

UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS



Presentado por:

ELIZABETH FLORES HUAMANÍ

**MODELO PREDICTIVO DE LOS ESTILOS DE VIDA
PARA ESTUDIANTES UNIVERSITARIOS BASADO
EN TÉCNICAS DE MACHINE LEARNING**

Asesor:

DRA. NORMA LORENA CATACORA FLORES

**TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE
INGENIERÍA DE SISTEMAS**

ANDAHUAYLAS – APURÍMAC – PERÚ

Octubre, 2022

APROBACIÓN DEL ASESOR

COPIA DEL ACTA DE SUSTENTACIÓN

APROBACIÓN DEL JURADO EVALUADOR

DEDICATORIA

Primeramente, esta tesis se la dedico a mi Dios siempre supo guiarme en el buen camino, por darme fuerzas para seguir adelante y no rendirme en los problemas que se presentaron, ayudándome a afrontar las adversidades sin perder nunca la dignidad ni desfallecer en el intento.

A mis padres y hermanos: Jorge, Paulina, Armando, Kely y Franco por su apoyo incondicional, por los recursos necesarios para estudiar. Me han dado todo lo soy como persona, mis valores, mis principios, el valor para conseguir mis objetivos.

Elizabeth Flores Huamaní

AGRADECIMIENTO

Agradezco a Dios por haberme ayudado hasta este momento de mi vida profesional y por lo que vendrá.

A mi asesora Dra. Norma Lorena Catacora Flores por su apoyo, dedicación y paciencia en este trabajo de investigación.

A los docentes y estudiantes de cada escuela profesional de la Universidad Nacional José María Arguedas, por haber colaborado en esta investigación.

A mis padres y hermanos por su amor y ejemplo, por desear siempre lo mejor para mí, gracias por cada consejo y por cada una de sus palabras que me guiaron durante mi vida.

A mis amigos, compañeros y todos aquellos con los que he compartido esta etapa de mi vida.

Elizabeth Flores Huamaní

ÍNDICE

APROBACIÓN DEL ASESOR	II
COPIA DEL ACTA DE SUSTENTACIÓN	III
APROBACIÓN DEL JURADO EVALUADOR	IV
DEDICATORIA	V
AGRADECIMIENTO	VI
ÍNDICE	VII
RESUMEN	XIII
ABSTRACT.....	XIV
CHUMASQA.....	XV
INTRODUCCIÓN	16
CAPÍTULO I	18
1. PLANTEAMIENTO DEL PROBLEMA	18
1.1. Descripción del Problema	18
1.2. Formulación del Problema	19
1.2.1. Problema General.....	19
1.2.2. Problemas Específicos	19
1.3. Objetivos	20
1.3.1. Objetivo General.....	20
1.3.2. Objetivos Específicos.....	20
1.4. Hipótesis de la Investigación	20
1.4.1. Hipótesis general.....	20
1.4.2. Hipótesis específicas	20
1.5. Justificación	20
1.6. Limitaciones.....	21
CAPÍTULO II	22
2. ANTECEDENTES	22

2.1. Antecedentes de investigación	22
2.1.1 Antecedentes Internacionales.....	22
2.1.2. Antecedentes Nacionales	23
CAPÍTULO III.....	25
3. MARCO TEÓRICO.....	25
3.1. Machine Learning	25
3.1.1. Tipos de sistemas de Machine Learning	25
3.1.2. Flujo de trabajo para usar Machine Learning	26
3.1.3. Técnicas de Machine Learning	27
3.1.4. K vecinos más cercanos (K-Nearest Neighbors)	27
3.1.5. Máquinas de soporte de vectores (SVM).....	29
3.1.6. Árboles de decisiones	29
3.1.7. Métodos de Machine Learning	31
3.1.8. Matriz de Confusión	31
3.1.9. Métricas de evaluación	32
3.1.10. Tiempo total de predicción	33
3.1.11. Metodología de KDD (Knowledge Discovery in Data)	34
3.2. Estilos de vida.....	35
3.2.1. Clasificación de los estilos de vida	36
3.2.2. Alimentación saludable.....	38
3.2.3. Beneficios de una alimentación saludable	39
3.2.4. Actividad Física	40
3.2.5. Beneficios de una actividad física saludable	42
3.2.6. Aspecto Psicológico.....	43
3.2.7. Las enfermedades de los adultos tienen que ver con los estilos de vida.....	44
3.3. Marco conceptual.....	45
3.4. Definición de variables	45
CAPÍTULO IV.....	48
4. METODOLOGÍA DE INVESTIGACIÓN.....	48
4.1. Operacionalización de variables	48
4.2. Tipo y Nivel de investigación	50
4.2.1. Tipo de investigación.....	50

4.2.2. Nivel de investigación.....	50
4.4. Población y Muestra	50
4.4.1. Población.....	50
4.4.2. Muestra	51
4.5. Procedimiento de la Investigación	51
4.6. Técnicas de Recolección de Datos.....	57
4.7. Diseño Estadístico para la Prueba de Hipótesis	58
4.7.1. Hipótesis de investigación	58
4.7.2. Hipótesis general.....	58
4.7.3. Hipótesis específica H1.....	58
4.7.4. Hipótesis específica H2.....	58
4.8.5. Hipótesis específica H3.....	59
CAPÍTULO V	60
RESULTADOS.....	60
5.1. Descripción de la solución desarrollada	60
5.2. Resultados para la hipótesis de la investigación	60
5.2.1. Hipótesis General.....	60
5.3. Resultados para las hipótesis específicas	61
5.3.1. Hipótesis Específica H1	61
5.3.2. Hipótesis Específica H2	81
5.3.3. Hipótesis Específica H3	95
CAPÍTULO VI.....	113
6. DISCUSIÓN	113
CONCLUSIONES	115
RECOMENDACIONES.....	116
REFERENCIAS.....	117
ANEXOS	123
Anexo 01. Recopilación de datos.....	123
Anexo 02. Generación del Modelo Predictivo.....	127
Anexo 03. Interpretación y Evaluación	129
Anexo 04. Participación de estudiantes universitarios en la encuesta	131
Anexo 05. Matriz de Consistencia	134

LISTA DE FIGURAS

Figura 1. Flujo de trabajo para usar Machine Learning.....	26
Figura 2. Ejemplo K vecinos más cercanos	28
Figura 3. Estructura de Máquina de Vectores de Soporte	29
Figura 4. Árbol de decisión para decidir sobre una actividad.....	30
Figura 5. Flujo de trabajo para usar Machine Learning.....	32
Figura 6. Proceso KDD.....	34
Figura 7. Pirámide de la Alimentación Saludable	39
Figura 8. Pirámide de Actividad Física.....	42
Figura 9. Alimentación: Reemplazando con valores ordinales	52
Figura 10. Actividad Física: Reemplazando con valores ordinales	52
Figura 11. Aspecto Psicológico: Reemplazando con valores ordinales	53
Figura 12. Alimentación: Matriz de correlación.....	54
Figura 13. Actividad Física: Matriz de correlación	55
Figura 14. Aspecto Psicológico: Matriz de correlación.....	56
Figura 15. Metodología KDD.....	60
Figura 16. Variables de entrada (features) de la Alimentación en Google Colab	62
Figura 17. Alimentación: Visualización de datos de las clases de la variable de salida.....	63
Figura 18. Alimentación (SVM): Modelo con SVM en Weka.....	64
Figura 19. Alimentación (SVM): Modelo en Weka parte 2	65
Figura 20. Alimentación: Matriz de confusión con el algoritmo de SVM	66
Figura 21. Alimentación (KNN): Modelo con KNN en Weka.....	67
Figura 22. Alimentación (KNN): Modelo en Weka parte 2	68
Figura 23. Alimentación: Matriz de confusión con KNN.....	70
Figura 24. Alimentación: Primera parte de Árbol de decisión	71
Figura 25. Alimentación: Segunda parte de Árbol de decisión	72
Figura 26. Alimentación: Tercera parte de Árbol de decisión.....	73
Figura 27. Alimentación: Cuarta parte del Árbol de decisión	74
Figura 28. Alimentación: Quinta parte del Árbol de decisión.....	75
Figura 29. Alimentación: Sexta parte del Árbol de decisión.....	76
Figura 30. Alimentación: Séptima parte del Árbol de decisión.....	77

Figura 31. Alimentación: Matriz de confusión con Árboles de decisiones	79
Figura 32. Variables de entrada (features) de la Actividad Física en Google Colab	82
Figura 33. Actividad Física: Visualización de datos de las clases de la variable de salida	83
Figura 34. Actividad Física (SVM): Modelo con SVM en Weka	84
Figura 35. Actividad Física (SVM): Modelo en Weka parte 2	85
Figura 36. Actividad Física: Matriz de confusión con SVM	86
Figura 37. Actividad Física (KNN): Modelo con KNN en Weka	87
Figura 38. Actividad Física (KNN): Modelo con KNN en Weka parte 2	88
Figura 39. Actividad Física: Matriz de confusión con KNN	89
Figura 40. Actividad Física: Modelo del Árbol de decisión	91
Figura 41. Actividad Física: Matriz de confusión con Árboles de decisiones	93
Figura 42. Variables de entrada (features) del aspecto Psicológico en Google Colab	96
Figura 43. Aspecto psicológico: Visualización de datos de las clases de la variable de salida	97
Figura 44. Aspecto Psicológico (SVM): Modelo con SVM en Weka	98
Figura 45. Aspecto Psicológico (SVM): Modelo en Weka parte 2	99
Figura 46. Aspecto Psicológico: Matriz de confusión con SVM	100
Figura 47. Aspecto Psicológico (KNN): Modelo con KNN en Weka	101
Figura 48. Aspecto Psicológico (KNN): Modelo con KNN en Weka parte 2	102
Figura 49. Aspecto Psicológico: Matriz de confusión con KNN	103
Figura 50. Aspecto Psicológico: Primera parte del Árbol de decisión	104
Figura 51. Aspecto Psicológico: Segunda parte del Árbol de decisión	105
Figura 52. Aspecto Psicológico: Tercera parte del Árbol de decisión	106
Figura 53. Aspecto Psicológico: Cuarta parte del Árbol de decisión	107
Figura 54. Aspecto Psicológico: Quinta parte del Árbol de decisión	108
Figura 55. Aspecto Psicológico: Matriz de confusión con Árboles de decisiones	110

LISTA DE TABLAS

Tabla 1. Matriz de Operacionalización de Variables.....	48
Tabla 2. Escuelas profesionales	50
Tabla 3. Comparación de resultados al aplicar el método de balanceo de datos	53
Tabla 4. Escala de valoración de Baremo	57
Tabla 5: Alimentación (SVM): Hábitos del estudiante que adopta un estilo de vida saludable...	65
Tabla 6. Alimentación (KNN): Hábitos del estudiante que adopta un estilo de vida saludable...	68
Tabla 7: Hábitos del estudiante que adopta un estilo de vida saludable	78
Tabla 8. Métricas de Evaluación en la Alimentación	79
Tabla 9. Alimentación: Resultados con y sin validación cruzada	80
Tabla 10. Actividad Física (SVM): Hábitos del estudiante que adopta un estilo de vida saludable	85
Tabla 11. Actividad Física (KNN): Hábitos del estudiante que adopta un estilo de vida saludable	88
Tabla 12. Actividad Física: Hábitos del estudiante que adopta un estilo de vida saludable.....	92
Tabla 13. Métricas de evaluación en la Actividad Física	94
Tabla 14. Actividad Física: Resultados con y sin validación cruzada.....	94
Tabla 15. Aspecto Psicológico (SVM): Hábitos del estudiante que adopta un estilo de vida saludable	99
Tabla 16. Aspecto Psicológico (KNN): Hábitos del estudiante que adopta un estilo de vida saludable	102
Tabla 17. Hábitos del estudiante que adopta un estilo de vida saludable	109
Tabla 18. Métricas de evaluación en el aspecto Psicológico	110
Tabla 19. Aspecto Psicológico: Resultados con y sin validación cruzada	111

RESUMEN

En el presente trabajo de investigación se inició a partir de una realidad problemática, estilos de vida no saludable que adoptan miles de estudiantes universitarios en el Perú y en el mundo durante la etapa universitaria, estos comportamientos inadecuados a largo plazo, llegan a causar enfermedades no transmisibles y degenerativas. Por tal razón, se planteó como objetivo general: Predecir los estilos de vida que adoptan los estudiantes universitarios con el modelo predictivo basado en Técnicas de Machine Learning. La investigación tuvo como hipótesis: El modelo predictivo basado en Técnicas de Machine Learning permite predecir los estilos de vida que adoptan los estudiantes universitarios. El tipo de investigación que se ha considerado fue aplicada y el nivel de investigación fue descriptivo, considerando una población de 2031 estudiantes, por lo que se tomó una muestra de 323 estudiantes, el instrumento de recolección de datos utilizado fue la encuesta.

La metodología que se utilizó para esta investigación fue la metodología de KDD (Knowledge Discovery in Databases), que sirvió como base para la secuencia de etapas para esta investigación.

Las etapas que incluyen son: 1. Recopilación de Datos, 2. Selección, Limpieza y Transformación, 3. Generación del modelo predictivo, 4. Interpretación y evaluación. En la primera etapa, se recolectó registros recolectados mediante una encuesta a los estudiantes universitarios. En la segunda etapa, se realizó el preprocesamiento de datos, el balanceo de datos y la selección de características. En la tercera etapa, se utilizó los algoritmos supervisados de Machine Learning para crear el modelo predictivo. En la cuarta etapa, se interpretaron y se evaluaron los resultados de las métricas de evaluación generadas por los algoritmos de SVM, KNN y Árboles de decisiones.

Para los resultados se logró predecir los estilos de vida que adoptan los estudiantes universitarios con el modelo predictivo basado en Técnicas de Machine Learning. Por lo tanto, se concluye que el modelo predictivo basado en Técnicas de Machine Learning sí permite predecir los estilos de vida que adoptan los estudiantes universitarios.

Palabras Clave: Estilos de vida, Machine Learning, diagnóstico, estudiantes universitarios.

ABSTRACT

In the present research work, it was started from a problematic reality, unhealthy lifestyles that thousands of university students in Peru and in the world adopt during the university stage, these long-term inappropriate behaviors, come to cause diseases not transmissible and degenerative. For this reason, the general objective was: Predict the lifestyles adopted by university students with the predictive model based on Machine Learning Techniques. The research had as a hypothesis: The predictive model based on Machine Learning Techniques allows predicting the lifestyles adopted by university students. The type of research that has been considered was applied and the level of research was descriptive, considering a population of 2031 students, for which a sample of 323 students was taken, the data collection instrument used was the survey.

The methodology used for this research was the KDD (Knowledge Discovery in Databases), methodology that served as the basis for the sequence of stages for this research.

The stages included are: 1. Data Collection, 2. Selection, Cleaning and Transformation, 3. Generation of the predictive model, 4. Interpretation and evaluation. In the first stage, records were collected through a survey of university students. In the second stage, data processing, data balancing and feature selection were performed. In the third stage, Machine Learning classification algorithms were used to create the predictive model. In the fourth stage, the results of the evaluation metrics generated by the classification algorithms of SVM, KNN and Decision Trees were interpreted and evaluated.

For the results, it was possible to predict the lifestyles adopted by university students with the predictive model based on Machine Learning Techniques. Therefore, it is concluded that the predictive model based on Machine Learning Techniques does allow predicting the lifestyles adopted by university students.

Keywords: Lifestyles, Machine Learning, diagnosis, university students.

CHUMASQA

Kay taqwiri y llamkaymi qallarin huk sasachakuykunata qawaripa, manam allin kawsanku chunka waranqakunata yachakukpak wakin Perú hinaman tukuy piti pachapak karqaku yachakukpak Paykuna manam allin chu kausanku unay killapi, paykuna tarikunqa tukuy unquykunata manam transmisibles nisqanta hinaspa degenerativas nisqanta. *Chaynata*, kunan kay llamkaymi taqwiriyninpi aypayta munan: Imaynam kawsan yachakukpak kay llankay yachachin hamun llapa kirmista Machine Learning nisqanta. Hinaman churakun huk tapukuyta: Lllankay yachan hamun llapa kirmista Machine Learning nisqanpi kay llankay atinqa imaynam kawsan yachakukpak. Qatipay hukrichkaq ruray aplicadanisqawan qatipakurqa hukrichkaq ruray descriptivonisqawan, runallaqta miraron 2031 yachanakunas, qinaspa akllasqa 323 yachakukpak. instrumentum huñuspam datukunaq kay encuesta nisqanta.

Taripanaqmi ruwakurqa uk KDD (Knowledge Discovery in Databases), nisqanwa, kay kirmista taqwiry allinmi.

Kirmista kan: 1. Datos Maskaq, 2. Akllaq, Pichaq hinapa Tiqraq nisqanta, 3. Yachan hina imaynam, 4. Yachaq hinaspa sanaychay. Ñaypaq kirmista, huñuspam datukunaq kay encuesta nisqanta. Iska kirmista, tiqraq datukunataq, chuya datukunaq hinaspa akllaq imaynan kasqan. Kimsa kirmista, llapa kirmista Machine Learning nisqanpi kay ruranataq yachan imaynam kawsan hinaspan. Tawa kirmista, yachaq hinaspan kay resultados nisqanta taqwiri y hinaspan métricas evaluación nisqanpi kay algoritmos de superisado nisqanta ruan SVM, KNN hinaspan Árboles de decisiones nisqanpi.

