



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Método para clasificar los estilos de conducción basado en técnicas de aprendizaje de máquina, empleando señales fisiológicas y sensorial capturada por dispositivos vestibulares. Caso de Estudio: Conductores del Área Metropolitana del Valle de Aburrá.

JUAN MANUEL LOMBANA MARTÍNEZ

Universidad Nacional de Colombia

Facultad de Minas

Medellín, Colombia

2022

Método para clasificar los estilos de conducción basado en técnicas de aprendizaje de máquina, empleando señales fisiológicas y sensorica capturada por dispositivos vestibles. Caso de Estudio: Conductores del Área Metropolitana del Valle de Aburrá.

Juan Manuel Lombana Martínez

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

Magister en Ingeniería de Sistemas

Director (a):

Ph.D. Albeiro Espinosa Bedoya

Universidad Nacional de Colombia

Facultad de Minas

Medellín, Colombia

2022

Declaración de obra original

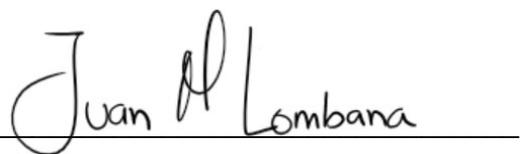
Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.



Juan Manuel Lombana

Fecha 20/02/2022

Resumen

Método para clasificar los estilos de conducción basado en técnicas de aprendizaje de máquina, empleando señales fisiológicas y sensorica capturada por dispositivos vestibles. Caso de Estudio: Conductores del Área Metropolitana del Valle de Aburrá.

La clasificación de los estilos de conducción usando aprendizaje de máquina es una propuesta novedosa al problema de salud pública y los impactos macroeconómicos que representa la pérdida de vidas debido a la accidentalidad vial. En este trabajo se propone el uso de relojes inteligentes para monitorear diversos aspectos del conductor y para las tareas de clasificación se han implementado algoritmos de máquinas de vectores de soporte (SVM), árboles de decisión y clasificador ingenuo de bayes. Como conclusión principal se muestra la correlación que tienen los patrones de sueño con el estilo de conducción, así como la pertinencia del uso de estos dispositivos para la construcción de un conjunto de datos de buena calidad.

Palabras clave: Estilo de conducción, muertes en accidentes de tránsito, aprendizaje de máquina, conjunto de datos, svm, árbol de decisión, naive-bayes.

Abstract

**Method for the classification of driving behaviors based on machine learning techniques, using physiological signals and sensors captured by wearable devices.
Case of Study: Metropolitan Area Drivers of Aburrá Valley.**

Machine learning driving behavior classification is a novel proposal to face public health problems and macroeconomy impacts that suppose road traffic injury deaths. This project proposes the use of smart watches to monitor different aspects of drivers, for classification tasks we had implemented support vector machines (SVM), decision trees and naïve bayes classifier. We show strong correlation between sleep patterns and driving behavior as main conclusion, also the relevance of using wearable devices to build high quality datasets.

Keywords: Driving behavior, road traffic deaths, machine learning, dataset, svm, decision tree, naive-bayes.

Contenido

	Pág.
Resumen	V
Lista de figuras.....	IX
Lista de tablas	X
Introducción	1
1. Marco Teórico de Referencia.....	3
1.1 Formulación de la hipótesis.....	3
1.2 Justificación.....	4
1.3 Objetivos	5
1.3.1 Objetivo general.....	5
1.3.2 Objetivos específicos	5
1.4 Marco teórico	5
1.4.1 Marco conceptual	5
1.5 Metodología	7
1.6 Alcance	7
1.6.1 Alcance del Conjunto de Datos.....	7
1.6.2 Alcance de la implementación	8
1.7 Estado del Arte.....	8
2. Construcción del conjunto de datos.....	10
2.1 Selección de los dispositivos	11
2.2 Diseño del experimento.....	12
2.2.1 Pulso cardíaco	12
2.2.2 Giroscopio	12
2.2.3 Aceleración lineal.....	13
2.2.4 Posición satelital.....	13
2.2.5 Sueño	13
2.2.6 Actividad física.....	13
2.3 Selección de los conductores	13
2.3.1 Cálculo del tamaño de la muestra.....	14
2.4 Arquitectura para recolección de datos	15
2.5 Características del conjunto de datos.....	16
2.5.1 Limpieza y Transformaciones de Datos	16
2.5.2 Agrupación de Datos	18

2.5.3	Análisis de Datos.....	18
3.	Método propuesto para clasificación de conductores.....	28
3.1	Árboles de Decisión.....	28
3.2	Máquinas de vectores de soporte (SVM).....	32
3.3	Clasificador Ingenuo de Bayes – Gaussiano.....	34
4.	Validación de los métodos de clasificación.....	36
4.1	Matriz de Confusión.....	36
4.1.1	Verdaderos Positivos.....	36
4.1.2	Verdaderos Negativos.....	37
4.1.3	Falsos Positivos.....	37
4.1.4	Falsos Negativos.....	37
4.2	Métricas de Validación.....	37
4.2.1	Exactitud (Accuracy).....	37
4.2.2	Precisión (Precision).....	38
4.2.3	Exhaustividad - (Recall).....	38
4.2.4	Valor F1 - (F1 Score).....	38
4.3	Validación de los Algoritmos Implementados.....	38
4.3.1	Árbol de Decisión.....	39
4.3.2	Máquina de vectores de soporte.....	40
4.3.3	Clasificador Ingenuo de Bayes.....	40
4.4	Comparación de los métodos.....	41
5.	Conclusiones y recomendaciones.....	44
5.1	Conclusiones.....	44
5.2	Recomendaciones.....	45
6.	Bibliografía.....	46

Lista de figuras

	Pág.
Figura 2-1 Ciclo de vida simplificado de la aplicación (Autoría propia).....	16
Figura 2-2 Ejemplo del conjunto de datos (Autoría propia)	17
Figura 2-3 Gráfico de cajas para el grupo 0 (1 de 6) (Autoría propia)	19
Figura 2-4 Gráfico de cajas para el grupo 0 (2 de 6) (Autoría propia)	20
Figura 2-5 Gráfico de cajas para el grupo 0 (3 de 6) (Autoría propia)	20
Figura 2-6 Gráfico de cajas para el grupo 0 (4 de 6) (Autoría propia)	21
Figura 2-7 Gráfico de cajas para el grupo 0 (5 de 6) (Autoría propia)	21
Figura 2-8 Gráfico de cajas para el grupo 0 (6 de 6) (Autoría propia)	22
Figura 2-9 Gráfico de cajas para el grupo 1 (1 de 6) (Autoría propia)	24
Figura 2-10 Gráfico de cajas para el grupo 1 (2 de 6) (Autoría propia)	24
Figura 2-11 Gráfico de cajas para el grupo 1 (3 de 6) (Autoría propia)	25
Figura 2-12 Gráfico de cajas para el grupo 1 (4 de 6) (Autoría propia)	25
Figura 2-13 Gráfico de cajas para el grupo 1 (5 de 6) (Autoría propia)	26
Figura 2-14 Gráfico de cajas para el grupo 1 (6 de 6) (Autoría propia)	26
Figura 3-1 Ejemplo de árbol de decisión (Bishop, 2006).....	29
Figura 3-2 Árbol de decisión resultante (Bishop, 2006).....	29
Figura 3-3 Estructura del árbol de decisión (Autoría propia)	31
Figura 3-4 Margen y vectores de soporte en una SVM (Géron, 2019).....	32
Figura 3-5 Margen suave en una SVM. (Géron, 2019).	33
Figura 3-6 SVM kernel RBF (Géron, 2019).....	33
Figura 4-1 Matriz de confusión del árbol de decisión (Autoría propia).....	39
Figura 4-2 Matriz de confusión para SVM (Autoría propia)	40
Figura 4-3 Matriz de confusión para el clasificador ingenuo de bayes (Autoría propia)..	41

Lista de tablas

	Pág.
Tabla 2-1 Dispositivos candidatos (Autoría propia).....	11
Tabla 2-2 Aspectos técnicos (Autoría propia)	11
Tabla 2-3 Características del conjunto de datos (Autoría propia).....	16
Tabla 2-4 Análisis Exploratorio para el grupo 0 (Autoría propia)	18
Tabla 2-5 Análisis para el grupo 1 (Autoría propia).....	22
Tabla 3-1 Precisión en validación cruzada para el árbol de decisión.	30
Tabla 3-2 Hiperparámetro de la SVM (Autoría propia).....	34
Tabla 3-3 Precisión en validación cruzada para la máquina de vectores de soporte. (Autoría propia).....	34
Tabla 3-4 Precisión en validación cruzada para el clasificador ingenuo de bayes. (Autoría propia)	35
Tabla 4-1 Reporte de métricas para el árbol de decisión (Autoría propia).....	39
Tabla 4-2 Reporte de clasificación para la SVM (Autoría propia).....	40
Tabla 4-3 Reporte de clasificación para el clasificador ingenuo de bayes (Autoría propia)	41
Tabla 4-4 Comparación de los métodos de clasificación (Autoría propia).....	42
Tabla 4-5 Métrica de exhaustividad para los métodos evaluados (Autoría propia).....	43

Introducción

La movilidad es uno de los elementos claves en la sociedad moderna, pues permite el desarrollo de los negocios, la educación y en general es una actividad que mantiene el dinamismo de las comunidades, pero desafortunadamente trae consigo la problemática de la accidentalidad vial. Los accidentes de tránsito acarrearán consigo un problema para la salud pública de las sociedades pues estos se han convertido desde hace varios años en una de las causas principales de mortalidad (OMS, 2018). Dado lo anterior el presente trabajo aborda el problema de la clasificación de estilos de conducción, pues este representa los hábitos que el conductor tiene cuando conduce un vehículo y se ha identificado como uno de los tres factores preponderantes en los accidentes de tránsito en conjunto con las condiciones de la vía y el comportamiento de los peatones (García de Quevedo N. F et al, 2018).

Como se observará en las secciones siguientes de este trabajo es de especial interés para las comunidades académicas, las organizaciones gubernamentales y el sector privado, contribuir con propuestas que mitiguen esta problemática, pues no solo es un tema de salud pública sino también un problema de índole macroeconómico, así pues, desde la computación y el aprendizaje de máquina se han realizado acercamientos con propuestas diversas.

En este trabajo particularmente se ha evaluado la pertinencia que tiene el uso de dispositivos vestibles como relojes inteligentes en las tareas de clasificación de estilos de conducción pues la fácil adaptación de este dispositivo como un elemento cotidiano de la vida del conductor, representa un método novedoso.

Para realizar las tareas de clasificación de estilos de conducción mediante aprendizaje de máquina se implementaron tres algoritmos supervisados: árbol de decisión, svm, naive-bayes; los cuales requieren de un conjunto de datos previamente etiquetado. En la

elaboración de este trabajo no se contaba con dicho conjunto de datos, por lo tanto, se desarrolló una aplicación para relojes inteligentes con el objetivo de recolectar información de los dispositivos a través del diseño e implementación de un experimento de recolección de datos con conductores en el Valle de Aburrá. La información de estos conductores se almacenó en la nube y posteriormente se realizaron tareas de limpieza, transformación y análisis estadístico hasta obtener un conjunto de datos de 426 sesiones de conducción. Estas se etiquetaron como estilos de conducción pasivo o agresivo y se asociaron con las diferentes variables de los sensores como pulso cardíaco, ciclos de sueño, actividad física, giroscopio y acelerómetro. Este conjunto de datos se dividió en dos subconjuntos: el 70% se usó para las tareas de entrenamiento y el 30% restante para validación.

El objetivo de este trabajo es desarrollar un método para clasificar los estilos de conducción basado en técnicas de aprendizaje de máquina empleando señales fisiológicas y sensorial capturada por dispositivos vestibles. Este se dividió en tres objetivos específicos que buscaban construir el conjunto de datos, desarrollar el método de clasificación y validar el método evaluando su precisión con métricas como la precisión y la exhaustividad.

El capítulo 1 aborda la formulación de la hipótesis y el marco teórico en donde se puede evidenciar claramente que la accidentalidad vial es una problemática urgente e importante tanto desde la perspectiva de salud como desde la perspectiva económica. El capítulo 2 lleva a cabo lo relacionado con la construcción del conjunto de datos, documentando el proceso de desarrollo para la construcción de la aplicación y el diseño del experimento de recolección de datos, hasta la limpieza y transformación de datos que generaron el conjunto de datos final usado en las tareas de clasificación. En el capítulo 3 se muestran las técnicas de aprendizaje de máquina implementadas y se realiza un entendimiento de cada una analizando las ventajas, desventajas y estrategias de aprendizaje. En el capítulo 4 se documenta el proceso de validación de las técnicas, cada una de sus métricas y se comparan con resultados de otros autores. Por último, en el capítulo 5 se presentan las conclusiones y recomendaciones.

1. Marco Teórico de Referencia

En el mundo anualmente 20 Millones de personas resultan heridas en accidentes de tránsito, y 1.17 Millones mueren en promedio (Mahafroz et al, 2015), a pesar de los esfuerzos continuos de diferentes actores como la Organización Mundial de la Salud (OMS), instituciones no gubernamentales, y las agencias de los gobiernos; las tasas de mortalidad se mantienen a nivel global año tras año, y en los países en vías de desarrollo con menores ingresos inclusive aumentan (OMS, 2018). La situación es preocupante pues en el año 2004 las muertes por accidentes de tránsito ocuparon la décima posición en las causas principales de mortalidad a nivel global (OMS, 2009), para el año 2018 se sitúa en la octava posición y es la causa de mortalidad principal para individuos entre 5 y 29 años (OMS, 2018).

1.1 Formulación de la hipótesis

Las causas de la accidentalidad son diversas, pero algunos estudios apuntan a que estos fenómenos no son aleatorios, ya que obedecen a una serie de factores relacionados con el conductor, el vehículo y la vía (García de Quevedo N. F et al, 2018). Para la OMS las naciones deben tomar acciones en factores como: reducir la conducción bajo efecto de psicoactivos, disminuir la velocidad de las vías, promover el uso de cinturón de seguridad, aumentar el uso de sillas para niños y reducir las distracciones al volante. A pesar de las medidas implementadas las muertes han alcanzado un techo de 1.25 Millones de muertes al año y se han mantenido así desde el año 2007, solo en algunos países con ingresos altos y medios se han logrado reducciones. (OMS, 2015). Los enfoques de sistemas seguros para las vías intentan disminuir el error humano, a través de mecanismos y legislaciones que permitan, prevenir y evitar los accidentes de tránsito. Una estrategia para la reducción y prevención de los accidentes es concientizar al conductor sobre su estilo de conducción, mostrar alertas tempranas ante condiciones de la vía y persuasión ante

maniobras peligrosas. Así pues, es necesario tener mecanismos que refuercen, mejoren y visibilicen los malos hábitos de conducción.

