_2, c_2; \ldots\ldots a_n, b_n, c_n]$

   4) Cálculo de las áreas de cada triángulo en el vector:

   $at = [a_1; a_2; \ldots. a_j]$ con *j* siendo el número de triángulos (en total se calcula el área de 107 triángulos por *frame*).

   5) Usando las formas triangulares se calculan los cambios en tamaño de las áreas faciales del *frame* actual (*t=f)* respecto al *frame* neutral (*t=0)*.

   $a_i = at_{i,f} - at_{i,0}$, obteniendo un vector de tamaño 107

   6) Concatenación de los vectores de características *mo y a* en el vector *br = [mo, a]* de dimensión 243 (68X2=136 de *mo* y *107 de a*). Este vector formará la matriz de características llamado: "***car_full***".

   7) Reducción de la dimensionalidad por medio de una exploración manual de los datos considerando los movimientos de las unidades de acción que están presentes en las emociones de aburrido e interesado. Seleccionando: cejas internas, cejas externas, parpados, nariz, esquina derecha del labio, esquina izquierda del labio, labio superior, labio inferior, mandíbula y esquinas de los labios (10 partes faciales). En cuanto a áreas

se seleccionan las de ojos y boca. De esta forma se reduce la dimensionalidad original de 243 a 22.

8) Para incluir todas las características de una manera más simple se utilizan dos operaciones de agrupación:

Agrupación Promedio $PooledCar = \frac{1}{p}\sum_{i=1}^{p} v_i$ y

Agrupación Máxima $PooledCar = \max(v_i)$, donde i es la i-*ésima* área distintiva a considerar y v_i es su subconjunto de puntos de referencia.

Lo que formarán las matrices de características: ***"car_avg"*** y *"**car_max**"*, respectivamente.

b. Obtención de la matriz de características de AUs

1) $PooledCar$ es el vector de características con el que se entrena un modelo *Fuzzy* para cada unidad de acción (ya sea con "*car_avg*" o con "*car_max*").

2) En conjunto todo el sistema es usado para describir los movimientos faciales de cada secuencia en términos de AUs, obteniendo un vector de características μ_{AU} con un valor de membresía para cada AU.

5. **Minería de datos** *(data mining)***:** consiste en elegir los algoritmos de aprendizaje computacional. Elegir la tarea de minería de datos y los algoritmos de minería que traten todos los criterios (clasificación, regresión, clusterización o modelos mixtos).

   a. Entrenar algoritmos de clasificación con los vectores de característica μ_{AU} y de características faciales generados a partir del etiquetado manual por cada etiquetador humano.

   b. Entrenar una Red Neuronal de Aprendizaje Profundo (DNN, *Deep Neural Network*) a partir de lo videos.

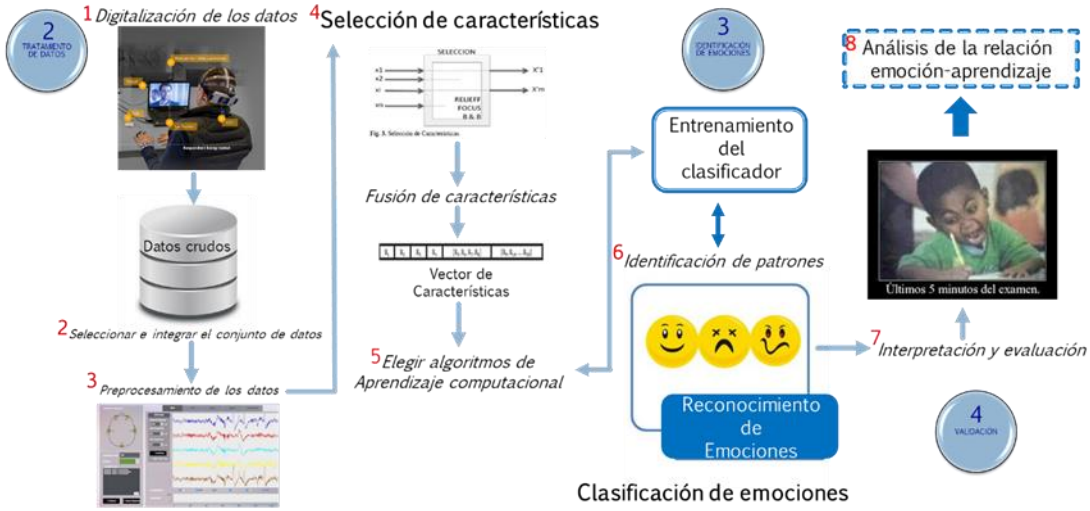
6. **Identificación de patrones** *(patterns)***:** buscar patrones de interés en una forma en particular, es decir el reconocimiento de emociones.
7. **Interpretación y evaluación de los patrones minados** *(interpretation / evaluation)***:** se realiza la interpretación y evaluación de los resultados de los algoritmos utilizados.

8. **Consolidación del conocimiento descubierto** *(knowledge)*: discusión y análisis de las tasas de reconocimiento de las emociones identificadas.

Los ocho pasos (del proceso KDD) antes explicados son identificados con su número en el proceso computacional para el reconocimiento automático de emociones que se pretende implementar como parte de nuestra metodología para el reconocimiento de emociones. Estos pasos se ilustran en la Figura 13. En esta también se visualizan las tres etapas de la metodología propuesta involucradas en este proceso (tratamiento de datos, identificación de emociones y validación).

Figura 13

Pasos del proceso KDD inmersos en las etapas de la metodología propuesta



4.3 Etapa 1: Investigación, Análisis y Selección

Esta etapa incluye las actividades de: investigación, análisis y selección de tecnologías de adquisición de datos fisiológicos, del modelo de emociones, del ambiente educativo y de los algoritmos de selección y clasificación. Después de la investigación y análisis correspondientes se han seleccionado los tres primeros elementos de esta etapa.

Las tecnologías de adquisición de datos fisiológicos y de comportamiento para crear la base de datos que se eligieron son las siguientes:

1. Cámara Web Logitech: para capturar video continuo durante los 35 minutos que dura el experimento (inicialmente fueron 38) en formato “MP4” (1920X1080, 28 f/s).
2. Cámara térmica ICI 9320P: para capturar matrices de temperaturas del rostro, en archivos “CSV” (240X320), generando 1 archivo por segundo. También se genera un video “AVI” (600 imágenes por minuto) de aproximadamente 3.5 minutos.
3. Sensor de pulso cardíaco implementado con Arduino: se genera un archivo “TXT” de las lecturas de la potencia de señal del sensor con un rango aproximadamente de 9.5 capturas por segundo (10Hz de frecuencia de muestreo).

Para el modelo de emociones se seleccionó el modelo discreto. Los tipos de emociones centradas en el aprendizaje que se intentan reconocer son: interés (placer) y aburrimiento.

El ambiente educativo seleccionado para el proceso de enseñanza-aprendizaje fue un MOOC de álgebra básica de Coursera trabajando únicamente con el primer tema (duración aproximada de 36 min.). Los alumnos de la muestra fueron estudiantes de carreras del área de ingeniería de la Universidad Politécnica de Puebla y de la BUAP, cuyas edades oscilan entre los 18 y 35 años.

4.4 Etapa 2: Tratamiento de Datos.

En esta etapa describimos cada uno de los pasos del tratamiento de datos:

1. El diseño del protocolo formal para la adquisición de los datos (que contiene la descripción del experimento, el consentimiento informado y el diseño de *tests* para auto evaluar emociones discretas y continuas).
2. Construcción de la base de datos crudos.
3. Preprocesamiento de los datos.

En los siguientes apartados describimos los detalles de estas actividades.

4.4.1 Diseño del Protocolo para la Adquisición de los Datos.

Se diseñó un protocolo para la recolección de datos considerando las tecnologías de adquisición de datos fisiológicos y de comportamiento ya seleccionados. También se diseñaron *tests* para evaluar los dos modelos de emociones como auto reportes. Se redactó la carta de consentimiento informado, la cual es proporcionada a cada alumno antes de iniciar con el experimento de captura de datos. En esta carta se le pide a cada alumno sus datos personales y se les explica en qué consiste el experimento. Además, se les informa que no tiene ningún tipo de consecuencia el decidir participar o no en el experimento. Se les hace saber que se guardará total confidencialidad de sus datos y que estos sólo serán usados para fines de investigación. Finalmente, se les solicita autoricen su participación firmando la carta.

En el apéndice A se presenta el diseño del protocolo del experimento; en el B, la carta de consentimiento informado y en el C se muestra el diseño de los *tests* de emociones.

Se utilizan dos *tests* de emociones que fueron diseñados con formularios de Google.

El primero es para evaluar emociones de forma discreta. Para éste se crea un cuestionario en el que se pide al alumno seleccione la ECA que está experimentando en ese momento, además del nombre de la emoción se le orienta con imágenes representativas de la misma. Este cuestionario fue validado por expertos del departamento de psicología de la Universidad Autónoma de Tlaxcala y por medio de pruebas previas al menos de 10 alumnos quienes corroboraron sus respuestas.

El segundo *test* es para evaluar emociones de forma continua. Para éste se utiliza la autoevaluación del maniquí *(Self-Assessment Manikin*, SAM), cuestionario propuesto por Lang en 1980 y revalidado en su publicación Bradley y Lang, (1994). En éste, los participantes auto reportan sus sentimientos momentáneos de placer (valencia), excitación y dominancia usando una escala de evaluación pictórica de 9 puntos. SAM provee una forma simple, rápida y no lingüística de evaluar el estado emocional a lo largo de las 3 dimensiones.

Dentro del experimento el alumno contesta los dos *tests* en tres momentos diferentes cada 10 minutos. A través de un temporizador se activa una alarma que indica al alumno

cambiarse de la ventana actual (en la cual está atendiendo el MOOC de álgebra) a la ventana donde está activo el *test*.

Una vez que contesta las dos secciones, envía sus respuestas las cuales son almacenadas en archivo "XLSX" y regresa a la ventana del MOOC para continuar.

4.4.2 Construcción de la Base de Datos.

Esta actividad se inició con capturas de prueba de 18 sujetos en el Laboratorio de Ingeniería del Lenguaje y del Conocimiento (LKE) ubicado en el edificio EMA7 de la Benemérita Universidad Autónoma de Puebla (BUAP) en Ciudad Universitaria. En esta etapa participaron alumnos de la Facultad de Computación de la Universidad. En la Figura 14 se pueden observar las instalaciones de este laboratorio en donde se hicieron las primeras capturas.

Las siguientes capturas se llevaron a cabo en la Universidad Politécnica de Puebla (UPPue). Este cambio fue sugerido después de conocer las condiciones físicas de uno de los laboratorios con los que cuenta la Universidad. Así se aprovecharon sus instalaciones buscando mejor comodidad para los alumnos. En la

Figura 15 se muestra una fotografía de las instalaciones del Laboratorio de Experiencia del Usuario de la UPPue, lugar donde se realizaron las demás capturas.

En la UPPue se hicieron capturas de 64 alumnos de diferentes ingenierías (tecnologías de la información, biotecnología, ciencias de la computación y financiera).

En la Tabla 5 se muestran estadísticas de las características del total de experimentos realizados, como número de participantes, edad, sexo, y nivel de estudios, de los 82 participantes, 50 son hombres y 32 mujeres con edad promedio de 21 años.

De cada experimento se recaba la información mostrada en la Tabla 6. Se toman como ejemplo los datos del experimento 1.

Figura 14

Instalaciones del Laboratorio del LKE en el edificio EMA7 de la BUAP

Figura 15

Instalaciones del Laboratorio de Experiencia de Usuario en la UPPue

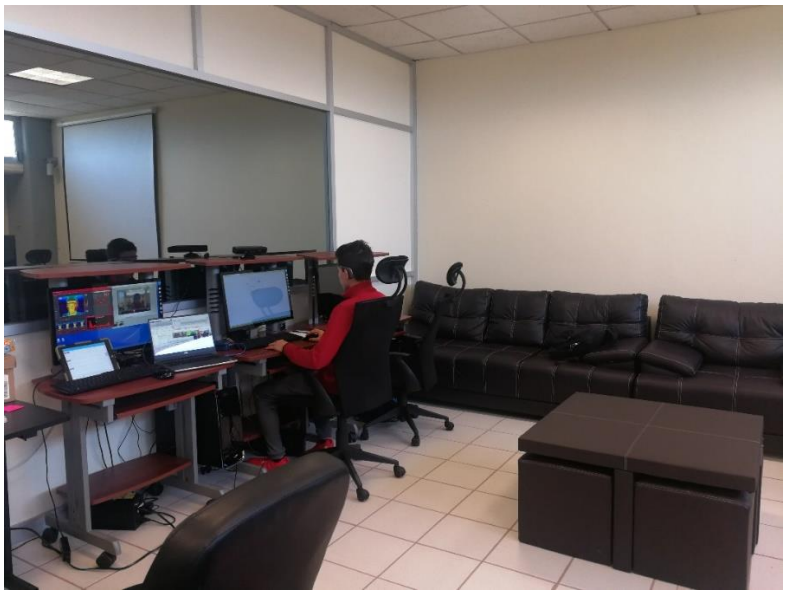


Tabla 5

Estadísticas de las características de la muestra

TOTAL DE ALUMNOS	EDAD PROMEDIO	HOMBRES	MUJERES	INGENIERÍAS	MAESTRÍA	DOCTORADO	PREPARATORIA
82	21	50	32	69	8	4	1

Tabla 6

Datos que se obtienen de cada uno de los experimentos

# EXPERIMENTO	**1**
FECHA	10/10/2018
NOMBRE	Alberto Remigio Alvarado
SEXO	M
EDAD	20
CARRERA	Ciencias de la Computación
HORA DE INICIO	08:40
HORA DE FINALIZACIÓN	09:19
TIEMPO PROGRAMADO (min)	38
TIEMPO REAL (min)	39
VALORES DE LOS AUTO REPORTES	
1ER TEST	08:50:33
EMOCIÓN	Interesado
PLACER	7
ACTIVACIÓN	5
DEPENDENCIA	8
2DO TEST	09:00:33
EMOCIÓN	Interesado
PLACER	7
ACTIVACIÓN	5
DEPENDENCIA	8
3ER TEST	09:19:01
EMOCIÓN	Interesado
PLACER	9
ACTIVACIÓN	6
DEPENDENCIA	8
DATOS DEL SENSOR DE RITMO CARDÍACO	
PULSO	SI, rango: 513-553, Media: 546
% DATOS DE PULSO	83%
CAPTURAS TOTALES	21672
RANGO DE MUESTREO	9.505263158
CAPTURAS POR MINUTO	571
DATOS DE LA CÁMARA TÉRMICA	
ARCHIVOS “CSV”	2074
TIEMPO ARCHIVO “AVI” (MINUTOS)	03:48
DATOS DEL VIDEO	
ARCHIVO DE VIDEO “WMV” (minutos)	39.00
OBSERVACIONES	3er. test contestado hasta el final

La primera parte son datos generales como número de experimento, fecha de la captura, hora de inicio, hora de finalización, datos del participante. La segunda parte corresponden a las respuestas de los *tests* de emociones, estadísticas de cada uno de los sensores utilizados y de las cámaras de video. Al final hay un apartado de observaciones para describir alguna situación especial que se presentó durante la captura y requiere ser documentada.