Kay resultados nisqanta lluksinmi imaynam yachan kawsan yachakukpak hamun llapa kirmista Machine Learning nisqanpi. Chaynapi kay Lllankay yachan imaynam kawsan yachakukpak hamun llapa kirmista Machine Learning nisqanpi.

Chanin simikuna: Imaynam kawsan, Machine Learning, musyaspa, yachakukpak.

INTRODUCCIÓN

La importancia de utilizar el Machine Learning en salud es que proporciona modelos predictivos de riesgos adversos para la salud, diagnóstico de enfermedades no transmisibles y degenerativas, niveles de estrés y detección de depresión y ansiedad en estudiantes universitarios. Además, el Machine Learning aparte de ser muy eficiente en sus predicciones, proporciona datos predictivos sobre la efectividad que un tratamiento tendrá en el perfil de un paciente, todo esto a través de datos de información sanitaria procedentes de historias clínicas, encuestas realizadas por especialistas expertos en la salud, aparatos médicos como el electroencefalograma y entre otros.

Los estilos de vida son comportamientos y actitudes en un individuo como su forma de llevar una alimentación saludable, si mantiene una actividad física adecuada y en el aspecto Psicológico si es saludable. En la actualidad los jóvenes universitarios ya sea del Perú y de países extranjeros suelen tener un estilo de vida no saludable, a causa de comportamientos inadecuados como en el aspecto de la Alimentación, en la Actividad física y en el aspecto Psicológico, las actitudes inadecuadas que mayormente presentan son: comer a deshoras, sedentarismo, carga académica, estrés académico, preocupación y entre otros. Estos comportamientos inadecuados a largo plazo, llegan a causar enfermedades no transmisibles y degenerativas, Según la Organización Mundial de la Salud, las enfermedades no transmisibles (ENT) matan a 41 millones de personas cada año, lo que representa el 71% de las muertes mundiales. (OMS, 2021).

La tesis que se presenta a continuación tiene como objetivo principal predecir los estilos de vida que adoptan los estudiantes universitarios con el modelo propuesto basado en técnicas de Machine Learning.

La hipótesis general que se planteó para esta tesis fue si el modelo predictivo basado en técnicas de Machine Learning permite predecir los estilos de vida que adoptan los estudiantes universitarios.

La presente tesis consta de seis capítulos, las cuales se explicarán a continuación con una breve introducción de cada capítulo en general:

CAPÍTULO I: Planteamiento del Problema: Dentro de este capítulo se encuentra todo acerca de la realidad problemática de la investigación, los objetivos de la investigación, las

hipótesis que sirve para hallar una posible propuesta de solución al problema, la justificación y algunas limitaciones.

CAPÍTULO II: Antecedentes: En este capítulo siguiente se encuentran los antecedentes internacionales y nacionales, son aquellas investigaciones similares a esta investigación de tesis.

CAPÍTULO III: Marco Teórico: En este capítulo se encuentran todas las definiciones de las características importantes de los temas estudiados que son: la variable independiente (técnicas de Machine Learning) y dependiente (los estilos de vida).

CAPÍTULO IV: Metodología de Investigación: En este capítulo siguiente se trata de los procedimientos, el método utilizado para esta investigación, como el tipo y nivel de investigación, la población estudiada.

CAPÍTULO V: Resultados: Consta de aquellos resultados que se obtuvieron al utilizar las técnicas de Machine Learning.

CAPÍTULO VI: Discusiones, conclusiones y recomendaciones: Por último, este capítulo se discute los resultados obtenidos por otros autores de aquellas investigaciones que proceden del capítulo de antecedentes.

CAPÍTULO I

1. PLANTEAMIENTO DEL PROBLEMA

1.1. Descripción del Problema

Según la OMS en 2007 las conductas tienen un impacto directo en la salud y en la calidad de vida humana. Algunas enfermedades, así como las no infecciosas y las degenerativas, son de larga duración, causadas no solo por una combinación de factores genéticos, fisiológicos y ambientales, sino también relacionadas con el estilo de vida. En todos los países del mundo, especialmente en la mayoría de los países de bajos y medianos ingresos, una de las enfermedades con mayor mortalidad y morbilidad son las enfermedades no transmisibles como: infartos, alcoholismo, obesidad, enfermedades cardiovasculares y respiratorias, etc. Además, la Organización Mundial de la Salud ha confirmado que el 70% de las muertes prematuras en adultos se deben a malas elecciones de estilo de vida que comienzan en la adolescencia. Por ello, se ha observado en los últimos años que los universitarios tienen un estilo de vida más exigente, lo que provoca cambios negativos y desequilibrios en la salud.

Perú tiene muchos casos de enfermedades crónico degenerativas. Por ello, organismos como el Ministerio de salud impulsan estrategias comunitarias para promover conductas saludables mediante el restablecimiento de la salud física, mental y social de las familias. Por otro lado, los estudiantes universitarios, al ingresar a la universidad, se enfrentan a un nuevo entorno laboral e intelectual, lo que afecta sus hábitos de vida y las largas jornadas de estudio que exigen los docentes. Debido a esto, muchos riesgos adversos para la salud aumentan con el tiempo y se desarrollan enfermedades crónicas no transmisibles y degenerativas.

En este estudio la población perteneció a la Universidad Nacional José María Arguedas, donde se observó que todos los universitarios presentan problemas de salud relacionados con factores de estilos de vida no saludables que perjudican su calidad de vida, rendimiento académico y también afectan sus proyectos de vida. Se informa sobre el estrés de los estudiantes a lo largo de sus carreras universitarias.

El mayor problema observado es que los estudiantes universitarios no logran diagnosticar adecuadamente el estilo de vida que pueden tener durante su vida universitaria, y que la falta de

conciencia sobre el estilo de vida es perjudicial para la salud y la vida a lo largo del tiempo y la Calidad de vida relacionada con la salud (CVRS) en adultos jóvenes.

Por un lado, los estudiantes universitarios inician la universidad con cambios que van a involucrar una inadecuada alimentación, llevándose a consolidarse con los malos hábitos alimenticios, con horarios de comida no saludables y entre otros.

Por otro lado, los estudiantes presentan inactividad física, lo cual esto conlleva a consolidar el sedentarismo, reduce la capacidad para crecer y desarrollarse plenamente, por ejemplo, el día a día de un estudiante universitario común es tomar su desayuno ir en transporte a la universidad estar 8 horas sentado en clases, volver en transporte al hogar y otras 4 horas estar sentado frente a laptop a realizar las tareas académicas.

Por último, los estudiantes universitarios suelen tener comportamientos no saludables en el aspecto Psicológico, se ha presenciado problemas de salud mental como la tensión, presión, estrés, ansiedad, preocupación causados por falta de sueño, descansos necesarios,

Por lo expuesto, es necesario diagnosticar si los estudiantes llevan un estilo de vida saludable para prevenir enfermedades no transmisibles y degenerativas en el transcurso del tiempo.

1.2. Formulación del Problema

1.2.1. Problema General

- ¿De qué manera se puede predecir los estilos de vida que adoptan los estudiantes universitarios?

1.2.2. Problemas Específicos

- ¿Cuál es el modelo para predecir los estilos de vida en la alimentación de los estudiantes universitarios?
- ¿Cuál es el modelo para predecir los estilos de vida en la actividad física de los estudiantes universitarios?
- ¿Cuál es el modelo para predecir los estilos de vida en el aspecto psicológico de los estudiantes universitarios?

1.3. Objetivos

1.3.1. Objetivo General.

- Predecir los estilos de vida que adoptan los estudiantes universitarios con el modelo predictivo basado en técnicas de Machine Learning.

1.3.2. Objetivos Específicos

- Establecer un modelo para predecir los estilos de vida en la alimentación de los estudiantes universitarios usando Máquina de Vectores de Soporte.
- Establecer un modelo para predecir los estilos de vida en la actividad física de los estudiantes universitarios usando Árboles de decisiones.
- Establecer un modelo para predecir los estilos de vida en el aspecto psicológico de los estudiantes universitarios usando Árboles de decisiones.

1.4. Hipótesis de la Investigación

1.4.1. Hipótesis general

- El modelo predictivo basado en técnicas de Machine Learning permite predecir los estilos de vida que adoptan los estudiantes universitarios.

1.4.2. Hipótesis específicas

- El modelo para predecir los estilos de vida en la alimentación está basado en la técnica de Máquina de Vectores de Soporte.
- El modelo para predecir los estilos de vida en la actividad física está basado en la técnica de Árboles de decisiones.
- El modelo para predecir los estilos de vida en el aspecto psicológico está basado en la técnica de Árboles de decisiones.

1.5. Justificación

En primer lugar, gracias a los datos recolectados se ha podido predecir los estilos de vida tanto en la alimentación, actividad física y en el aspecto psicológico de los estudiantes de la UNAJMA que participaron en este estudio.

En segundo lugar, la importancia del uso del Machine Learning en este estudio, porque permitió realizar la predicción de manera más eficiente sobre los estilos de vida, además porque permite procesar grandes datos de información que en este caso fueron recolectados mediante una encuesta.

Se justifica académicamente, para el conocimiento de la comunidad estudiantil y de otros investigadores la aplicación de algoritmos de Machine Learning para determinar cuál es el algoritmo supervisado más óptimo para la predicción de los estilos de vida tanto en la alimentación, actividad física y en el aspecto psicológico con registros de estudiantes universitarios, lo cual esto ha sido posible gracias a los modelos usando los algoritmos supervisados, a las métricas de evaluación y el tiempo de predicción más eficiente de cada algoritmo supervisado.

Asimismo, se ha creado un modelo predictivo que ha permitido predecir los estilos de vida. Por esta razón, gracias a este modelo predictivo que tuvo resultados aceptables tanto en la matriz de confusión y en las métricas de evaluación halladas, se puede predecir los estilos de vida con registros de otros jóvenes de otras universidades y de otros institutos superiores.

Para que los estudiantes universitarios eviten adoptar hábitos inadecuados evitando a largo plazo enfermedades cardiovasculares y degenerativas, ya que en esta investigación se identifica cuáles son los factores más principales en un estilo de vida no saludable.

Además, la investigación permitirá que otros investigadores utilicen los modelos predictivos en nuevos conjuntos de datos o los integren en aplicaciones móviles y sitios web.

1.6. Limitaciones

Una de las limitaciones que se tuvo en esta investigación fueron las limitaciones Bibliográficas para la actividad física y el aspecto psicológico, debido a que no se encontraban libros actuales y necesarios para referenciarlos en este trabajo de Investigación.

Otra limitación muy necesaria fue en la recolección de datos debido ya que algunos de los estudiantes no responden a las encuestas, fue necesario motivarlos para que puedan responder a la encuesta brindada.

CAPÍTULO II

2. ANTECEDENTES

2.1. Antecedentes de investigación

La investigación en este artículo se refiere a estudios previos similares al tema de investigación para servir al proceso y discusión de los resultados de la investigación. A continuación, se presenta un resumen de las experiencias internacionales y nacionales:

2.1.1 Antecedentes Internacionales

La autora Ovalle (2020), en su tesis titulada **“Estudio de factibilidad de un modelo de clasificación de estudiantes universitarios según el estado crítico de estrés, ansiedad y sintomatología depresiva”**, de la Universidad de Chile propuso un ensayo clínico aleatorizado (ECA) de propósito general para clasificar a los estudiantes mediante modelos matemáticos en función de su estado de depresión, ansiedad y estrés. A diferencia de este estudio de estudiantes de la Facultad Técnica de la Universidad de Chile, el 30% de los estudiantes sufría de estrés, ansiedad y depresión, se utilizó el método EMA con un cuestionario para determinar depresión, ansiedad, estrés, higiene del sueño, características. y el contexto de la App Ánimi; en cuanto a los resultados de las variables de estudio, destacando que la higiene del sueño y la formación personal, situacional y académica son factores importantes. Además, los algoritmos de bosque aleatorio, regresión logística y máquina de vectores de soporte predijeron correctamente las variables estudiadas y, por lo tanto, lograron los objetivos planteados. En conclusión, al demostrar la viabilidad del estudio y el éxito del software, también puede proporcionar pautas para el futuro diseño de RCT y proporcionar valor público.

El autor Colula (2018), en su tesis titulada **“Método de Clasificación para detección de estrés empleado electroencefalograma”**, La Universidad Autónoma de Puebla en México tiene como objetivo desarrollar e implementar un método para la detección de estrés mediante electroencefalografía, donde el objetivo específico es la selección de algoritmos de aprendizaje automático adecuados para la clasificación de señales de EEG. Para lograr los objetivos propuestos, se obtuvieron electroencefalogramas (EEG) de 25 participantes que portaban la diadema Emotiv Epoc plus. Además, los participantes fueron evaluados en la "Escala de estrés percibido" (PSS), que mide el estrés general. Los participantes eran estudiantes universitarios, y

una vez adquiridas las señales EEG de cada participante, se analizaron con un software desarrollado en Matlab. El software implementa tres familias de wavelets: coiflet, Daubichies y biorthogonal. Entonces, como el voltaje se presenta a la alta frecuencia de la onda beta, se toman los coeficientes de onda correspondientes a la frecuencia beta. Estos coeficientes se dividen en dos categorías según los resultados de las pruebas de PSS: bajo voltaje y alto voltaje. Una vez que los coeficientes están etiquetados, se clasifican utilizando weka y los siguientes algoritmos de aprendizaje automático: Support Vector Machines (SVM), Random Forest y Naive Bayes. Los mejores resultados se obtuvieron utilizando descomposiciones de coiflet de primer orden de la señal y bosque aleatorio para la clasificación, alcanzando un 70,25% de precisión. En resumen, el algoritmo de bosque aleatorio es el mejor algoritmo para distinguir señales de estrés alto y bajo de señales EEG de estrés.

2.1.2. Antecedentes Nacionales

Los autores como: Velezmoro & Bazán (2020), en su investigación de tesis titulada **“Sistema web basado en algoritmos de inteligencia artificial para la predicción de la depresión en adolescentes de la ciudad de Cajamarca”**, de Ingeniería de Sistemas Computacionales, Universidad Privada del Norte, en la ciudad de Cajamarca, propone identificar un algoritmo de inteligencia artificial para predecir la depresión en jóvenes cajamarquinos. El uso de algoritmos seleccionados ayudó a predecir la depresión temprana en la juventud. En 2020, en la ciudad de Cajamarca se implementó un sistema en red basado en algoritmos de inteligencia artificial para predecir la depresión en jóvenes. El método utilizado, con un enfoque cuantitativo y un diseño preexperimental, tuvo en cuenta un conjunto de 115 expedientes, de los cuales solo 47 correspondían a estudiantes diagnosticados con depresión. Los resultados mostraron que el 27,70% de los jóvenes presentaba depresión, el 27,70% insomnio, falta de efecto psicomotor, el 25,50% ansiedad física moderada, presentaba más de 4 síntomas somáticos en el tracto gastrointestinal y la pérdida de peso de más de 50 gramos. En conclusión, este estudio concluyó que el algoritmo J48 fue más preciso para predecir la depresión adolescente en función de la relación del área bajo la curva ROC (0,993) a la relación de los valores de precisión (0,917). De esta forma, se ha desarrollado un sistema de red con todas las características del algoritmo seleccionado.

El autor Vilca & Bazán (2019), en su investigación de tesis **“Predicción del nivel de estrés en estudiantes universitarios utilizando técnicas de Machine Learning”**, de la Escuela de Ingeniería de Sistemas en Puno establecieron el objetivo de entrenar un modelo computarizado de aprendizaje automático para predecir el nivel de estrés de los universitarios, los investigadores enviaron una prueba (instrumento) utilizando la red social de la universidad y el correo electrónico institucional. El método en el que se basa es el método IBM Knowledge denominado CRISP-DM, el cual tiene los siguientes pasos: entendimiento del negocio, entendimiento de datos, preparación de datos, modelado, evaluación y distribución, por otro lado, el algoritmo supervisado de árboles de decisiones en los resultados del estudio logró una precisión del 97 %, demostrando que se adapta bien a este tipo de datos y concluyendo que estas técnicas de aprendizaje automático se pueden aplicar a problemas como el estrés.

CAPÍTULO III

3. MARCO TEÓRICO

3.1. Machine Learning

También llamado aprendizaje de máquinas, más conocido en el mundo actual como Machine learning.

El aprendizaje automático es un término aplicado a las máquinas relacionadas con la Inteligencia Artificial que a menudo se encargan del diagnóstico, la predicción y el reconocimiento, según Dark (2019) “Estas máquinas a menudo aprenden de los datos que se les proporcionan. Los datos comúnmente conocidos como datos de entrenamiento pueden ser datos de muestra o datos históricos que ayudan a entrenar el sistema.” (pág. 9).