Las cifras de accidentalidad vial son alarmantes y desde la computación se ha intentado dar solución a diferentes aspectos de esta problemática (Cheng Zang et al, 2016), especialmente en los últimos años las técnicas de aprendizaje de máquina han tomado un papel importante en la clasificación de los estilos de conducción, para predecir y clasificar comportamientos agresivos por parte de los conductores en las vías (Yu-Ting Pang et al, 2018), así como diferentes esfuerzos de la industria automotriz en la construcción de sistemas de asistencia en la conducción, sensórica e IoT en los automóviles hasta el desarrollo de sistemas de conducción semiautónoma y autónoma (Yeong-kang Lai et al, 2018).

Este trabajo abordará el problema de clasificar en agresivo o pasivo el estilo de conducción de un conductor durante la sesión de conducción, usando las señales fisiológicas recibidas por un reloj inteligente (pulso, acelerómetro, giroscopio etc.) que vestirá continuamente durante algunos días, especialmente mientras conduce.

1.2 Justificación

La gran cantidad de vidas perdidas en accidentes de tránsito representa un problema de salud pública para las naciones lo cual ha incentivado el desarrollo de estudios que calculan para algunos países y regiones los años potenciales de vida perdidos (APVP), así como el impacto económico de estos eventos, evidenciando que los accidentes de tránsito son un problema para el desarrollo económico, especialmente para los países en vías de desarrollo, pues se estima que el 3% del producto interno bruto se pierde a causa de los accidentes de tránsito (OMS, 2015). Esto demuestra el efecto devastador que tiene la accidentalidad vial en la capacidad productiva de las naciones, pues la población más vulnerable a muertes en accidentes de tránsito se encuentra en los años más productivos de su vida. (Porras Sandra et al, 2017). Diferentes técnicas han sido usadas como aproximaciones de solución a la problemática de la accidentalidad, en el caso de las técnicas de aprendizaje de máquina se han enfocado dichos esfuerzos en la clasificación del conductor y la correlación con los datos de sensores en los vehículos (Derick A. Johnson et al, 2011); también se han desarrollado sistemas para la detección de peatones

(Obaida M. Al-Hazaimeh et al, 2019), ciclistas (Zhang et al, 2020), señales de tránsito y cambios de carril (J Yang et al, 2017). Como se observará durante el desarrollo de este trabajo, es pertinente el uso de dispositivos vestibles como relojes inteligentes para identificar la correlación entre las variables fisiológicas y los estilos de conducción. Este campo representa una oportunidad de investigación en pro de la mejora en los hábitos de conducción aportando activamente a la disminución de la accidentalidad vial y sus nefastas consecuencias.

1.3 Objetivos

1.3.1 Objetivo general

Desarrollar un método para clasificar los estilos de conducción basado en técnicas de aprendizaje de máquina empleando señales fisiológicas y sensorica capturada por dispositivos vestibles.

1.3.2 Objetivos específicos

- Construir un conjunto de datos de conductores, con las variables fisiológicas y sensorica entregadas por los dispositivos vestibles.
- Desarrollar un método de clasificación usando técnicas y algoritmos de aprendizaje de máquina.
- Validar el método de clasificación probando su precisión, realizando comparación con otras técnicas de clasificación de estilos de conducción.

1.4 Marco teórico

1.4.1 Marco conceptual

- **Accidente de Tránsito:** “Evento generalmente involuntario, generado al menos por un vehículo en movimiento que causa daños a personas y bienes involucrados en él, igualmente afecta la normal circulación de los vehículos que se movilizan por la vía o vías comprendidas en el lugar o dentro de la zona de influencia del hecho” (Ley 769, 2002).

- **Salud Pública:** “Conjunto de políticas que buscan garantizar de una manera integrada, la salud de la población por medio de acciones de salubridad dirigidas tanto de manera individual como colectiva, ya que sus resultados se constituyen en indicadores de las condiciones de vida, bienestar y desarrollo del país. Dichas acciones se realizarán bajo la rectoría del Estado y deberán promover la participación responsable de todos los sectores de la comunidad” (Ley 1122, 2007).
- **Años Potenciales de Vida Perdidos:** “Medida resumen de la carga de la enfermedad que cuantifica el volumen de años que se han dejado de vivir por muertes ocurridas antes de una determinada edad. Esta medida se calcula definiendo de manera arbitraria un límite potencial a la vida. Sobre esta base se estiman años perdidos como la diferencia entre el límite potencial de la vida menos la edad de muerte de cada defunción” (Ministerio de Salud de Colombia, 2016).
- **Producto Interno Bruto:** “Resultado final de la actividad productiva de las unidades de producción residentes. Se mide desde el punto de vista del valor agregado, de la demanda o las utilidades finales de los bienes y servicios y de los ingresos primarios distribuidos por las unidades de producción residentes” (DANE, 2018).
- **Aprendizaje de Máquina:** Ciencia (y arte) de programar computadores así estos pueden aprender de los datos (Géron, 2019). Hacer que los computadores modifiquen o adapten sus acciones para ser más precisos, midiendo su precisión en que tan bien se escogen las acciones que reflejan las acciones correctas (Marsland, 2015).
- **Clasificación:** El problema de clasificación consiste en tomar un vector de entradas y decidir a cuál clase de un conjunto N de clases pertenece la entrada actual, basado en un proceso previo de entrenamiento que toma como bases ejemplares de cada clase (Marsland, 2015).
- **Dispositivo Vestible:** Dispositivo tecnológico que está puesto sobre el cuerpo humano, de diversos tipos y finalidades como dispositivos para la medición de la salud, utilitarios como ropa tecnológica, dispositivos para entretenimiento y de consumo (MESUT ÇİÇEK, 2015).

1.5 Metodología

Para la realización de este proyecto se requiere la construcción de un conjunto de datos con señales fisiológicas y sensórica de los dispositivos vestibles; para esto se validarán y evaluarán algunos de los dispositivos disponibles en el mercado, con el objetivo de seleccionar un dispositivo sobre el cual se desarrollará una aplicación para recolectar la información que conformarán el conjunto de datos. Este último tendrá información relacionada con el pulso, datos del sueño, acelerómetro y giroscopio entre otros. Se diseñará un experimento para seleccionar la cantidad de conductores, así como determinar las condiciones y el tiempo durante el cual cada conductor llevará el dispositivo y la frecuencia con la que se recolectará la información para conformar el conjunto de datos.

Posteriormente se realizarán tareas de preprocesamiento, limpieza, extracción de características, así como división del conjunto de datos en entrenamiento y validación. Con el conjunto de datos se realizará el proceso de entrenamiento de algoritmos de aprendizaje de máquina para clasificar los estilos de conducción de cada conductor, para cada algoritmo implementado se obtendrán métricas para identificar la eficacia y/o precisión de dicho algoritmo en función de la tarea de clasificación del estilo de conducción. Finalmente se construirá y presentará un informe con los resultados, hallazgos y análisis realizados.

1.6 Alcance

1.6.1 Alcance del Conjunto de Datos

Si bien este trabajo desarrollará un método para la clasificación de los estilos de conducción y se comparará con trabajos y resultados alrededor del mundo en dicha materia, los datos usados para su desarrollo estarán enmarcados en la zona geográfica comprendida por los municipios del Área Metropolitana del Valle de Aburrá, Antioquia, Colombia. Por ende, algunos de los patrones y hallazgos estarán condicionados por las limitaciones topográficas y propias del sistema vial en dicha zona, así como la normatividad y regulación vigente al momento del desarrollo del experimento.

1.6.2 Alcance de la implementación

Este proyecto desarrollará el servicio de clasificación exponiéndolo como un Servicio Web REST que retornará la clasificación para la muestra consultada, no realizará almacenamiento, consulta o análisis posterior a las clasificaciones realizadas.

1.7 Estado del Arte

Los accidentes de tránsito, las causas de estos y las estrategias de prevención son una temática que genera alto interés en la comunidad científica pues se evidencia un número alto de trabajos, artículos y congresos alrededor de la problemática, así mismo la OMS destina un grupo de trabajo exclusivo en esta área que publica recomendaciones a los gobiernos, para disminuir las cifras de muertes y heridos ocasionados en dichos accidentes. Los trabajos de investigación se orientan principalmente en tres frentes: los relacionados con el vehículo, relacionados con el conductor y por último el estado de las vías, para cada uno de estos aspectos los gobiernos y las instituciones locales desarrollan el ámbito legal y regulatorio sobre el cual operan las estructuras de tránsito, siendo estas últimas en muchos casos las encargadas de reportar las cifras oficiales de muertes y heridos por accidentes de tránsito así como también orquestar el aparato de atención a emergencias sobre el cual cae el peso de dicho problema de salud pública.

Con respecto a los vehículos, se encontraron trabajos relacionados con análisis de sensorial y su correlación con aspectos climáticos (Goberville N et al, 2019), predicción de la velocidad y administración energética (Lui K et al, 2019) y detección de colisiones (Yeong-kang Lai et al, 2018). Se evidencia en estos trabajos el uso de simulación computarizada, así como el uso de técnicas de aprendizaje de máquina e inclusive la aplicación de técnicas de aprendizaje profundo como LSTM y aprendizaje por refuerzo; durante el desarrollo de estos trabajos se evidenció el uso de cámaras, sensores OBD y sensores LiDAR.

Para las vías se ha encontrado un ecosistema rico en investigaciones referentes a la detección de peatones (Obaida M. Al-Hazaimeh et al, 2019), ciclistas (Zhang et al, 2020), señales de tránsito y cambio de carril (J Yang et al, 2017), en estos trabajos se evidencia un uso fuerte de técnicas de aprendizaje de máquina como máquinas de vectores de soporte (SVM), bosques aleatorios y técnicas de agrupación como dbscan,

también el uso de redes neuronales convolucionales (CNN); en conjunto con cámaras, teléfonos inteligentes y sensores láser.

En los aspectos relacionados con el conductor se han encontrado dos tópicos, el primero relacionado con identificar el conductor basado en datos de sensores como teléfonos inteligentes y sensor OBD (Cheng Zang et al, 2016), con el objetivo de realizar personalizaciones o ajustes a las preferencias del vehículo en función de los parámetros del conductor; desde otra perspectiva algunos trabajos buscan clasificar el estilo de conducción del conductor. En muchos de estos trabajos la taxonomía más común es determinar el estilo del conductor en Pasivo/Agresivo (Derick A. Johnson et al, 2011) (Wang H et al, 2022), pero en otros trabajos las tareas de clasificación se han orientado a determinar patrones de conducción anormal (Thi-Hau Nguyen et al, 2020), así como determinar si el conductor tiene conductas distractoras durante la sesión de conducción (Yu-Ting Pang et al, 2018). En el desarrollo de estos trabajos se evidencia el uso de técnicas estadísticas como modelos gaussianos, pca y métodos de ondícula (wavelet); también el uso de algoritmos como DTW y técnicas de aprendizaje de máquina como máquinas de vectores de soporte (SMV), bosques aleatorios, k-means, naive bayes, redes neuronales convolucionales (CNN), LSTM, RCNN.

2. Construcción del conjunto de datos.

Los algoritmos de aprendizaje de máquina usualmente se separan en dos grandes grupos, los algoritmos supervisados y no supervisados, ambos grupos requieren de un conjunto de datos para realizar las tareas de aprendizaje, en el primer grupo (supervisados), los algoritmos usan el conjunto de datos para encontrar un modelo que describa el comportamiento de una variable dependiente con respecto a las variables independientes, para lograr esto cada algoritmo implementa técnicas específicas, la optimización, el descenso del gradiente o coeficientes de entropía son las más usadas, con estas técnicas se establecen unos pesos que acompañan a las variables independientes formando un modelo que al ser ejecutado realizará una predicción (clasificación o regresión) de la variable dependiente. En el segundo grupo (no supervisados), el conjunto de datos no cuenta con variable dependiente y lo que se busca es la mejor agrupación de los datos usando diferentes métricas para medir la “cercanía” entre ellos, diferentes medidas de distancia son usadas para identificar estos grupos. Este conjunto de algoritmos genera una variable adicional al conjunto de datos que indica cual es el grupo al que pertenece cada dato después de ser analizado por el algoritmo, es importante recordar que, si ingresan nueva información al conjunto de datos, el algoritmo no supervisado debe volver a ejecutarse para formar los grupos nuevamente.

Dicho lo anterior resalta la importancia que tienen los datos, tanto su calidad como cantidad para realizar un buen ejercicio de aprendizaje de máquina y que las técnicas aplicadas entreguen un resultado confiable. Para el caso particular de este trabajo, no se cuenta con un conjunto de datos previo, por lo cual es necesario construirlo, para esto se ha desarrollado una aplicación para dispositivos vestibles y se ha diseñado un experimento con el objetivo de recolectar información en tiempo real durante las sesiones de conducción, adicional a esto se han recolectado datos de sueño y actividad física. Estos

datos serán analizados posteriormente y serán el punto de partida para la aplicación de técnicas de aprendizaje de máquina.

2.1 Selección de los dispositivos

El primer paso en la construcción del conjunto de datos es la selección del dispositivo vestible que permita la captura de las señales fisiológicas emitidas por los sensores; para lograr una selección objetiva se han evaluado los dispositivos de la **Tabla 2-1**.

Tabla 2-1 Dispositivos candidatos (Autoría propia)

	DISPOSITIVO
A	Fitbit Versa 2
B	Fitbit Charge 4
C	Fitbit Inspire 2
D	Samsung Galaxy Watch Active 2
E	Samsung Galaxy Watch
F	Samsung Galaxy Fit

Se realiza una comparación de los aspectos técnicos más relevantes **Tabla 2-2**, asignando un punto por cada característica presente, el dispositivo seleccionado será aquel que obtenga el mayor puntaje.