4.4.2.1 Base de Datos Sin Etiquetar

La base de datos sin etiquetar contiene los datos crudos capturados por cada dispositivo utilizado y almacenados en sus correspondientes archivos. Los datos del sensor de pulso se almacenan en 1 archivo de datos, *txt*. Para los datos de la cámara térmica el número de archivos varía dependiendo del tiempo programado de captura, si el tiempo programado es de 40 minutos se almacenan 2400 archivos *csv*, 2400 archivos *jpg* y 1 archivo *avi* (4801 archivos) y el video de la cámara web se almacena en 1 archivo *wmv*. Por experimento se generan un total de 4802 archivos. Los archivos de datos se almacenan en un disco duro externo de 4TB. El resumen de las estadísticas de los datos crudos capturados se muestra en la Tabla 7.

Tabla 7

Estadísticas de archivos de datos crudos que forman la base de datos

SENSOR DE PULSO "*txt*"	CÁMARA TÉRMICA "*csv/jpg/avi*"	CÁMARA WEB "*wmv*"	ARCHIVOS POR EXPERIMENTO	ARCHIVOS TOTALES (82 EXPERIMENTOS)
1	4801	1	4803	393,846

Con la finalidad de verificar el contenido de los archivos almacenados, de manera general, se ha hecho una revisión de cada uno de ellos. Por ejemplo, para el caso de los archivos de video se ha verificado que contengan la grabación del experimento correspondiente y que su tamaño corresponda a los minutos que duró el experimento. Se ha verificado que el número de archivos de la cámara térmica correspondan al número de

minutos capturados. Y se ha verificado si las capturas de la potencia de señal del sensor de pulso corresponden a medidas correctas relacionadas con el pulso de los seres humanos para que puedan ser consideradas como útiles.

Un resumen de la cantidad de experimentos que cuentan con archivos de datos completos para cada uno de los dispositivos utilizados se muestra en la Tabla 8. En esta se puede ver también la cantidad de experimentos considerados como completos, estos son aquellos experimentos que cuenta con información de los 3 dispositivos utilizados. Así, hay un 78% de experimentos con archivos completos de la cámara térmica, un 59% con datos útiles del sensor de pulso, un 96% con video de la cámara web completo. Treinta y cuatro de los experimentos (41%) cuentan con archivos completos de la cámara térmica, del sensor de pulso y del video.

Tabla 8

Número de experimentos con archivos completos por dispositivo utilizado

CÁMARA TÉRMICA	SENSOR DE PULSO	VIDEO	TEMPERATURA, VIDEO Y PULSO
64	48	79	34

4.4.2.2 Base de Datos Etiquetada.

Se realizó el proceso de etiquetamiento del video visible y de las imágenes térmicas.

1) Etiquetamiento manual:

El primer planteamiento es a través de un etiquetamiento manual realizado por humanos. Para esto se capacitaron a dos personas involucradas en el proyecto. Para la capacitación se utilizó el Manual del FACS (P. Ekman et al., 2002) tanto de manera teórica como práctica, realizando algunos ejercicios de forma simultánea con los dos etiquetadores humanos. Mediante observación desde el segundo cero hasta el último segundo se fueron recorriendo los videos a etiquetarse. Se eliminaron videos que no contenían la grabación completa del experimento y los de aquellos alumnos que usaban lentes. Cada etiquetador analizó 57 videos, la información registrada se almacena en un archivo de Excel cuyo esquema

se muestra en la Tabla 9, donde t_0 es el tiempo de inicio de la emoción, t_f, el momento en el que culmina la emoción y $etiqueta$ es el nombre de la emoción discreta identificada por observación de las AUs.

Tabla 9

Etiquetamiento manual de videos

to		***tf***		***Etiqueta***
min	seg	min	seg	Interesado/Aburrido
0	0	2	33	Interesado
2	33	2	35	Aburrido
2	35	17	10	Interesado
17	10	18	0	Aburrido
18	0	19	37	Interesado
19	37	19	42	Aburrido
19	42	23	35	Interesado
23	35	25	0	Aburrido
25	0	46	18	Interesado

Para poder llevar a cabo el etiquetamiento manual se identificaron las AUs más representativas de las emociones *interesado* y *aburrido* para considerarlas en la identificación de la emoción. La selección de las AUs asociadas a cada emoción se determinó tomando como referencia el número de AUs que utilizan diferentes autores para identificar emociones básicas. La mayoría de los autores utilizan solo algunas de las primeras 27 AUs. Por ejemplo en *The Emotional Intelligence Academy (https://www.eiagroup.com)* y en el trabajo de Wegrzyn, et al. (2017) utilizan 17 AUs para identificar emociones básicas. En la empresa *iMotions* utilizan 14 (Farnsworth, 2019). De acuerdo con el consenso de los etiquetadores humanos y tomado como referencia los trabajos mencionados, para esta investigación se seleccionaron 16 AUs asociadas a las emociones *interesado* y *aburrido*. La presencia de estas AUs en la expresión facial son la referencia para realizar su etiquetamiento. En la Tabla 10 se enlistan las 16 AUs asociadas a cada emoción (*interesado/aburrido*).

Tabla 10

Unidades de acción asociadas a las emociones de interesado y aburrido

Acción	AU	Interesado	Aburrido	Criterio
Levanta ceja interior	1	X		Asombro

Levanta ceja exterior	2	X			Atención
Baja cejas	4			X	Impaciente, desagrado
Levantar párpado superior	5	X			Asombro
Parpados tensos	7	X			Mirada fija
Arruga nariz	9			X	Disgusto
Levanta labio superior	10			X	Descontento
Acentúa pliegue naso labial	11			X	Disgusto
Jalar las esquinas de los labios unilateral	12 L/R			X	Desprecio, rechazo
Hoyuelos	14	X			Agrado
Bajar esquinas de labios	15			X	Desprecio
Tensar labios hacia enfrente	23			X	Inhibición
Presionar los labios	24	X			Atención
Separa labios	25	X			Placer
Abrir la boca	26			X	Sueño
Chuparse los labios	27	X			Pensativo

2) Auto etiquetamiento:

El auto etiquetamiento se realizó considerando la etiqueta seleccionada por el alumno en el momento en el que contesta el *test* de emociones. Por cada video se selecciona el minuto completo en el que se encuentra a la mitad de éste la hora registrada en la que el alumno contesta el *test* de emociones. Por lo tanto, para cada experimento obtenemos 3 videos de un minuto de duración cada uno etiquetado por la emoción discreta asignada por el propio alumno. Esta información es registrada en archivos de Excel con los mismos datos que en el etiquetamiento manual, un ejemplo es mostrado en la Tabla 11.

Tabla 11

Auto etiquetamiento de videos

hora	***t0***		***tf***		***emoción***
	min	***seg***	***min***	***seg***	
13:15:44	15	14	16	14	Aburrido
13:23:57	23	27	24	27	Interesado
13:33:46	33	16	34	16	Interesado

Para las dos propuestas de etiquetamiento, se hace la implementación automática de etiquetado para video visible, imágenes térmicas y ritmo cardiaco (Arroyo et al., 2009). Con el video visible se generan fragmentos de video y fotogramas individuales de cada fragmento

para el rango de tiempo especificado por cada etiqueta. El video se genera a un rango de 1 *frame* por segundo con la idea de reducir el número de *frames* a procesar.

En el caso de las imágenes térmicas el etiquetamiento es sobre el conjunto de *frames* correspondientes al rango de tiempo que se etiqueta. Recordando que tenemos 1 archivo *csv/jpg* por cada segundo de grabación, 2400 *frames* en un experimento que dura 40 minutos.

El etiquetado del ritmo cardiaco también se realizará considerando los dos esquemas de etiquetamiento con los datos capturados a una frecuencia 10 capturas por segundo.

4.4.3 Preprocesamiento de los Datos

En este paso, si es necesario, se ejecutan estrategias para eliminación de ruido, tratamiento de datos aislados o faltantes, eliminación de inconsistencias y duplicados.

El preprocesamiento en video visible que se realiza fue descrito en el capítulo de la metodología propuesta. Aquí se muestra los procedimientos implementados.

En video visible:

1. Mejorar el contraste de la imagen, usando ecualización del histograma.

En la Figura 16 , en *a)* se puede observar una imagen original obtenida del video y en *b)* la misma imagen con el contraste mejorado. La finalidad es resaltar las diferentes partes de la cara para facilitar la identificación de los puntos de referencia faciales.

Figura 16

Preprocesamiento de la imagen original para mejorar el contraste

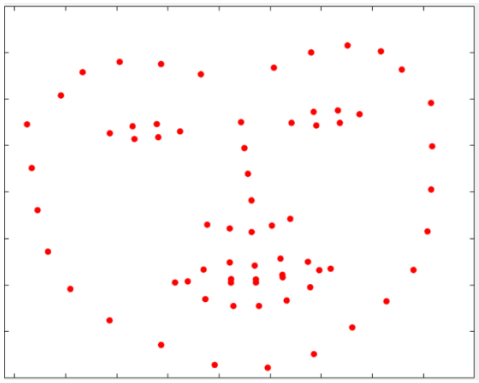
a) Imagen original b) imagen con contraste

2. Detección y segmentación automática de la cara, eliminando el fondo.

a) Detección de puntos de referencia de la cara: La identificación de los puntos de referencia de ojos, cejas, boca, nariz y contorno de la cara se hace con un algoritmo del AAM, que recuperan 68 pares de puntos que describen la forma de la cara. En la Figura 17, se pueden ver los puntos de referencia facial para un *frame* del video.

Figura 17

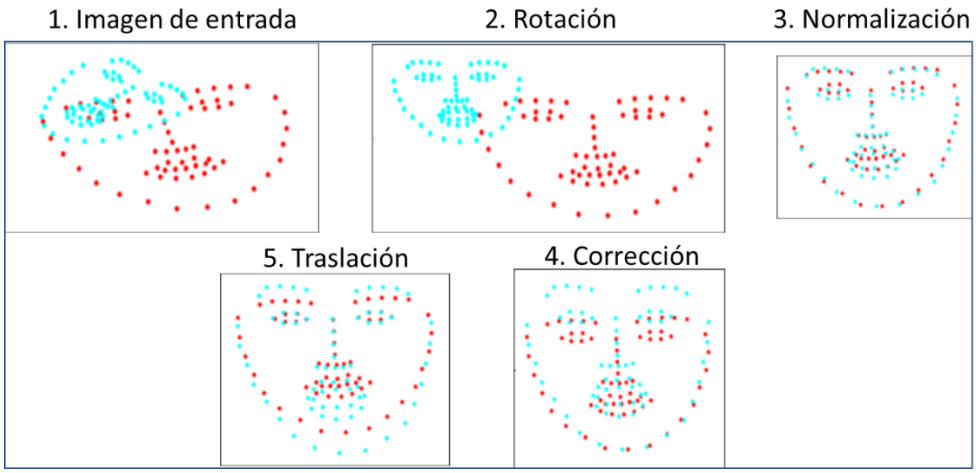
Puntos de referencia facial (facial landmark) obtenidos de un frame del video



b) Alineación de puntos de referencia facial: como los puntos de referencia facial pueden ser afectados por la orientación, localización y escala (tamaño) y esto a su vez reflejarse en la eficiencia de los algoritmos de reconocimiento de expresiones, se utiliza un método para alinear los puntos de referencia a través de un modelo heurístico basado en transformaciones afines. Las transformaciones aplicadas se pueden ver en la Figura 18, la imagen 1 corresponde a la imagen de entrada, la 2 a la rotación para alinear la imagen de entrada con la imagen neutral que para nosotros es el *frame* #1 de cada video etiquetado. La imagen 3 corresponde a la normalización de los puntos de referencia, entre [0,1]. En la imagen 4, se hace una corrección, para mantener las deformaciones causadas por los movimientos faciales y en la imagen 5, se hace la translación final para alinear estado neutral con la expresión facial (imagen de entrada).

Figura 18

Alineación de puntos de referencia facial

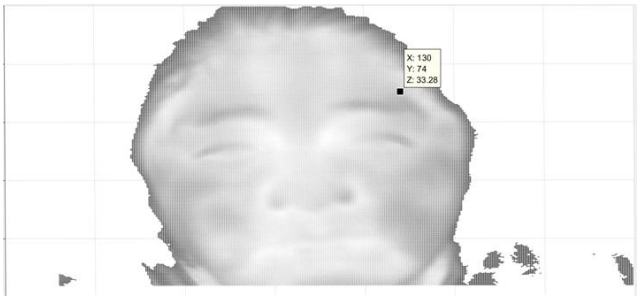


En imágenes térmicas:

1. Se propone primero hacer la segmentación del rostro. La Figura 19 muestra un ejemplo de un *frame* segmentando el rostro del estudiante.

Figura 19

Segmentación del rostro en imágenes térmicas



En la potencia de señal del sensor de ritmo cardíaco:

1. Se realizó la limpieza de datos eliminando valores atípicos a partir de la distancia intercuartil con la cual se definen valores de corte inferiores y superiores, internos y externos.

2. Los valores eliminados se llenan con ceros.

3. Se propone hacer el etiquetado de cada archivo de datos de acuerdo con las mismas etiquetas asignadas para el video, bajo los dos esquemas de etiquetado.

4.4.4 Implementación de Algoritmos para la Extracción, Selección e Integración de Características

En video visible:

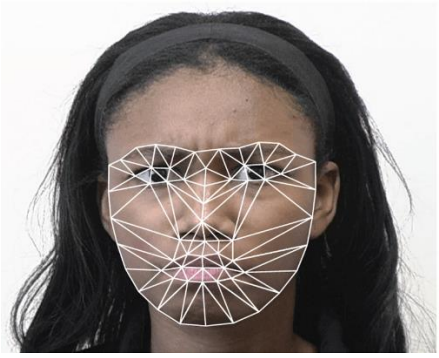
1. Extracción de características:

a. Cálculo de la magnitud y orientación de los puntos de referencia del *frame* neutral (t=0) y del *frame* final (t=f). Se genera el vector *mo*.

b. Obtención de una forma triangular de los puntos de referencia facial. De acuerdo con el esquema planteado en la metodología se genera el vector de áreas (*at)* resultado de triangular la cara a partir de los puntos de referencia facial. En la Figura 20 se muestra el resultado de este proceso.

Figura 20

Triangulación del rostro, para obtención de sus características geométricas



c. Usando las formas triangulares se calculan los cambios en tamaño de las áreas faciales del *frame* actual (t=f) respecto al *frame* neutral (t=0). Se genera el vector *a.*

d. Se concatenan los vectores de características *mo* y *a* obteniendo el vector:

br = [mo, a] de dimensión 243

(con el que se genera el vector de características "*car_full")*

e. Para incluir todas las características de una manera más simple se realiza la agrupación de los puntos de referencia facial de las 10 áreas de la cara seleccionadas, bajo dos esquemas diferentes: promedio y máximo generando el vector de características, *PooledCar*, y se agregan las áreas de los ojos y boca. Así, se crean los vectores de características "*car_avg*" y *car_max*" (con 22 características).

f. Se reduce la dimensionalidad del vector de características *br* de 243 a 22.

2. Obtención del vector de características de AUs

a. Los vectores de características "*car_avg*" o *car_max*" se utilizan para entrenar un modelo de inferencia *fuzzy* para el reconocimiento de las AUs.

b. El modelo *fuzzy* genera como salida un vector de características AUs con un valor de membresía para cada AU. En la Tabla 12, se pueden observar los vectores μ_{AU} para los primeros 10 *frames* de uno de los videos procesados. Los valores de las columnas corresponden al valor de membresía de cada AU (considerando 16 AUs: 1, 2, 4, 5, 7, 9, 10, 11 ,12, 14, 15, 23, 24, 25, 26 y 27). El primer vector corresponde al estado neutral *t(0).* Un valor de cero indica que esa AU no está presente. Un valor negativo representa un movimiento contrario al de esa AU, por ejemplo, la AU5 corresponde a *"levantar el parpado"*, un valor negativo indica que el parpado no está levantado o el ojo está entre cerrado con el valor de membresía obtenido.