3.1.1. Tipos de sistemas de Machine Learning

Según Russell (2018), existen diferentes tipos de sistemas de aprendizaje automático según el tipo y el alcance del seguimiento. Se pueden dividir en 4 categorías amplias en función de si han sido entrenados por humanos:

- Machine Learning Supervisado
- Machine Learning Sin supervision
- Machine Learning Semi-supervisado
- Machine Learning de Refuerzo.

a) Machine Learning Supervisado

Cuando los datos utilizados por un algoritmo y las soluciones que requiere se denominan "características". Tareas supervisadas de clasificación de grupos de aprendizaje automático. El programa anterior es un ejemplo porque se entrenó con muchos correos electrónicos.

b) Machine learning No Supervisado

En este tipo de sistemas no presentan etiquetas brindadas por el individuo, al contrario los registros o datos deben de buscar ser etiquetados de acuerdo por ejemplo al parecido que lleven, y con el uso de algoritmos no supervisados. Es decir, que no hay una intervención humana para

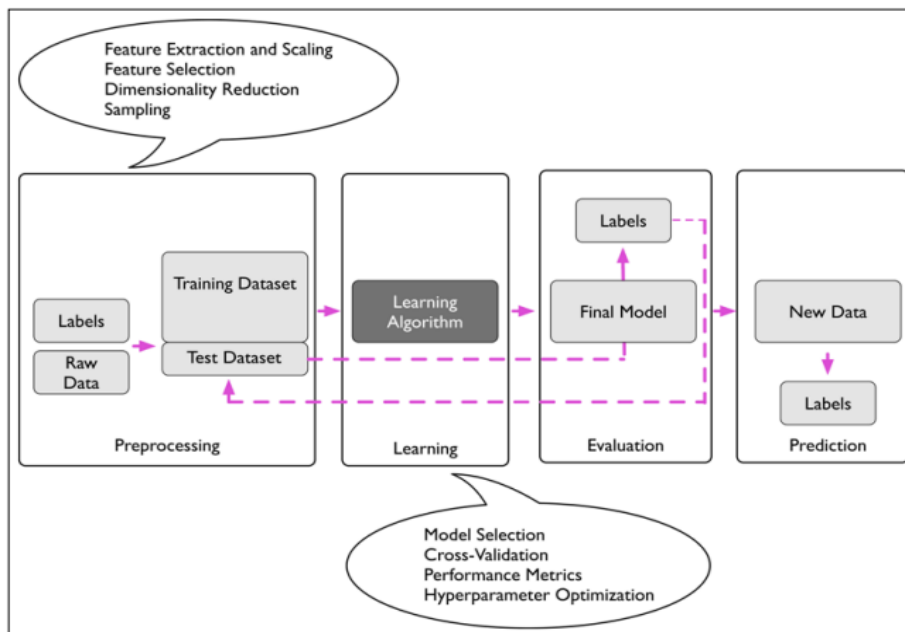
poner las etiquetas o sea las clases del target, sino por sí mismos los datos de a variable de entrada busquen ser etiquetados o el clustering. (pág. 11).

3.1.2. Flujo de trabajo para usar Machine Learning

Para los autores Raschka y Mirjalili (2017) existe un flujo de trabajo para usar Machine Learning, es decir, la forma de cómo crear una serie de tareas o procesos por los que se atraviesa un trabajo desde un inicio hasta un final.

En el siguiente diagrama se muestra un flujo de trabajo típico para usar el aprendizaje automático en modelado predictivo, que luego se explicará más adelante cada proceso:

Figura 1. Flujo de trabajo para usar Machine Learning



Fuente: Extraído (Raschka & Mirjalili, 2017, pág. 11)

- a) Preprocesamiento:** Esta etapa consiste en la extracción de características y escalado, es decir, que los datos estén en la misma escala para un rendimiento óptimo, así como también, se analiza y modifica si algunas de las características seleccionadas están altamente correlacionadas. También, se realiza la selección de características. Por un lado, si los datos están redundantes hasta cierto punto se realiza la reducción de dimensionalidad y muestreo sirve para mejorar el rendimiento del modelo predictivo,

además aquí se analiza si los datos contienen características irrelevantes llamado ruido cuando el conjunto de datos tiene una baja relación señal-ruido. (pág. 12).

- b) Aprendizaje:** En esta etapa se va a entrenar y seleccionar el mejor rendimiento modelo. También, se realiza lo que es la validación cruzada, las métricas de rendimiento como la precisión de la clasificación y el uso de técnicas de optimización de hiper parámetros que ayudan a afinar el desempeño del modelo.
- c) Evaluación:** Con el modelo entrenado se usa el dataset de prueba para estimar qué tan bien se desempeña en estos datos invisibles para estimar el error de generalización.
- d) Predicción:** Una vez que el rendimiento es bueno se puede utilizar este modelo para predecir nuevos datos futuros. (pág. 13).

3.1.3. Técnicas de Machine Learning

También llamado específicamente como algoritmos de clasificación de Machine Learning, herramientas de Machine Learning. Según el autor Russell (2018) menciona a los algoritmos supervisados más importantes:

- K-nearest neighbors (KNN, vecinos más cercanos)
- Máquinas de soporte de vectores (SVM)
- Árboles de decisiones y bosques aleatorios.
- Regresión Logística.
- Redes Neuronales. (pág. 13).

3.1.4. K vecinos más cercanos (K-Nearest Neighbors)

Según el autor Bobadilla (2020) menciona que es un algoritmo que se puede utilizar para resolver diferentes enfoques de aprendizaje automático, como la regresión o la clasificación. La operación de la regresión es simple, se realizan predicciones basadas en las muestras más cercanas a la que se pretende predecir. La principal ventaja de KNN es su simplicidad. Sus principales desventajas son: las predicciones resultan lentas y la precisión de sus resultados es moderada. (pág. 64).

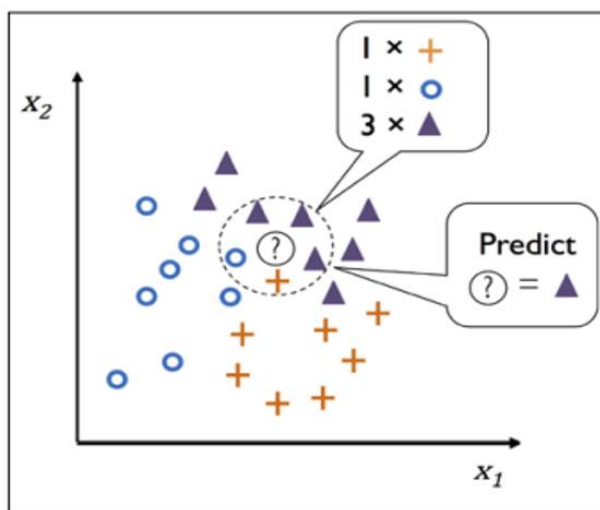
Según los autores Raschka y Mirjalili (2017) definen KNN como un aprendizaje perezoso ya que memoriza los datos del entrenamiento, en los siguientes pasos se resumen como es que el KNN realiza sus predicciones:

- 1). Elija el número de k y una medida de distancia.
- 2). Encuentra los k vecinos más cercanos de la muestra que queremos clasificar.
- 3). Las calificaciones de las clases se otorgan por mayoría de votos. (pág. 102).

- **Ejemplo de K vecinos más cercanos**

Raschka y Mirjalili (2017) describieron un ejemplo de caso en la siguiente figura de qué manera se asigna la clase de triángulo a un nuevo punto de datos (?), la etiqueta que se muestra se basan en la votación mayoritaria entre sus cinco vecinos más cercanos.

Figura 2. Ejemplo K vecinos más cercanos



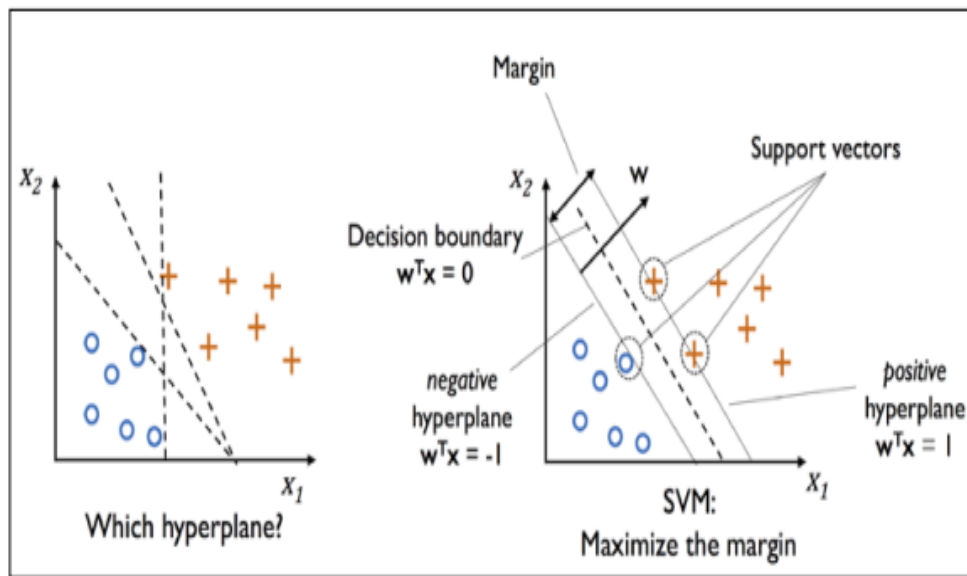
Fuente: Extraído (Raschka & Mirjalili, 2017, pág. 102)

En base a la medida de trayecto elegida, KNN encuentra los k ejemplares en el conjunto de datos de preparación más cercano (más similar) al punto que queremos clasificar. Dentro de círculo interlineado se encuentran 3 triángulos, 1 círculo y una cruz, como la mayoría son tres son triángulos. El clustering del nuevo punto de datos se establece luego por mayoría de votos entre sus k colindantes más cercanos, en este caso el valor predicho (?) sería un triángulo. (pág. 103).

3.1.5. Máquinas de soporte de vectores (SVM)

Otro algoritmo de aprendizaje poderoso y ampliamente utilizado es el SVM, que puede considerarse una extensión del perceptrón. Trayendo el algoritmo del perceptrón, minimizamos los errores de clasificación errónea. Sin embargo, en la SVM nuestro objetivo de optimización es maximizar el margen. El margen se define como el recorrido entre el hiperplano de apartamiento (final de decisión) y el entrenamiento de los especímenes más cercanas a este hiperplano, que son los llamados vectores de soporte. (Raschka & Mirjalili, 2017, pág. 77).

Figura 3. Estructura de Máquina de Vectores de Soporte



Fuente: Extraído (Raschka & Mirjalili, 2017, pág. 77)

En la figura 3 muestra cómo el hiperplano de separación logra separar las clases del entrenamiento, cuando el margen sea mayor mejor será la clasificación de las clases.

3.1.6. Árboles de decisiones

Es un modelo predictivo que construye un gráfico de construcciones lógicas a partir de una base de datos y tiene como objetivo aprender racionalmente a partir de observaciones y construcciones lógicas. También, Martinez (2020) ha afirmado lo siguiente:

Este clasificador evalúa la ganancia de información sobre cada atributo, basándose en la reducción de entropía. El atributo que ofrece la mayor ganancia de información será el siguiente

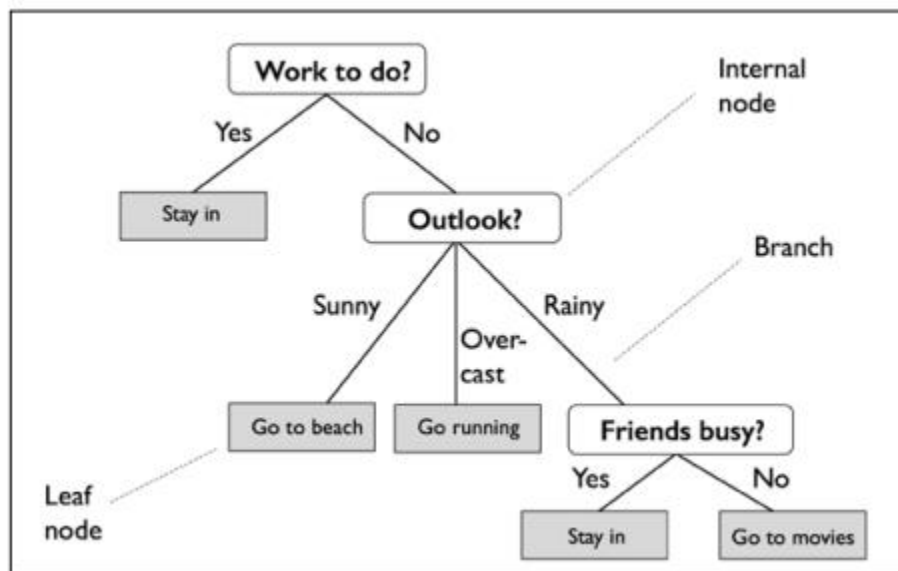
nodo. Entonces, la parte superior del árbol representará el atributo que proporciona la mayor cantidad de información de todos y cada rama representa una decisión. (pág. 49).

Por otro lado, el árbol de decisiones de nombre sugiere, podemos pensar en este modelo como desglosando nuestros datos tomando una decisión basada en hacer una serie de preguntas.

a) Ejemplo de Árboles de decisiones

Los autores Raschka y Mirjalili (2017) describieron un ejemplo como se presenta en la figura 4 en el que se usa un árbol de decisión para concluir sobre una acción en un día en específico:

Figura 4. Árbol de decisión para decidir sobre una actividad



Fuente: Extraído (Raschka & Mirjalili, 2017, pág. 89)

Las características del conjunto de entrenamiento, el modelo de árbol de decisión específico para el conjunto de entrenamiento aprende una serie de preguntas para obtener las etiquetas de clase de las muestras. El diagrama anterior ilustra el concepto de árboles de decisión basados en variables categóricas. Con un algoritmo de decisión, comenzamos en la raíz del árbol y dividimos los datos con la función que resulte en la mayor ganancia de información (IG). En un proceso iterativo, este proceso de división se puede repetir en cada nodo secundario hasta que la página esté limpia. Esto significa que las muestras de cada nodo pertenecen a la misma clase. (pág. 89).

3.1.7. Métodos de Machine Learning

Según la autora Gonzáles (2019) en su libro menciona a dos tipos de metodologías de Machine Learning basados en el formato de sus salidas:

a) Algoritmos de Clasificación

Si el resultado deseado son etiquetas discretas, es decir. el resultado es un conjunto limitado de respuestas. De manera similar, si se entrena un modelo para predecir una de dos clases de objetos, se denomina clasificación binaria. Finalmente, cuando necesitamos predecir más de dos clases objetivo, se denomina clasificación multiclase, multiclase o multivariante. (pág. 31).

b) Algoritmos de Regresión

Su utilización es para pronosticar bienes o servicios que son continuos. Así mismo, se exhibe mediante un conjunto que puede establecer de modo flexible a cargo de los ingresos del modelo. Finalmente, los valores pronosticados pueden usarse para determinar relaciones lineales entre atributos. (pág. 38).

3.1.8. Matriz de Confusión

Es una matriz cuadrada que reporta los conteos de los verdaderos predicciones positivas, negativas, verdaderas, positivas falsas y negativas falsas de un clasificador, como se muestra en la figura 5. (Sebastián, 2015, pág. 190).

Figura 5. Flujo de trabajo para usar Machine Learning

		Predicted class	
		<i>P</i>	<i>N</i>
Actual Class	<i>P</i>	True Positives (TP)	False Negatives (FN)
	<i>N</i>	False Positives (FP)	True Negatives (TN)

Fuente: (Sebastián, 2015, pág. 190)

En la Figura 5 se visualiza diferentes métodos que están coligados con la Matriz de Confusión, se definen de la siguiente manera:

- **True Positives (TP):** Cuando los valores reales son unos 1 (Verdaderos) y la predicción son unos 1 (Verdadero).
- **True Negatives (TN):** Cuando los valores reales son ceros 0 (Falso) y la predicción son ceros 0 (Falso).
- **False Positives (FP):** Cuando los valores reales son ceros 0 (Falso) y la predicción son unos 1 (Verdadero), esto indica que la predicción ha sido equivocada.
- **False Negatives (FN):** Cuando los valores reales son unos 1 (Verdadero) y la predicción son ceros 0 (Falso), esto indica que la predicción ha sido equivocada.

3.1.9. Métricas de evaluación

Según la autora Gonzáles (2019) en su libro menciona algunas métricas de evaluación:

a) Exactitud (accuracy)

Se calcula con la siguiente fórmula:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

TP son aquellos valores de Verdaderos Positivos, TN son aquellos valores de Verdaderos Negativos, FP son aquellos valores que son Falsos Positivos y FN son aquellos valores de Falsos Negativos.

b) Precisión (precision)

Se calcula con la siguiente fórmula:

$$\text{Precision} = \frac{TP}{TP + FP}$$

La abreviación TP significa aquellos valores de Verdaderos Positivos, la abreviación TN aquellos valores Verdaderos Negativos y FP son los Falsos Positivos.

c) Sensibilidad (recall)

Se calcula con la siguiente fórmula:

$$\text{recall} = \frac{TP}{TP + FN}$$

La abreviación TP significa aquellos valores Verdaderos Positivos, FN son los Falsos Negativos.

d) Puntaje de F1

Se calcula el F1 con la siguiente fórmula:

$$F1 = 2 \cdot \frac{\text{precisión} + \text{recall}}{\text{precisión} + \text{recall}}$$

Donde *precisión* viene de la métrica de evaluación de precisión y *recall* viene de la métrica de evaluación de recall.

3.1.10. Tiempo total de predicción

Es el tiempo que demora en predecir en promedio el modelo de todas las predicciones.