Tabla 2-2 Aspectos técnicos (Autoría propia)

CARACTERÍSTICA	A	B	C	D	E	F
MONITOREO DE ACTIVIDAD	1	1	1	0	1	0
GPS INTEGRADO	0	1	0	1	1	0
ACELERÓMETRO	1	1	1	1	1	1
GIROSCOPIO	0	0	0	1	1	0
BARÓMETRO	0	0	0	1	1	0
SENSOR LUZ	1	0	0	1	1	0
ALTÍMETRO	1	1	0	0	1	0
RITMO CARDIACO CONTINUO	1	1	1	1	1	1
SEGUIMIENTO DEL SUEÑO	1	1	1	1	1	1
WIFI	1	0	0	1	1	0
ENTORNO DE DESARROLLO	1	0	0	1	1	0
PUNTAJE TOTAL	8	6	4	9	11	3

El dispositivo seleccionado es el “E: Samsung Galaxy Watch”, para la construcción del conjunto de datos se han adquirido tres (3) de estos dispositivos que serán usados durante la ejecución del experimento.

2.2 Diseño del experimento

Para la construcción del conjunto de datos debe desarrollarse una aplicación que recolecte datos durante la sesión de conducción, esta aplicación monitoreará de forma continua las siguientes variables:

- Pulso Cardíaco
- Giroscopio
- Aceleración Lineal
- Posición Satelital

Adicionalmente usando aplicaciones existentes en el sistema operativo del dispositivo, se recolectarán muestras de las siguientes variables:

- Pulso Cardíaco
- Sueño
- Actividad Física

2.2.1 Pulso cardíaco

Al iniciar la sesión de conducción, la aplicación inicia el sensor de pulso cardíaco, el cual reporta cada 30 segundos el último dato registrado en pulsos por minuto (bpm), este sensor también reporta cada 10 minutos una muestra del pulso y se almacena por la aplicación de salud del sistema operativo. Estos dispositivos en estudios relacionados han demostrado una precisión superior al 95% (Hwang et al, 2019), lo cual permite concluir que son una medición confiable.

2.2.2 Giroscopio

Al iniciar la sesión de conducción, la aplicación inicia el sensor del giroscopio, este reporta cada 6 segundos el último dato registrado, este sensor calcula la velocidad angular en grados por segundo (deg/s) para los 3 ejes coordenados (x,y,z).

2.2.3 Aceleración lineal

Al iniciar la sesión de conducción, la aplicación inicia el sensor de aceleración lineal, este reporta cada 2 segundos el último dato registrado, este sensor reporta datos sobre el cambio en la aceleración en metros por segundo al cuadrado (m/s^2) para los 3 ejes coordenados (x,y,z) usando el acelerómetro del dispositivo y excluyendo el valor de la gravedad.

2.2.4 Posición satelital

Al iniciar la sesión de conducción, la aplicación inicia el sensor de posición satelital, este reporta cada 60 segundos o cada 100 metros, este sensor reporta datos de posición usando satélites GPS, entregando posición en latitud, longitud y altitud.

2.2.5 Sueño

Para la ejecución del experimento se le pide a cada participante que use el reloj durante sus horas de sueño, permitiendo que la aplicación de salud recolecte datos relacionados a este como son: la eficiencia (cantidad de horas dormido dividido por la cantidad de horas que está en cama), fases del sueño (despierto, sueño ligero, sueño profundo, REM) y cantidad de horas de sueño diaria. Estos datos son procesados automáticamente por el dispositivo capturando los movimientos que realiza el participante mientras está en cama.

2.2.6 Actividad física

El dispositivo usado captura de forma automática la actividad física del usuario, monitoreando la cantidad de tiempo que pasa en reposo, pasos, y entrenamiento físico. Estos datos son almacenados en la aplicación de salud del sistema operativo.

2.3 Selección de los conductores

Para la construcción del conjunto de datos se implementa una técnica de muestreo no probabilística por conveniencia es decir los conductores tienen alguna proximidad con el investigador y aceptan ser incluidos en el experimento entregando los datos requeridos (Tamara Otzen et al, 2017). Como requisitos para la selección de un conductor se

establece que este debe tener disponible un automóvil para realizar como mínimo 2 sesiones de conducción diarias durante 7 días consecutivos, residir en el área metropolitana del valle de Aburrá, vestir el dispositivo suministrado durante el día incluyendo las sesiones de conducción y las horas de sueño.

2.3.1 Cálculo del tamaño de la muestra

Usando como base la encuesta origen destino realizada por el área metropolitana del valle de Aburrá en el año 2017, se estima que la población de conductores de automóviles particulares es de 546.768 y que se realizan 830.261 viajes diarios en automóvil, también se establece que en promedio un viaje dura 35 minutos con una distancia promedio de 7.8 km (AMVA, 2017). Para este experimento la población objetivo son la cantidad de viajes. Para el cálculo de la muestra se usará la ecuación 2.1 de estimación de proporción poblacional (Guillen Serra et al, 2006).

$$n = \frac{Z_{1-\alpha}^2 * p * q}{e^2} \quad (2.1)$$

Siendo:

$1 - \alpha$: Nivel de confianza deseado.

$Z_{1-\alpha}$: Valor de la distribución normal para el nivel de confianza $1 - \alpha$.

p : proporción esperada.

q : $1 - p$.

e : precisión deseada.

n : Tamaño de la muestra.

Para este experimento se han usado los siguientes valores:

$$1 - \alpha = 0.95$$

$$Z_{1-\alpha} = 1.6448$$

$$p = 0.5$$

$$q = 0.5$$

$$e = 0.06$$

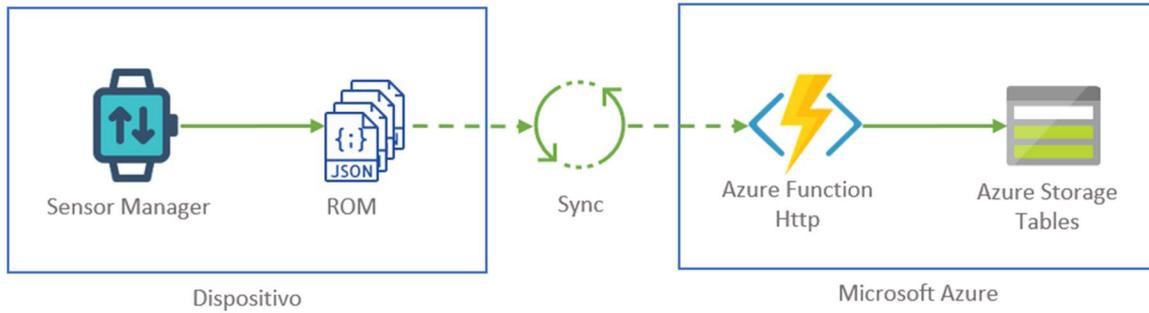
$$188 = \frac{(1.6448)^2 * 0.5 * 0.5}{(0.06)^2}$$

Se han usado los valores 0.5 para p y q usando el criterio de máxima indeterminación, esto entrega el tamaño máximo de la muestra necesaria y una precisión del 0.06 el cual

representa un intervalo de confianza del 6% (Guillen Serra et al, 2006). Al aplicar estos valores en la fórmula se ha obtenido un tamaño muestral de 188 sesiones de conducción.

2.4 Arquitectura para recolección de datos

Para construir el conjunto de datos con las condiciones previamente descritas se propone crear una aplicación para el dispositivo, usando el entorno de desarrollo Microsoft Visual Studio 2019, y las herramientas de desarrollo Tizen proporcionadas por Samsung, estas permiten usar la plataforma .net core y xamarin para acceder a los recursos de hardware del dispositivo, además de entregar un entorno de desarrollo bien documentado y accesible. La aplicación le permite al conductor en tiempo real indicar el inicio de la sesión de conducción, en este punto se encenderán los sensores y se persiste en el almacenamiento un archivo de texto plano usando serialización JSON con la trama de datos de cada sensor, de forma predeterminada la sesión de conducción tiene una duración de 20 minutos, al cumplirse este tiempo la aplicación detiene de forma automática los sensores, esto con el objetivo de no agotar la batería del dispositivo. Al iniciar y finalizar la sesión de conducción el dispositivo emite una vibración lo cual le comunica al conductor que la aplicación ha reconocido la acción, en caso de que el conductor esté realizando una sesión de conducción de más de 20 minutos, este podrá iniciar una sesión nueva de forma consecutiva, si la sesión de manejo termina antes de este tiempo, los datos mostrarán velocidades cercanas a cero y posiciones satelitales con muy poca variación, lo cual permite fácilmente descartar estos datos. Como parte del experimento se le pide a cada conductor que por lo menos una vez al día realice la acción de sincronización, esto le permite a la aplicación subir los archivos almacenados y enviarlos a un servicio en la nube, para esto se ha usado Microsoft Azure como proveedor de nube pública, específicamente los servicios de Azure Function como servicio HTTP y el servicio de Azure Table Storage para almacenar en una estructura No-SQL los archivos serializados en JSON. El ciclo de vida se explica de forma simplificada en la **Figura 2-1**

Figura 2-1 Ciclo de vida simplificado de la aplicación (Autoría propia)

2.5 Características del conjunto de datos

Durante el proceso de recolección de datos se obtienen cerca de 500.000 archivos json con información de los sensores durante las sesiones de conducción. Por cada conductor que participa en el experimento se obtiene un archivo .zip que contiene diferentes archivos de texto plano separado por comas (csv) con la información de las aplicaciones de salud del dispositivo. Estos archivos se consolidan usando las librerías pandas y numpy de Python, para generar un conjunto de datos unificado que permite realizar las tareas de limpieza y transformación de datos necesarias para la ejecución del experimento.

2.5.1 Limpieza y Transformaciones de Datos

Como resultado del proceso de recolección se ha construido un conjunto de datos que relaciona las sesiones de conducción que tienen una duración entre 18 y 22 minutos con los datos de sueño de como mínimo 4 días consecutivos, de la sesión de conducción se han calculado, la cantidad de veces que el conductor frena, los cambios de aceleración, generando 3 grupos (Bajo, Medio y Alto) y los coeficientes en la eficiencia del sueño para cada uno de los 4 días.

Como parte de este proceso se han descartado las sesiones de conducción que tienen la duración suficiente y las sesiones de conducción de los primeros 3 días pues no presentan suficiente información de sueño para realizar el experimento.

El conjunto de datos resultante tiene la siguiente estructura:

Tabla 2-3 Características del conjunto de datos (Autoría propia)

Número de Filas	426
Número de Columnas	30

Formato del Archivo	Archivo Texto Plano	
Separador	; (Punto y Coma)	
Columnas	Nombre	Tipo de Dato
	DriverId	Texto
	SessionID	Entero
	Ef1	Decimal (6)
	Ef2	Decimal (6)
	Ef3	Decimal (6)
	Ef4	Decimal (6)
	Awake1	Entero
	REM1	Entero
	Deep1	Entero
	Ligth1	Entero
	Awake2	Entero
	REM2	Entero
	Deep2	Entero
	Ligth2	Entero
	Awake3	Entero
	REM3	Entero
	Deep3	Entero
	Ligth3	Entero
	Awake4	Entero
	REM4	Entero
	Deep4	Entero
	Ligth4	Entero
	SessionMinutes	Entero
	SessionBrake	Entero
	Acc_Change_Low	Entero
	Acc_Change_Med	Entero
	Acc_Change_High	Entero
	EfAcc	Decimal (6)
	Brakes/Min	Decimal (6)
	Acc_High/Min	Decimal (6)

Figura 2-2 Ejemplo del conjunto de datos (Autoría propia)

DriverId	SessionID	Ef1	Ef2	Ef3	Ef4	Awake1	REM1	Deep1	Ligth1	Awake2	...	Ligth4	SessionMinutes	SessionBrake	Acc_Change_Low	Acc_Change_Med	Acc_Change_High	EfAcc	Brakes/Min
31edc2a3-a0d4-4088-8fc7-680ee67b2e30	267	0.849448	0.730515	0.493485	0.796334	109	204	311	100	142	...	42	19	10	11	2	15	0.464	0.526
31edc2a3-a0d4-4088-8fc7-680ee67b2e30	268	0.756144	0.727273	0.692456	0.850088	129	156	159	85	172	...	36	20	26	7	8	23	0.395	1.300
0f6d37e3-4363-4807-80e8-2ba1fab5e408	93	0.951531	0.9399314	0.923507	0.968326	19	125	223	25	23	...	54	19	13	31	7	4	0.905	0.684
0c54e512-b818-44e7-a57c-f1ef6d7713da	54	0.871570	0.806769	0.445049	0.805455	117	177	510	107	122	...	33	20	19	7	6	12	0.520	0.950
23487ef6-d415-4ece-9a95-26d0d7c5f6be	187	0.828383	0.596014	0.590129	0.794872	104	223	191	88	50	...	33	20	14	7	4	16	0.407	0.700

2.5.2 Agrupación de Datos

Con los datos capturados por los dispositivos se ha ejecutado el algoritmo K-Means, con K=2, se obtienen dos grupos, a cada fila se le asigna la etiqueta entregada por el algoritmo. A continuación, se detallan los grupos de datos haciendo un análisis de los mismos.

2.5.3 Análisis de Datos

Para cada variable del conjunto de datos se realiza una descripción que incluye análisis de cuartiles, media y desviación estándar.