Tabla 12

Ejemplo de vectores de características $\boldsymbol{\mu_{AU}}$

# FRAME							UNIDADES DE ACCIÓN FACIAL						
	1	2	3	4	5	6	7	23	24	25	26	27
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0.086920353	-0.087781499	0	0.191393858	-0.115772166	0	0.13792959	0.095066791	-0.086976008	-0.016134692	-0.009234107	-0.095820488
3	0.400204895	0.469842737	0	0.326715145	-0.064400575	0	0.142108596	0.002124084	-0.090291585	0.027900686	-0.0184323	-0.056019226
4	0.383735395	0.511373163	0	0.33365076	-0.004739833	0	0.196583807	0.015750755	0.294621725	0.037929306	-0.019396031	-0.053118764
5	0.363436106	1.034559102	0	0.454895283	-0.040631021	0	0.046459533	0.0404546	-0.166587478	0.023634897	-0.02152879	-0.043505756
6	0.516778277	1.02091871	0	0.448931474	0.023652363	0	0.079359396	0.319357178	0.165060629	-0.031243383	-0.02580574	-0.026299897
7	0.694182616	1.117885516	0	0.498037026	0.10990929	0	-0.021940848	0.162751288	0.140385646	-0.029428976	-0.025450557	-0.027897326
8	0.748497556	0.649206871	0	0.357454036	0.014655282	0	0.133615624	0.282843494	-0.44856777	0.001421282	-0.022706122	-0.038698736
9	0.748497556	0.649206871	0	0.357454036	0.014655282	0	0.133615624	0.282843494	-0.44856777	0.001421282	-0.022706122	-0.038698736
10	0.493736697	1.043212336	0	0.457062675	0.121652715	0	-0.046137855	0.140860713	-0.850049925	-0.043865249	-0.019294491	-0.052444062

El vector de características *"AUs"* se utiliza para entrenar los algoritmos de clasificación. Se propone integrar al vector los datos obtenidos de la extracción de características de archivos del sensor de ritmo cardíaco y de la cámara térmica para probar diferentes propuestas de fusión de datos, así como diferentes algoritmos de clasificación.

En ritmo cardíaco:

Se seleccionan 27 características sobre datos del ritmo cardíaco: media, mediana, desviación estándar, desviación media absoluta, cuartil 25, cuartil 75, rango inter-cuartil, medida de asimetría, curtosis, entropía de Shannon, entropía del espectro, radio, magnitud y valor del espectro de poder y coeficientes cepstrales de las frecuencias de Mel.

4.5 Etapa 3: Identificación de Emociones

La etapa de identificación de emociones consta de tres apartados: selección de algoritmos de clasificación, entrenamiento de algoritmos y pruebas de clasificación de emociones centradas en el aprendizaje (aburrido e interesado). El proceso de identificación de emociones se realizó con características obtenidas de video visible. Como se explicó en la etapa anterior se hizo la extracción de características de videos etiquetados por humanos obteniendo 4 vectores de características:

car_full: formado por 243 características.

AUs: formado por características de 16 AUs.

car_avg: formado por 22 características (agrupación promedio de magnitud y orientación de los puntos de referencia facial de 10 descriptores faciales y 2 áreas).

car_max: formada por 22 características (agrupación máxima de magnitud y orientación de los puntos de referencia facial de 10 descriptores faciales y 2 áreas).

Para la preparación de los vectores de características se obtiene un *frame* de cada 15 por cada video etiquetado. Para las observaciones del etiquetador #1, se obtienen las estadísticas resumidas en la Tabla 13 en cuanto a número de videos y vectores de características.

Tabla 13

Estadísticas de las observaciones del etiquetador humano #1

EMOCIÓN	ABURRIDO	INTERESADO
Capturas (# de videos)	46*	57
Videos generados por etiqueta	238	291
# de vectores (*frames*) generados	14,560	63,768

* En 11 videos no se identificó la emoción "*aburrido*"

De acuerdo con la Tabla 13 para cada tipo de características (*full, max, avg y AUs*) se generan un total de 14,560 vectores correspondientes a la emoción de "*aburrido"* y 63,768 para la emoción de *"interesado"*.

Una vez que se procesan todos los videos seleccionados para el entrenamiento, se hace una limpieza y ajuste a los vectores de datos, que consiste en:

a) Eliminación del *frame* neutral para los vectores de características "*avg", max y "full"*.
b) Eliminación de AUs no clasificadas por el sistema de inferencia *fuzzy*.
c) Ajuste de valores fuera de rango de los vectores de AUs.
d) Separación de los vectores de características para entrenamiento y para clasificación.
e) Concatenación de los vectores de características de la emoción *"aburrido",* con vectores de características de la emoción *"interesado"* para entrenamiento y prueba.

4.5.1 Selección de Algoritmos de Clasificación

La elección de los algoritmos de clasificación a probar está en función del análisis del estado del arte que se presenta en el capítulo 3, en el que se identifican los algoritmos más utilizados para el reconocimiento de ECA en las investigaciones más recientes. En primer lugar, están las redes neuronales, en segundo la clasificación SVM, le siguen algunos algoritmos de aprendizaje no supervisado de clusterización, después están los algoritmos más tradicionales como regresión logística, Naïve Bayes, clasificador Bayesiano, KNN, ensambles y lógica difusa. En dos de los trabajos más recientes utilizan CNN (Gupta et al., 2018), (Nezami et al., 2020).

Inicialmente se realizaron pruebas de entrenamiento con la mayoría de los clasificadores antes mencionados sin considerar técnicas de clusterización, lógica difusa, redes neuronales y CNN. En una segunda etapa se entrenó una red neuronal artificial y en la última etapa una CNN.

4.5.1.1 Selección de Algoritmos de Clasificación Tradicionales

A partir de las primeras pruebas con los algoritmos de clasificación tradicionales se identificaron aquellos que daban mejores resultados al procesar los datos, esto ayudó a seleccionar tres de ellos para profundizar en los entrenamientos. Los algoritmos seleccionados fueron SVM, KNN y ensambles de árboles. Estos fueron entrenados bajo varios esquemas de clasificación cambiando los porcentajes de datos de entrenamiento y de prueba, el número de *folds* para validación cruzada y utilizando para todos los entrenamientos 16 características obtenidas de los valores de membresía de las *AUs* que fueron seleccionadas previamente de manera directa. Para cada tipo de algoritmo también se ejecutó un proceso para optimización de sus hiperparámetros. Los algoritmos seleccionados, las configuraciones para su entrenamiento y sus hiperparámetros con los rangos válidos para cada uno de ellos se resumen en la Tabla 14 y en seguida se describen brevemente las características de estos algoritmos.

Tabla 14

Esquemas de Clasificación

Esquema de Clasificación	Hiperparámetros	Rango
SVM ***Folds*: 3, 5 y 10** **% datos de entrenamiento: 60, 70 y 80** **% datos de prueba: 40, 30 y 20** **Características: 16**	Restricción de caja	0.001 – 1000
	Escala *kernel*	0.001 - 1000
	Función *kernel*	Gaussiana, Lineal, Cuadrática, Cúbica
	Datos estandarizados	Verdadero, falso
KNN ***Folds*: 3, 5 y 10** **% datos de entrenamiento: 60, 70 y 80** **% datos de prueba: 40, 30 y 20** **Características: 16**	Número de vecinos	1-2745
	Métrica de distancia	City Block, Chebyshev, Correlación, Coseno, Euclidiana, Hamming, Jaccard, Mahalanobis, Minkowski (cúbica), Spearman.
	Peso de la distancia	Igual, Inversa, Cuadrado Inverso

Ensamble ***Folds*: 3, 5 y 10** **% datos de entrenamiento: 60, 70 y 80** **% datos de prueba: 40, 30 y 20** **Características: 16**	Método de ensamble	Bag, GentleBoost, LogitBoost, AdaBoost, RUSBoost.
	Número máximo de *splits*	1-54829
	Número de modelos de aprendizajes	10-500
	Rango de aprendizaje	0-1
	Modelo de aprendizaje	Discriminante, KNN, árboles.

A. **Máquina de vectores de soporte** es una técnica de aprendizaje computacional que encuentra la mejor separación posible entre clases. La SVM encuentra el hiperplano que maximiza el margen de separación entre clases. Los parámetros elegibles para ajustar una SVM se enlista en la Tabla 14 y son los siguientes:
   1. Restricción de caja: es un parámetro que controla la penalización máxima impuesta a las observaciones que violan el margen, lo que ayuda a evitar el sobreajuste. Si aumenta la restricción de caja, el clasificador SVM asigna menos vectores de soporte. Sin embargo, aumentar la restricción de la caja puede conducir a tiempos de entrenamiento más largos.
   2. Función *kernel*: es la función usada para procesar la matriz de predictores.
   3. Escala *kernel*: es un escalar positivo que divide todos los elementos de la matriz de predictores para luego aplicar la función *kernel*.
   4. Datos estandarizados: permiten estandarizar o no los predictores usando sus correspondientes medias y desviaciones estándar ponderadas.

B. **KNN** es un método que busca en las observaciones más cercanas a la que se está tratando de predecir y clasifica el punto de interés basado en la mayoría de los datos que le rodean. Los parámetros elegibles para el clasificador KNN se enlistan en la Tabla 14 y son los siguientes:
   1. Número de vecinos más cercanos a encontrar en la matriz de predictores para clasificar cada punto al predecir.
   2. Métrica de distancia es la función utilizada para calcular de la distancia de los vecinos más cercanos.
   3. Peso de la distancia especifica la función de peso: ninguna, *1/distancia*, *1/distancia*^{2}.

C. **Ensambles** son un modelo predictivo compuesto por una combinación ponderada de varios modelos de clasificación. En general, la combinación de varios modelos de clasificación aumenta el rendimiento predictivo. Los parámetros elegibles para ensambles se enlistan en la Tabla 14 y son los siguientes:

1. Método de ensamble es el método de agregación de modelos de aprendizaje para generar el ensamble. Puede ser alguno de los siguientes:
    a) *Bagging* (o agregación *Bootstrap*): se utiliza para poder combinar modelos de aprendizaje bajo el mismo esquema de clasificación, por ejemplo, para poder combinar muchos árboles de decisión. Cada modelo se entrena con subconjuntos del conjunto de entrenamiento. Estos subconjuntos se forman eligiendo muestras aleatoriamente (con repetición) del conjunto de entrenamiento. Los resultados se combinan, para problemas de clasificación, igual que en la votación por mayoría, con el voto suave para los modelos que den probabilidades. Para problemas de regresión, normalmente se utiliza la media aritmética (Breiman, 1996) y (Breiman, 2001).
    b) En *boosting* (reforzamiento): cada modelo intenta arreglar los errores de los modelos anteriores. Por ejemplo, en el caso de clasificación, el primer modelo tratará de aprender la relación entre los atributos de entrada y el resultado. Seguramente cometerá algunos errores. Así que el segundo modelo intentará reducir estos errores. Esto se consigue dándole más peso a las muestras mal clasificadas y menos peso a las muestras bien clasificadas. Para problemas de regresión, las predicciones con un mayor error cuadrático medio tendrán más peso para el siguiente modelo (Freund y Schapire, 1997). Hay muchas implementaciones de ensambles que usan *boosting*. Algunos son *AdaBoost* (*Adaptative Boosting*), *GentleBoost, LogitBoost* y *RUSBoost*.
    c) *AdaBoost* (refuerzo adaptativo): entrena modelos de aprendizaje secuencialmente y su método de agregación se basa en minimizar el error de

clasificación ponderado, entonces incrementa el peso de las observaciones mal clasificadas y reduce el peso de las observaciones correctamente clasificadas (Freund y Schapire, 1997).

d) *LogitBoost:* trabaja de forma similar al *AdaBoost* sólo que este minimiza la desviación binomial asignando menos peso a las observaciones muy mal clasificadas (observaciones con valores negativos muy grandes) puede dar mejor exactitud promedio con clases pobremente separables. En *LogitBoost* el aprendizaje ajusta un modelo de regresión a los valores de respuesta (Friedman et al., 2000).

e) *GentleBoost:* combina características de *AdaBoost* y *LogitBoost*. Minimiza la pérdida exponencial (como *AdaBoost*) pero su optimización numérica la hace ajustando modelos de regresión como *LogitBoost* (Friedman et al., 2000).

f) *RUSBoost:* es especialmente efectivo para clasificación de clases desbalanceadas. RUS son las siglas de *Random Under Sampling*. El algoritmo toma *N*, el número de miembros en la clase con menos datos de entrenamiento como unidad básica para el muestreo. Las clases con más miembros son sub muestreadas tomando solamente *N* observaciones de cada clase. En otras palabras, si hay *K* clases, entonces, para cada modelo de aprendizaje en el ensamble *RUSBoost* toma un subconjunto de datos con *N* observaciones para cada una de la *K* clases. El proceso *boosting* sigue el mismo procedimiento que el *AdaBoost* para volver a ponderar y construir el ensamble (Seiffert et al., 2008).

2. Número máximo de *splits* es el número máximo de divisiones de decisión o ramas en árboles.
3. Número de modelos de aprendizaje es el número de modelos de aprendizaje débiles usados en el ensamble.
4. Rango de aprendizaje: escalar numérico entre 0 y 1, utilizado para contracción. Elegir un valor para el rango de aprendizaje menor a 1, como por ejemplo 0.1 es una

elección común. Entrenar un ensamble usando un valor de contracción mayor requiere más iteraciones de aprendizaje, pero puede lograr una mejor exactitud.

5. Modelos de aprendizaje son los nombres de los modelos de aprendizaje utilizados para generar el ensamble.

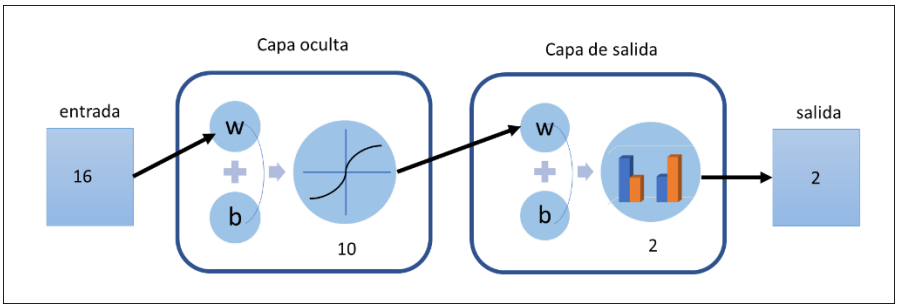
En los siguientes apartados se presentan los resultados del entrenamiento usando los vectores de características "*AUs*". La organización de los datos y los resultados del entrenamiento con los esquemas de clasificación seleccionados se muestran por cada uno de los experimentos realizados.

4.5.1.2 Selección de la Red Neuronal Artificial

El problema de clasificación para las dos clases (*aburrido* e *interesado*) se aborda con una red neuronal de propagación hacia delante de dos capas con una función de transferencia sigmoide (función logística, esta puede tomar cualquier valor de entrada y producir un número entre 0 y 1 sobre una curva en forma de "*S*") en la capa oculta y una función de transferencia *softmax* (que selecciona el valor de la clase considerando la probabilidad más grande) en la capa de salida. El número de neuronas ocultas es de 10. El número de neuronas de salida es de dos, que corresponden a cada una de las categorías de clase. La red se define a partir de la matriz de entrada de *N* vectores de *AUs* con 16 columnas y de la matriz de etiquetas de *N* vectores con dos columnas, una para cada clase. La arquitectura de la red se muestra en la Figura 21.