- **Time:** Calcula el tiempo de realización de pequeños fragmentos de código. Este módulo proporciona una forma de cronometrar pequeños fragmentos de código Python. Además, proporciona varias funciones relacionadas con el tiempo. Para hallar el tiempo en las funciones en cualquier código es necesario importar el módulo y usar la función correspondiente de time. (Python Software Foundation, 2022).

```
import time
```

```
start_time = time.time()
```

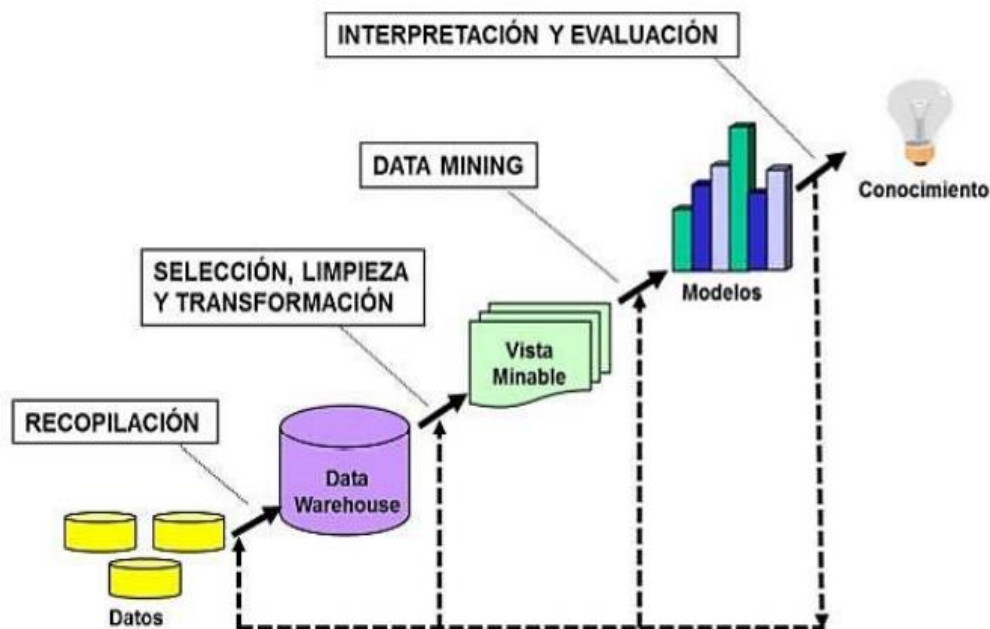
```
total_time = time.time() - start_time
```

3.1.11. Metodología de KDD (Knowledge Discovery in Data)

Es un camino secuencial metodológico para encontrar conocimiento en un conjunto de datos sin procesar, el cual tiene como fin hallar un "modelo" legal, útil y comprensible que refiera pautas de acuerdo a la investigación.

Para el autor Lara (2014) el proceso de descubrimiento del conocimiento lo clasifica en una secuencia iterativa de los sucesivos pasos:

Figura 6. Proceso KDD



Fuente: (Lara, 2014, pág. 44)

- **Recopilación:** Esta etapa es donde se recolecta los datos ya sea de páginas web mediante APIs generalmente en formato json, o la empresa es la que proporciona los datos recolectados durante varios años para su posterior proceso, o para el caso de la mayoría de investigadores les corresponde recolectar los datos mediante encuestas, o en el área de medicina utilizan ciertos aparatos médicos para los datos, en esta etapa los datos están en bruto, no están estructurados nos dan en formato csv, avro o parquet dependiendo al tipo de datos ya sea binario o secuencial.
- **Selección, Limpieza y Transformación:** Se realiza una elección de datos a aquellos datos más distinguidos para la tarea de estudio que provienen de la base de datos, en la limpieza de datos se elimina el ruido y los datos frágiles y la transformación de datos es donde los datos se le da unos cambios más adecuados a la realidad y consolidan en formularios apropiados.
- **Data Mining:** Este es el proceso básico de aplicar técnicas sutiles para extraer esquemas de datos.
- **Interpretación y Evaluación:** En la evaluación se identifica los esquemas dogmáticamente encantadores que simbolizan el juicio basado en medidas de utilidad, y en la Interpretación abarca la presentación del juicio donde se manejan métodos de visualización y representación del juicio para mostrar el juicio desarraigado a los usuarios. (Han, Kamber, & Pei, 2012, pág. 6).

3.2. Estilos de vida

Para los autores Moiso, Mestorino, y Oscar (2007) el estilo de vida es una representación de vida que se nace de patrones de conducta determinables por la interacción entre distintas peculiaridades personales, actividades mutuas, situaciones socioeconómicas de subsistencia y el medio ambiente. Así es como los individuos consiguen autoimponerse a que empeoren o mejoren la fortaleza a través de la forma en que viven.

Conductas y modos de vida sanas, entradas y posición social, nivel educativo, encargo y contextos de trabajo, vía a servicios de salud y a un medio físico adecuado. Combinados todos ellos, crean diferentes situaciones de vida que practican un claro impacto sobre salud, son considerados como resultados intermedios de salud. (págs. 172,270).

3.2.1. Clasificación de los estilos de vida

Los autores Uriarte y Vargas (2018) clasifican en dos tipos a los estilos de vida en saludables y no saludables.

- **Estilo de vida saludable:** Son aquellos comportamientos y actitudes que de alguna u otra manera hacen bien al organismo del ser humano, que a pesar de que transcurran los años de vida de la persona, esa forma de vivir no le traiga enfermedades, porque siempre mantuvo un estilo de vida saludable.

Para los autores como: Bezares, Cruz, Acosta, y Ávila (2020), OMS (2018), señalan que existen aspectos que ayudan a determinar el estilo de vida saludable las cuales son:

- a) **Hábitos de alimentación,** es toda eficacia y costumbre de consumo de suministros, de tal forma que estos cuiden una nutrición conveniente o que no alteren la salud.
 - b) **La actividad física,** es cualquier movimiento corporal procedente del sistema músculo esquelético que requiera de consumo energético, posee diversos bienes en la salud.
 - c) **El descanso y sueño,** una buena calidad del sueño no directamente está conexas con una excelente salud, sino asimismo con mayor ejercicio sabio en estudiantes. Por otra parte, los estándares de sueño y reposo en alumnos universitarios igualmente se ven trastornados mayormente por un ritmo de vida desordenado y obligaciones de trabajo. (Sánchez y de Luna, 2015, como se citó en Bezares, Cruz, Acosta, & Ávila, 2020, pág. 51).
 - d) **Socialización,** las destrezas mutuas alcanzan estudiarse en disímiles dimensiones, concebir fermentaría a ayudar al progreso de habilidades que aviven los estilos de vida saludable, de manera reunida o propia.
 - e) **Empatía,** asertividad, el análisis de la ira, etc., por su entorno interactivo, adaptativa y trama con la conducta social asimilado, podrían cooperar a los modos de vida saludable cuando estas destrezas tienen un progreso efectivo.
- **Estilo de vida no saludable:** Son aquellas costumbres inadecuadas que los individuos padecen en su vida diaria, su forma de vivir, un ejemplo claro con una

pregunta sencilla: ¿Tiene una vida sedentaria?, ¿Qué alimentos suele comer?, ¿Qué hace el individuo después de varias horas de estudio? Al transcurrir los años esa forma inadecuada de vivir le traiga como consecuencia enfermedades como las enfermedades cardiovasculares, la diabetes, la obesidad. (pág. 30).

Para los autores Bezares, Cruz, Acosta, & Ávila (2020) los aspectos que determinan el estilo de vida no saludable son:

- **Alimentación inadecuada**, el dispendio de azúcares, aguas muy dulces como las gaseosas, dispendio de grasa y comidas rápidas, las verduras y frutas la consumen raras veces o algunos ni las comen, suena irreal pero mucha gente las sobra cuando les sirven por ejemplo en un evento familiar, no les importa el tema de alimentación adecuada están más en sus asuntos o no dominan del tema todo esto hace identificar que aquellos individuos no tengan una alimentación adecuada. (pág. 45).
- **El sedentarismo**, es la inactividad física, aquellos individuos que no desempeñan al menos 30 minutos continuos de movimiento moderado o intensa tres veces a la semana y fuera del trabajo o actividades usuales. (pág. 48).
- **Alcoholismo y tabaquismo**, son hábitos inadecuados mayormente en el lugar joven, como el dispendio de tabaco, alcohol y otras drogas. En cuanto al dispendio de alcohol, igualmente se han encontrado aceptaciones prevalencias en la ciudad universitaria. (Rangel et al., 2017, como se citó en Bezares, Cruz, Acosta, & Ávila, 2020, pág. 51).
- **Nivel socioeconómico**, puede influir si se relaciona a poseer o no un estilo de vida conveniente, la situación social que tiene que ver con los ingresos de cada persona, el nivel educativo u oficio que posean se relacionan con situaciones de la mayoría de estudiantes, ya que su importante preocupación es comprar provisiones baratas y que tomen menos tiempo de preparación y dispendio, así que los alimentos que contiene aditivos y también los de altos niveles de grasas saturadas. (Sánchez y De Luna, 2015, como se citó en Bezares, Cruz, Acosta, & Ávila, 2020, pág. 53).

3.2.2. Alimentación saludable

Para una alimentación saludable los autores Azcona (2013), Moiso, Mestorino, y Oscar (2007), mencionan las siguientes variables que determinan una alimentación saludable:

- a) **Verduras y frutas:** de 3 a 5 raciones de vegetales diarias y respecto para las frutas lo recomendado es de 2 a 4 raciones. El consumo de frutas debe ser variado entre verduras, legumbres, cereales, alimentos fortificados y entre otros. Para el autor Gázquez, y otros (2017) se debe elegir alimentos basándose en la pirámide alimenticia como se muestra en la figura 7. (pág. 406).
- b) **Grasas y aceites:** Para el autor Azcona (2013), los autores Ruiz y Cenarro (2016), es recomendable consumir los alimentos menos de 60 g/día, es mucho mejor el uso de aceite de oliva ya que es mucho más beneficioso para nuestra salud. (pág. 322)
- c) **Información nutricional:** Son todos los datos básicos e importantes que necesitamos saber de un alimento empaquetado que se puede encontrar en los supermercados.
- d) **Bebidas alcohólicas:** Consumir bebidas alcohólicas en forma moderada.
- e) **Beber agua:** Beber agua de bebida es decir 1,5 a 2,5 litros por día esto equivale al menos 8 vasos al día.
- f) **Horario de comida:** Mantener un hábito alimentario respetando los horarios de comidas.
- g) **Dulces, golosinas y azúcar:** Consumir dulces, golosinas y azúcar con moderación.
- h) **Peso saludable:** Mantener un peso ideal según el Índice de Masa Corporal (IMC) y el Índice Cintura-Cadera (ICC). (págs. 628, 640).

Figura 7. Pirámide de la Alimentación Saludable



Fuente: Extraído de la Red Peruana de Alimentación y Nutrición (RPAN, 2019).

El triángulo de la pirámide nutricional nos guía en cuanto a la frecuencia y la cantidad que debemos consumir de los diferentes alimentos. En el primer nivel se ubican los alimentos que tienen mayor peso en la dieta diaria. A medida que subes de nivel, aparecen otros alimentos que debes comer en menor cantidad. (RPAN, 2019).

3.2.3. Beneficios de una alimentación saludable

Para los autores Gázquez, y otros (2017) uno de los beneficios de llevar una o una dieta saludable es que previene las enfermedades degenerativas, esto debido a que, se basa en un consumo sobrio de lácteos y carnes; de pescado y aceite de oliva y el aporte de alcohol es importantemente en forma de vino. (pág. 223).

Por otro lado, el autor Jiménez (2015) ha afirmado lo siguiente:

En una investigación titulada *Vegetales and fruit in the prevention of chronic diseases* del año 2012 en *European Journal of Nutrition*. En este trabajo, epidemiólogos, fisiólogos y nutricionistas de varias universidades alemanas analizaron la evidencia científica sobre los beneficios de las frutas y verduras en la prevención de diversas enfermedades: obesidad, diabetes tipo 2, hipertensión, enfermedad coronaria, ictus, cáncer, osteoporosis, enfermedad, demencia, síndrome del intestino irritable, artritis, asma y enfermedad pulmonar obstructiva crónica. (pág. 127).

Otros estudios de intervención más específicos, como en niños o mujeres embarazadas, han encontrado algún beneficio de comer frutas y verduras sobre la densidad ósea, pero no siempre y con resultados mixtos. El consumo regular de productos lácteos y leche tiene muchos beneficios, ya que se asocian con un riesgo reducido de la mayoría de las enfermedades. Con base en el estado actual de la ciencia, la posición de la ASN es que el consumo de alimentos de granos ricos en fibra o mezclas de granos integrales y el consumo moderado de sal están asociados con un menor riesgo de obesidad, diabetes tipo 2 y enfermedades cardiovasculares. (págs. 223, 256).

3.2.4. Actividad Física

La Organización Mundial de la Salud (2018) define la actividad física como cualquier movimiento físico producido por el sistema musculoesquelético que requiere gasto de energía. Debido a que la actividad física es beneficiosa para la salud, se reconoce como un determinante esencial de un estilo de vida saludable. (Bezares, Cruz, Acosta, & Ávila, 2020, pág. 48).

Por otro lado, la Organización Mundial de la Salud (2020) considera algunas recomendaciones para niveles saludables de actividad física para adultos de 18 a 64 años:

- a) Actividades físicas aeróbicas moderadas:** Realizar actividades físicas aeróbicas moderadas durante 2 horas y 30 minutos a 5 horas por semana.
- b) Actividades físicas aeróbicas intensas:** Al menos 1 hora y 15 minutos a 2 horas y 30 minutos de cardio de intensidad vigorosa por semana, o puedes hacer actividad física de moderada a vigorosa durante la semana.

- c) **Actividades de fortalecimiento muscular:** Realice actividades de fortalecimiento muscular de intensidad moderada o alta 2 o más días a la semana que involucren a todos los grupos musculares principales.
- d) **Actividades físicas de flexibilidad:** Para la mayoría de las personas, de 15 a 30 segundos de estiramiento estático por día es suficiente.(Márquez & Garatachea, 2013, pág. 510).
- e) **Televisión, laptop y videojuegos:** Según los autores Márquez Rosa y Garatachea Vallejo (2013), es importante reducir el tiempo de inactividad, como el tiempo dedicado a ver la televisión, simplemente estar parado durante una hora puede perder 1-2 kg de grasa por año. (pág. 13).

Los expertos en salud dicen que el uso de dispositivos en el hogar debe limitarse a dos horas o menos al día, a menos que esté relacionado con el trabajo o las tareas escolares. (NHLBI, 2013).
- f) **Caminar:** Para los autores (Márquez & Garatachea, 2013, pág. 49) caminar diariamente al menos 30 minutos al día es muy provechoso para la salud ya que ayuda a combatir la depresión.

En la figura 8 se muestra la pirámide de actividad física:

Figura 8. Pirámide de Actividad Física



Fuente: Extraído de la Fundación Española del Corazón (Madaria, 2018)

Se muestra la forma y ejemplos de cómo combinar los diferentes tipos de actividades físicas para llegar a obtener una actividad física saludable.

3.2.5. Beneficios de una actividad física saludable

Para los autores García, D. et al, (2012) como se citó en Bezares, Cruz, Acosta, y Ávila, (2020) ellos mencionan que los beneficios de llevar una actividad física saludable son muchos por esto se ha dividido según los siguientes niveles:

- **Nivel cardiovascular:** Dentro de este beneficio se tiene:
 - a) Neovascularización y aumento de la densidad capilar, aumento del gasto cardíaco
 - b) Disminución de la frecuencia cardíaca al inicio.
 - c) Disminución de la presión arterial en pacientes hipertensos.

- d) Reducir el riesgo de enfermedades cardíacas, presión arterial alta y muerte prematura

- **Nivel metabólico:**

- a) Bajar los triglicéridos
- b) Aumento de la sensibilidad a la insulina
- c) Reducir el riesgo de diabetes
- d) Reducir la grasa corporal

- **Nivel pulmonar:**

- a) Mejorar la función respiratoria
- b) Aumentar la capacidad pulmonar y el consumo máximo de oxígeno

- **Nivel músculo esquelético:**

- a) Aumento activación del metabolismo del calcio y fósforo en el hueso
- b) Aumentar la fuerza de los huesos
- c) Mayor resistencia a la presión y la carga.

- d) **Nivel social**

- a) Aceptación social
- b) Mejores relaciones interpersonales. (pág. 49).

3.2.6. Aspecto Psicológico

La salud mental es un estado de dicha en el que un individuo se da cuenta de su potencial y es capaz de hacer frente a las presiones estándar de la vida, ser productivos y favorecer a sus comunidades. Por otro lado, es la base del bienestar particular y del ejercicio productivo de las comunidades. (OMS, 2018).