Tabla 2-4 Análisis Exploratorio para el grupo 0 (Autoría propia)

	EF1	EF2	EF3	EF4	AWAKE1	REM1	DEEP1	LIGHT1
COUNT	222.000	222.000	222.000	222.000	222.000	222.000	222.000	222.000
MEAN	0.918	0.923	0.928	0.947	29.590	88.869	201.437	26.293
STD	0.032	0.046	0.119	0.035	22.438	35.961	76.804	16.897
MIN	0.761	0.701	0.258	0.764	4.000	24.000	99.000	0.000
25%	0.907	0.907	0.942	0.931	20.250	66.000	163.250	16.250
50%	0.922	0.931	0.954	0.953	25.500	88.000	190.500	24.000
75%	0.938	0.949	0.967	0.970	30.000	106.750	219.750	32.000
MAX	0.988	1.000	1.000	1.000	144.000	349.000	739.000	110.000

	AWAKE2	REM2	DEEP2	LIGHT2	AWAKE3	REM3	DEEP3	LIGHT3
COUNT	222.000	222.000	222.000	222.000	222.000	222.000	222.000	222.000
MEAN	30.068	98.203	231.856	34.856	25.221	182.928	226.554	59.108
STD	22.020	41.197	72.134	15.703	16.904	40.798	26.990	17.922
MIN	0.000	12.000	51.000	1.000	0.000	59.000	151.000	3.000
25%	20.000	67.000	187.000	25.250	17.000	157.000	207.000	47.000
50%	27.000	101.000	230.500	34.000	22.000	183.000	225.500	59.500
75%	35.000	124.750	284.750	43.000	29.750	209.750	244.000	71.000
MAX	174.000	197.000	429.000	99.000	122.000	314.000	305.000	111.000

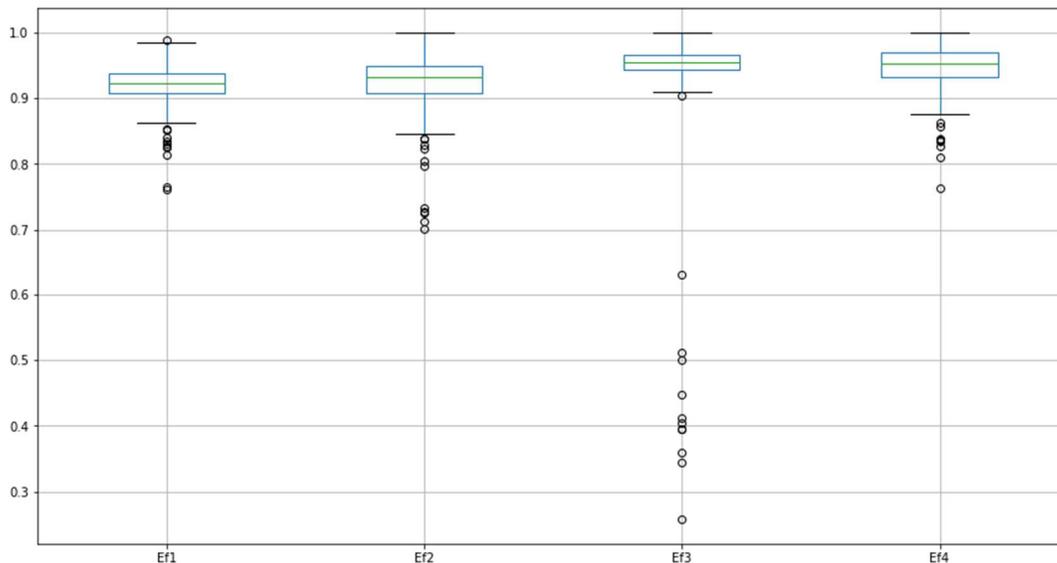
	AWAKE4	REM4	DEEP4	LIGHT4	MINUTES	BRAKE
COUNT	222.000	222.000	222.000	222.000	222.000	222.000
MEAN	27.234	168.180	241.288	59.450	19.541	12.685
STD	20.023	77.712	26.097	29.412	0.621	2.271
MIN	0.000	8.000	69.000	1.000	18.000	8.000
25%	14.250	119.000	237.000	39.250	19.000	11.000
50%	23.000	163.000	242.000	60.000	20.000	13.000
75%	35.000	216.500	248.000	75.000	20.000	14.000
MAX	108.000	378.000	399.000	161.000	21.000	24.000

	ACC_CHANGE LOW	ACC_CHANGE MED	ACC_CHANGE HIGH	EFACC	BRAKES/MIN	ACC_HIGH/MIN
COUNT	222.000	222.000	222.000	222.000	222.000	222.000
MEAN	17.559	8.833	3.464	0.877	0.653	0.177
STD	9.595	3.340	3.967	0.117	0.135	0.201
MIN	0.000	0.000	0.000	0.324	0.000	0.000
25%	10.000	7.000	2.000	0.854	0.579	0.100
50%	16.000	9.000	3.000	0.905	0.650	0.150
75%	23.000	11.000	4.000	0.940	0.722	0.200
MAX	49.000	18.000	26.000	1.000	1.200	1.238

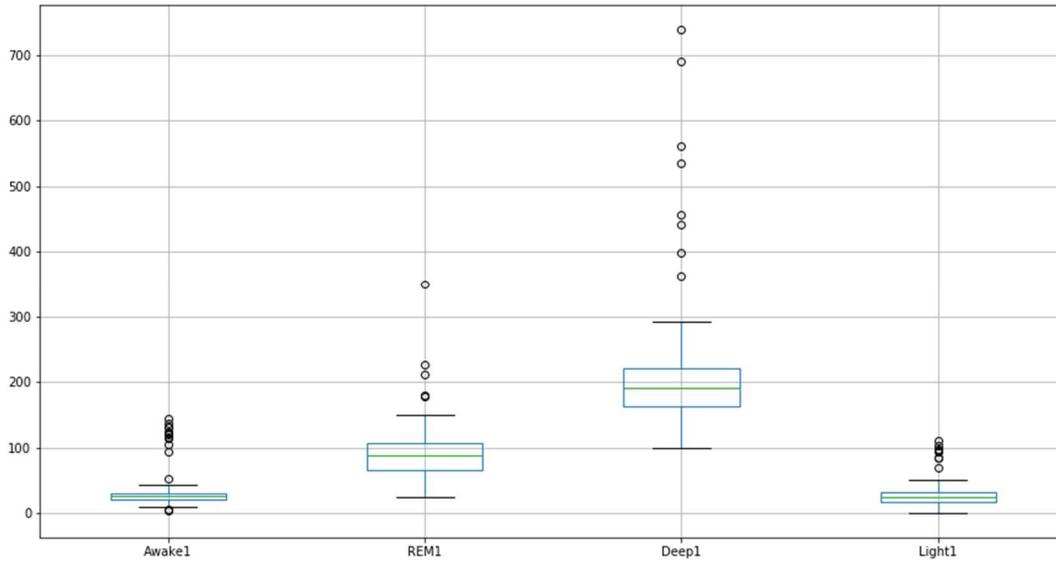
Para cada variable se ha realizado un gráfico de cajas que permite comprobar gráficamente características de las variables bajo estudio.

La eficiencia del sueño se ha calculado para cada noche, usando los datos del sueño, como se observa en el gráfico siguiente, se obtienen cajas compactas con la media centrada, y bigotes cortos, con algunos datos atípicos en la parte inferior, lo que corresponde a una distribución normal de cola izquierda.

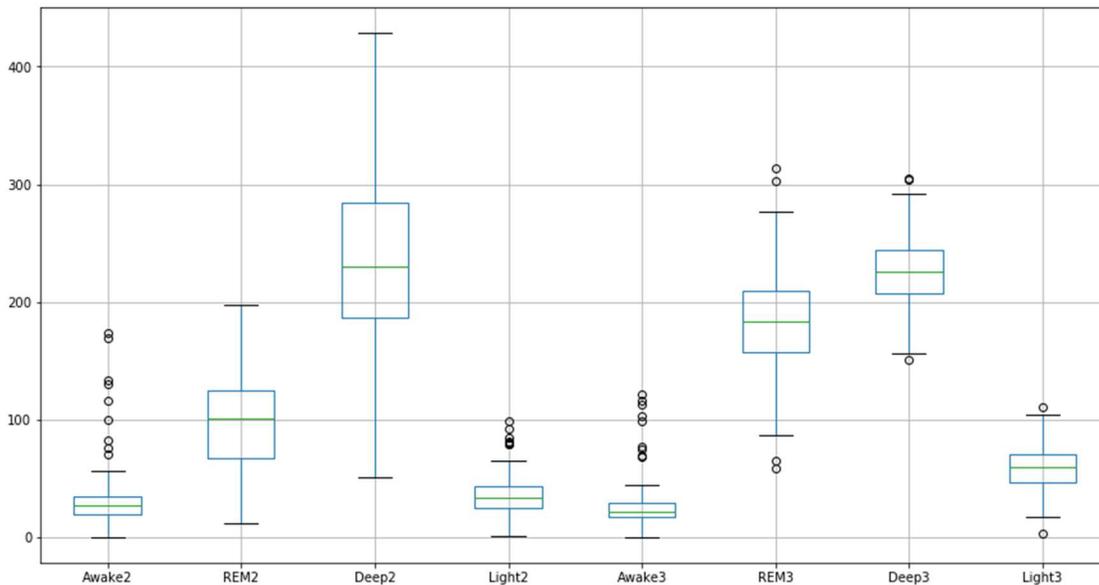
Figura 2-3 Gráfico de cajas para el grupo 0 (1 de 6) (Autoría propia)



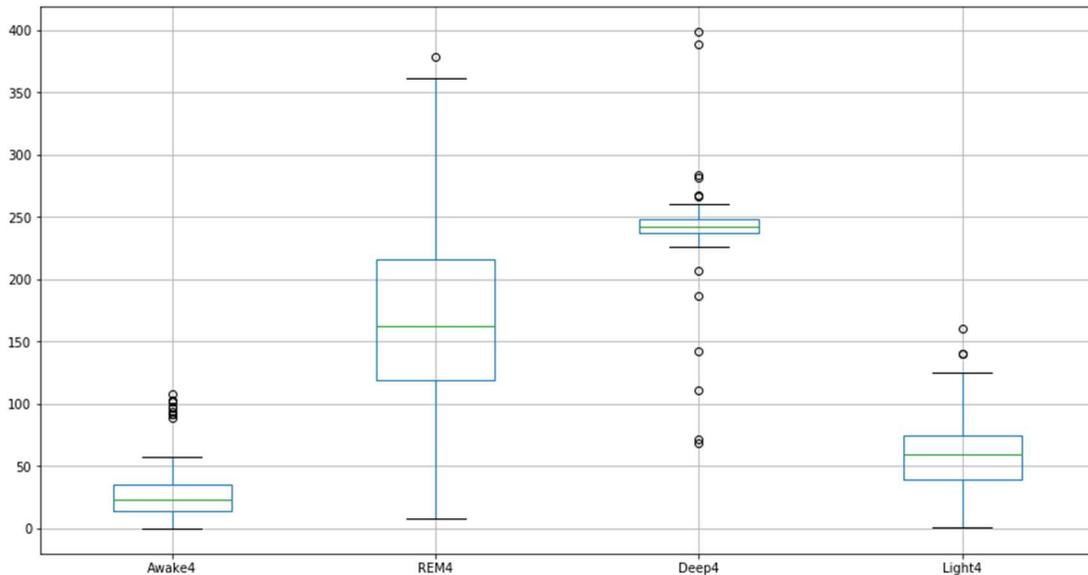
Los datos del sueño de la noche uno, muestran sueño profundo predominante, pocos minutos despiertos y algunos datos atípicos en la parte superior lo que corresponde con una distribución normal con cola derecha.

Figura 2-4 Gráfico de cajas para el grupo 0 (2 de 6) (Autoría propia)

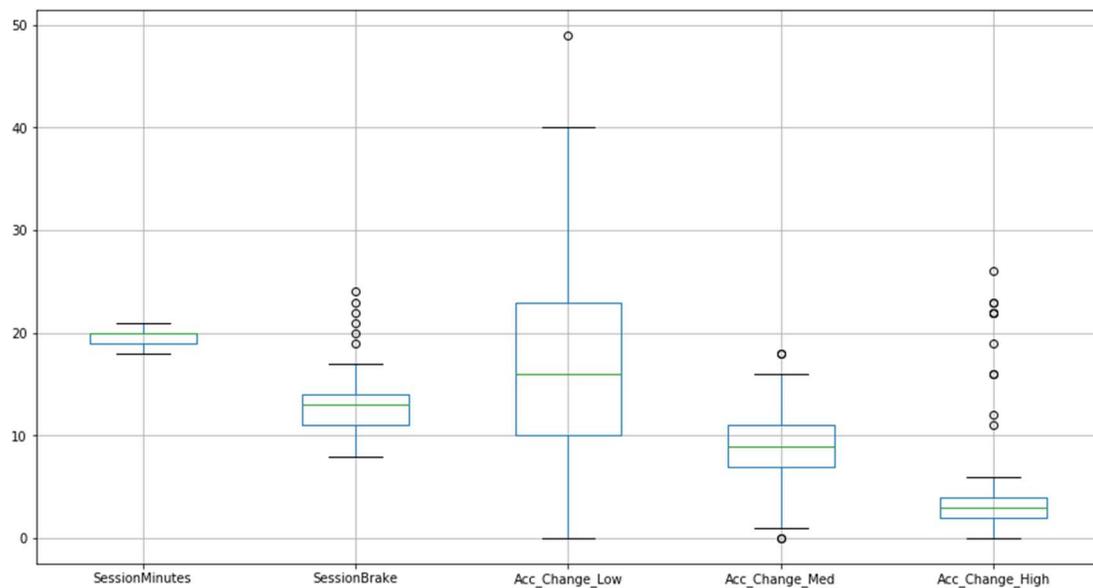
Los datos de sueño, de las noches dos y tres, muestran medias muy similares, lo que indica condiciones de sueño similares y estas corresponden con las mostradas en la noche uno.

Figura 2-5 Gráfico de cajas para el grupo 0 (3 de 6) (Autoría propia)

En la noche 4, el dato predominante es la fase REM, los datos de sueño profundo presentan atípicos en la parte superior e inferior, al igual que en las noches anteriores se evidencian pocos minutos despiertos.