Figura 21

Arquitectura de la red neuronal artificial



4.5.2 Entrenamiento de Algoritmos de Clasificación

En los siguientes apartados se describe cómo se realizó el proceso de entrenamiento de los algoritmos de clasificación tradicionales, de la red neuronal artificial y de la red neuronal convolucional, los resultados que se obtuvieron y la selección de los modelos que obtuvieron mejores resultados para llevar a cabo la etapa de pruebas de la clasificación de las emociones.

4.5.2.1 Entrenamiento de Algoritmos de Clasificación Tradicionales

El entrenamiento de los algoritmos de clasificación se realizó con validación cruzada de 3, 5 y 10 *folds*, para cada una de estas configuraciones los datos de entrenamiento y prueba se fueron variando entre 60%-40%, 70%-30% y 80%-20% respectivamente, para todos los entrenamientos se utilizaron 16 características previamente seleccionadas.

La optimización de hiperparámetros para cada esquema de clasificación se hizo con el algoritmo de optimización Bayesiana, que intenta minimizar una función *f(x)* escalar para *x* en un dominio acotado. Los elementos del algoritmo son:

- ✓ Un modelo del proceso Gaussiano de *f(x).*
- ✓ Un procedimiento de actualización Bayesiana para modificar el modelo del proceso Gaussiano en cada nueva evaluación de *f(x).*
- ✓ Y una función de adquisición *a(x)* que se maximiza para determinar el siguiente punto de *x* a evaluarse (Gelbart, M. y Snoek, 2014), (Snoek, et al., 2014).

En todas las corridas el algoritmo se ejecutó con una función de adquisición de mejora esperada por segundo plus. La función evalúa la cantidad esperada de mejora en la función objetivo, ignorando los valores que causan un incremento en el objetivo y establece 1 segundo en la ponderación de tiempo para la función de adquisición. Plus significa que la función de adquisición modifica su comportamiento cuando estima que esta sobre explotando un área (Bull, 2011). El algoritmo de optimización se corrió en 100 iteraciones y sin límite de tiempo de entrenamiento.

Se entrenaron los esquemas de clasificación propuestos en la Tabla 14. Se ejecutaron experimentos con SVM, KNN y ensambles corriéndose todas las versiones que Matlab® tiene implementadas para cada algoritmo (6 versiones diferentes para SVM, 6 para KNN y 5 para ensambles). Posteriormente se ejecutó el proceso de optimización Bayesiana para cada algoritmo. Se analizaron los resultados de todas las pruebas de entrenamiento realizadas (180) y de las 54 mejores se seleccionaron 9, las que dieron mejores resultados para exactitud. En la Tabla 15 se resumen los datos de los entrenamientos que obtuvieron mejores resultados.

El esquema de clasificación con mayor exactitud en el entrenamiento es el de ensambles con 91.80%, le sigue KNN con 91.10% y SVM con 91.00%. Las mejores configuraciones para el entrenamiento son con 70% de datos de entrenamiento y 30% de prueba con 5 y 10 *folds*. En la Tabla 15 también se observa que los entrenamientos con 60 y 40 por ciento para datos de entrenamiento y prueba y 5 *folds* obtienen los siguientes mejores resultados con una *exactitud* de 91.50%.

Con estas nueve mejores configuraciones de entrenamiento se generaron los modelos de clasificación para validarlos con los datos de prueba.

Tabla 15

Algoritmos de clasificación con mejores resultados de entrenamiento

#	Clasificador	Entrenamiento / Prueba (%)	Datos de Entrenamiento	*Folds*	Exactitud de Entrenamiento
1	SVM (Optimizada)	70/30	54830	5	90.80%
2	KNN *Weighted*	70/30	54830	5	90.80%
3	**Ensamble (Optimizado)**	**70/30**	**54830**	**5**	**91.80%**
4	SVM Gaussiana Fina	70/30	54830	10	90.40%
5	KNN (Optimizada)	70/30	54830	10	91.10%
6	**Ensamble (Optimizado)**	**70/30**	**54830**	**10**	**91.80%**
7	SVM (Optimizada)	60/40	46997	5	91.00%
8	KNN (Optimizada)	60/40	46997	5	90.50%

9	**Ensamble (Optimizado)**	**60/40**	**46997**	**5**	**91.50%**

4.5.2.2 Entrenamiento de la Red Neuronal Artificial

Para el entrenamiento de la red neuronal, los datos para entrenamiento, validación y prueba se fueron variando entre 60%-20%-20%, 70%-15%-15%, 80%-10%-10% y 90%-5%-5% respectivamente. Se utilizaron los mismos vectores de 16 características. La red se entrenó con el algoritmo de gradiente de conjugación escalado de propagación hacia atrás (*backpropagation*). El conjunto de entrenamiento se divide en tres subconjuntos:

1. Datos de entrenamiento: usado para calcular el gradiente y actualizar los pesos de la red y sesgos.
2. Datos de validación: el error sobre el conjunto de datos de validación es monitoreado durante el proceso de entrenamiento. Se usa para validar que la red esté generalizando y pare el entrenamiento antes de que suceda un sobreajuste. El error de validación normalmente disminuye durante la fase inicial de entrenamiento al igual que el error del conjunto de entrenamiento. Sin embargo, cuando la red comienza a sobre ajustar los datos, el error en el conjunto de validación generalmente comienza a aumentar. Los pesos y sesgos de la red deben mantenerse en el mínimo del error del conjunto de validación.
3. Datos de prueba: son usados para una prueba completamente independiente de generalización de la red. El error del conjunto de prueba no es usado durante el entrenamiento, pero es usado para comparar diferentes modelos de clasificación. También es útil para trazar el error del conjunto de prueba durante el proceso de entrenamiento. Si el error en el conjunto de prueba alcanza un mínimo en un número de iteración significativamente diferente al error del conjunto de validación, esto podría indicar una división deficiente del conjunto de datos.

La división de los datos se hace usando índices aleatorios. La ejecución de la red neuronal en la validación se evaluó con el aumento del error de entropía cruzada. Se corrieron

diferentes configuraciones para el entrenamiento, variando el número de capas ocultas y cambiando las proporciones para datos de entrenamiento, validación y prueba. Finamente se seleccionan aquellas con mejores valores para las métricas de evaluación. Los resultados obtenidos para exactitud general, precisión para la emoción de *aburrido* y para *interesado* en la etapa de entrenamiento de la red neuronal se muestran en la Tabla 16. El mejor valor para exactitud de 84% se logra con las proporciones de 70-15-15, 80-10-10 y 90-5-5 para datos de entrenamiento, validación y prueba de la red. La mejor precisión de entrenamiento para la emoción aburrido es de 72% y para la emoción de interesado de 85.20%. Los dos modelos con mejores resultados serán utilizados en la etapa de pruebas para la clasificación.

Tabla 16

Resultados del entrenamiento de la red neuronal artificial

#	Datos de entrenamiento/ validación/prueba (%)	Datos de entrenamiento validación /prueba	Exactitud de Entrenamiento	Precisión (*aburrido*)	Precisión (*interesado*)
1	60-20-20	37598 / 12532 / 12532	83%	64%	84%
2	70-15-15	43864 / 9399 / 9399	**84%**	**72%**	85%
3	80-10-10	50130 / 6266 / 6266	**84%**	70%	85%
4	90-5-5	56396 / 3133 / 3133	**84%**	67%	**85.20%**

4.5.3 Clasificación de Emociones Centradas en el Aprendizaje.

En este punto se describe como se llevó a cabo el proceso de clasificación de ECA con los algoritmos de clasificación tradicionales que fueron seleccionados y con la configuración de la red neuronal que obtuvo mejores resultados.

4.5.3.1 Clasificación con los Algoritmos de Clasificación Tradicionales

En la Tabla 17 se muestran las configuraciones de los modelos de clasificación seleccionados para reconocer las ECA con los datos de prueba.

Se realizan las pruebas de clasificación con los 9 modelos seleccionados en el proceso de entrenamiento. Los resultados obtenidos son validados en la siguiente etapa y se presentan en el capítulo 7.

Tabla 17

Modelos de clasificación para reconocimiento de las ECA

#	Clasificador	Entrenamiento / prueba (%)	Datos de prueba	*Folds*	Configuración de hiperparámetros
1	SVM (Optimizada)	70/30	23498	5	Función *Kernel*: Gaussiana Escala *Kernel*: 0.35499 Nivel límite de caja: 111.2311 Datos estandarizados: falso Iteración: 11
2	KNN *Weighted*	70/30	23498	5	Número de vecinos: 10 Métrica de distancia: Euclidiana Peso de la distancia: Cuadrado inverso Datos estandarizados: verdadero Iteración: 5
3	Ensamble (Optimizado)	70/30	23498	5	Método de ensamble: *GentleBoost* Número máximo de *splits*: 53004 Número de modelos de aprendizaje: 20 Rango de aprendizaje: 0.071038 Iteración: 78
4	SVM Gaussiana	70/30	23498	10	Función *Kernel*: Gaussiana Escala *Kernel*: 1 Nivel límite de caja: 1 Método multiclase: 1 vs 1 Datos estandarizados: verdadero Iteración: 67
5	KNN (Optimizada)	70/30	23498	10	Número de vecinos: 3. Métrica de distancia: *City block* Peso de la distancia: Inversa. Datos estandarizados: verdadero. Iteración: 7
6	Ensamble (Optimizado)	70/30	23498	10	Método de ensamble: *GentleBoost* Número máximo de *splits*: 41986 Número de modelos de aprendizaje: 40 Rango de aprendizaje: 0.0043161 Iteración: 9
7	SVM (Optimizada)	60/40	31331	5	Función *Kernel*: Gaussiana Escala *Kernel*: 0.8483 Nivel límite de caja: 3.5126 Método multiclase:1 vs 1 Datos estandarizados: verdadero Iteración: 15
8	KNN (Optimizada)	60/40	31331	5	Número de vecinos: 32 Métrica de distancia: Correlación Peso de la distancia: Cuadrado inverso Datos estandarizados: verdadero. Iteración: 8
9	Ensamble (Optimizado)	60/40	31331	5	Método de ensamble: *LogitBoost* Número máximo de *splits*: 2543

	Número de modelos de aprendizaje: 62 Rango de aprendizaje: 0.81805 Iteración: 28

4.5.3.2 Clasificación con la Red Neuronal Artificial

Las configuraciones con las que se corren las pruebas con la red neuronal se muestran en la Tabla 18. En ésta se observa que para la etapa de entrenamiento se usaron el 80% de los datos, el 20% restante son los que se usan para la etapa de pruebas en la clasificación de la ECA. Se corrieron las pruebas con las dos configuraciones y los resultados obtenidos son presentados en el siguiente capítulo.

Tabla 18

Configuraciones de la red neuronal artificial para la clasificación de ECA

#	Entrenamiento y Prueba (%)	Datos entrenamiento y prueba	Capa oculta	Capa de salida	Algoritmos		
					División de datos	Entrenamiento	Evaluación de la ejecución
1	80% (70-15-15) y 20%	62633 / 15665	Función de transferencia sigmoide Número de neuronas: 10	Función de transferencia *softmax* Número de neuronas: 2	Aleatorio	Gradiente de conjugación escalado de propagación hacia atrás	Error de entropía cruzada
2	80% (90-5-5) y 20%						

Capítulo 5. Etapa 4: Validación y Análisis de los Resultados

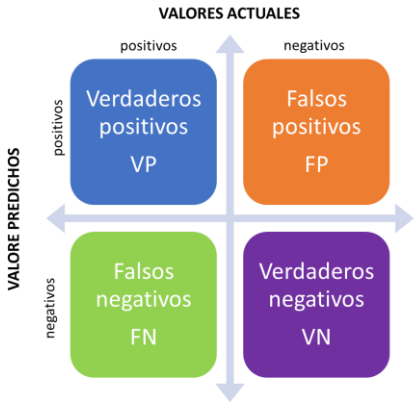
Para la validación de los resultados se utilizan las métricas de evaluación como exactitud, precisión y los datos de la matriz de confusión. El análisis se hace considerando los modelos con los cuatro mejores resultados. Se hace una interpretación de ellos y finalmente se realiza una discusión de los alcances.

5.1 Métricas de Evaluación

1. **Matriz de confusión** es un método de evaluación de rendimiento de un modelo de clasificación. La matriz compara los valores reales con los predichos por el modelo de aprendizaje. Esto ayuda a ver qué tan bien está funcionando el modelo. Su forma general se puede ver en la Figura 22. La variable objetivo tiene dos valores: positivo o negativo. Las columnas representan los valores reales de la variable objetivo. Las filas representan los valores predichos de la variable objetivo (Acevedo, 2020).

Figura 22

Matriz de confusión



Cada celda de la matriz representa aciertos y errores del modelo de clasificación con los siguientes significados:

Verdadero positivo (VP o TP "*True Positive"*)

Significa que el valor predicho coincide con el valor real ya que el valor real es positivo y el modelo predijo un valor positivo (predijo positivo y es verdad).

Verdadero negativo (VN o TN *"True Negative"***)**

Significa que el valor predicho coincide con el valor real pero el valor real es negativo y el modelo predijo un valor negativo (predijo negativo y es verdad).

Falso positivo (FP o *False Positive***): error de tipo 1**

Significa que el valor predicho no coincide con el valor real porque éste es negativo pero el modelo predijo un valor positivo. Se le conoce como error tipo 1 (predijo positivo, pero es falso).

Falso negativo (FN o *False Negative***): error de tipo 2**

Significa que el valor predicho no coincide con el valor real porque éste es positivo pero el modelo predijo un valor negativo. También conocido como error tipo 2 (predijo negativo, pero es falso).

A partir de los valores de la matriz de confusión se obtiene diferentes métricas que sirven para evaluar mejor la matriz y minimizar los falsos negativos contra los falsos positivos. En los siguientes puntos se resumen estas métricas.

2. **Exactitud** (*accuracy*) es básicamente el número total de predicciones correctas dividido por el número total de predicciones, es decir, de todas las muestras cuántas se predijeron correctamente. A partir de los valores de la matriz de confusión la exactitud, se calcula con la ecuación 1.

$$exactitud = \frac{VP+VN}{Total\ de\ muestras}$$ **Ecuación 1**

3. **Precisión** de una clase define cuán confiable es un modelo en responder si una muestra pertenece a esa clase. De todas las muestras positivas que se han predicho correctamente cuántas de ellas son realmente positivas. A partir de las variables de la matriz de confusión la precisión, se calcula con la ecuación 2.

$$precisión = \frac{VP}{VP+FP}$$ **Ecuación 2**

4. **Sensibilidad** o exhaustividad (*recall*) determina de todas las muestras positivas cuántas se predijeron correctamente. A partir de los valores de la matriz de confusión, la sensibilidad se calcula con la ecuación 3.

$$sensibilidad = \frac{VP}{VP+FN}$$ **Ecuación 3**

5. **Medida F** permite comparar dos modelos de baja precisión y alta exhaustividad (*recall*), utiliza la media armónica para castigar los valores extremos**.** A partir de los valores de la matriz de confusión la medida F, se calcula con la ecuación 4.

$$medida\ F = \frac{2*recall*precisión}{recall+precisión}$$ **Ecuación 4**

6. **Curva ROC** (acrónimo de *Receiver Operating Characteristic*)*,* más que una métrica, es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación. Otra interpretación de este gráfico es la representación de la razón o proporción de verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la razón o proporción de falsos positivos (FPR = Razón de Falsos Positivos) también según se varía el umbral de discriminación, valor a partir del cual se decide que un caso es un positivo (Fogarty et al., 2005).