Por otro lado, los autores Cabanyes y Monge (2017) mencionan que los individuos que conservan una buena salud mental exponen las siguientes características:

- a) **Personalidad:** Los aspectos de la personalidad que afectan la salud mental incluyen la asertividad y la empatía, los estilos cognitivos y los patrones de pensamiento, y el optimismo. Además, si la persona es Optimista y extrovertido, confiado y seguro, asertivo y paciente, persistente y flexible. Estos aspectos se forman sobre la personalidad en el asunto educativo (modo educativo, orientación y contenidos) y con la confluencia de diferentes elementos del entorno (amigos, medios, estilo, etc. vida social y cultural). (pág. 129).
- b) **Manejo de estrés o ansiedad:** Con respecto a la ansiedad en exposiciones, ansiedad en trabajos finales, ansiedad en exámenes, para afrontar el estrés o la ansiedad se puede buscar apoyo interpersonal, acomodarse a la situación, pensamiento positivo, ejecutar un plan de acción. (pág. 287).
- c) **Manejo de la depresión:** Durante las horas de clases el estudiante tiene una actitud de alegría, motivado, interés por las clases, expresión de atento con ganas de aprender las clases. (Barradas, 2014, págs. 161, 162).
- d) **Sueño:** La falta de sueño suficiente provoca el deterioro de la salud física y mental La necesidad de dormir de cada persona en conjunto e individualmente depende de factores biológicos, psicológicos y sociales. En general, la mayoría de los adultos necesitan de 7 a 8 horas de sueño cada noche. (pág. 249).
- e) **Actividades para sentirse mejor:** Para los autores (Uriarte & Vargas, 2018) la manera de crear espacios para sentirse mejor en la universidad es: bailar, jugar algún deporte, hacer ejercicios. (págs. 36,37).
- f) **Sobrellevar la sobrecarga académica:** El estudiante planifica sus actividades, busca ayuda con las dificultades del material de estudio, controla la exigencia del curso, organización de horario, busca la manera para no tener dificultad para concentrarte, (pág. 130).

3.2.7. Las enfermedades de los adultos tienen que ver con los estilos de vida

En el país latinoamericano según el autor Moiso, Mestorino, y Oscar (2007) precisamente en las provincias de Argentina de niveles de ingreso mayoritario, muestran un patrón de pérdida

característico de la mayoría de los países desarrollados, aquí predomina la morbilidad relacionada con el modo de vida, como los padecimientos cardiovasculares y las neoplasias. (págs.202, 270).

La inactividad física ha sido reconocida como un posible factor que contribuye a los cambios musculares, lo que conduce a enfermedades crónicas, metabólicas y cardiovasculares. (Castro et al., 2018, como se citó en Bezares, Cruz, Acosta, & Ávila, 2020, pág. 48).

3.3. Marco conceptual

- **Sckit learn**

Es una simple y eficiente herramienta en el mundo de la ciencia de datos y en el análisis de datos, construido en Numpy, Scipy y matplotlib, es de código abierto, utilizable comercialmente bajo la licencia BSD. (Gonzáles, 2019).

- **Matplotlib**

Es una biblioteca de Python que tiene el propósito de cualquier tipo de gráficos desde gráfico de líneas, barras, graficar puntos y cuadros, permite hacer diagramas de dispersión y gráficos circulares, histogramas y muchos más. (Gonzáles, 2019).

3.4. Definición de variables

- **Verduras y frutas:** Son aquellas porciones de verduras y porciones de frutas que un estudiante universitario consume al día. Al mencionar la porción se refiere a 80 gramos de fruta o puede ser un vegetal.
- **Restricción de grasas:** Es la restricción de grasas o aceites que el individuo realiza al preparar o consumir los alimentos.
- **Lectura de Información nutricional:** Las comidas empaquetadas llevan una tabla de información nutricional es muy importante saber interpretarlas para una alimentación saludable, ya que ahí se encuentra toda la información nutricional del alimento que se va a consumir.
- **Bebidas alcohólicas:** Tomar bebidas alcohólicas en muchas cantidades es dañino para la salud e incluso no es un alimento saludable, según expertos es recomendable consumir

moderadamente o evitarlas y es mejor reemplazarlas por vino. (Gázquez, y otros, 2017, pág. 223).

- **Beber agua:** Se recomienda beber 8 vasos de agua al día, si eres deportista lo mejor es que bebas unos cuantos vasos más de agua.
- **Horario de comida:** Los estudiantes universitarios es común que no respeten los horarios de comida ya sea por exámenes parciales, distancia a la universidad, exposiciones finales, etc. pero es recomendable que a pesar de tener deberes es importante respetar el horario de comida.
- **Dulces, golosinas y azúcar:** Según expertos en nutrición recomiendan que es mejor no consumir nada de golosinas y peor las empaquetadas, ya que contiene mucha grasa saturada y mucha azúcar.
- **Peso saludable:** Lo recomendable es tener un peso normal, y para esto se toma en cuenta el ICC y el IMC, conocer el peso de un individuo es una señal de que hábitos de alimentación lleva.
- **Aeróbicas moderadas:** Son actividades físicas aeróbicas moderadas según la OMS recomienda a un individuo adulto realizar entre 2 horas y media 5 horas a la semana. Ejercicios como: caminar rápidamente, natación, manejar bicicleta, correr, jugar futbol y voley.
- **Aeróbicas intensas:** Son actividades físicas aeróbicas intensas como: cuclillas, planchas, saltos de tijera, abdominales, ejercicios con el abdomen o cintura, bailes.
- **Fortalecimiento muscular:** Son actividades físicas de fortalecimiento muscular como: alzar pesos, estiramientos, trabajos de albañilería o agricultores, alzar bolsas pesadas.
- **Flexibilidad:** Son actividades físicas de flexibilidad de los músculos y tendones.
- **Televisión, laptop y videojuegos:** Según la OMS es recomendable pasar solo 2 horas frente al televisor, pero a causa de trabajo o estudios puede extenderse, en lo posible evitar una vida sedentaria.
- **Caminar:** Es recomendable caminar 30 minutos diarios según OMS, pero tratar de caminar más tiempo es mejor.
- **Personalidad:** Es aquel comportamiento en el que un individuo actúa ante situaciones difíciles.

- **Manejo de estrés:** Hay personas que controlan el estrés más que otras y otras no ya sea por traumas de la vida, pero en caso de universitarios es muy común que, en exposiciones finales, o exámenes estén nerviosos o estresados, pero es muy importante manejar estas situaciones.
- **Depresión:** Existe universitarios que a pesar que estén en clases no se concentran ya que su mente está en otros pensamientos como problemas personales, una forma de identificar estas situaciones es simplemente identificar el estado de ánimo en clases que presenten estos individuos.
- **Sueño:** Se debe descansar 8 horas al día para que nuestra mente y cuerpo adquiera un descanso saludable, existe estudiantes universitarios que sufren de enfermedades mentales ellas deben de dormir más una hora. (Cabanyes & Monge, 2017, pág. 249).
- **Crear espacios:** Es muy importante que los estudiantes universitarios creen espacios para sentirse mejor y desestresen de las largas horas de estudio para una salud mental saludable.
- **Sobrecarga académica:** En la universidad los estudiantes llegan a tener sobrecarga académica, es muy importante que cada estudiante universitario tome medidas para evitar la sobrecarga académica ya que por eso se produce estrés académico lo cual no es saludable para la salud mental.
- **Estilos de vida:** Los modos de vida en salud, son comportamientos colectivos de salud, configurados a raíz de los diversos sufragios que realizan las personas establecidas por las pertinencias de existencia que les brinda el contexto en el que se desenvuelven. En otras palabras, es el modo de cómo vive un individuo o lleva su salud física, su alimentación, su higiene personal, salud mental, ocio, social y espiritual. Para muchos autores los estilos de vida pueden ser saludables o no saludables e incluso hay autores que los clasifican en tres o cinco tipos de estilos de vida claro siempre en ese intervalo llegando a muy saludable. (Bezares, Cruz, Acosta, & Ávila, 2020, pág. 16).

CAPÍTULO IV

4. METODOLOGÍA DE INVESTIGACIÓN

4.1. Operacionalización de variables

Tabla 1. Matriz de Operacionalización de Variables

	Variables de entrada	Indicadores
H1: Alimentación	- Verduras y frutas	Matriz de Confusión
	- Restricción de Grasas	
	- Lectura de Información nutricional	
	- Bebidas alcohólicas	
	- Beber agua	
	- Horario de comida	
	- Dulces, golosinas y azúcar	
	- Peso saludable	$Precisión = \frac{TP}{TP + FP}$ $exactitud = \frac{TP + TN}{TP + TN + FP + FN}$ $tiempo_{total} = tiempo_{final} - tiempo_{inicial}$
	Variable de salida	
	Estilos de vida (Diagnóstico):	
	-Saludable	
	-No saludable	
H2: Actividad Física	Variables de entrada	Indicadores
	- Aeróbicas moderadas	Matriz de Confusión
	- Aeróbicas intensas	
	- Fortalecimiento muscular	
	- Flexibilidad	
	- Televisión, laptop y videojuegos	
	- Caminar	
		$Precisión = \frac{TP}{TP + FP}$ $exactitud = \frac{TP + TN}{TP + TN + FP + FN}$ $tiempo_{total} = tiempo_{final} - tiempo_{inicial}$

Variable de salida	
Estilos de vida (Diagnóstico):	
-Saludable	
-No saludable	
<hr/>	
H3: Aspecto psicológico	<div> <div> Variables de entrada <ul style="list-style-type: none"> - Personalidad - Manejo de estrés - Depresión - Sueño - Crear espacios - Sobrecarga académica </div> <div> Indicadores <p>Matriz de Confusión</p> $Precisión = \frac{TP}{TP + FP}$ $exactitud = \frac{TP + TN}{TP + TN + FP + FN}$ $tiempo_total = tiempo_final - tiempo_inicial$ </div> </div>
Variable de salida	
Estilos de vida (Diagnóstico):	
-Saludable	
-No saludable	
<hr/>	

Fuente: Elaboración Propia

Para la primera hipótesis específica, se quiere conocer cuáles de las variables de entrada tienen mayor efecto en la variable de salida Diagnóstico cuyas dos clases de la variable de salida son: Saludable o no saludable.

Para la segunda hipótesis específica, se quiere conocer cuáles de las variables de entrada tienen mayor efecto en la variable de salida Diagnóstico cuyas dos clases de la variable de salida son: Saludable o no saludable.

Para la tercera hipótesis específica, se quiere conocer cuáles de las variables de entrada tienen mayor efecto en la variable de salida Diagnóstico cuyas dos clases de la variable de salida son: Saludable o no saludable.

4.2. Tipo y Nivel de investigación

4.2.1. Tipo de investigación

El tipo de investigación es predictiva, ya que en esta investigación se va a predecir los comportamientos futuros de individuos, llegar a conocer las causas que inciden en la ocurrencia de un fenómeno se puede anticipar sus efectos. (Muñoz, 2015, pág. 85).

4.2.2. Nivel de investigación

El nivel de investigación es descriptivo, ya que en esta investigación se buscó describir comportamientos, características, tendencias y perfiles de personas o una determinada población. (Hernández, Fernández, & Bapstista, 2014, pág. 92).

4.4. Población y Muestra

4.4.1. Población

La Población que se utilizó durante el desarrollo del estudio estuvo compuesta por el total de 2031 estudiantes universitarios de las seis escuelas profesionales de Ingeniería de Sistemas, Ingeniería Agroindustrial, Ingeniería Ambiental, Administración de Empresas, Contabilidad, Educación Primaria Intercultural de la Universidad Nacional José María Arguedas del periodo académico 2021-II. Comprende el espacio temporal compuesto por los meses de octubre del año 2021 a octubre del año 2022.

Tabla 2. Escuelas profesionales

ESCUELA PROFESIONAL DE:	Población	Muestra
EPIS	358	53
EPIA	330	54
EPIAM	314	54
EPAE	421	54
EPC	307	54
EPEPI	301	54
Total	2031	323

Fuente: Elaboración Propia

4.4.2. Muestra

La muestra calculada fue de 323 estudiantes universitarios encuestados a las seis escuelas profesionales de la Universidad Nacional José María Arguedas que son de Ingeniería de Sistemas, Ingeniería Agroindustrial, Ingeniería Ambiental, Administración de Empresas, Contabilidad y Educación Primaria Intercultural. (Martínez, 2012, pág. 328).

Donde:

$$n = \frac{N*Z^2*P*Q}{(N-1)E^2+Z^2*P*Q}; \frac{2031*1.96^2*0.5*0.5}{(2031-1)0.05^2+1.96^2*0.5*0.5} = 323 \text{ estudiantes}$$

n = tamaño de muestra buscado.

N = tamaño de la Población o Universo

Z = Parámetro estadístico que depende del Nivel de Confianza (NC)

E = Error de estimación máximo aceptado

P = Probabilidad de que ocurra el evento estudiado (éxito)

$Q = (1-P)$ = Probabilidad de que no ocurra el evento estudiado.

4.5. Procedimiento de la Investigación

Etapas I: Recopilación de datos

Se realizó la recolección de datos a través de una encuesta trayendo como instrumento un cuestionario para conseguir la información necesaria sobre las actitudes y comportamientos que adoptan los estudiantes universitarios de las distintas facultades acerca de la alimentación, actividad física y el aspecto psicológico. Para mayor detalle se encuentra en el Anexo 01.

Etapas II: Selección, Limpieza y Transformación

Para esta etapa y las demás se utilizaron las herramientas informáticas de Inteligencia Artificial como el entorno gratuito de Google Colab con el lenguaje de programación de Python y algunas librerías de Scikit Learn. Dentro de esta etapa se hizo el preprocesamiento de datos, el balanceo de datos y la selección de características, para lo cual se explica en los siguientes subtemas:

- **Preprocesamiento de datos**

Primeramente, se verificaron los datos faltantes, se eliminaron aquellas columnas que no fueron relevantes, a las variables de entrada del conjunto de datos de Alimentación, Actividad Física y el aspecto Psicológico se reemplazaron con valores ordinales, a la variable de salida también llamado objeto o Target se reemplazó con valores binario de “0” para “No saludable” y “1” para “Saludable” de igual manera a las variables de entrada:

Figura 9. Alimentación: Reemplazando con valores ordinales

```
df['v_frutas'].replace(('5 porciones de verduras y 4 frutas', '4 porciones de verduras y 3 frutas', '3 porciones de verduras y 2 frutas'), (0,1,2), inplace = True)
df['r_grasas'].replace(('Nada', 'Poco', 'Mucho', 'Demasiado'), (0,1,2,3), inplace = True)
df['l_etiquetas'].replace(('Nunca', 'Casi nunca', 'A veces', 'Frecuentemente', 'Siempre'), (0,1,2,3,4), inplace = True)
df['b_alcoholicas'].replace(('Nunca', 'Una vez al año', 'Menos de 12 veces al año', 'Una vez al mes', 'Más de una vez al mes'), (0,1,2,3,4), inplace = True)
df['agua'].replace(('Menos de 6 vasos', 'Entre 6 a 8 vasos', 'Más de 8 vasos'), (0,1,2), inplace = True)
df['comer_deshoras'].replace(('Presentación de proyectos Finales', 'Exámenes Parciales', 'Tengo que tomar un examen'), (0,1,2), inplace = True)
df['c_dulces'].replace(('Más de 7 dulces', '4 a 7 dulces', 'Menos de 3 dulces', '0 dulces'), (0,1,2,3), inplace = True)
df['peso'].replace(('Obesidad', 'Sobrepeso', 'Peso normal', 'Debajo de su peso normal'), (0,1,2,3), inplace = True)
#Target: Reemplazando con valores binarios
df['Diagnostico'].replace(('No Saludable', 'Saludable'), (0,1), inplace = True)
```

Fuente: Elaboración propia

Para el conjunto de datos de Actividad Física:

Figura 10. Actividad Física: Reemplazando con valores ordinales

```
df['a_moderadas'].replace(('Menos de 2 h y 30 min', '2 h y 30 min a 5 h', 'Más de 5 h'), (0,1,2), inplace = True)
df['a_intensas'].replace(('Menos de 1 h y 15 min', '1 h y 15 min a 2 h y 30 min', 'Más de 2 h y 30 min'), (0,1,2), inplace = True)
df['f_muscular'].replace(('Menos de 2 días', 'Igual a 2 días', 'Más de 2 días'), (0,1,2), inplace = True)
df['f_tendones'].replace(('Ninguno', 'Menos de 15 s', '15 a 30 s', 'Más 30 s'), (0,1,2,3), inplace = True)
df['f_televisor'].replace(('2 h', 'Menos de 4 h', '4 a 7 h', '7 a 10 h', 'Más de 10 h'), (4,3,2,1,0), inplace = True)
df['t_caminar'].replace(('Menos de 30 min', '30 min', 'Más de 30 min'), (0,1,2), inplace = True)
#Target: Reemplazando con valores binarios
df['Diagnostico'].replace(('No Saludable', 'Saludable'), (0,1), inplace = True)
```

Fuente: Elaboración propia

Para el conjunto de datos del aspecto Psicológico:

Figura 11. Aspecto Psicológico: Reemplazando con valores ordinales

```
df['personalidad'].replace(('Intolerancia', 'Flexibilidad', 'Prepotencia', 'Empatía'), (0,1,2,3), inplace = True)
df['ansiedad'].replace(('Exposiciones finales', 'Trabajos finales', 'Exámenes', 'Trabajos grupales'), (0,1,2,3), inplace = True)
df['a_horas_clases'].replace(('Fatiga o cansancio', 'Expresión de desvelado u ojeroso', 'Expresión de deprimido', 'Desmotivado'), (0,1,2,3), inplace = True)
df['h_sueño'].replace(('Menos de 7 h', '7 a 8 h', 'Más de 8 h'), (0,1,2), inplace = True)
df['c_espacios'].replace(('No me alcanza tiempo', 'Ausencia de lugares para sentirte mejor', 'Jugar videojuegos', 'Bailar', 'Ocio'), (0,1,2,3,4), inplace = True)
df['s_academica'].replace(('Dejas a última hora las tareas', 'Dificultad para concentrarte', 'Dificultad con el material'), (0,1,2), inplace = True)
#TARGET: Reemplazando con valores binarios
df['Diagnostico'].replace(('No Saludable', 'Saludable'), (0,1), inplace = True)
```

Fuente: Elaboración propia

Los reemplazos de valores ordinales que fueron asignadas a las variables de entrada fueron a criterio de otras investigaciones de tesis estudiadas sobre estilos de vida, recopilación de libros relacionados al tema de investigación que tienen reconocimiento internacional (ISBN), y otros expertos en el tema.