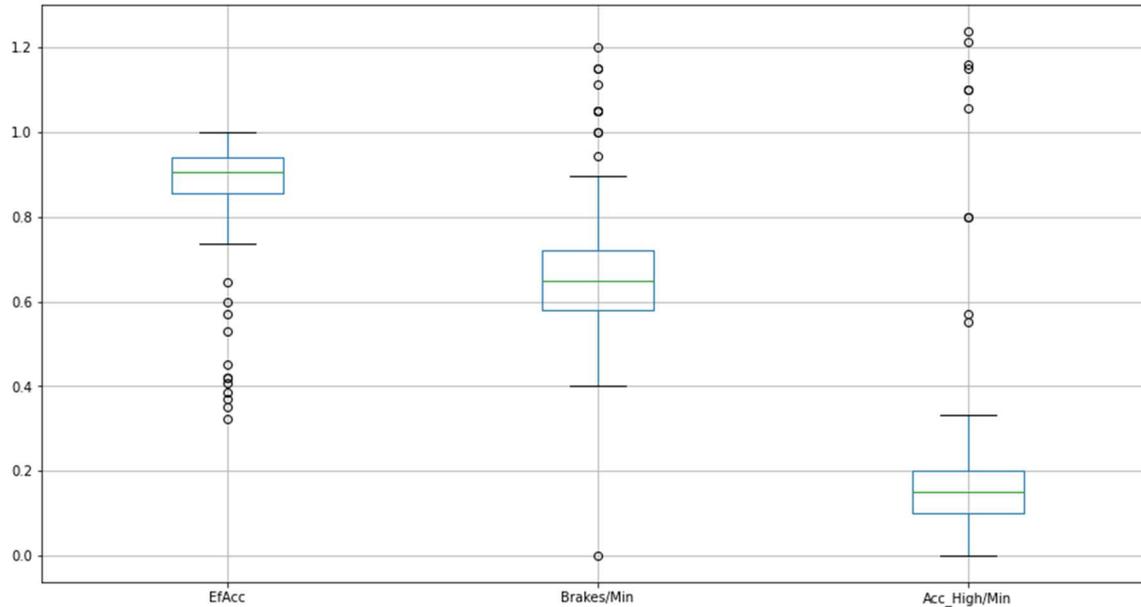
Figura 2-6 Gráfico de cajas para el grupo 0 (4 de 6) (Autoría propia)

En la gráfica siguiente se muestran datos conducción, como la cantidad de veces que se detecta frenado, y los cambios en aceleración capturados por el acelerómetro, predomina los cambios de aceleración lentos con una media bien centrada en la caja. Los cambios de aceleración altos son pocos y se evidencian datos atípicos en la parte superior.

Figura 2-7 Gráfico de cajas para el grupo 0 (5 de 6) (Autoría propia)

Con las variables anteriores se ha calculado coeficientes de eficiencia en el acelerado, también se ha calculado las veces que los conductores de este grupo frenan por minuto, y los cambios de aceleración altos por minuto.

Figura 2-8 Gráfico de cajas para el grupo 0 (6 de 6) (Autoría propia)



En función de los datos visualizados se muestra que el grupo cero tiene conductores que frenan poco y hacen cambios de aceleración graduales, esto asociado a noches con sueño profundo y REM, con altos valores en la eficiencia del sueño, consistentes en las 4 noches con valores medios superiores al 90%.

Se repetido el análisis para el grupo 1, las variables analizadas son las mismas del grupo 0, en este caso el algoritmo K-Means clasificó 204 sesiones de conducción en este grupo.

Tabla 2-5 Análisis para el grupo 1 (Autoría propia)

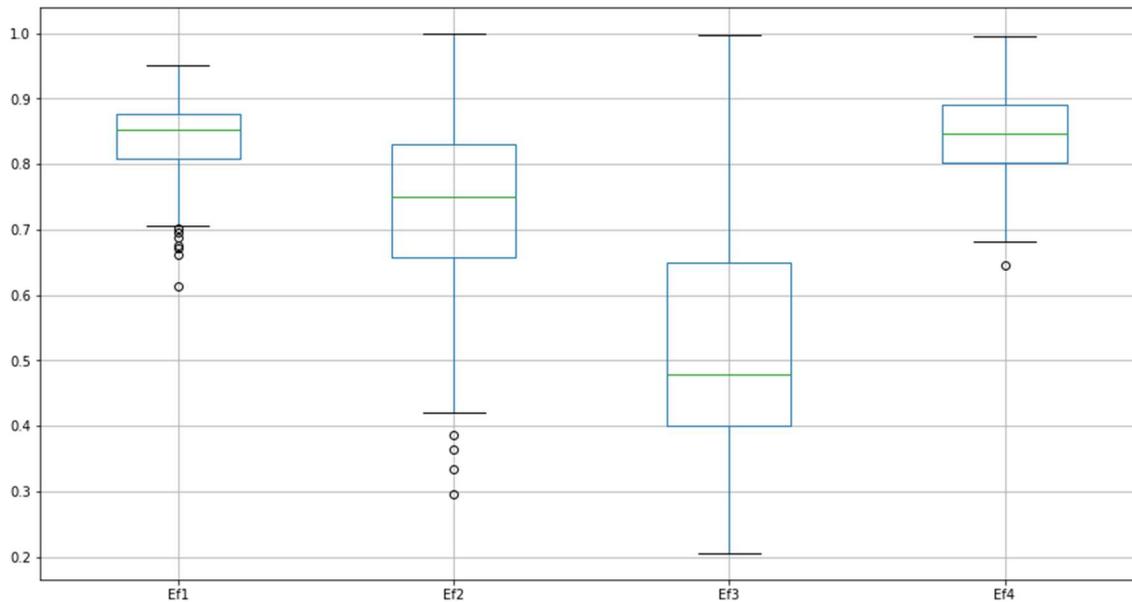
	EF1	EF2	EF3	EF4	AWAKE1	REM1	DEEP1	LIGHT1
COUNT	204.000	204.000	204.000	204.000	204.000	204.000	204.000	204.000
MEAN	0.838	0.740	0.540	0.848	111.461	188.240	357.892	78.730
STD	0.058	0.133	0.201	0.066	25.241	87.707	197.296	33.351
MIN	0.614	0.295	0.205	0.645	19.000	2.000	1.000	2.000
25%	0.808	0.657	0.400	0.803	107.000	134.750	204.000	55.000
50%	0.852	0.750	0.478	0.846	117.000	181.500	330.500	81.000
75%	0.877	0.830	0.650	0.890	126.250	246.000	510.000	101.250
MAX	0.952	1.000	0.997	0.996	165.000	438.000	931.000	158.000

	AWAKE2	REM2	DEEP2	LIGHT2	AWAKE3	REM3	DEEP3	LIGHT3
COUNT	204.000	204.000	204.000	204.000	204.000	204.000	204.000	204.000
MEAN	93.461	139.877	202.260	68.417	82.657	148.025	247.760	59.417
STD	42.784	46.556	51.432	18.099	36.075	71.858	35.684	27.329
MIN	0.000	2.000	79.000	10.000	0.000	0.000	166.000	1.000
25%	64.750	111.000	167.000	59.000	59.000	96.500	222.000	39.750
50%	90.000	143.000	204.000	71.000	84.000	147.000	247.000	61.000
75%	116.000	169.250	238.000	80.250	104.250	199.000	272.250	78.500
MAX	235.000	261.000	390.000	111.000	173.000	318.000	372.000	133.000

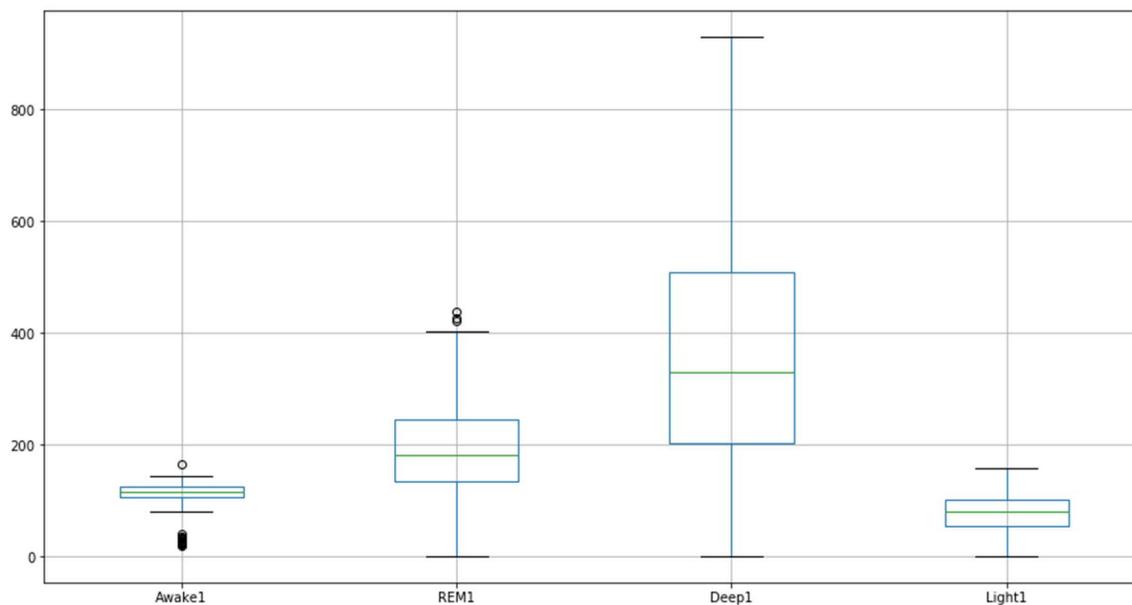
	AWAKE4	REM4	DEEP4	LIGHT4	MINUTES	BRAKE
COUNT	204.000	204.000	204.000	204.000	204.000	204.000
MEAN	93.431	102.069	201.000	45.426	19.931	18.515
STD	19.747	38.397	91.152	18.225	0.874	5.401
MIN	5.000	20.000	5.000	1.000	18.000	2.000
25%	92.000	77.750	136.750	33.000	19.000	14.000
50%	97.000	96.000	207.500	44.000	20.000	19.000
75%	103.000	120.250	254.250	56.250	20.000	22.000
MAX	122.000	301.000	470.000	117.000	22.000	33.000

	ACC_CHANGE LOW	ACC_CHANGE MED	ACC_CHANGE HIGH	EFACC	BRAKES/MIN	ACC_HIGH/MIN
COUNT	204.000	204.000	204.000	204.000	204.000	204.000
MEAN	9.716	5.010	17.275	0.463	0.928	0.866
STD	3.749	2.989	5.779	0.139	0.268	0.289
MIN	3.000	0.000	2.000	0.152	0.105	0.100
25%	8.000	2.000	14.000	0.390	0.726	0.700
50%	9.000	5.000	17.000	0.435	0.950	0.857
75%	11.000	7.000	21.000	0.514	1.101	1.053
MAX	28.000	13.000	31.000	0.946	1.650	1.550

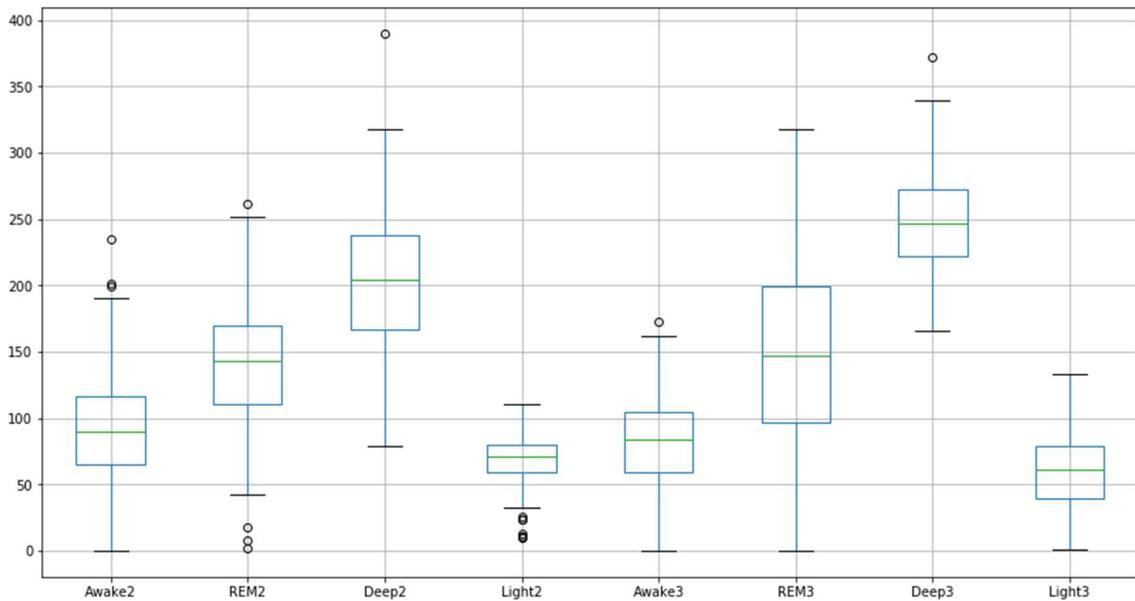
En contraste con el grupo 0 los valores de eficiencia del sueño están por debajo del 90%, en algunos casos como en la noche 3 llegando a valores por debajo del 50%.

Figura 2-9 Gráfico de cajas para el grupo 1 (1 de 6) (Autoría propia)

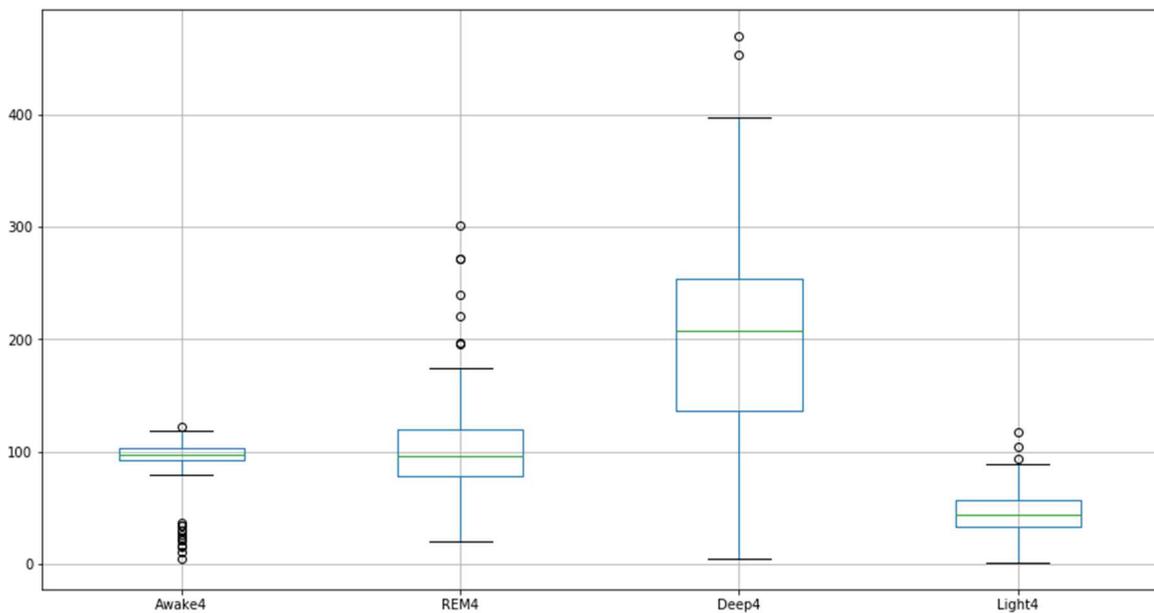
Para la noche uno, se observan valores sueño profundo similares al grupo 0, con valores predominantes en el sueño profundo.