De estas métricas se utilizaron exactitud y precisión para evaluar los resultados obtenidos ya que son las más utilizadas por los trabajos analizados en el estado del arte. También se calcula especificidad y sensibilidad para graficar la curva ROC. Todas las métricas son obtenidas de la matriz de confusión.

5.2 Evaluación de Resultados

En la Tabla 19 se muestran los resultados alcanzados para las métricas exactitud y precisión por clase para cada una de las pruebas ejecutadas con los modelos de clasificación tradicionales. Los resultados de la etapa de pruebas para la clasificación de ECA con la red neuronal con las dos configuraciones seleccionadas se muestran en la Tabla 20.

En la Figura 23 se pueden identificar visualmente los 3 mejores resultados de clasificación para las emociones de aburrido e interesado. El conjunto de métricas con mejor evaluación para los modelos de clasificación seleccionados son las de las redes neuronales artificiales. Estas logran una exactitud general de 84%, una precisión para la emoción de aburrido de 69% y una precisión para la emoción de interesado de 86%. A continuación, le sigue el ensamble de árboles, este modelo de clasificación alcanza una exactitud de 79%, una precisión para la emoción de aburrido de 36% y una precisión para la emoción de interesado de 83%. Después está el modelo KNN (fila 8, de la Tabla 19), con una exactitud de 76.4% y una segunda mejor precisión para la clase interesado de 83.5%.

Tabla 19

Resultados de los modelos de clasificación evaluados

#	Clasificador	Exactitud	Aburrido Precisión	Interesado Precisión
1	SVM (Optimizada)	70.43%	19.83%	81.69%
2	KNN *Weighted*	69.19%	18.39%	81.36%
3	Ensamble (Optimizado)	75.15%	23.71%	82.11%
4	SVM Gaussiana	74.61%	18.22%	81.37%
5	KNN (Optimizada)	67.39%	19.59%	81.71%
6	Ensamble (Optimizado)	75.65%	24.58%	82.18%
7	SVM (Optimizada)	76.24%	25.61%	82.24%
8	KNN (Optimizada)	76.40%	31.66%	**83.48%**
9	Ensamble (Optimizado)	**79.03%**	**35.77%**	82.98%

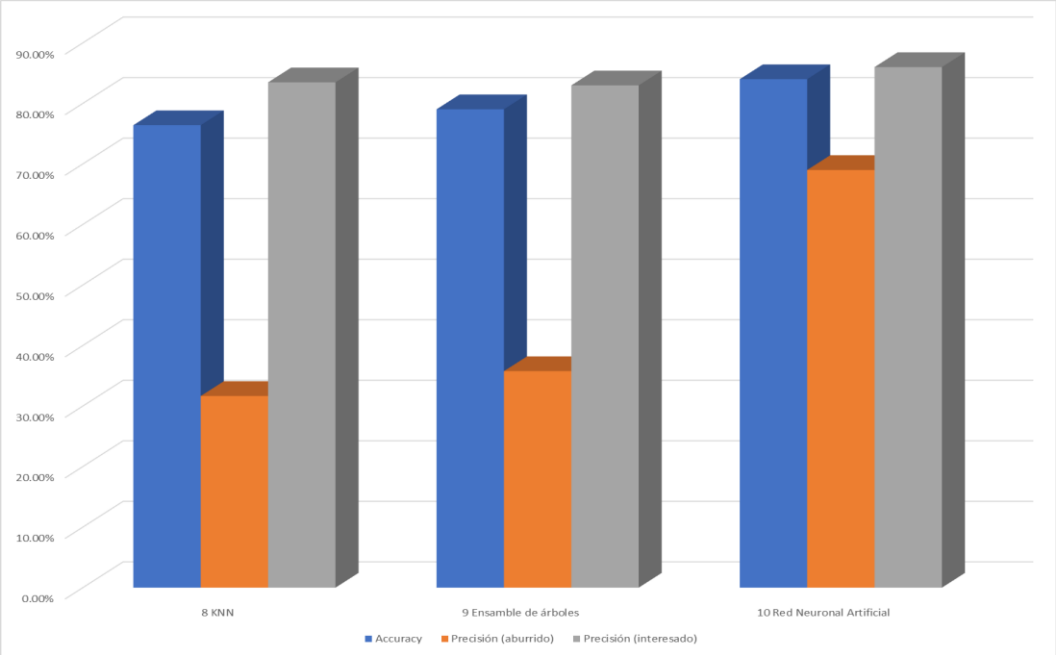
Tabla 20

Resultados obtenidos por la red neuronal artificial en la clasificación de ECA

#	**Datos de prueba (%)**	**Datos de prueba**	**Exactitud**	**Precisión (*aburrido*)**	**Precisión (*interesado*)**
1	20	15665	**84%**	**69%**	85%
2	20	15665	**84%**	65%	**86%**

Figura 23

Métricas de los tres mejores modelos de clasificación evaluados



La matriz de confusión para la red neuronal artificial que logra mejores resultados se muestra en la Figura 24, donde la clase positiva es aburrido y la clase negativa es interesado. Se alcanza una tasa de VP=4.7%, de FN=13.4, de FP=2.6% y de VN=79.3%.

La matriz de confusión para el modelo de ensambles de árboles se observa en la Figura 25. Se obtienen los porcentajes de VP=3%, de FN=16%, de FP=5% y VN=76%. La matriz de confusión para el modelo KNN se muestra en la Figura 26. Con los datos de la matriz de cada modelo se calculan la razón de verdaderos negativos o especificidad de acuerdo con la ecuación 5, y la razón de verdaderos positivos o sensibilidad *(recall*) de acuerdo con la ecuación 6 (Hastie, et al., 2009). En la

Tabla 21 se muestra un resumen de estas métricas que serán utilizadas más adelante en la interpretación de los resultados.

$$especificidad = \frac{VN}{VN+FP} \qquad \textbf{Ecuación 5}$$

$$sensibilidad = \frac{VP}{VP+FN}$$ Ecuación 6

Figura 24

Matriz de confusión para la red neuronal artificial

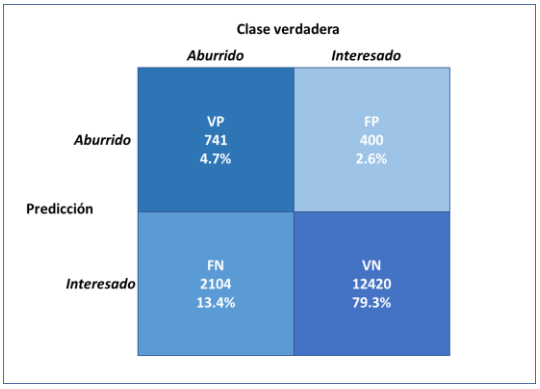


Figura 25

Matriz de confusión para el modelo de ensamble de árboles

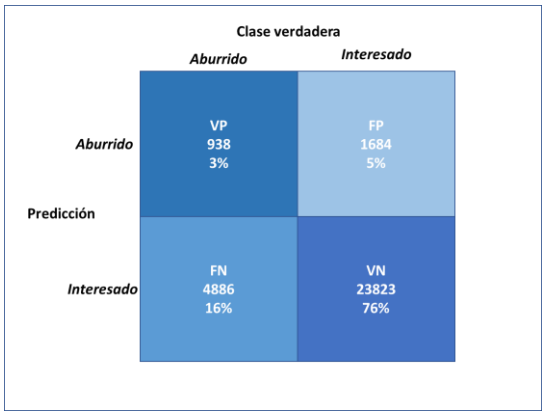


Figura 26

Matriz de confusión para el modelo de KNN

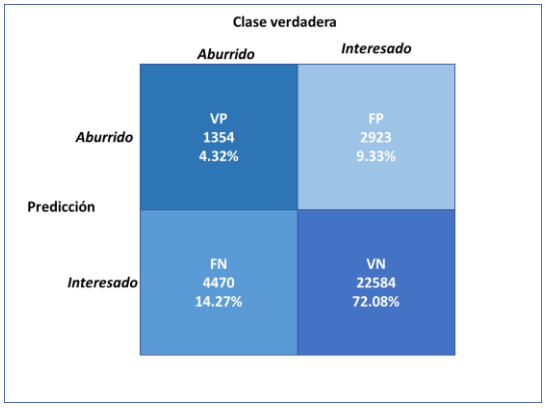


Tabla 21

Resumen de los resultados para las principales métricas de evaluación

#	MODELO DE CLASIFICACIÓN	EXACTITUD	PRECISIÓN ABURRIDO	PRECISIÓN INTERESADO	ESPECIFICIDAD (razón de VN)	SENSIBILIDAD (razón de VP)
1	Red neuronal	84%	69%	86%	26%	97%
2	Ensamble de árboles	79%	36%	83%	16%	93%
3	KNN	76%	32%	83.5%	23%	89%

5.3 Interpretación y Discusión de los Resultados

Los modelos de clasificación con mejores resultados fueron la red neuronal entrenada con 80% de los datos y probada con el 20%. El segundo y tercer mejor resultado se obtiene de los modelos de ensamble de árboles y KNN entrenados con un conjunto de datos del 60% y probados con el 40% de datos y con una validación cruzada de 5 *folds*. Para todas las pruebas la clase positiva fue aburrido y la clase negativa fue interesado. Los resultados obtenidos y su interpretación para la red neuronal son los siguientes:

- Exactitud = 84% representa qué tan frecuentemente es correcta la red neuronal.
- Precisión(aburrido)=69% indica que si predice aburrido qué tan frecuentemente esto es correcto. El segundo mejor resultado se obtiene del ensamble de árboles con precisión=36%.

- Precisión(interesado)=86% indica que si predice interesado qué tan frecuentemente esto es correcto. El segundo mejor resultado se obtiene del modelo KNN con precisión=83.5%.
- Sensibilidad (*recall*) = 26% indica qué tan frecuentemente se clasifica de manera correcta la clase positiva (aburrido) con la red neuronal.
- Especificidad = 97% indica qué tan frecuentemente se clasifica de manera correcta la clase negativa (interesado) con la red neuronal.

Con estas razones de evaluación se observa que, aunque el modelo del algoritmo KNN está en el último lugar de los resultados, clasifica de manera correcta la clase aburrido (23%) casi con la misma proporción como lo hace la red neuronal en un 26%, pero obtiene la menor razón de verdaderos positivos (clasificación correcta de la clase interesado) de 89%.

Las métricas de evaluación de la red neuronal son los mejores resultados que se logran indicando un buen reconocimiento de la emoción interesado, alcanzando un 86% de precisión con una especificidad del 97%. Para la emoción aburrido la mejor precisión es de 69%, una sensibilidad del 26% y una exactitud general=84%.

En la Figura 27 se puede observar gráficamente la relación de las razones de especificidad-sensibilidad (considerando como clase positiva interesado) para la red neuronal. En la Figura 28 se muestra la gráfica para el modelo de ensamble de árboles. En ambos casos se observa una razón de verdaderos positivos relevante y una razón de verdaderos negativos bajo indicador de una buena clasificación para la clase interesado e inferior para la clase aburrido.

Figura 27

Gráfica de relación especificidad-sensibilidad para la red neuronal

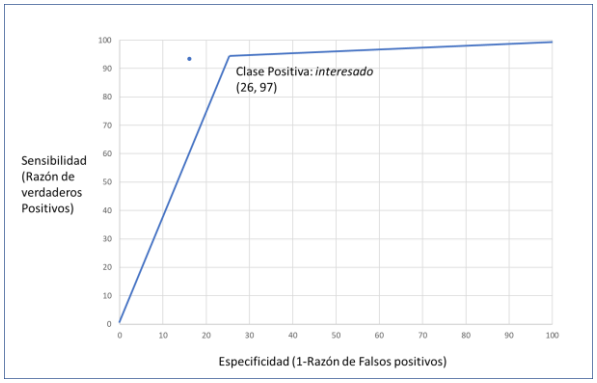
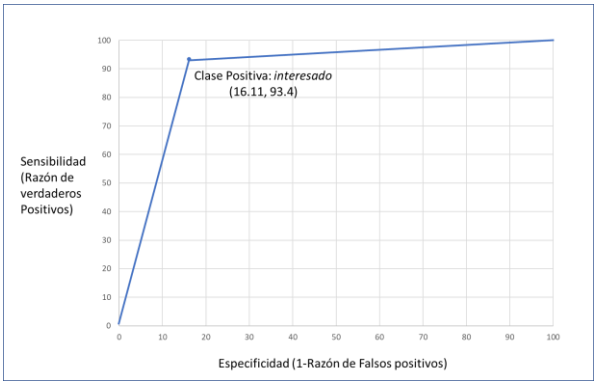


Figura 28

Gráfica de relación especificidad-sensibilidad para el modelo de ensamble de árboles



5.4 Comparación de Resultados con Otros Trabajos

No es sencillo hacer una comparación directa con los trabajos relacionados, ya que existen varias diferencias, por ejemplo, en las emociones reconocidas, en la forma de medir las emociones, en la manera de adquirir los datos, en el tipo de dispositivos para adquirirlos, en el tipo de datos, en la forma de etiquetarlos, en las características seleccionadas, en los modelos de clasificación utilizados y en las métricas de evaluación con las que cada trabajo justifica sus resultados. En la Tabla 22 se muestra una comparación con los resultados

obtenidos por los trabajos más recientes listados en el estado del arte. Se observa que la exactitud general de 84% alcanzada por la red neuronal de este trabajo sólo es superada por el trabajo de González-Hernández et al. (2017) en un 4% al reconocer cuatro ECA con una CNN (utilizando el dispositivo *Emotiv Epoc* para etiquetamiento). Con este trabajo no es posible hacer una comparación por emoción ya que no presenta métricas individuales.

Por lo que respecta a la identificación de la emoción interesado la mejor precisión la obtienen la red neuronal con 86% y el modelo KNN con 83.5% evaluadas en este trabajo. Ambos porcentajes están por arriba del resultado del mejor trabajo analizado de Nezami et al. (2020) que con una CNN logra una precisión para la emoción de interesado de 72.38%, única emoción reconocida. En cuanto a la identificación de la emoción aburrido la mejor precisión es de 69% alcanzada por la red neuronal de este trabajo arriba del 64% alcanzado por el trabajo de Bosch, et al (2016a) con el modelo de K-*means* y del 53.7% del trabajo de Gupta et al. (2018) quienes utilizan una red convolucional recurrente de termino-largo.

En cuanto a aspectos generales por analizar se observa primero, que la mayoría de las bases son creadas a partir de una cantidad no muy grande de alumnos (de 8 hasta 137) con un máximo de 9068 imágenes a diferencia, por ejemplo, de las 14,560 que en este trabajo se generan para la emoción aburrido y de 63,768 para la emoción de interesado (con la participación de 86 estudiantes).

En segundo lugar, el etiquetado de las imágenes en la mayoría de los trabajos es realizado por observadores humanos bajo la clasificación discreta, solo en esta propuesta se hace bajo el FACS con lo que se genera un modelo de *AUs* asociadas a la emoción de aburrido e interesado. En otros trabajos las imágenes son etiquetadas usando el software de identificación de emociones básicas del dispositivo *Emotiv Epoc* y *Emotiv Insight*.

En tercer lugar, en cuanto a la selección de características sólo el trabajo de Bosch, et al, (2016a) utiliza *AUs* de la misma manera que en este trabajo, pero se basan en las *AUs* de las emociones básicas. En el de Zatarain-Cabada, et al. (2017a) utilizan los descriptores

generados por LBP y el resto de los trabajos emplean filtros convolucionales sobre las imágenes ya que sus propuestas son desarrolladas con algún tipo de CNN.