- **Balanceo de datos**

En esta subetapa se procedió a separar los datos de entrenamiento con un tamaño de 80 %, para los datos de prueba se le asignó un tamaño de 20 %, estas mismas asignaciones se realizó para los conjuntos de datos de Alimentación, Actividad Física y aspecto Psicológico.

Dentro del Balanceo de datos se utilizó el método llamado Sobremuestreo para mantener una igualdad en los datos de entrenamiento, que consiste en aumentar las instancias correspondientes a la clase minoritaria replicándose hasta un grado constante, ya que este método no conduce a la pérdida de información. (González, 2022).

Tabla 3. Comparación de resultados al aplicar el método de balanceo de datos

	Antes de aplicar el balanceo de datos		Después de aplicar el balanceo de datos	
	No saludable	Saludable	No saludable	Saludable
Alimentación	158	100	158	158
Actividad Física	140	118	140	140
Aspecto Psicológico	151	107	151	151

Fuente: Elaboración propia.

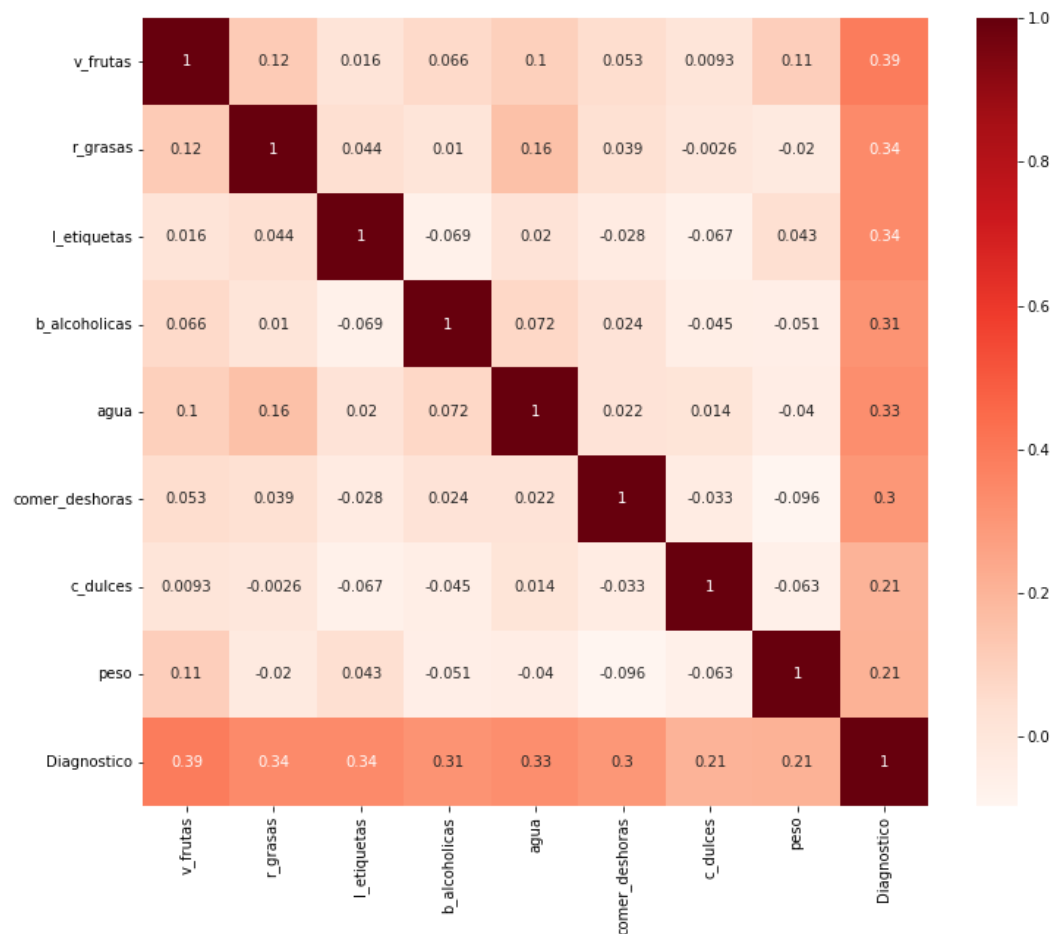
En la tabla 3 se visualiza que hubo una equivalencia de las instancias de las clases de No saludable y saludable, para que a la máquina que está aprendiendo no le entrenemos más de una clase que la otra.

- **Selección de Características**

Se realizó la selección de características aplicadas a las variables de entrada para seleccionar aquellas variables de entrada que tienen mayor correlación con la variable de salida que en este caso es Diagnóstico, a través del método de filtros de correlación, para lo cual se procedió a crear un nuevo bloque de código con el módulo en Python llamado “corr()” y luego graficar a través de la herramienta “seaborn” la matriz de correlación de las variables para cada conjunto de datos.

a) Matriz de Correlación para el conjunto de datos de Alimentación

Figura 12. Alimentación: Matriz de correlación



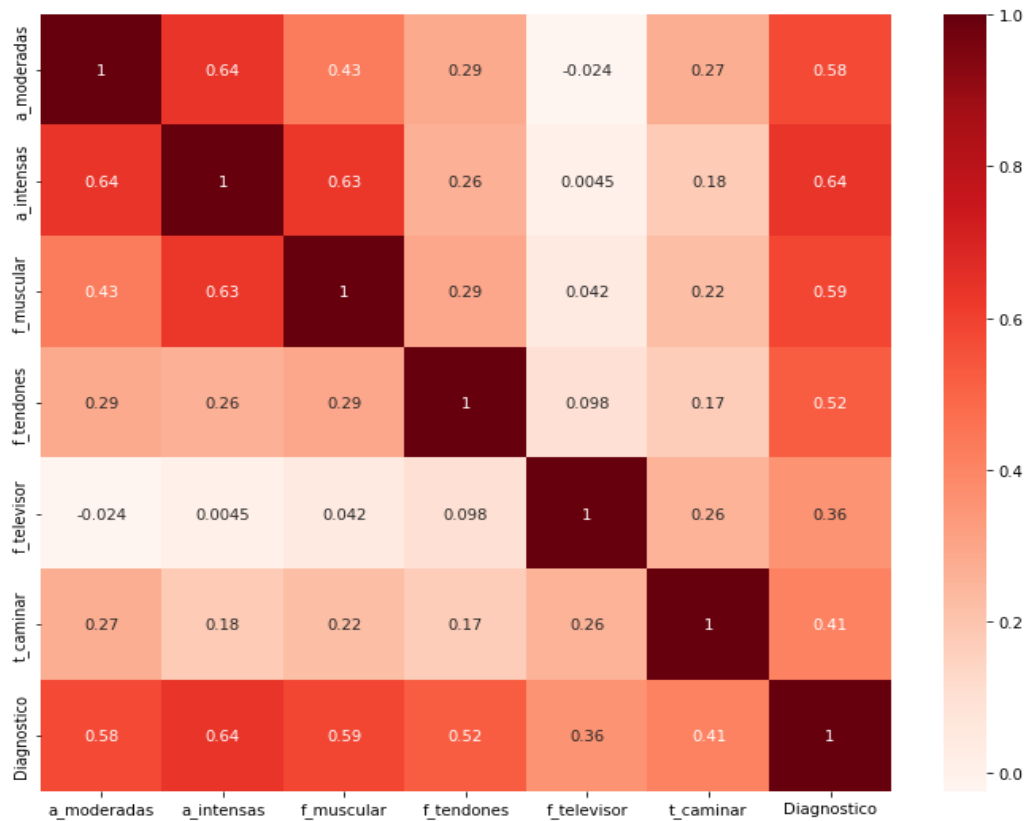
Fuente: Elaboración propia.

Seguidamente, se procedió a realizar el filtro de los atributos que posean un grado de correlación mayor a 0.3. Las variables de entrada que tienen mayor correlación con la variable de salida llamada Diagnóstico de acuerdo a la figura 13 son:

- v_frutas = verduras y frutas
- r_grasas = restricción de grasas
- l_etiquetas = lectura de etiquetas
- b_alcoholicas = bebidas alcohólicas
- agua = beber agua
- comer_deshoras = comer a deshoras

b) Matriz de Correlación para el conjunto de datos de Actividad Física

Figura 13. Actividad Física: Matriz de correlación



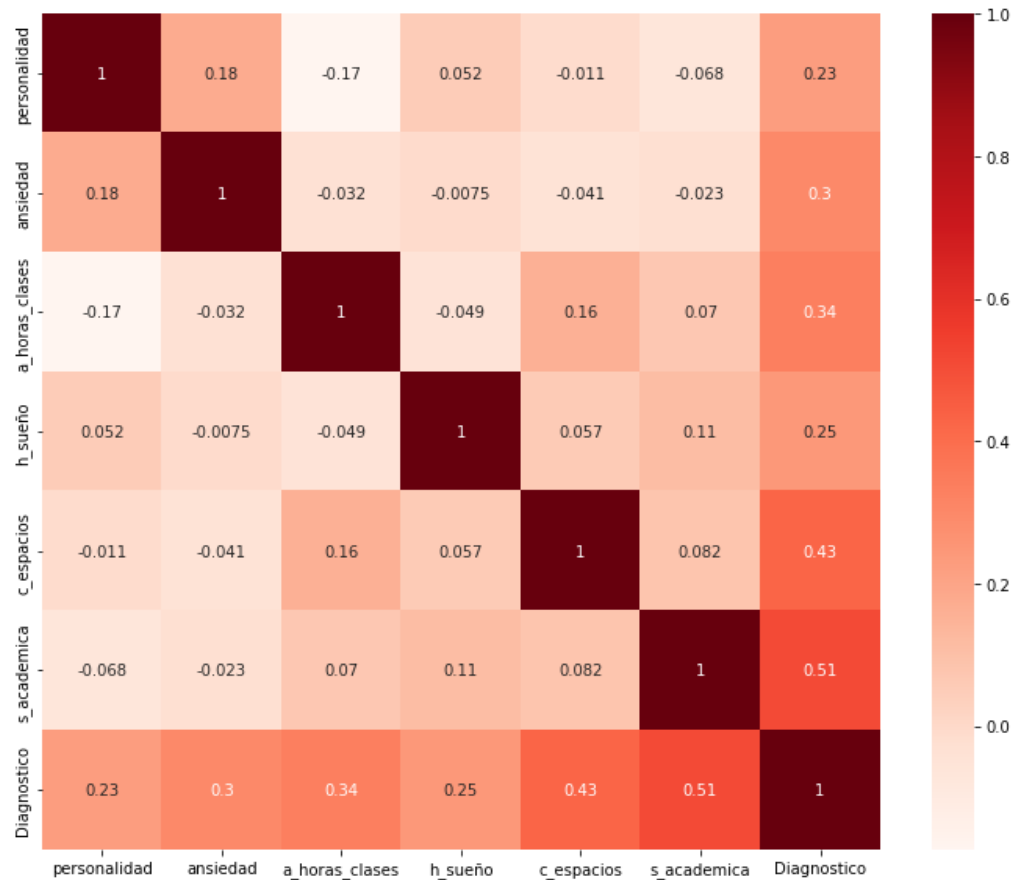
Fuente: Elaboración propia.

Asimismo, se procedió a realizar el filtro de los atributos que posean un grado de correlación mayor a 0.5. Las variables de entrada que tienen mayor correlación con la variable de salida llamada Diagnóstico de acuerdo a la figura 14 son:

- a_moderadas = Aeróbicas moderadas
- a_intensas = Aeróbicas intensas
- f_muscular = Fortalecimiento muscular
- f_tendones = Flexibilidad

c) Matriz de Correlación para el conjunto de datos del aspecto Psicológico

Figura 14. Aspecto Psicológico: Matriz de correlación



Fuente: Elaboración propia.

Al obtener los resultados de la matriz de correlación se procedió a realizar el filtro de los atributos que posean un grado de correlación mayor a 0.4. Las variables de

entrada que tienen mayor correlación con la variable de salida llamada Diagnóstico de acuerdo a la figura 15 son:

- s_academica = Sobrecarga académica
- c_espacios = Crear espacios

Etapas III: Generación del modelo predictivo

Aquí se implementó el algoritmo supervisado de Árboles de Decisiones, KNN y Máquinas de Vectores de Soporte para entrenar al modelo propuesto y luego generar los valores predictivos. Para mayor detalle se encuentra en el [Anexo 02](#).

Etapas IV: Interpretación y evaluación

Se hizo la interpretación y evaluación de métricas de evaluación de Machine Learning para evaluar las predicciones del modelo predictivo, para luego identificar cuál de los algoritmos supervisados de Machine Learning fue el más adecuado para la predicción de los estilos de vida. Para mayor detalle se encuentra en el [Anexo 03](#).

4.6. Técnicas de Recolección de Datos

Se utilizó un instrumento de medición para los Estilos de vida de los estudiantes universitarios de la Universidad Nacional José María Arguedas. Para la escala de calificación se manejó como técnica la encuesta. Este constituye de 20 ítems divididos en 3 aspectos: Alimentación (1,2,3,4,5,6,7,8), Actividad Física (9,10,11,12,13,14) y aspecto Psicológico (15,16,17,18,19,20).

Para cada respuesta de cada ítem se han asignado valores numéricos entre 0 al 2, 0 al 3 y de 0 al 4, seguidamente, se hizo una suma de estos valores según a las respuestas brindadas por los estudiantes.

Seguidamente, se realizó la escala de valoración o Baremo para determinar la escala de valoración como se detalla en la tabla 4.

Tabla 4. Escala de valoración de Baremo

Baremo: Escala de Valoración	
No Saludable	3-11

Saludable	12-20
-----------	-------

Fuente: Elaboración Propia

Para interpretar la tabla 4 se describe que si la suma de un registro dio como resultado entre el intervalo del 3 al 11 se consideró como saludable, y si la suma de un registro dio como resultado entre el intervalo de 12 al 20 se consideró como saludable.

De esta manera se ha diagnosticado los estilos de vida que adoptan los estudiantes universitarios con el conjunto de datos recolectados, porque para aplicar Machine Learning de tipo supervisado es necesario que exista una variable de salida que en este estudio es el diagnóstico para luego generar un modelo predictivo de los estilos de vida para estudiantes universitarios.

4.7. Diseño Estadístico para la Prueba de Hipótesis

4.7.1. Hipótesis de investigación

4.7.2. Hipótesis general

H₀: El modelo predictivo basado en técnicas de Machine Learning no permite predecir los estilos de vida que adoptan los estudiantes universitarios.

H₁: El modelo predictivo basado en técnicas de Machine Learning permite predecir los estilos de vida que adoptan los estudiantes universitarios.

4.7.3. Hipótesis específica H1

H₀: El modelo para predecir los estilos de vida en la alimentación no está basado en la técnica de Máquina de Vectores de Soporte.

H₁: El modelo para predecir los estilos de vida en la alimentación está basado en la técnica de Máquina de Vectores de Soporte.

4.7.4. Hipótesis específica H2

H₀: El modelo para predecir los estilos de vida en la actividad física no está basado en la técnica de Árboles de decisiones.

H₁: El modelo para predecir los estilos de vida en la actividad física está basado en la técnica de Árboles de decisiones.

4.8.5. Hipótesis específica H3

H₀: El modelo para predecir los estilos de vida en el aspecto psicológico no está basado en la técnica de Árboles de decisiones.

H₁: El modelo para predecir los estilos de vida en el aspecto psicológico está basado en la técnica de Árboles de decisiones.

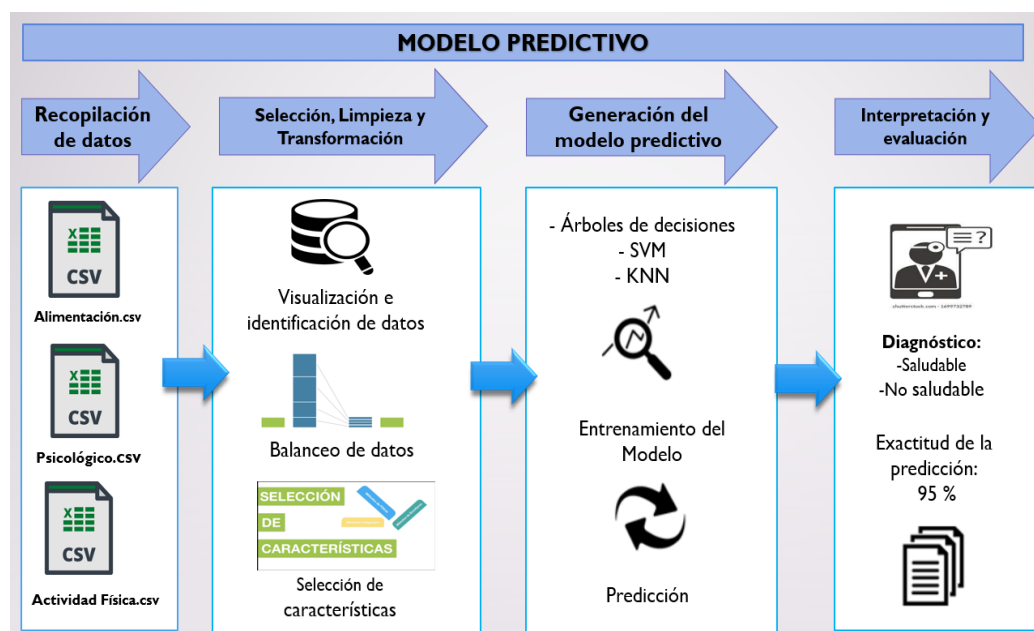
CAPÍTULO V

RESULTADOS

5.1. Descripción de la solución desarrollada

El presente trabajo de investigación está basado en la metodología de Descubrimiento de conocimiento en bases de datos KDD (Knowledge Discovery in Databases), ya que sirvió para tener el orden y la secuencia de las etapas tal como se muestra en la siguiente figura 9:

Figura 15. Metodología KDD



Fuente: Adaptado a la Metodología KDD (Lara, 2014, pág. 44).