Figura 2-10 Gráfico de cajas para el grupo 1 (2 de 6) (Autoría propia)

Para las noches dos y tres se observan valores de sueño diferentes al grupo cero, en este caso se muestran valores de minutos despiertos muy cercanos a los valores de sueño REM.

Figura 2-11 Gráfico de cajas para el grupo 1 (3 de 6) (Autoría propia)

Para la cuarta noche de sueño los valores predominantes son el sueño profundo, pero los datos de minutos despierto presentan un promedio elevado con respecto a los datos presentado en el grupo cero.

Figura 2-12 Gráfico de cajas para el grupo 1 (4 de 6) (Autoría propia)

Los datos relacionados con la conducción muestran un número más alto de cambios altos en la aceleración, así como un uso más frecuente del frenado.

Figura 2-13 Gráfico de cajas para el grupo 1 (5 de 6) (Autoría propia)

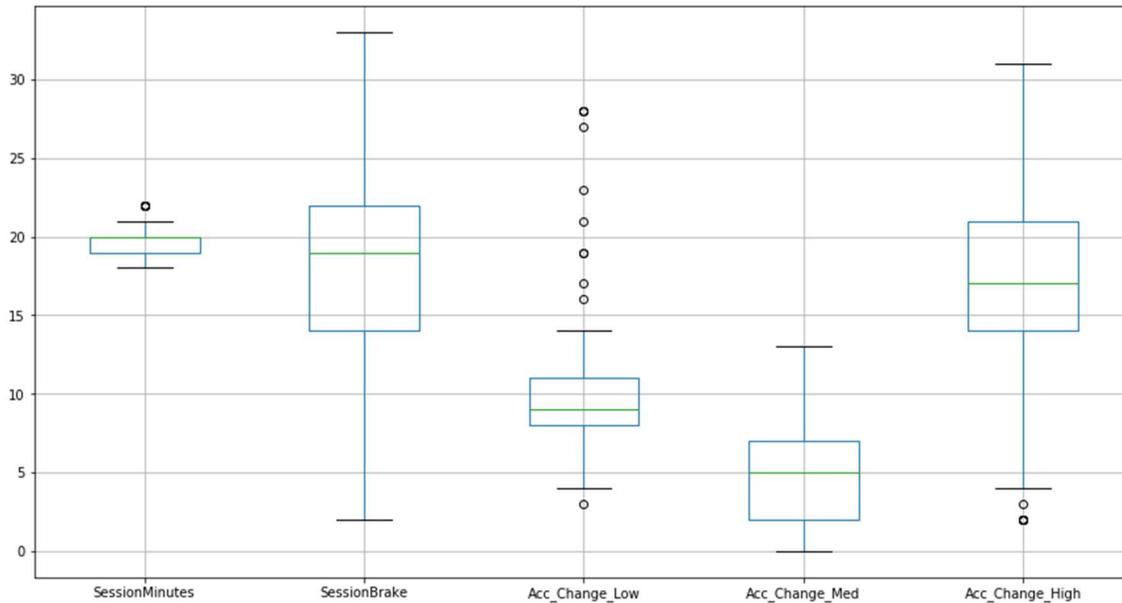
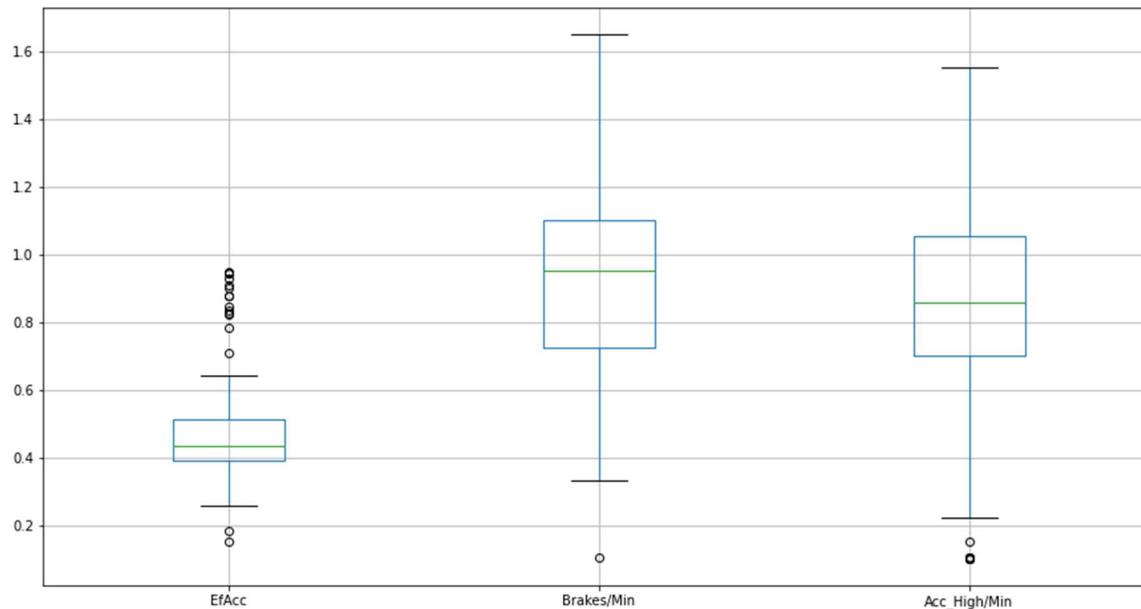


Figura 2-14 Gráfico de cajas para el grupo 1 (6 de 6) (Autoría propia)



Como se observa en las figuras y se corrobora en las tablas de datos, con algunas excepciones las variables presentan condiciones de normalidad, pues se observan cajas

pequeñas con bigotes simétricos y algunas anomalías, esto muestra homogeneidad en el conjunto de datos y permite concluir que en la mayoría de las variables para el grupo el valor más probable es la media.

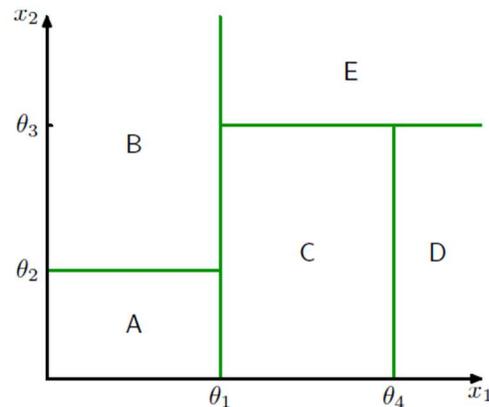
3. Método propuesto para clasificación de conductores.

Como se muestra en el capítulo 2, se ha recolectado información del sueño y variables asociadas a la conducción, adicionalmente se han identificado y diferenciado dos grupos asociando los datos del sueño con los patrones de conducción, el objetivo de este capítulo es tomar la información para entrenar algoritmos de aprendizaje de máquina que reconozcan dichos patrones para predecir un estilo de conducción.

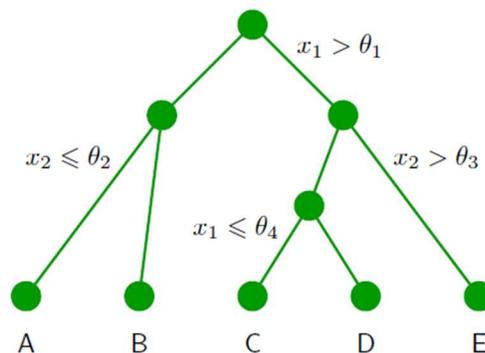
Se han puesto a prueba tres algoritmos de aprendizaje de máquina para clasificación, árboles de decisión, máquinas de vectores de soporte y un clasificador ingenuo de bayes. Se han seleccionado estas tres técnicas pues presentan mecanismos de aprendizaje diferentes, usando aproximaciones matemáticas distintas para entrenar, evaluando en cada uno de ellos las métricas como precisión, sensibilidad, F1-Score. Usando los valores de Falsos Negativos, Verdaderos Negativos, Falsos Positivos y Verdaderos Positivos para calcular dichas métricas, además se han entrenado usando un 70% del conjunto de datos descrito en el capítulo 2, el 30% restante se ha usado para validar las métricas con datos que no han ingresado al proceso de entrenamiento, entregando así un reporte de métricas confiables que permiten validar la capacidad de generalización de los algoritmos implementados.

3.1 Árboles de Decisión

Los árboles de decisión son una técnica de aprendizaje de máquina que divide el espacio de entrada usando una estructura secuencial de decisiones, de esta forma el modelo resultante es una combinación de regiones bien separadas con la cual se realizan las predicciones. (Bishop, 2006) Como ejemplo se muestra la **Figura 3-1** de un espacio de entrada de dos dimensiones que se ha particionado en 5 regiones.

Figura 3-1 Ejemplo de árbol de decisión (Bishop, 2006)

De esta forma el árbol resultante será una secuencia de decisiones binarias basadas en los límites definidos en cada una de las dimensiones, creando subregiones hasta llegar a los nodos finales denominados hojas, los cuales realizarán la predicción como se muestra en la **Figura 3-2**.

Figura 3-2 Árbol de decisión resultante (Bishop, 2006)

Las fronteras que dividen cada región del espacio de entrada, y por consiguiente los nodos en el árbol resultante, son determinadas por un algoritmo que evalúa cada característica en función de una métrica que busca minimizarse, esta puede ser la medida de impureza Gini o la métrica de Entropía. (Géron, 2019)

Para este ejercicio se ha usado la medida de impureza Gini y se ha realizado un ejercicio de validación cruzada, entrenando 5 árboles de decisión independientes con subconjuntos de datos distintos, probando así la robustez de la técnica frente al conjunto de datos.

La precisión del clasificador evaluada con datos de entrenamiento es del **89%**

Usando validación cruzada, se obtienen las siguientes medidas de precisión.

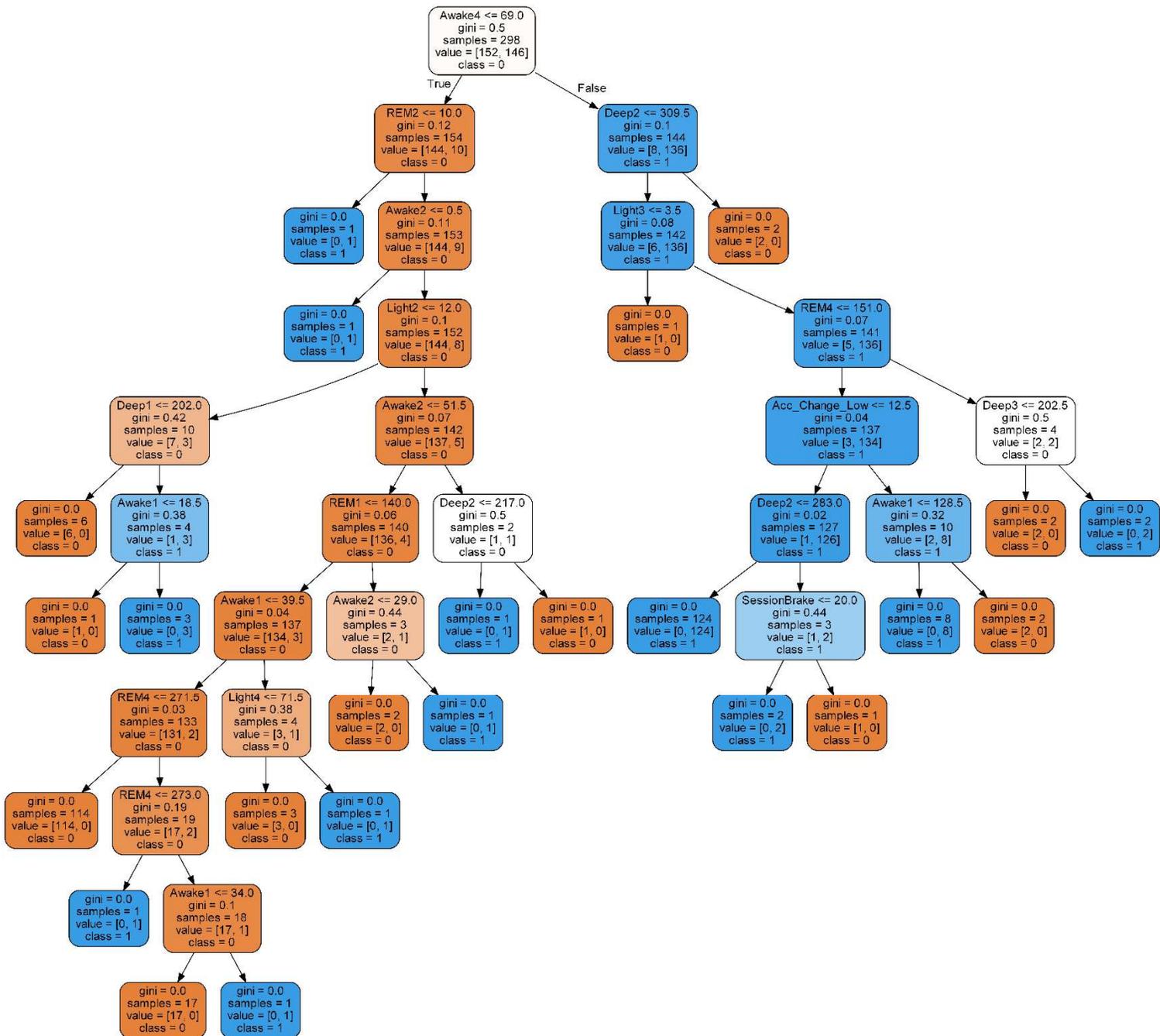
Tabla 3-1 Precisión en validación cruzada para el árbol de decisión.

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
0.900	0.8500	0.9333	0.8644	0.8983

Estas medidas muestran un clasificador robusto, y de buen desempeño para las dos clases.

El árbol resultante tiene las siguientes características.