Tabla 22

Comparación de resultados con trabajos relacionados

Modelo de Clasificación	Base de Datos		Etiquetado	Características	Exactitud/ precisión
Red neuronal artificial	Base de datos propia 14,560 imágenes aburrido 63,768 imágenes de interesado	Aburrido, interesado	Observadores humanos de acuerdo con FACS.	Membresía de las *AUs* 86 estudiantes	84% exactitud 86% Precisión interesado 69% Precisión aburrido
Ensamble de árboles	Base de datos propia 14,560 imágenes aburrido 63,768 imágenes de interesado	Aburrido, interesado	Observadores humanos de acuerdo con FACS.	Membresía de las *AUs* 86 estudiantes	79% exactitud 83% Precisión interesado 36% Precisión aburrido
KNN	Base de datos propia 14,560 imágenes aburrido 63,768 imágenes de interesado	Aburrido, interesado	Observadores humanos de acuerdo con FACS.	Membresía de las *AUs* 86 estudiantes	83.5% Precisión interesado
CNN (Nezami et al., 2020)	Base de datos propia *ER* 2290 imágenes interesado 2337 imágenes no interesado	Interesado	Estudiantes de psicología	Imagen (filtros convolucionales) 22 estudiantes	72.38% exactitud
LRCN (Red convolucional recurrente de termino-largo) (Gupta et al., 2018)	Base de datos propia *DAiSEE* 9068 imágenes de aburrido 9068 imágenes de interesado	Aburrido, interesado, confundido, frustrado	Estudiantes de psicología	Imagen 112 estudiantes	53.7% exactitud aburrido 57.9% exactitud interesado
CNN (González-Hernández et al., 2017)	Base de datos propia (*Emotiv Epoc*) 17 imágenes de aburrido 519 imágenes de interesado	Aburrido, interesado, excitado, frustrado	Sistema del dispositivo Emotiv y validadas por expertos humanos	Imágenes (filtros convolucionales) 8 estudiantes	88% exactitud promedio
CNN (González-Hernández et al.2017)	Base de datos propia (*Emotiv Insight*) 1040 imágenes de aburrido	Aburrido, interesado, excitado, enfocado, relajado, enganchado	Sistema del dispositivo Emotiv y validadas por expertos humanos	Imagen (filtros convolucionales) 8 estudiantes	74% exactitud promedio

	150 imágenes de interesado				
Reconocedor de apariencia basado en LBP (Zatarain-Cabada, Barrón-Estrada, et al., 2017)	Base de datos propia 13 imágenes de aburrido 56 imágenes de interesado	Aburrido, enganchado, excitado, frustrado	Sistema del dispositivo Emotiv y validadas por expertos humanos	256 características (descriptores generados por LBP) 8 estudiantes	80% exactitud
K *means* / *Bayes net* (Bosch, D'mello, et al., 2016a)	Base de datos propia 1305 instancias de aburrido 1228 instancias de interesado	Interés, aburrimiento, frustración. deleite, confusión	Observadores humanos entrenados en BROMP	Probabilidad de AUs (con el software FACET) 137 estudiantes	64% exactitud aburrido (K-*means*) 64% exactitud interesado (*Bayes net*)

Capítulo 6. Conclusiones y Trabajo a Futuro

En este capítulo se hace una descripción de los alcances y objetivos logrados, las limitaciones enfrentadas y los resultados obtenidos. En cuanto a los alcances se desarrolló una base de datos con capturas de 2 dispositivo de adquisición de señales fisiológicas, el ritmo cardiaco y la temperatura. Se obtuvieron datos del comportamiento de 82 alumnos a través de grabaciones video. Con esto se cumplió con el objetivo de crear una base de datos de información fisiológica y de comportamiento obtenida de alumnos aprendiendo en ambientes en línea reales.

Se planteó la metodología para la identificación de emociones con el objetivo de guiar las etapas necesarias para el proceso de clasificación. Como parte de la metodología se desarrolló un protocolo formal para la captura de datos de alumnos realizando actividades de aprendizaje en un entorno educativo, la cual es una propuesta que puede guiar la captura de datos en diferentes entornos y con el uso de diferentes dispositivos.

El análisis de los datos capturados permitió seleccionar el video como mejor entrada para entrenar algoritmos de clasificación, concluyendo que la fusión del resto de los datos requiere de un proceso adicional que ocupará más tiempo por lo que se considera como trabajo futuro.

Se probaron y evaluaron los algoritmos de aprendizaje computacional (SVM, KNN, ensambles de árboles binarios y redes neuronales artificiales) para el reconocimiento de las emociones interesado y aburrido utilizando grados de membresía de *AUs* obtenidos de las imágenes de video haciendo uso del modelo de inferencia *fuzzy* propuesto por Vargas (2017). Se seleccionan y proponen como características para el proceso de identificación de ECA 16 *AUs* del modelo de codificación de acción facial propuesto por Paul Ekman y etiquetadas por observadores humanos para cada una de las dos emociones. Con el proceso de entrenamiento se probaron modelos de clasificación de los cuales se seleccionaron los que arrojaron mejores resultados para realizar el proceso de pruebas y validación. En esta última

etapa se obtuvo una exactitud de 84% con la red neuronal, valor no muy lejano al obtenido por una CNN con una exactitud de 88% (usando *Emotiv Epoc*, clasificando 6 ECA) y del 74% (usando *Emotiv Insight*, clasificando 4 ECA) pero con bases de datos pequeñas y sin reportar precisiones por emoción lo que no permite realizar un análisis más detallado. En el análisis de emociones de manera individual se obtiene una buena precisión para el reconocimiento de la emoción interesado con un 86% con la red neuronal y del 83.5% con el modelo de clasificación KNN muy por arriba del mejor trabajo analizado que alcanza el 64%. También se logra una buena precisión de reconocimiento para la emoción de aburrido de 69% con la red neuronal, arriba del 64% obtenido por el mejor de los trabajos analizados con un modelo de *Bayes net*. Con esto se cumplió el objetivo de clasificar las ECA de interesado y aburrido utilizando algoritmos de aprendizaje computacional.

Esta información se resume en la Tabla 23 donde se observa la matriz de concordancia de los objetivos específicos, describiendo cada una de sus variables asociadas de una manera sintética.

Tabla 23

Matriz de concordancia

	Objetivo 1	**Objetivo 2**	**Objetivo 3**
Descripción	Identificar algoritmos de aprendizaje computacional para la selección de características y para la clasificación de emociones.	Diseñar e implementar una metodología para el reconocimiento automático de ECA a partir del uso de las tecnologías de adquisición de datos propuestas y de la evaluación de algoritmos de selección de características y de clasificación.	Probar y validar la exactitud de reconocimiento con métricas que permitan comparar los resultados con los de trabajos relacionados.
Pregunta específica	¿Qué algoritmos de aprendizaje computacional se identificaron para la selección de características y para la clasificación de emociones?	¿Qué etapas es necesario definir en la metodología para el reconocimiento automático de emociones centradas en el aprendizaje haciendo uso de tecnologías de adquisición de datos y de algoritmos de selección de características y de clasificación?	¿Qué métricas permiten validar la exactitud de reconocimiento y comparar los resultados con los de trabajos relacionados?

Concepto (s)	Algoritmos de aprendizaje computacional, selección de características, clasificación de emociones.	Reconocimiento automático de emociones, emociones centradas en el aprendizaje, tecnologías de adquisición de datos, algoritmos de selección de características, algoritmos de clasificación.	Exactitud, Métricas
Teoría (s)	Aprendizaje computacional	Metodología	Métricas
Método	Cualitativo / cuantitativo	Cualitativo	Cuantitativo
Resultados	Para la selección de características se procesaron las imágenes faciales y se usó el modelo de inferencia *fuzzy* de (Vargas, 2017) para obtener valores de membresía para las *AUs* definidas para cada emoción. Para clasificación de emociones se analizaron, probaron y se seleccionaron los modelos que obtuvieron mejores resultados: KNN, ensamble de árboles, SVM y redes neuronales artificiales.	Se propone una metodología para el reconocimiento automático de ECA a partir del uso de las tecnologías de adquisición de datos: cámara de video, cámara térmica y sensor de ritmo cardiaco para la creación de la base de datos. Y se identifican algoritmos de selección de características (16 *AUs*) y de clasificación (KNN, ensamble de árboles, SVM y redes neuronales artificiales).	Los resultados obtenidos por los modelos de clasificación se evalúan con las métricas obtenidas de la matriz de confusión de: exactitud y precisión. Obteniendo los siguientes mejores resultados. Exactitud general= 84% con la red neuronal. Precisión para la emoción interesado= 86% con la red neuronal. Precisión para la emoción de aburrido= 69% con la red neuronal. Estos últimos arriba de lo encontrado en el estado del arte.

6.1 Trabajos a Futuro

En cuanto al reto de identificar la emoción de aburrido, la diferencia en cantidad de datos respecto a la emoción de interesado parece marcar la deficiencia en el reconocimiento de la emoción aburrido. Una limitación importante que influye en el número de datos capturados de “aburrido” es la complejidad de detectar acciones faciales que representen esta emoción. La limitante se podría subsanar en un trabajo futuro generando entornos de aprendizaje que propicien el aburrimiento de manera intencional, así como utilizar la metodología propuesta en entornos que provoquen de manera deliberada cada una de las otras ECA, lo que evitaría el trabajo de etiquetamiento humano y sobre todo evaluaciones subjetivas en la interpretación de las emociones.

Otra propuesta es la experimentación con técnicas de clasificación computacional como aprendizaje profundo, probando otros modelos y configuraciones. También fusionar las características geométricas del rostro y los vectores de características que las agrupan.

Otra propuesta es trabajar con la integración de todos los datos capturados, provenientes de diferentes fuentes o procesarlos de manera individual como se hace actualmente para el caso de las imágenes de video.

También se propone como trabajo futuro colaborar con desarrolladores de tutores inteligentes utilizando el modelo que arroja mejores resultados en esta investigación. El objetivo es proporcionar información del estado emocional de los estudiantes que permita generar una retroalimentación en el proceso de aprendizaje de cada alumno. Una vez obtenido esto, es posible plantear un trabajo donde de manera automática se presente un análisis de la relación entre cada estado emocional del alumno con su nivel de aprendizaje, considerando las transiciones de sus emociones dentro de este proceso.

Referencias Bibliográficas

Acevedo, N. (2020). Matriz de confusión en Machine Learning. Explicado paso a paso. Retrieved June 10, 2021, from https://nataliaacevedo.com/matriz-de-confusion-en-machine-learning-explicado-paso-a-paso/

Aifanti, N., Papachristou, C., & Delopoulos, A. (n.d.). The MUG facial expression database. *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, 1–4. https://doi.org/10.1371/journal.pone.0009715

Almohammadi, K., Hagras, H., Yao, B., Alzahrani, A., Alghazzawi, D., & Aldabbagh, G. (2017). A type-2 fuzzy logic recommendation system for adaptive teaching. *Soft Computing*, *21*(4), 965–979. https://doi.org/10.1007/s00500-015-1826-y

Aneja, D., Colburn, A., Faigin, G., Shapiro, L., & Mones, B. (2016). Modeling Stylized Character Expressions via Deep Learning. *Asian Conference on Computer Vision. Springer*, *1*, 136–153.

Arana-Llanes, J. Y., González-Serna, G., Pineda-Tapia, R., Olivares-Peregrino, V., Ricarte-Trives, J. J., & Latorre-Postigo, J. M. (2017). EEG lecture on recommended activities for the induction of attention and concentration men tal states on e-learning students. *Journal of Intelligent & Fuzzy Systems*.

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009). Emotion sensors go to school. *Frontiers in Artificial Intelligence and Applications*, *200*(1), 17–24. https://doi.org/10.3233/978-1-60750-028-5-17

Barrón-Estrada, M. L., Zatarain-Cabada, R., Aispuro-Medina, B. G., Valencia-Rodríguez, E. M., & Lara-Barrera, A. C. (2016). Building a Corpus of Facial Expressions for Learning-Centered Emotions. In *Research in Computing Science.* (Vol. 129, pp. 45–52). México.

Barrón, M. L., Zatarain, R., & Hernández, Y. (2014). Intelligent Tutor with Emotion Recognition and Student Emotion Management for Math Performance. *Revista Electrónica de Investigación Educativa*, *16*, 88–102.

Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2016). EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 5562–5570. https://doi.org/10.1109/CVPR.2016.600

Berger, M. (2006). A panoramic view of riemannian geometry. *The Mathematical Intelligencer*, *28*(2), 73–74. https://doi.org/10.1007/bf02987161

Bixel, R., & D´Mello, S. (2013). Towards Automated Detection and Regulation of Affective States During Academic Writing. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *LNAI 7926 - Artificial Intelligence in Education.* (pp. 904–907). Heidelberg Dordrecht London NewYork: Lecture Notes in Artificial Intelligence Series Editors. Springer.

Bosch, N., & D'Mello, S. (2015). The Affective Experience of Novice Computer Programmers. *International Journal of Artificial Intelligence in Education*, *27*(1), 181–206. https://doi.org/10.1007/s40593-015-0069-5

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., … Zhao, W. (2015). Automatic Detection of Learning-Centered Affective States in the Wild. *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, 379–388. https://doi.org/10.1145/2678025.2701397

Bosch, N., D'Mello, S. K., Baker, R. S., Ocumpaugh, J., Shute, V., Ventura, M., … Zhao, W. (2016a). Detecting student emotions in computer-enabled classrooms. *IJCAI International Joint Conference on Artificial Intelligence*, *2016-Janua*, 4125–4129.

Bosch, N., D'mello, S. K., Ocumpaugh, J., Baker, R. S., & Shute, V. (2016b). Using Video to Automatically Detect Learner Affect in Computer-Enabled Classrooms. *ACM Transactions on Interactive Intelligent Systems*, *6*(2), 1–26. https://doi.org/10.1145/2946837

Botelho, A. F., Baker, R. S., & Heffernan, N. T. (2017). Improving Sensor-Free Affect Detection Using Deep Learning. *Artificial Intelligence in Education. Springer*, *LNAI 10331*(ISBN 978-3-319-61424-3), 40,52. https://doi.org/10.1007/978-3-319-61425-0

Bradley, M., & Lang, P. J. (1994). Measuring Emotion: The Self-Assessment Semantic Differential Manikin and the. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(I), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Breiman, L. (1996). Bagging Predictors. In *Machine Learning* (pp. 123–140).

Breiman, L. (2001). Random Forest. In *Machine Learning* (pp. 5–32).

Breuer, R., & Kimmel, R. (2017). A Deep Learning Perspective on the Origin of Facial Expressions, 1–16. Retrieved from http://arxiv.org/abs/1705.01842

Bull, A. D. (2011). Convergence rates of efficient global optimization algorithms. Retrieved from https://arxiv.org/abs/1101.3501v3

Cao, Y., Faloutsos, P., & Pighin, F. (2003). Unsupervised learning for speech motion editing. *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 225–231. Retrieved from http://dl.acm.org/citation.cfm?id=846276.846308

Chu, W. S., De La Torre, F., & Cohn, J. F. (2017). Learning Spatial and Temporal Cues for Multi-Label Facial Action Unit Detection. *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge*, 25–32. https://doi.org/10.1109/FG.2017.13

Cornelius, R. R. (1996). *The science of Emotion*. (P. HAll, Ed.) (1996th ed.). Nueva Jersey EUA.