5.2. Resultados para la hipótesis de la investigación

5.2.1. Hipótesis General

H₀: El modelo predictivo basado en técnicas de Machine Learning no permite predecir los estilos de vida que adoptan los estudiantes universitarios.

H₁: El modelo predictivo basado en técnicas de Machine Learning permite predecir los estilos de vida que adoptan los estudiantes universitarios.

De acuerdo a los resultados obtenidos en las hipótesis específicas 1, 2 y 3 el modelo predictivo basado en técnicas de Machine Learning sí permite predecir los estilos de vida

que adoptan los estudiantes universitarios en cuanto a la Alimentación el algoritmo supervisado más adecuado fue Máquina de Vectores de Soporte, en cuanto a la Actividad Física el algoritmo supervisado más adecuado fue Árboles de decisiones y en cuanto al aspecto Psicológico el algoritmo supervisado más adecuado fue Árboles de decisiones.

5.3. Resultados para las hipótesis específicas

5.3.1. Hipótesis Específica H1

H₀: El modelo para predecir los estilos de vida en la alimentación no está basado en la técnica de Máquina de Vectores de Soporte.

H₁: El modelo para predecir los estilos de vida en la alimentación está basado en la técnica de Máquina de Vectores de Soporte.

Primeramente, al tener los datos de Alimentación en formato Excel se procedió a cargar los datos a Google Colab, a estos datos se les conoce como características o variables de entrada como se muestra en la figura 16.

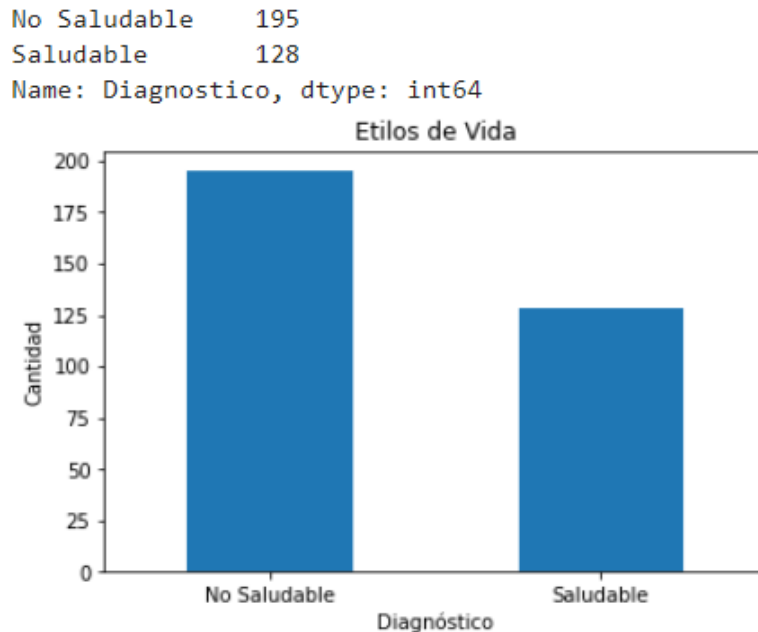
Figura 16. Variables de entrada (features) de la Alimentación en Google Colab

edad	sexo	e_profesional	c_academico	trabaja	e_civil	v_frutas	r_grasas	l_etiquetas	b_alcoholicas	agua	comer_deshoras	c_dulces	peso	Diagnostic
Entre 18 a 21 años	Femenino	EPC	V	No	Soltero	4 porciones de verduras y 3 frutas	Demasiado	Siempre	Cada fin de semana	Entre 6 a 8 vasos	Vivo lejos a la universidad	0 dulces	Peso normal	Saludable
Entre 18 a 21 años	Masculino	EPIA	VI	Sí	Soltero	5 porciones de verduras y 4 frutas	Poco	Nunca	Una vez al mes	Entre 6 a 8 vasos	Falta de apoyo económico	0 dulces	Debajo de su peso normal	Saludable
Entre 18 a 21 años	Masculino	EPIS	IV	No	Soltero	5 porciones de verduras y 4 frutas	Poco	Nunca	Nunca	Entre 6 a 8 vasos	Tengo que trabajar y estudiar	Menos de 3 dulces	Debajo de su peso normal	Saludable
Entre 18 a 21 años	Femenino	EPAE	VIII	No	Soltero	3 porciones de verduras y 2 frutas	Demasiado	Siempre	Menos de 12 veces al año	Más de 8 vasos	Vivo lejos a la universidad	Menos de 3 dulces	Peso normal	Saludable
Entre 18 a 21 años	Masculino	EPEPI	IV	No	Soltero	4 porciones de verduras y 3 frutas	Demasiado	Casi nunca	Una vez al año	Entre 6 a 8 vasos	Vivo lejos a la universidad	4 a 7 dulces	Peso normal	Saludable
Entre 18 a 21 años	Masculino	EPIA	I	Sí	Soltero	4 porciones de verduras y 3 frutas	Poco	Nunca	Menos de 12 veces al año	Más de 8 vasos	Exámenes Parciales	4 a 7 dulces	Debajo de su peso normal	No Saludable
Entre 18 a 21 años	Femenino	EPAE	V	No	Soltero	4 porciones de verduras y 3 frutas	Demasiado	A veces	Una vez al año	Menos de 6 vasos	Presentación de proyectos Finales	Menos de 3 dulces	Peso normal	No Saludable
Más de 21 años	Masculino	EPIS	X	Sí	Soltero	5 porciones de verduras y 4 frutas	Demasiado	Nunca	Nunca	Menos de 6 vasos	Vivo lejos a la universidad	Más de 7 dulces	Sobrepeso	No Saludable
Entre 18 a 21 años	Masculino	EPAE	V	Sí	Soltero	5 porciones de verduras y 4 frutas	Mucho	A veces	Nunca	Más de 8 vasos	Presentación de proyectos Finales	Más de 7 dulces	Sobrepeso	No Saludable
Entre 18 a 21 años	Femenino	EPC	VI	Sí	Soltero	3 porciones de verduras y 2 frutas	Poco	Casi nunca	Menos de 12 veces al año	Menos de 6 vasos	Tengo que trabajar y estudiar	Menos de 3 dulces	Sobrepeso	No Saludable
Entre 18 a 21 años	Masculino	EPIAM	III	No	Soltero	4 porciones de verduras y 3 frutas	Mucho	Siempre	Nunca	Más de 8 vasos	Tengo que trabajar y estudiar	Menos de 3 dulces	Peso normal	Saludable
Entre 18 a 21 años	Femenino	EPAE	VIII	No	Soltero	5 porciones de verduras y 4 frutas	Nada	Nunca	Una vez al año	Más de 8 vasos	Vivo lejos a la universidad	0 dulces	Debajo de su peso normal	Saludable
Entre 18 a 21 años	Masculino	EPC	II	No	Soltero	3 porciones de verduras y 2 frutas	Demasiado	A veces	Una vez al año	Menos de 6 vasos	Tengo que trabajar y estudiar	4 a 7 dulces	Peso normal	No Saludable

Fuente: Elaboración propia

Se realizó una visualización de datos con el lenguaje de Python y la librería Pandas acerca de las clases que existe en la variable de salida llamada “Diagnóstico”.

Figura 17. Alimentación: Visualización de datos de las clases de la variable de salida



Fuente: Elaboración propia

Como se visualiza en la figura 17 se reportaron 195 datos que pertenecen a la clase de no saludable que pertenecen a aquellos estudiantes universitarios que adoptan un estilo de vida no saludable y 128 datos que pertenecen a la clase de saludable que pertenecen a aquellos estudiantes que adoptan un estilo de vida saludable.

Consecutivamente, se muestra el modelo con la técnica de Máquina de Vectores de Soporte realizada en el programa de Weka:

Figura 18. Alimentación (SVM): Modelo con SVM en Weka



Fuente: Elaboración propia

Se puede visualizar que cada aspa de color azul son las predicciones correctas que dieron como “Saludable” y cada aspa de color rojo son las predicciones correctas que dieron como “No Saludable”, los cuadros azules que se encuentran en la parte de las predicciones “No Saludable”, son las predicciones que fallaron, o sea que predijeron como “Saludable” pero en realidad debieron de predecir como “No Saludable”, así mismo los cuadros rojos que se encuentran en la parte de lo “Saludable” son las predicciones que fallaron, o sea predijeron como “No saludable” pero en realidad debieron de predecir como “Saludable”.

Al hacer click en cada aspa se genera una interfaz, en este caso se hizo click al de color azul como se demuestra:

Figura 19. Alimentación (SVM): Modelo en Weka parte 2

```

Weka: Instance info

Plot : weka.classifiers.functions.SMO (cate_balanced_food_s
Instance: 311
    v_frutas : 5 porciones de verduras y 4 frutas
    r_grasas : Poco
    l_etiquetas : Frecuentemente
    b_alcoholicas : Una vez al mes
    agua : Menos de 6 vasos
    comer_deshoras : Falta de apoyo econ mico
    prediction margin : 1.0
    predicted Diagnostico : Saludable
    Diagnostico : Saludable

```

Fuente: Elaboración propia

Se muestra el perfil de un estudiante que adopta hábitos para un estilo de vida saludable en la Alimentación que el algoritmo supervisado de Máquina de Vectores de Soporte ha logrado predecir. Asimismo, se muestra más casos en la tabla 5:

Tabla 5: Alimentación (SVM): Hábitos del estudiante que adopta un estilo de vida saludable

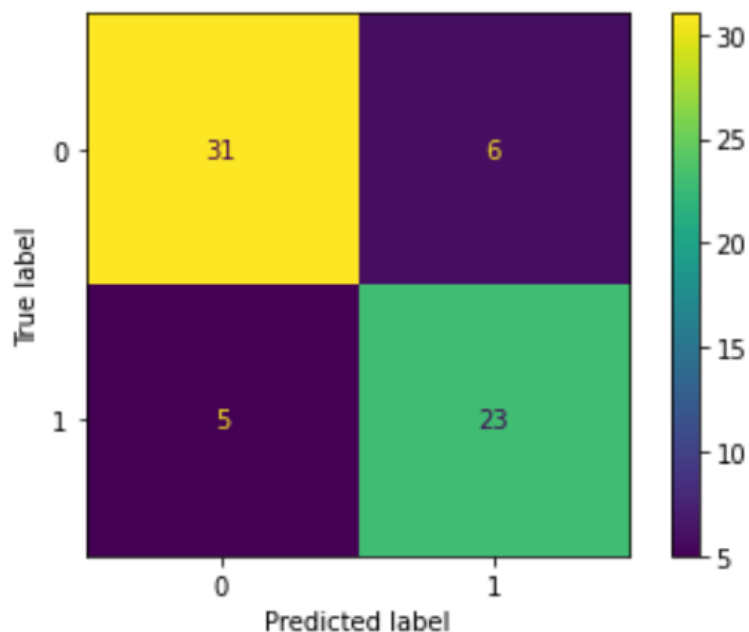
Reglas del modelo de Máquina de Vectores de Soporte de los estudiantes que adoptan un estilo de vida saludable	Variable de salida (Diagnóstico)
Verduras y frutas = 4 porciones de verduras y 3 frutas Restricción de grasas = Mucho Lectura de etiquetas = Siempre Bebidas alcohólicas = Una vez al mes Agua = Más de 8 vasos Comer a deshoras = Tengo que trabajar y estudiar	Saludable
Verduras y frutas = 5 porciones de verduras y 4 frutas Restricción de grasas = Demasiado	Saludable

Lectura de etiquetas = Casi nunca Bebidas alcohólicas = Nunca Agua = Menos de 6 vasos Comer a deshoras = Falta de apoyo económico	
Verduras y frutas = 4 porciones de verduras y 3 frutas Restricción de grasas = Poco Lectura de etiquetas = Frecuentemente Bebidas alcohólicas = Una vez al año Agua = Entre 6 a 8 vasos Comer a deshoras = Falta de apoyo económico	Saludable

Fuente: Elaboración Propia.

A continuación, en la figura 20 se muestra la matriz de confusión para los datos de alimentación con la técnica de Máquina de Vectores de Soporte:

Figura 20. Alimentación: Matriz de confusión con el algoritmo de SVM

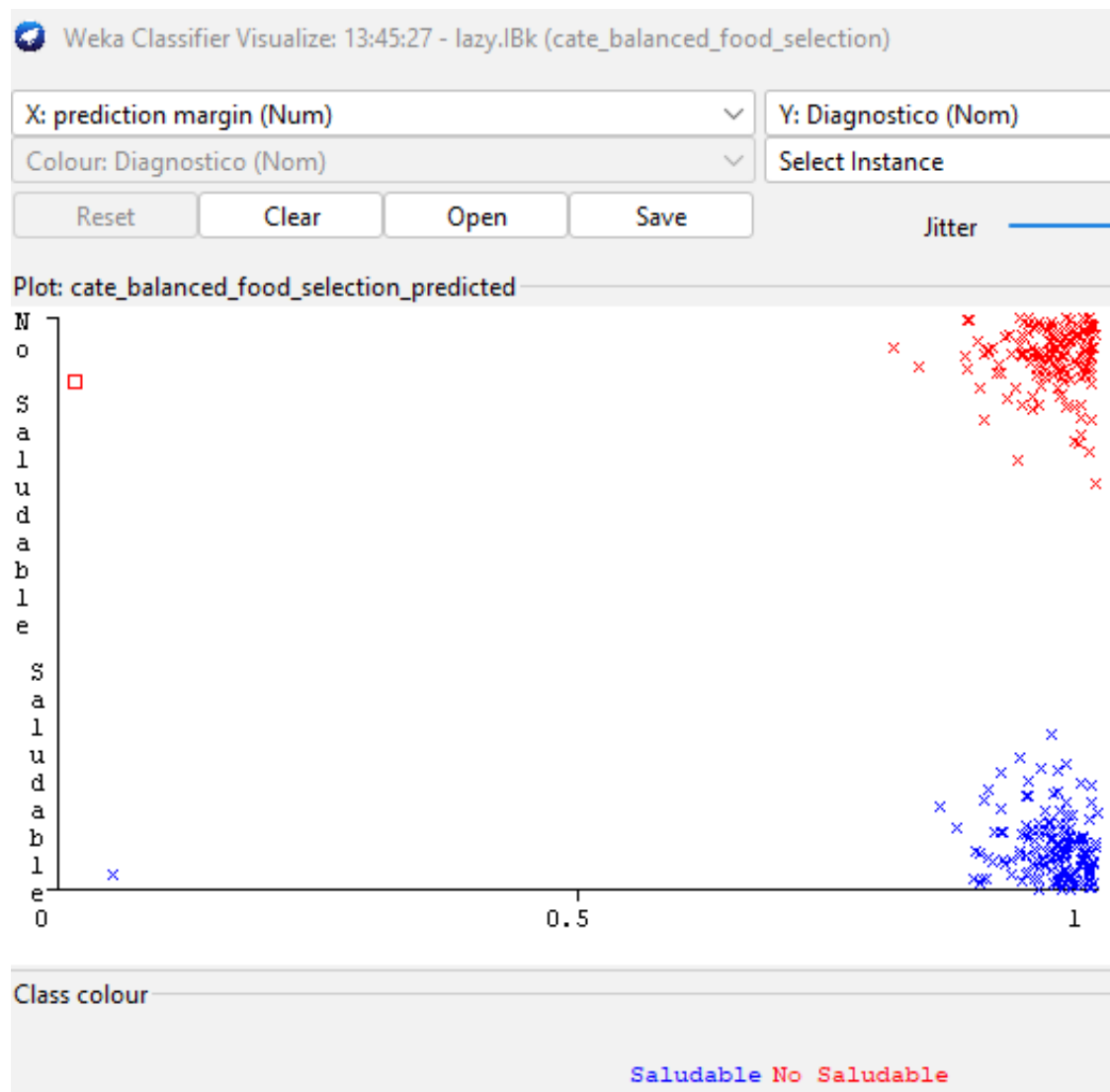


Fuente: Elaboración propia

Conforme a la figura 20, los valores de verdaderos positivos que resultaron es de 31 datos, para los verdaderos negativos se hallaron 23 datos, los falsos positivos son 5 datos y los falsos negativos son 6 datos. Esto quiere decir, que al comparar los valores predichos con los valores verdaderos la mayoría de los datos predichos sí coinciden con los datos verdaderos, el porcentaje del error en esta matriz de confusión equivale al 16.92 %, lo cual es un porcentaje mínimo y menos del 50 %.