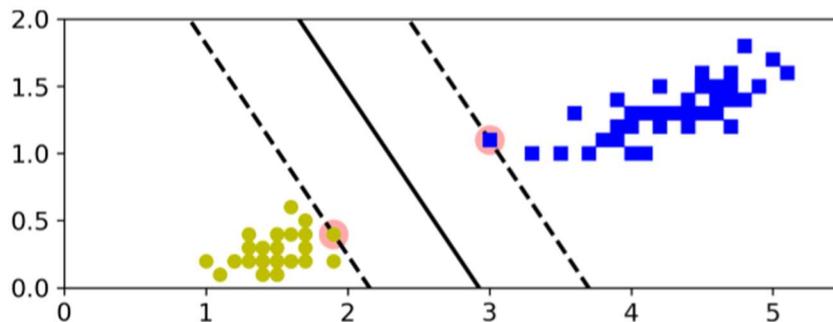
Figura 3-3 Estructura del árbol de decisión (Autoría propia)



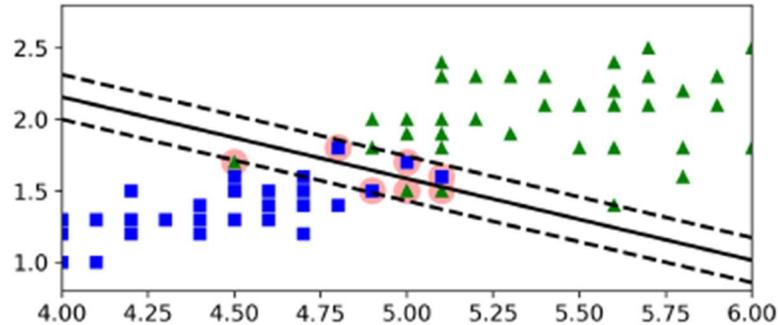
3.2 Máquinas de vectores de soporte (SVM)

Las máquinas de vectores de soporte son una técnica popular del aprendizaje de máquina moderno pues entregan en muchas oportunidades un rendimiento impresionante en tareas de clasificación con conjuntos de datos de tamaño razonable (Marsland, 2015). Las máquinas de vectores de soporte usan como estrategia de aprendizaje optimización numérica, específicamente resolviendo un problema de Programación Cuadrática que busca seleccionar los vectores de soporte que maximicen el margen en la frontera de decisión (Géron, 2019). En la **Figura 3-4**, pueden observarse los elementos mencionados, los cuadrados son los elementos de la clase 1, los elementos circulares pertenecen a la clase 2, en ambos grupos se encuentra sombreado un elemento que es el vector de soporte elegido para dicho grupo, sobre este elemento se traza una línea punteada, la distancia entre las líneas punteadas es el margen, el cual busca maximizarse, la línea continua en medio representa la frontera de decisión de la técnica.

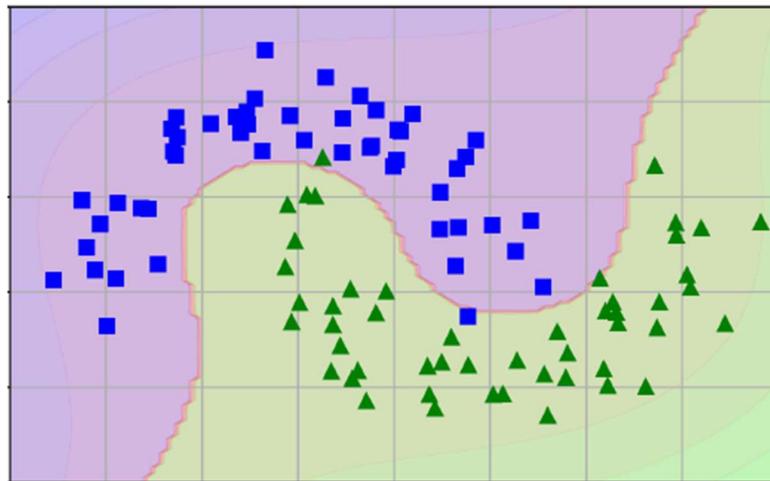
Figura 3-4 Margen y vectores de soporte en una SVM (Géron, 2019)



En la **Figura 3-4** se observa que los vectores de soporte seleccionados están en el límite de la frontera de decisión, es decir, no hay vectores en el margen, pero este escenario es posible encontrarlo y es referenciado en la literatura como “Margen Suave” (Géron, 2019), esto es necesario en muchas aplicaciones de esta técnica, pues un margen estricto podría causar que la técnica no generalice correctamente o que la solución no sea factible pues los datos están mezclados en la frontera de decisión como se observa en la **Figura 3-5**.

Figura 3-5 Margen suave en una SVM. (Géron, 2019).

El margen suave es un balance entre la mayor distancia posible y el menor número de vectores invadiendo el margen (violaciones del margen), en las implementaciones más comunes como la de la librería `skit-learn`, el hiperparámetro es “C” y es un número decimal, cuando este número es pequeño se amplía el margen, pero incrementan el número de nodos que violan dicho margen. Las SVM predeterminadamente trazan una frontera de decisión lineal, cuando se desea aplicar la técnica a conjuntos de datos que no son linealmente separables, deben aplicarse transformaciones en las variables independientes, las implementaciones más comunes aplican transformaciones polinomiales o con funciones de similitud como la RBF (Gaussian Radial Basis Function), esta transformación puede especificarse vía el hiperparámetro “kernel” en la librería `skit-learn`.

Figura 3-6 SVM kernel RBF (Géron, 2019)

Como se observa en la **Figura 3-6** (Géron, 2019), la SVM implementada encuentra una frontera no lineal para el conjunto de datos, adicionalmente se pueden observar algunas violaciones al margen, lo que permite que la técnica generalice bien.

Para este ejercicio usaron los siguientes hiperparámetros:

Tabla 3-2 Hiperparámetro de la SVM (Autoría propia)

HIPERPARÁMETRO	VALOR
C	1.0
KERNEL	RBF
DEGREE	3
GAMMA	Scale
TOL	0.001

La precisión del clasificador evaluada con datos de entrenamiento es del **94%**

Adicionalmente se realizó el entrenamiento usando validación cruzada, entrenando 5 máquinas de vectores de soporte para comprobar la robustez de la técnica.

Tabla 3-3 Precisión en validación cruzada para la máquina de vectores de soporte. (Autoría propia)

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
0.9666	0.9666	0.91666	0.9152	0.9322

3.3 Clasificador Ingenuo de Bayes – Gaussiano

Esta técnica hace parte de los algoritmos de aprendizaje supervisados, además de pertenecer a la familia de los clasificadores probabilísticos pues usa el teorema de las probabilidades a posteriori de bayes para realizar clasificaciones (Marsland, 2015). Se considera una técnica ingenua pues el cálculo de las probabilidades a posteriori se realiza suponiendo que las características evaluadas en el conjunto de datos tienen independencia condicional (Marsland, 2015). Dicha suposición es importante pues minimiza el esfuerzo computacional requerido para entrenar la técnica, pero es difícil encontrarla en los conjuntos de datos, pues lo usual es que las características presenten algún tipo de dependencia entre ellas, aun así, otro supuesto fuerte que acompaña a esta técnica es la estimación de los parámetros para el cálculo de las probabilidades a priori, es común encontrar la implementación Gaussiana la cual asume normalidad en la distribución de los datos, el promedio y la desviación son estimados usando la máxima verosimilitud. A pesar de los supuestos fuertes que acompañan esta técnica, resulta ser un buen clasificador (Zhang, 2004).

Como se ha mostrado en la sección **2.5.3** es válido suponer normalidad en las variables del conjunto de datos, pues el cuartil del 50% está muy próximo a la media teórica y en los gráficos de cajas pueden observarse bigotes simétricos con pocos datos de anomalías.

La precisión del clasificador evaluada con datos de entrenamiento es del **94%**. Adicionalmente se realizó el entrenamiento usando validación cruzada, entrenando 5 máquinas de vectores de soporte para comprobar la robustez de la técnica.

Tabla 3-4 Precisión en validación cruzada para el clasificador ingenuo de bayes. (Autoría propia)

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
0.9167	0.9333	0.9333	0.9830	0.9152

4. Validación de los métodos de clasificación

El proceso de validación de las técnicas de aprendizaje de máquina es vital para verificar la capacidad predictiva de las mismas, este proceso varía según el tipo de algoritmo implementado, para este caso al tratarse de un ejercicio de clasificación se realiza usando métricas como precisión, sensibilidad y F1-Score; tomando los valores de falsos negativos, falsos positivos, verdaderos positivos y verdaderos negativos de la matriz de confusión.

4.1 Matriz de Confusión

La matriz de confusión es un método que permite validar si el modelo entrenado para un resolver un problema de clasificación lo hace o no de buena manera (Marsland, 2015). En términos sencillos la matriz de confusión es una matriz cuadrada que permite comparar los resultados obtenidos versus los resultados esperados en términos de la clasificación, para construirla se ubican los valores esperados sobre el eje y o las filas, y sobre las columnas o eje x, se ubican los valores de la predicción emitida por el modelo. De esta forma se cuentan los valores cuya predicción es igual al valor esperado en la clasificación y se ubican sobre la diagonal principal de la matriz.

Para este trabajo se ha desarrollado un modelo de clasificación binario, es decir de dos clases. El grupo 0 corresponde a los conductores pasivos y el grupo 1 corresponde a los conductores agresivos. De esta forma la matriz de confusión considera los siguientes conceptos.

4.1.1 Verdaderos Positivos

En inglés referenciados como True Positives o TP, son las sesiones de conducción que el modelo clasifica en el grupo 0 y que efectivamente en el conjunto de datos están etiquetados como grupo 0.

4.1.2 Verdaderos Negativos

En inglés referenciados como True Negatives o TN, son las sesiones de conducción que el modelo clasifica en el grupo 1 y que efectivamente en el conjunto de datos están etiquetados como el grupo 1.

4.1.3 Falsos Positivos

En inglés referenciados como False Positives o FP, son las sesiones de conducción que el modelo clasifica en el grupo 0 y que en el conjunto de datos de prueba están etiquetados con el grupo 1.

4.1.4 Falsos Negativos

En inglés referenciados como False Negatives o FN, son las sesiones de conducción que el modelo clasifica en el grupo 1 y que en el conjunto de datos de prueba están etiquetados con el grupo 0.

4.2 Métricas de Validación

Para los problemas de clasificación se usan métricas que aportan perspectivas diferentes del clasificador, como la sensibilidad del clasificador hacia una clase y la exactitud global la técnica con respecto a todas las clases. Estas métricas se construyen usando los valores de la matriz de confusión.

4.2.1 Exactitud (Accuracy)

Es la suma del número de verdaderos positivos y verdaderos negativos divididas por el número total de muestras. (Marsland, 2015)

$$Accuracy = \frac{TP+FP}{TP+FP+TN+FN} \quad (4-1)$$

Esta métrica como lo muestra la ecuación 4-1, indica la ratio de aciertos del modelo sobre el total de muestras del conjunto de prueba.

4.2.2 Precisión (Precision)

Es la ratio de verdaderos positivos dividido por el total de elementos positivos de la matriz de confusión. (Marsland, 2015)

$$Precision = \frac{TP}{TP+FP} \quad (4-2)$$

Esta métrica como lo muestra la ecuación 4-2, indica la proporción de predicciones positivas en el conjunto de prueba que resultaron acertadas con respecto a todas las predicciones positivas hechas por el modelo.

4.2.3 Exhaustividad - (Recall)

Es la ratio de verdaderos positivos dividido por la suma de verdaderos positivos y falsos negativos. (Marsland, 2015)

$$Recall = \frac{TP}{TP+FN} \quad (4-3)$$

Esta métrica como lo muestra la ecuación 4-3, indica la proporción de predicciones positivas en el conjunto de prueba que resultaron acertadas con respecto al total de elementos potencialmente positivos.

4.2.4 Valor F1 - (F1 Score)

Métrica que pondera la precisión y la exhaustividad para tener un elemento simple de comparación.

$$F_1 = 2 \frac{Precision * Recall}{Precision + Recall} \quad (4-4)$$

4.3 Validación de los Algoritmos Implementados

Para cada uno de los algoritmos implementados se ha realizado un proceso de validación en un conjunto de prueba que no fue usado para entrenar dichos algoritmos, dicho conjunto de datos representa el 30% del conjunto total de datos obtenido.

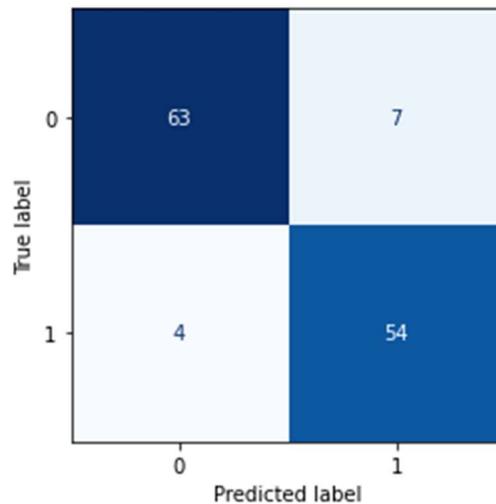
El conjunto de datos de validación tiene 128 sesiones de conducción, dividido de la siguiente forma: 70 sesiones están etiquetadas en el grupo 0 (estilos de conducción pasivo) y 58 sesiones tienen la etiqueta 1 (estilo de conducción agresivo).

En el proceso de validación se ha construido la matriz de confusión, además se ha generado un reporte con las métricas de precisión, exhaustividad, valor F1, exactitud.

4.3.1 Árbol de Decisión

Para el árbol de decisión se obtiene la matriz de confusión mostrada en la **Figura 4-1**. La diagonal principal muestra 117 sesiones de conducción que son correctamente clasificadas, y solo 11 valores que representan errores al momento de la clasificación.

Figura 4-1 Matriz de confusión del árbol de decisión (Autoría propia)



Adicionalmente se ha generado el reporte de la **Tabla 4-1** usando la función “classification_report” de la librería sklearn.