Cowie, R., Douglas - Cowie, E., Tsapatsoulis, N., Votis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human computer interaction. *IEEE Signal Processing Magazine*, *18(1)*(January), 32–80. https://doi.org/10.1109/79.911197

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

Darwin, C. (1890). *the Expression of the Emotions in Man and Animals*. (D. Francis, Ed.), *The American Journal of the Medical Sciences* (Second, Vol. 232). London: Cambridge

University Press. https://doi.org/10.1097/00000441-195610000-00024

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(15). https://doi.org/10.1073/pnas.1322355111

Ekman, P., Friesen, W., & Hager, J. (2002). *Facial Action Coding System. The Manual*. Salt Lake City, United States of America.

Ekman, Paul. (2003). *Emotions Revealed. Recognizing Faces and Feelings to Improve Communication and Emotional Life*. (T. Books, Ed.) (1st ed.). New York: Henrry Holt and Company.

Farnsworth, B. (2019). *Facial Action Coding System (FACS) – A Visual Guidebook*. (iMotions, Ed.). Boston, United States. Retrieved from https://imotions.com/blog/facial-action-coding-system/

Fayyad, U., Piatetsky, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*, 37–54. https://doi.org/10.1609/AIMAG.V17I3.1230

Fogarty, J., Baker, R. S., & Hudson, S. E. (2005). Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. *Proceedings - Graphics Interface*, 129–136.

Freitas-Magalhães, A. (2018). *Facial Action Coding System 3.0: Manual of Scientific Codification of the Human Face (english edition).* Porto: FEELab Science Books.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Computer and System Sciences*, *55*, 119–139.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, *28*(2), 337–407.

Fuentes, C., Herskovic, V., Rodríguez, I., Gerea, C., Marques, M., & Rossel, P. O. (2016). A systematic literature review about technologies for self-reporting emotional information. *Journal of Ambient Intelligence and Humanized Computing*, 1–14. https://doi.org/10.1007/s12652-016-0430-z

Gelbart, M., J., & Snoek, R. P. (2014). Adams. Bayesian Optimization with Unknown Constraints. Retrieved from https://arxiv.org/abs/1403.5607

Ghimire, D., Jeong, S., Lee, J., & Park, S. H. (2017). Facial expression recognition based on local region specific features and support vector machines. *Multimedia Tools and Applications*, *76*(6), 7803–7821. https://doi.org/10.1007/s11042-016-3418-y

Ghimire, D., & Lee, J. (2013). Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. *Sensors (Switzerland)*, *13*(6), 7714–7734. https://doi.org/10.3390/s130607714

González-Hernández, F., Zatarain-Cabada, R., Barrón-Estrada, M. L., & Rodríguez-Rangel, H. (2017). Recognition of learning-centered emotions using a convolutional neural network. *Journal of Intelligent & Fuzzy Systems*.

Gower, J. (1975). *Generalized procrustes analysis.* Psychometrika.

Graesser, A. C., & D'Mello, S. (2012a). *Emotions During the Learning of Difficult Material*. *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 57). https://doi.org/10.1016/B978-0-12-394293-7.00005-4

Graesser, A. C., & D'Mello, S. (2012b). Moment-to-moment emotions during reading. *Reading Teacher*, *66*(3), 238–242. https://doi.org/10.1002/TRTR.01121

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2018). DAiSEE: Towards User Engagement Recognition in the Wild, *14*(8), 1–12. Retrieved from http://arxiv.org/abs/1609.01885

Happy, S. L., Patnaik, P., Routray, A., & Guha, R. (2017). The Indian Spontaneous Expression Database for Emotion Recognition. *IEEE Transactions on Affective Computing*, *8*(1), 131–142. https://doi.org/10.1109/TAFFC.2015.2498174

Hasani, B., & Mahoor, M. H. (2017). Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2017-July*, 2278–2288. https://doi.org/10.1109/CVPRW.2017.282

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2da ed.). Springer.

Hill, D. (2014). *Emotionomics*. (G. E. Patria, Ed.) (1st ed.). México.

Hjortsj, C.-H., Ekman, P., Friesen, W. V, Hager, J. C., Facs, F., & Facs, F. M. (2019). Sistema de Codificación Facial, 1–8.

Kaliouby, R. el, & Picard, R. W. (2019). Affectiva Database. Retrieved April 8, 2019, from https://www.affectiva.com

Karpouzis, K., & Votsis, G. (1999). Emotion recognition using feature extraction and 3-d models. *Proceedings of IMACS …*. Retrieved from http://www.image.ece.ntua.gr/physta/conferences/537.pdf

Kim, D. H., Baddar, W. J., Jang, J., & Ro, Y. M. (2019). Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*, *10*(2), 223–236. https://doi.org/10.1109/TAFFC.2017.2695999

Ko, K.-E., Yang, H.-C., & Sim, K.-B. (2009). Emotion recognition using EEG signals with relative power values and Bayesian network. *International Journal of Control, Automation and Systems*, *7*(5), 865–870. https://doi.org/10.1007/s12555-009-0521-0

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and Emotion*, *24*(8), 1377–1388. https://doi.org/10.1080/02699930903485076

Li, S., & Deng, W. (2018). Deep Facial Expression Recognition: A Survey. *Ground AI*, *1*, 1–25. Retrieved from http://arxiv.org/abs/1804.08348

Livingstone, S. R., & Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American english. *PLoS ONE*, *13*(5), 14–18. https://doi.org/10.1371/journal.pone.0196391

Lopatovska, I., & Arapakis, I. (2011). Theories, methods and current research on emotions in

library and information science, information retrieval and human-computer interaction. *Information Processing and Management*, *47*(4), 575–592. https://doi.org/10.1016/j.ipm.2010.09.001

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., & Ambadar, Z. (2010). The extended Cohn-Kanade dataset (CK+): a complete facial expression dataset for action unit and emotion-specified expression Conference on. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on* (pp. 94–101). Iain Matthews: IEEE Computer Society Conference on.

Lyons, M. J., Gyoba, J., & Kamachi, M. (1997). *Japanese Female Facial Expressions (JAFFE), Database of digital images*. Retrieved from http://www.kasrl.org/jaffe_info.html

M. Harley, J., Bouchet, F., & Azebedo, R. (2013). Aligning and Comparing Data on Emotions Experienced during Learning with MetaTutor. *LNAI 7926 - Artificial Intelligence in Education.*, (July), 61–70.

Martín de Serrano, D. I., Conde, Á., & Cabello, C. (2006). Técnicas de Reconocimiento Automático de Emociones. *Teoría de La Educación. Educación y Cultura En La Sociedad de La Información*, *7*(2), 107–127. https://doi.org/201017296007

Mavadati, S. M., Member, S., Mahoor, M. H., Bartlett, K., Trinh, P., & Cohn, J. F. (2013). DISFA : A Spontaneous Facial Action Intensity Database. *IEEE Transactions on Affective Computing*, *6*(1), 1–13.

Mehmood, R., & Lee, H. (2017). Towards Building a Computer Aided Education System for Special Students Using Wearable Sensor Technologies. *Sensors*, *17*(317), 1–22. https://doi.org/10.3390/s17020317

Mena-Chalco, Jesus Marcondes, R., & Velho, L. (2008). *Banco de Dados de Faces 3D: IMPA-FACE3D*. Brasil. Retrieved from http://www.visgraf.impa.br/Data/RefBib/PS_PDF/tr07-2010/tr-bernardo.pdf

Mitchell, T. M. (2009). *Machine learning*. *IJCAI International Joint Conference on Artificial Intelligence* (1ra.). New York: McGraw-Hill Science/Engineering/Math.

https://doi.org/10.1007/978-3-540-75488-6_2

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, *10*(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

Monkaresi, H., Bosch, N., Calvo, R. A., & D'Mello, S. K. (2016). Automated Detection of Engagement using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing*, 1–14. https://doi.org/10.1109/TAFFC.2016.2515084

Morales-Vargas, E., Reyes-Garcia, C. A., & Peregrina-Barreto, H. (2017). Reconocimiento de expresiones faciales con base en la dinámica de puntos de referencia faciales. In *Research in Computing Science* (Vol. 140, pp. 9–18).

Morales-Vargas, E., Reyes-García, C. A., & Peregrina-Barreto, H. (2019). On the use of action units and fuzzy explanatory models for facial expression recognition. *PLoS ONE*, *14*(10), 1–13. https://doi.org/10.1371/journal.pone.0223563

Nezami, O. M., Dras, M., Hamey, L., Richards, D., Wan, S., & Cécile Paris. (2020). Automatic Recognition of Student Engagement usingDeep Learning and Facial Expression. In *Lecture Notes in Computer Science*. https://doi.org/https://dx.doi.org/10.1007/978-3-030-46133-1_17

Nye, B., Karumbaiah, S., Tokel, S. T., Core, M. G., Stratou, G., Auerbach, D., & Georgila, K. (2017). Analyzing Learner Affect in a Scenario-Based Intelligent Tutoring System. *Artificial Intelligence in Education. Springer*, 544–547.

Paul, E., & Rosenberg, E. L. (2012). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. (E. Paul & E. L. Rosenberg, Eds.) (Second). California, San Francisco: Oxford Scholarship Online. https://doi.org/10.1093/acprof:oso/9780195179644.001.0001

Picard, R. W. (1995). Affective Computing. *MIT Press*, (321), 1–16. https://doi.org/10.1007/BF01238028

Picard, R. W. (1997). *Affective Computing*. *MIT*. https://doi.org/10.1007/BF01238028

Picard, R. W. (2003). Affective computing: Challenges. *International Journal of Human Computer Studies*, *59*(1–2), 55–64. https://doi.org/10.1016/S1071-5819(03)00052-1

Rivera, H., Valadão, C., Caldeira, E., Krishnan, S., & Bastos-Filho, T. F. (2019). Development of a Toolkit for Online Analysis of Facial Emotion. *XXVI Brazilian Congress on Biomedical Engineering. Springer.*, *70/2*(October), 83–87. https://doi.org/10.1007/978-981-13-2517-5

Sawyer, R., Smith, A., Rowe, J., Azevedo, R., & Lester, J. (2017). Enhancing Student Models in Game-based Learning with Facial Expression Recognition. *Association for Computing Machinery. ACM*, 1–10.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. https://doi.org/10.1177/0539018405058216

Seiffert, C., Khoshgoftaar, T., Hulse, J., & Napolitano, A. (2008). RUSBoost: Improving clasification performance when training data is skewed. In *19th International Conference on Pattern Recognition* (pp. 1–4).

Sneddon, I., Mcrorie, M., Mckeown, G., & Hanratty, J. (2012). Belfast Induced Natural Emotion Database. *IEE Transactions on Affective Computing*, *3*(1), 32–41.

Snoek, J., Larochelle, H., & Rian. P., A. (2014). Practical Bayesian Optimization of Machine Learning Algorithms. Retrieved from https://arxiv.org/abs/1206.2944

Soleymani, M., Member, S., & Lee, J. (2012). DEAP : A Database for Emotion Analysis Using Physiological Signals, *3*(1), 18–31.

Speybroeck, N. (2012). Classification and regression trees. *International Journal of Public Health. Springer*, *57*(1), 243–246.

Steidl, S. (2009). *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Universität Erlangen-Nürnberg.

Thomaz, C. E. (2012). FEI Face Database. Retrieved April 4, 2019, from https://fei.edu.br/~cet/facedatabase.html

Valstar, M. F., & Pantic, M. (2010). Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database.

Vargas, E. M. (2017). *Granular fuzzy model with hyperboxes for facial expression recognition*.

Walecki, R., Rudovic, O., Pavlovic, V., Schuller, B., & Pantic, M. (2017). Deep structured learning for facial action unit intensity estimation. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 5709–5718. https://doi.org/10.1109/CVPR.2017.605

Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., & Kissler, J. (2017). Mapping the emotional face. How individual face parts contribute to successful emotion recognition. *PLoS ONE*, *12*(5), 1–15. https://doi.org/10.1371/journal.pone.0177239

Xiao, X., Pham, P., & Wang, J. (2017). Dynamics of Affective States During MOOC Learning. *Artificial Intelligence in Education. Springer*, 586–589.

Yacoub, S., Simske, S., Lin, X., & Burns, J. (2003). Recognition of Emotions in Interactive Voice Response Systems. *Speech Communication*, (September), 1–4.

Zafeiriou, S., Kollias, D., Nicolaou, M. A., Papaiooannou, A., & Kotsia, I. (2017). Aff-Wild : Valence and Arousal ' in-the-wild ' Challenge. *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1980–1987.

Zatarain-Cabada, R., Barrón-Estrada, M. L., González-Hernández, F., Oramas-Bustillos, R., Alor-Hernández, G., & Reyes-García, C. A. (2017a). Building a Corpus and a Local Binary Pattern Recognizer for Learning-Centered Emotions. *Advances in Computational Intelligence*, *II*, 524–535. https://doi.org/10.1007/978-3-319-58088-3

Zatarain-Cabada, R., Barron-Estrada, M. L., González-Hernández, F., & Rodríguez-Rangel, H. (2017b). Building a Face Expression Recognizer and a Face Expression Database for an Intelligent Tutoring System. *Proceedings - IEEE 17th International Conference on Advanced Learning Technologies, ICALT 2017*, (2161-377X/17), 391–393. https://doi.org/10.1109/ICALT.2017.141

Zatarain-Cabada, R., Barrón-Estrada, M. L., & Ríos-Félix, J. M. (2017c). Affective Learning

System for Algorithmic Logic Applying Gamification. *Chapter in Lecture Notes in Computer Science · October 2017*, *576*(August). https://doi.org/10.1007/978-3-319-58088-3

Zataraín, R., Barrón, M. L., González, F., & Reyes-García, C. A. (2017). An Affective and Web 3.0 Based Learning Environment for a Programming Language. *Telematics and Informatics*. https://doi.org/10.1016/j.tele.2017.03.005

Zatarain, R., Barrón, M. L., Luis, O. J., & Martínez, J. A. (2014). Reconocimiento automático y aspectos éticos de emociones para aplicaciones educativas. *Komputer Sapiens*, *II*, 27–31.

Zhao. (n.d.). Oulu-CASIA NIR&VIS facial expression database. Retrieved April 4, 2019, from http://www.cse.oulu.fi/wsgi/CMV/Downloads/Oulu-CASIA

Apéndice A. Protocolo del Experimento para la Recolección de Datos

PROTOCOLO DE EXPERIMENTO PARA RECOLECCIÓN DE DATOS

Introducción

Este documento tiene como objetivo describir el experimento que se llevará a cabo para la recolección de datos fisiológicos y de comportamiento de alumnos de la Facultad de Computación mientras interactúan con una computadora en un proceso de aprendizaje a través del uso de una plataforma en línea. El experimento forma parte de la primera etapa del proyecto de tesis doctoral "Reconocimiento de emociones en un ambiente educativo con aprendizaje computacional" del Doctorado en Ingeniería del Lenguaje y del Conocimiento de la Facultad de Computación.

En el contenido se hace una descripción del experimento, se define el tipo de población y los instrumentos de recolección de datos, se enumeran los dispositivos de captura de señales que se utilizaran y sus características. También se enlistan las herramientas de software que se usarán y se describen los pasos para la ejecución del experimento. Al final se anexa el formato de consentimiento informado que se les proporcionara a los alumnos para autorizar su participación en el experimento y se hace mención de la Ley de protección de datos que se debe respetar.