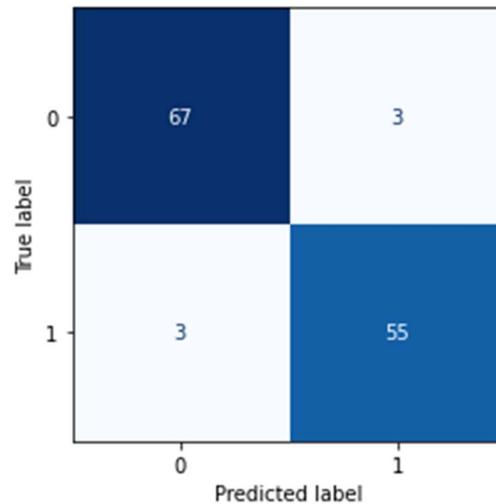
Tabla 4-1 Reporte de métricas para el árbol de decisión (Autoría propia)

	PRECISION	RECALL	F1-SCORE	SUPPORT
0 (PASIVO)	0.94	0.90	0.92	70
1 (AGRESIVO)	0.89	0.93	0.91	58
ACCURACY			0.91	128
MACRO AVG	0.91	0.92	0.91	128
WEIGHTED AVG	0.92	0.91	0.91	128

4.3.2 Máquina de vectores de soporte

Para la máquina de vectores de soporte se obtiene la matriz de confusión mostrada en la **Figura 4-2**. La diagonal principal muestra 122 sesiones de conducción que son correctamente clasificadas, y solo 6 valores que representan errores al momento de la clasificación.

Figura 4-2 Matriz de confusión para SVM (Autoría propia)



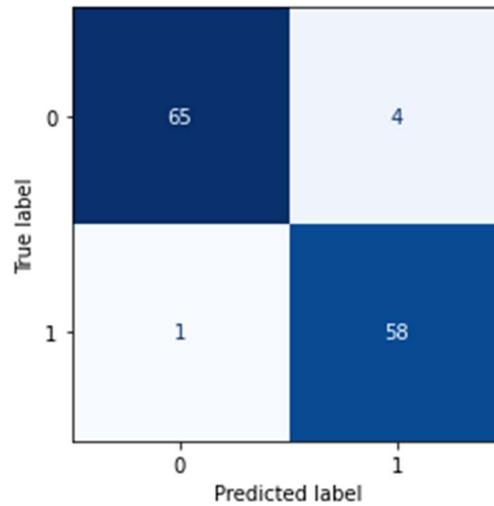
Adicionalmente se ha generado el reporte de la **Tabla 4-2** usando la función “classification_report” de la librería sklearn.

Tabla 4-2 Reporte de clasificación para la SVM (Autoría propia)

	PRECISION	RECALL	F1-SCORE	SUPPORT
0 (PASIVO)	0.96	0.96	0.96	70
1 (AGRESIVO)	0.95	0.95	0.95	58
ACCURACY			0.95	128
MACRO AVG	0.95	0.95	0.95	128
WEIGHTED AVG	0.95	0.95	0.95	128

4.3.3 Clasificador Ingenuo de Bayes

Para la máquina de vectores de soporte se obtiene la matriz de confusión mostrada en la **Figura 4-3**. La diagonal principal muestra 123 sesiones de conducción que son correctamente clasificadas, y solo 5 valores que representan errores al momento de la clasificación.

Figura 4-3 Matriz de confusión para el clasificador ingenuo de bayes (Autoría propia)

Adicionalmente se ha generado el reporte de la **Tabla 4-3** usando la función “classification_report” de la librería sklearn.

Tabla 4-3 Reporte de clasificación para el clasificador ingenuo de bayes (Autoría propia)

	PRECISION	RECALL	F1-SCORE	SUPPORT
0 (PASIVO)	0.98	0.94	0.96	70
1 (AGRESIVO)	0.94	0.98	0.96	58
ACCURACY			0.96	128
MACRO AVG	0.96	0.96	0.96	128
WEIGHTED AVG	0.96	0.96	0.96	128

4.4 Comparación de los métodos

Para la comparación de estos métodos se usará las métricas F1 y Exhaustividad promediada en todas las clases, adicionalmente se compararán los métodos desarrollados en este trabajo con otros métodos encontrados en la literatura que tienen como objetivo el reconocimiento de los estilos de conducción como el desarrollado por Wang H et al, y el desarrollado por Yu-Ting Pang et al.

Tabla 4-4 Comparación de los métodos de clasificación (Autoría propia)

Método	F1-Score
Árbol de decisión	0.9100
SVM	0.9500
Clasificador ingenuo de bayes	0.9600
Dedistractednet	0.9736
Wang H – CNN	0.9002
Wang H – LSTM	0.9042
Wang H – GRU	0.8731

Como se observa en la **Tabla 4-4** los 3 primeros modelos resaltados en color amarillo corresponden a los desarrollados en este trabajo, el siguiente modelo llamado “Dedistractednet” no reconoce los estilos de conducción en Pasivo/Agresivo si no que reconoce y clasifica al conductor en 7 clases diferentes, estableciendo una relación del estilo de conducción con las conductas distractoras al conducir. Una de las clases entrenadas es el conductor sin distractores y donde la posición de conducción es correcta, las otras 6 clases corresponden a conductas distractoras que afectan el estilo de conducción. Los modelos desarrollados por Wang H et al, responden al igual que este trabajo a la clasificación de estilos de conducción en Pasivo/Agresivo usando técnicas de aprendizaje profundo, pues las entradas usadas por estos modelos son imágenes del conductor y datos de algunos sensores instalados en el vehículo. Para los métodos desarrollados por Wang H et al, se ha tomado el valor más alto reportado para la métrica F1-Score de todas las configuraciones de reglas de ensamble probadas en cada una de las técnicas y como se muestra en la **Tabla 4-5** se ha tomado el valor de exhaustividad (recall) asociado a dicho al valor de F1-Score reportado en la **Tabla 4-4**.

Puede observarse que los clasificadores desarrollados en este trabajo muestran un buen rendimiento en la clasificación de los estilos de conducción comparados con métodos pares. De los métodos desarrollados en este trabajo el clasificador ingenuo de bayes a pesar de sus supuestos fuertes logra el puntaje más alto en la métrica F1-Score, aun así el árbol de decisión tiene un rendimiento superior al 90% y este tiene particularmente a favor que es una técnica de caja blanca, pues puede conocerse mediante su diagrama (**Figura 3-3**) cada una de los límites de decisión aprendidos de los datos, lo que permite explicar y razonar acerca de cada una de las predicciones realizadas por la técnica.

Tabla 4-5 Métrica de exhaustividad para los métodos evaluados (Autoría propia)

Método	Recall
Árbol de decisión	0.9100
SVM	0.9500
Clasificador ingenuo de bayes	0.9600
Dedistractednet	0.9736
Wang H – CNN	0.8867
Wang H – LSTM	0.8967
Wang H – GRU	0.8833

5. Conclusiones y recomendaciones

5.1 Conclusiones

La problemática de la accidentalidad vial es compleja y tiene muchas aristas, este trabajo propone como estrategia de mitigación el uso de dispositivos vestibles en este caso relojes inteligentes con el objetivo de clasificar el estilo de conducción de una sesión de conducción. Hemos encontrado que dicho estilo de conducción está fuertemente correlacionado con la eficiencia del sueño, contrario a como sucedió con el pulso cardiaco, pues esta correlación resultó ser una tarea más compleja y por ende fue descartada como variable del conjunto de datos. Para representar mejor el estilo de conducción hemos usado la taxonomía Pasivo/Agresivo, usando los cambios en la aceleración como el elemento central de análisis, así pues, cambios repentinos en la aceleración se traducen en comportamientos agresivos y cambios graduales representan un estilo de conducción pasivo. La obtención de un conjunto de buena calidad ha sido fundamental para lograr un buen desempeño en las tareas de aprendizaje de máquina, el primer paso hacia este objetivo fue diseñar detalladamente una metodología para la recolección de datos, esto incluye el desarrollo de una aplicación móvil para el reloj inteligente que recolecte la mayor cantidad de información de los sensores y la almacene de tal forma que pudiese recuperarse vía internet para consolidarla, también fue importante validar con algunos conductores el proceso de recolección de datos y verificar que todos los mecanismos de transmisión de datos hacia el almacén en la nube funcionaban correctamente, esto permitió que la ejecución de la fase de recolección de datos se realizara sin ningún contratiempo, obteniendo la mayor cantidad de información de los conductores. En la fase de entrenamiento de las técnicas de aprendizaje de máquina fue vital seleccionar técnicas que usaran estrategias de aprendizaje diferentes pues esto permitió abordar el problema desde diferentes perspectivas validando así la robustez del conjunto de datos construido.

5.2 Recomendaciones

Como propuestas de trabajos futuros a desarrollar relacionados con esta temática se plantea usar los datos del giroscopio para establecer como el cambio angular en el movimiento de los brazos puede ser un descriptor complementario del estilo de conducción, validar la pertinencia los datos de actividad física como regresores complementarios para las tareas de clasificación del estilo de conducción, relacionar los datos del conductor, su estilo de conducción y las condiciones de tráfico en tiempo real. Por último, las tres técnicas implementadas usan estrategias de aprendizaje totalmente diferentes y para cada una las métricas obtenidas son superiores al 90%, particularmente el árbol de decisión al ser una técnica de caja blanca representa un elemento fundamental para continuar explorando las diferentes variables independientes y medir su impacto sobre la clasificación del estilo de conducción.

6. Bibliografía

- AMVA, Á. (2017). *Encuesta Origen Destino*. Medellín: Área Metropolitana del Valle de Aburrá.
- Bishop. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Cheng Zang et al. (2016). Driver Classification Based on Driving Behaviors. *IUI '16: Proceedings of the 21st International Conference on Intelligent User Interfaces* (págs. 80-84). Sonoma, CA, USA: ACM.
- DANE. (2018). *PIB - FAQ*. Recuperado el 20 de 07 de 2020, de https://www.dane.gov.co/files/faqs/faq_pib.pdf
- Derick A. Johnson et al. (2011). Driving Style Recognition Using a Smartphone as a Sensor Platform. *International IEEE Conference on Intelligent Transportation Systems* (págs. 1609-1615). Washington: IEEE.
- García de Quevedo N. F et al. (2018). Determinación de los componentes entrópicos de la accidentalidad: el trinomio vehículo/usuario/camino en la metrópoli de Guadalajara, México. *Revista Tecnura*, 22(55), 51-65.
doi:<https://doi.org/10.14483/22487638.13245>
- Géron, A. (2019). *Hands On Machine Learning with Scikit Learn, Keras and Tensorflow*. O'Reilly.
- Goberville N et al. (2019). Analysis of LiDAR and Camera Data in Real-World Weather Conditions for Autonomous Vehicle Operations. *SAE Technical Paper*.
- Guillen Serra et al. (2006). Técnicas de muestreo. Sesgos más Frecuentes. En *Métodos estadísticos para enfermería nefrológica* (págs. 121-132). SEDEN.
- Hwang et al. (2019). Assessing Accuracy of Wrist-Worn Wearable Devices in Measurement of Paroxysmal Supraventricular Tachycardia Heart Rate. *Korean Circulation Journal*, 437-445.
- J Yang et al. (2017). Traffic Sign Recognition Based on Deep Convolutional Neural Network. *Springer Nature Singapore Pte Ltd*, 685-695.

- Ley 1122. (2007). Colombia: Congreso de la Republica de Colombia.
- Ley 769. (2002). Código Nacional de Transito. Colombia: Congreso de la Republica de Colombia.
- Lui K et al. (2019). Vehicle Velocity Prediction and Energy Management Strategy Part 1: Deterministic and Stochastic Vehicle Velocity Prediction Using Machine Learning. *SAE Technical Papper*.
- Mahafroz et al. (2015). Prevalence pattern of road traffic accidents in developing countries: a systematic review. *International Journal of Medical Science and Public Health.*, 4(10), 1324-1333.
- Marsland, S. (2015). *Machine Learning An algorithmic perspective, Second Edition*. Boca Raton, FL: Taylor & Francis Group, LLC.
- MESUT ÇIÇEK. (2015). WEARABLE TECHNOLOGIES AND ITS FUTURE APPLICATIONS. *International Journal of Electrical, Electronics and Data Communication*, 3(4), 45-50.
- Ministerio de Salud de Colombia. (Febrero de 2016). Años de Vida Potencial Perdidos en los accidentes de transporte terrestre en Colombia, 2005 -2013. *Boletín ASIS - Análisis de Situación de Salud*, 2(4), 1-7.
- Obaida M. Al-Hazaimh et al. (2019). Automated Pedestrian Recognition Based on Deep Convolutional Neural Networks. *International Journal of Machine Learning and Computing*, 9(5).
- OMS. (2009). *Global status report on road safety: time to action*. Geneva: Organización Mundial de la Salud.
- OMS. (2015). *Global status report on road safety*. Geneva: Organización Mundial de la Salud.
- OMS. (2018). *Global status report on road safety 2018*. Geneva: Organización Mundial de la Salud.
- Porras Sandra et al. (2017). Años potenciales de vida perdidos por incidentes viales de motociclistas, Medellín, 2009-2012: un análisis espacial por sitio de la ocurrencia. *Revista Médica de Risaralda*, 23(1), 22-29.
- Tamara Otzen et al. (2017). Técnicas de Muestreo sobre una Población a Estudio. *Int, J Morphol*, 227-232.
- Thi-Hau Nguyen et al. (2020). Dynamic Basic Activity Sequence Matching Method in Abnormal Driving Pattern Detection Using Smartphone Sensors. *MDPI*.

- Wang H et al. (2022). A Recognition Method of Aggressive Driving Behavior Based on Ensemble Learning. *Sensors*, 22(2), 644.
- Yeong-kang Lai et al. (2018). Intelligent Vehicle Collision Warning System Based on A Deep Learning Approach. *IEEE International Conference on Consumer Electronics*. Taiwan: IEEE.
- Yu-Ting Pang et al. (2018). An Advanced Deep Framework for Recognition of Distracted Driving Behaviors. *IEEE 7th Global Conference on Consumer Electronics* (págs. 802-803). Taiwan: IEEE.
- Zhang et al. (2020). Cyclist detection and tracking based on multi-layer laser scanner. *Human-centric Computing and Information Sciences - Open Springer*, 1-18.
- Zhang, H. (2004). The Optimality of Naive Bayes. *Seventeenth International Florida Artificial Intelligence Research Society Conference*. 2, pág. 562. Miami Beach, FL: AAAI Press.