Contenido

1. **Descripción:** El experimento consistirá en la captura de señales de comportamiento y señales fisiológicas de alumnos. Este se llevará a cabo en el laboratorio de experimentos del Doctorado LKE, ubicado en el edificio EMA7 piso 3 de la BUAP. El laboratorio es una habitación en color blanco con dos mesas de trabajo. En la primera mesa se localiza la computadora portátil con la que los alumnos trabajaran y frente a ella se encuentran colocadas las tres cámaras de grabación (la térmica, la Web y el Kinect). En la segunda mesa se localiza una computadora de escritorio con la que se controlan las grabaciones del Kinect. En esta mesa también hay una computadora portátil con la que se controlan las grabaciones de la cámara Web, de la cámara térmica y las lecturas del sensor de ritmo cardíaco. Frente a las mesas de trabajo se localiza un espejo que cubre toda la pared, el cual colinda con la habitación de la cámara Gesell. En el pasillo de antesala al laboratorio se utiliza como área de estar, para quienes desean esperar a sus compañeros.
   Los alumnos que participarán en el experimento son estudiantes de la facultad de computación de la misma Universidad, a quienes se les pide realicen una actividad de aprendizaje haciendo uso de un sistema tutorial inteligente de álgebra básica. Los registros grabados serán almacenados, procesados y utilizados para reconocer las emociones que los alumnos presentan durante el proceso de aprendizaje en el que participan.
   Los datos de su comportamiento corresponderán a imágenes de expresiones faciales y de movimiento de cabeza y manos. Las señales fisiológicas corresponderán a mediciones del ritmo cardíaco y de la temperatura de áreas faciales.
   La captura de señales de comportamiento se hará a través de una cámara de video y de las cámaras del Kinect 360.

La captura de señales fisiológicas se hará través de un sensor de pulso cardíaco el cual será colocado en el dedo anular de la mano izquierda. Se utilizará una cámara térmica para la captura de la temperatura facial.

Cada 10 minutos se aplicará un test de dos secciones para obtener una autoevaluación por parte del alumno. En la primera sección se les pedirá seleccionen el tipo de emoción que ellos perciben que sienten (interés, aburrimiento, frustración o confusión), para obtener una medición de la emoción desde el enfoque discreto. En la segunda sección se les solicita seleccionen un valor para cada una de las tres variables involucradas en la emoción presentada (valencia, activación y dominancia), para obtener una medición desde el enfoque continuo.

2. **Definición de la población muestra:** Estará formada por al menos 50 alumnos de nivel superior del área de ingeniería.
3. **Instrumentos de recolección de información:** Se utilizarán dos técnicas; el experimento (grabaciones) y los cuestionarios. Ambas técnicas para obtener datos cuantitativos. Se utilizarán algoritmos aprendizaje computacional para el procesamiento de los datos.
4. **Dispositivos de adquisición de señales:** Para la captura de señales de comportamiento se utilizarán los siguientes dispositivos:
   - Cámara de video Web Logitech Full HD 1080p y 10MP.
   - Cámara térmica ICI 9320P. Resolución térmica 320 X 240 píxeles. Resolución visual 2160 X 1440.
   - Sensor de pulso cardíaco implementado con Arduino.
   - Kinect 360 para Windows
5. **Herramientas de Software por utilizar:**
   La actividad de aprendizaje se realizará en línea utilizando el MOOC libre de álgebra de Coursera:

a) Dirección electrónica: ***https://www.coursera.org/learn/algebra-basica/***

b) Datos de la cuenta:

   Usuario: *yeseniaglez@hotmail.com*

   Contraseña: yesenia0

c) Descripción del contenido y duración de cada actividad

Numeración y lenguaje algebraico	
Expresiones algebraicas (video)	15 min
Lenguaje algebraico (video)	6 min
Tutorial para resolver los cuestionarios (video)	10 min
Cuestionario practico: Lenguaje algebraico 10 preguntas	5 min
Tiempo Total	**36 min**

6. **Ejecución del experimento:**
   a. Al llegar el alumno se le saluda cordialmente.
   b. Se le da una explicación del experimento.
   c. Se le da una explicación de las definiciones y características de las emociones que se identificarán.
   d. Se le explican las dos secciones del test que contestarán durante el experimento.

e. Se le pregunta si tiene alguna duda.
f. Se le da el Formato de Consentimiento Informado para que firme la autorización de la grabación de sus datos.
g. Se le pide sentarse frente a la computadora.
h. Se le explica el uso del tutorial y del cuestionario de autoevaluación.
i. Se le coloca en el dedo anular izquierdo el sensor de ritmo cardíaco.
j. Se inicializan el Kinect, la cámara de video, la cámara térmica y el sensor de ritmo cardíaco.
k. Se inicia la grabación de los datos.
l. Una vez terminada la captura de los datos se le pregunta al alumno cómo se siente y se le agradece su participación ofreciéndole algún dulce, galletas o botella de agua.

7.- Ley de protección de datos

Seremos respetuosos de la Ley Federal de Protección de Datos Personales en Posesión de los Particulares. DOF: 05/07/2010 para proteger la privacidad de las personas respecto al tratamiento que demos a su información personal.

Sus disposiciones son aplicables a todas las personas físicas o morales, del sector público y privado, tanto a nivel federal como estatal, que lleven a cabo el tratamiento de datos personales en el ejercicio de sus actividades, por lo tanto empresas como bancos, aseguradoras, hospitales, escuelas, compañías de telecomunicaciones, asociaciones religiosas, y profesionistas como abogados, médicos, entre otros, se encuentran obligados a cumplir con lo que establece esta ley.

Un dato personal, de acuerdo con el artículo 3 fracción V de esta Ley es toda aquella información que permita identificar a una persona.

Apéndice B. Carta de Consentimiento Informado

No Folio:____

Carta de Consentimiento Informado

Adquisición de datos fisiológicos y de comportamiento en ambientes de aprendizaje con medios no invasivos en estudiantes de la Facultad de Computación de la BUAP.

El objetivo de esta investigación es recolectar datos correspondientes a grabaciones de video tradicional, video de imágenes térmicas y lecturas del ritmo cardíaco de estudiantes mientras interactúan con un tutorial de álgebra básica con fines de investigación.

En caso de aceptar participar en este estudio se te pedirá como sujeto voluntario que proporciones tu nombre y firmes tu consentimiento explícito para participar en el proceso de captura y permitir utilizar tus datos con fines de investigación.

En la sesión se te colocará un sensor de ritmo cardíaco en el dedo anular de la mano izquierda para efectuar el registro de tu ritmo cardíaco. Este procedimiento no es invasivo, doloroso ni atenta contra tu integridad. Simultáneamente se realizarán grabaciones de video con una cámara canon de video tradicional, con una cámara térmica y con un Kinect, mientras participas en un proceso de enseñanza-aprendizaje de álgebra básica utilizando un MOOC de Coursera.

Tu decisión de participar en el estudio es completamente voluntaria.

- No habrá ninguna consecuencia desfavorable para ti, en caso de no aceptar la invitación.
- No tendrás que hacer gasto alguno durante el estudio.
- No recibirás pago por tu participación.
- La información obtenida en este estudio, será mantenida con estricta confidencialidad.
- Si consideras que no hay dudas ni preguntas acerca de tu participación, puedes, si así lo deseas, firmar esta Carta de Consentimiento Informado.

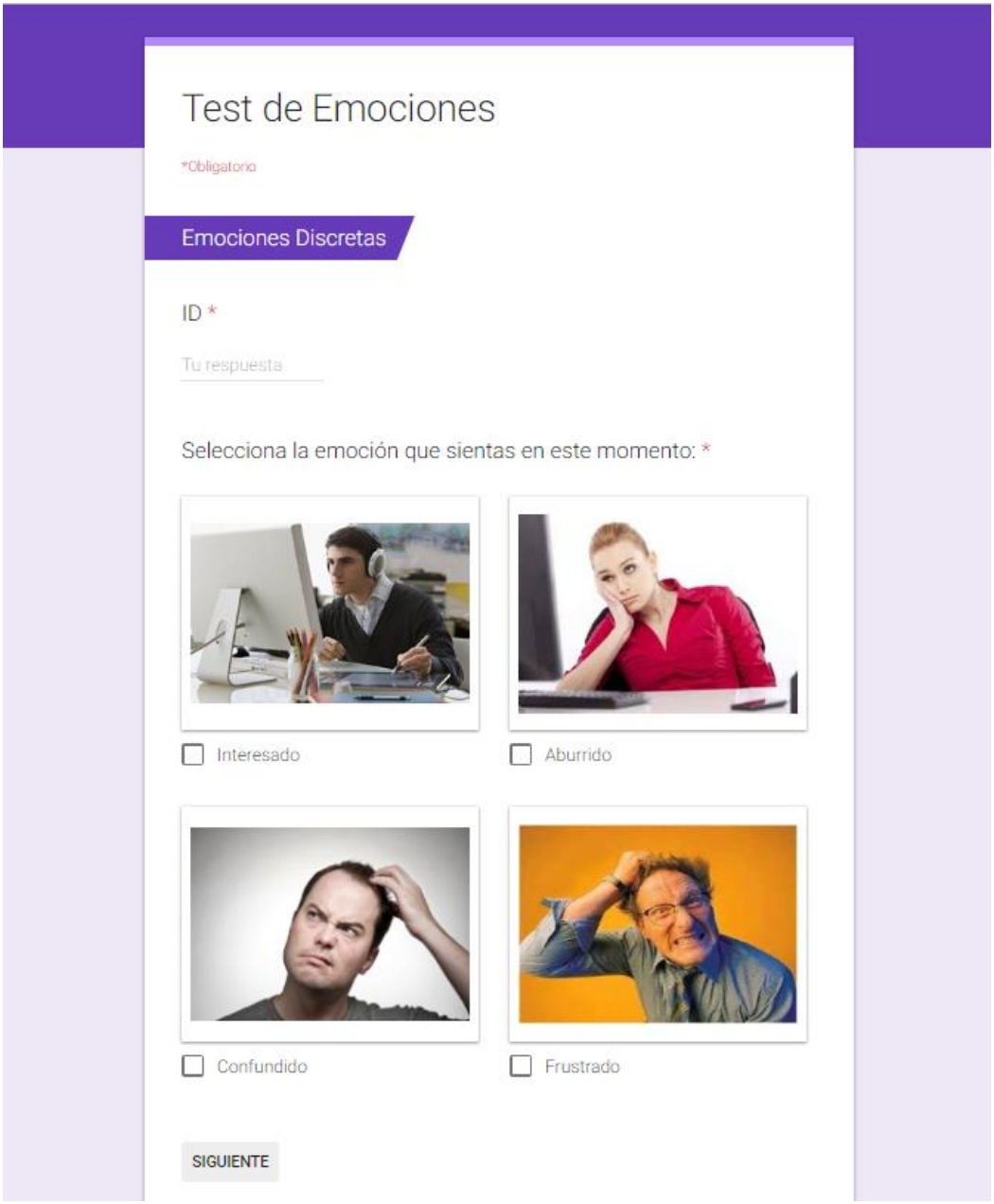
Yo, _____de género, _____ he comprendido la información anterior y mis preguntas han sido respondidas de manera satisfactoria. He sido informado (a) y entiendo que los datos obtenidos en el estudio pueden ser publicados o difundidos con fines científicos. Convengo en participar en este estudio de investigación.

_____ _____

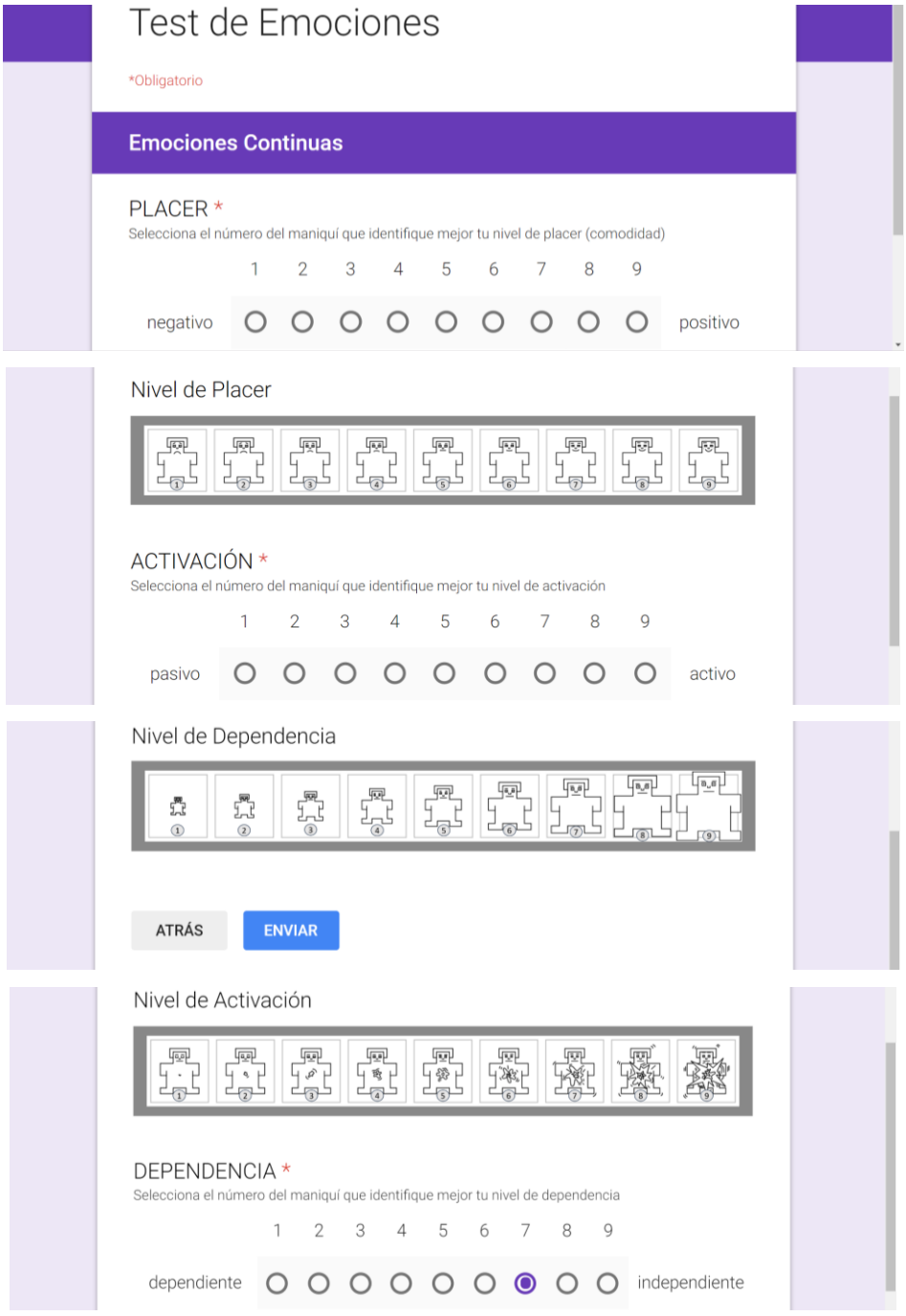
Firma del participante Fecha

Cualquier duda o comentario, dirigirse con:
Yesenia N. González Meneses.
yesenia.gonzalezm@alumno.buap.mx
Investigador responsable

Apéndice C. *Tests* de Emociones



(parte 1)


Test de Emociones
*Obligatorio
Emociones Continuas
PLACER *
Selecciona el número del maniquí que identifique mejor tu nivel de placer (comodidad)
1 2 3 4 5 6 7 8 9
negativo
positivo
Nivel de Placer
ACTIVACIÓN *
Selecciona el número del maniquí que identifique mejor tu nivel de activación
1 2 3 4 5 6 7 8 9
pasivo
activo
Nivel de Dependencia
ATRÁS
ENVIAR
Nivel de Activación
DEPENDENCIA *
Selecciona el número del maniquí que identifique mejor tu nivel de dependencia
1 2 3 4 5 6 7 8 9
dependiente
independiente

(parte 2)