Seguidamente, se realizó el modelo con la técnica de K vecinos más cercanos con el programa de Weka:

Figura 21. Alimentación (KNN): Modelo con KNN en Weka

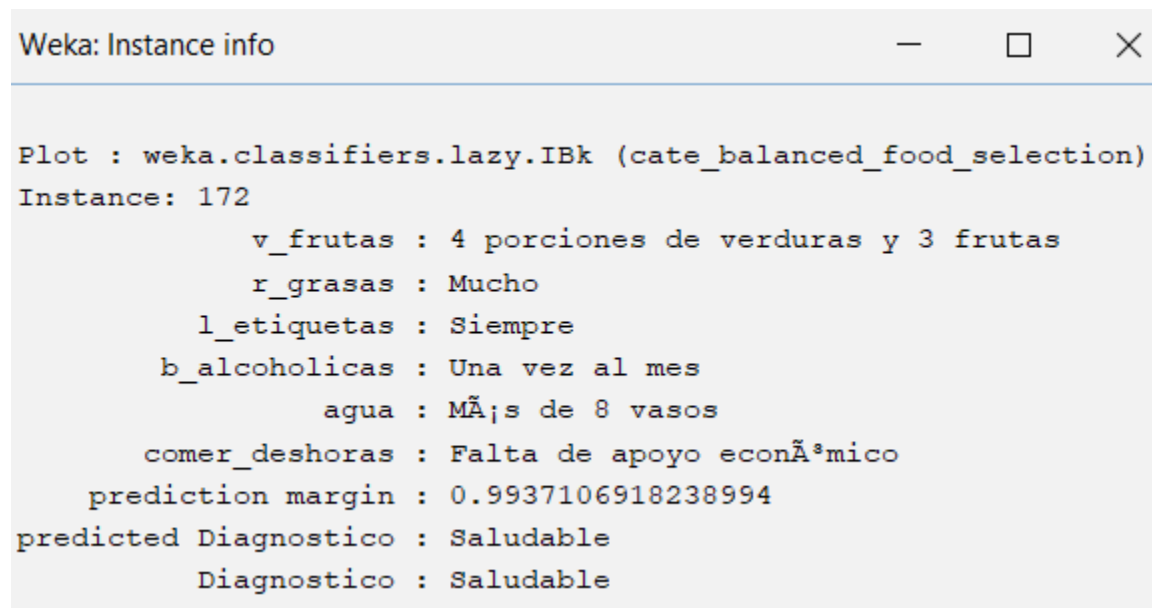


Fuente: Elaboración propia

Es factible visualizar que cada aspa de color azul son las predicciones correctas que dieron como “Saludable” y cada aspa de color rojo son las predicciones correctas que dieron como “No Saludable”, los cuadros azules son las predicciones que fallaron, o sea que predijeron como “Saludable” pero en realidad debieron de predecir como “No Saludable”, así mismo los cuadros rojos son las predicciones que fallaron, o sea predijeron como “No saludable” pero en realidad debieron de predecir como “Saludable”. Asimismo, se visualiza que cada clase busca estar junto a su vecino más cercano.

Al hacer click en cada aspa se genera una interfaz como se especifica en la figura 16:

Figura 22. Alimentación (KNN): Modelo en Weka parte 2



```

Weka: Instance info

Plot : weka.classifiers.lazy.IBk (cate_balanced_food_selection)
Instance: 172
      v_frutas : 4 porciones de verduras y 3 frutas
      r_grasas : Mucho
      l_etiquetas : Siempre
      b_alcoholicas : Una vez al mes
      agua : Más de 8 vasos
      comer_deshoras : Falta de apoyo económico
      prediction margin : 0.9937106918238994
      predicted Diagnostico : Saludable
      Diagnostico : Saludable
  
```

Fuente: Elaboración propia

El perfil de un estudiante que adopta hábitos para un estilo de vida saludable en la Alimentación donde la técnica de K vecinos más cercanos ha logrado predecir. Asimismo, se muestra más casos en la tabla 6:

Tabla 6. Alimentación (KNN): Hábitos del estudiante que adopta un estilo de vida saludable

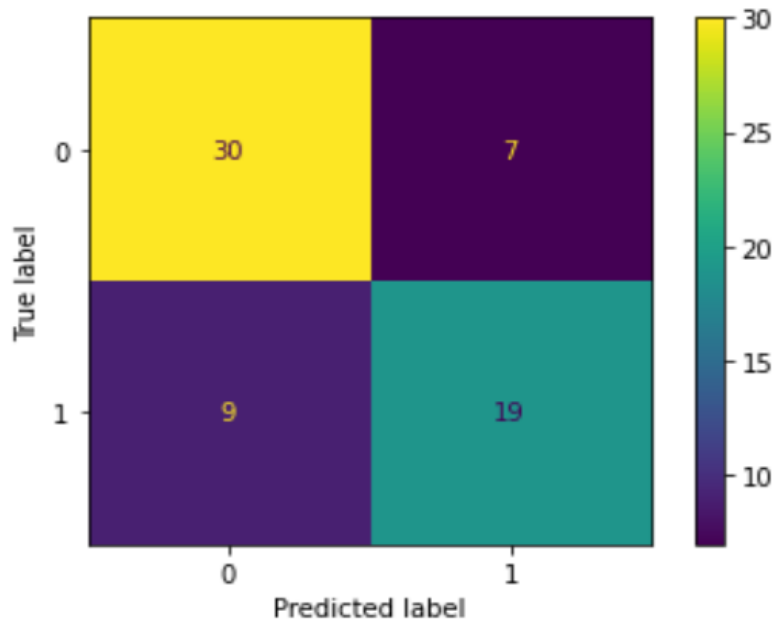
Reglas del modelo de K vecinos más cercanos de los estudiantes que adoptan un estilo de vida saludable	Variable de salida (Diagnóstico)
--	----------------------------------

Verduras y frutas = 5 porciones de verduras y 4 frutas Restricción de grasas = Demasiado Lectura de etiquetas = Frecuentemente Bebidas alcohólicas = Nunca Beber agua = Más de 8 vasos Comer a deshoras = Vivo lejos a la universidad	Saludable
Verduras y frutas = 5 porciones de verduras y 4 frutas Restricción de grasas = Demasiado Lectura de etiquetas = Siempre Bebidas alcohólicas = Una vez al mes Agua = Más de 8 vasos Comer a deshoras = Presentación de proyectos finales	Saludable
Verduras y frutas = 4 porciones de verduras y 3 frutas Restricción de grasas = Poco Lectura de etiquetas = A veces Bebidas alcohólicas = Menos de 12 veces al año Agua = Más 8 vasos Comer a deshoras = Vivo lejos a la universidad	Saludable

Fuente: Elaboración propia

A continuación, en la figura 23 se muestra la matriz de confusión para los datos de alimentación con la técnica de K vecinos más cercanos:

Figura 23. Alimentación: Matriz de confusión con KNN

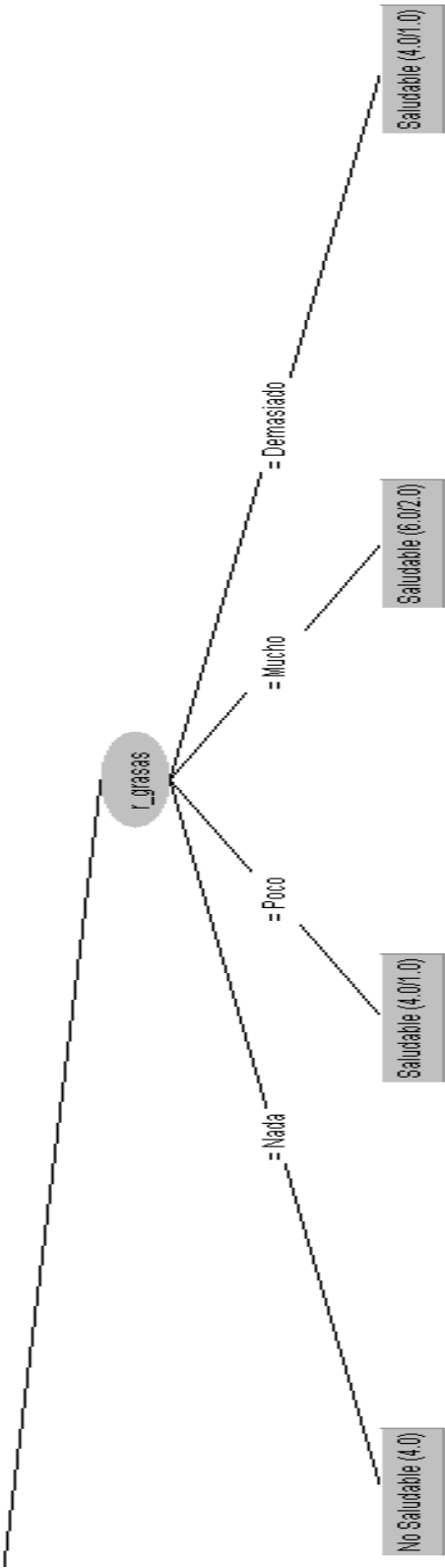


Fuente: Elaboración propia

Conforme a la figura 23, los valores de verdaderos positivos que resultaron es de 30 datos, para los verdaderos negativos se hallaron 19 datos, los falsos positivos son 9 datos y los falsos negativos son 7 datos. Esto quiere decir, que al comparar los valores predichos con los valores verdaderos la mayoría de los datos predichos sí coinciden con los datos verdaderos, el porcentaje del error en esta matriz de confusión equivale al 24.62 %, lo cual es un porcentaje mínimo y menos del 50 %.

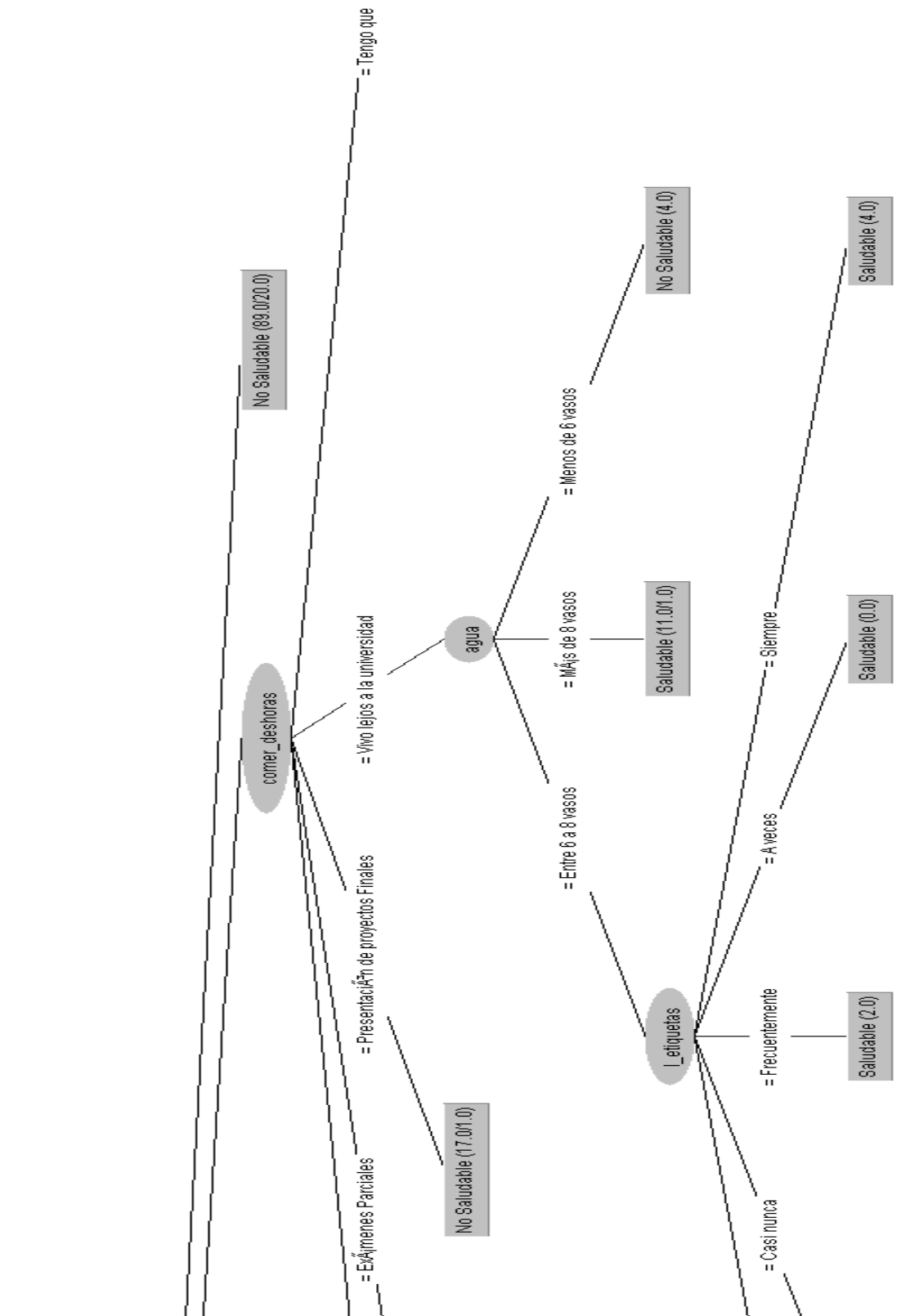
Seguidamente, conforme a la figura 24 la construcción del árbol de decisiones con los datos de alimentación, para analizar la información que contiene el árbol de decisión y como se utilizaría si un estudiante está adoptando un estilo de vida saludable o no. Se presenta en siete páginas el modelo del árbol de decisión para apreciarlo mejor.

Figura 24: Alimentación: Primera parte de Árbol de decisión



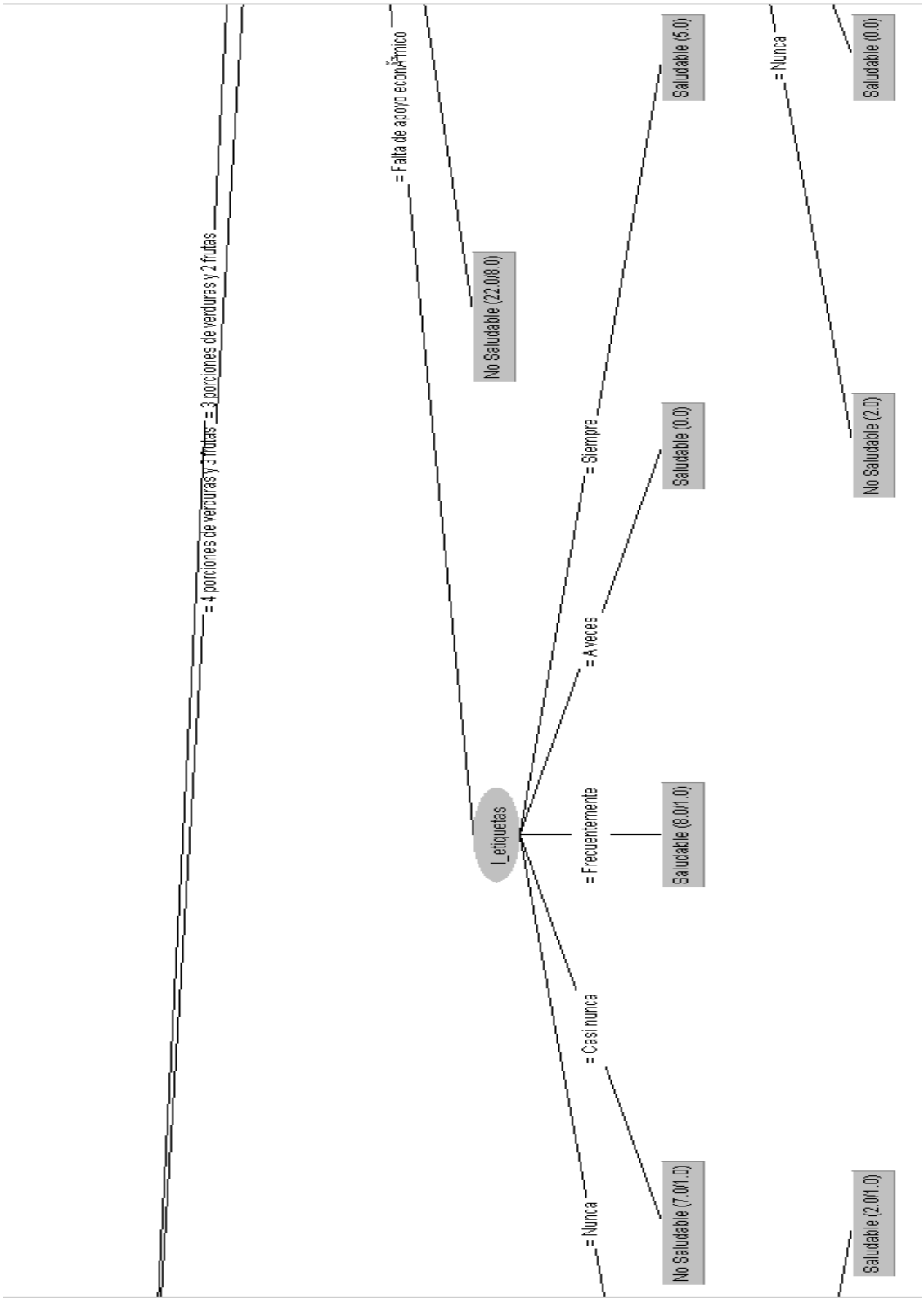
Fuente: Elaboración Propia.

Figura 25: Alimentación: Segunda parte de Árbol de decisión



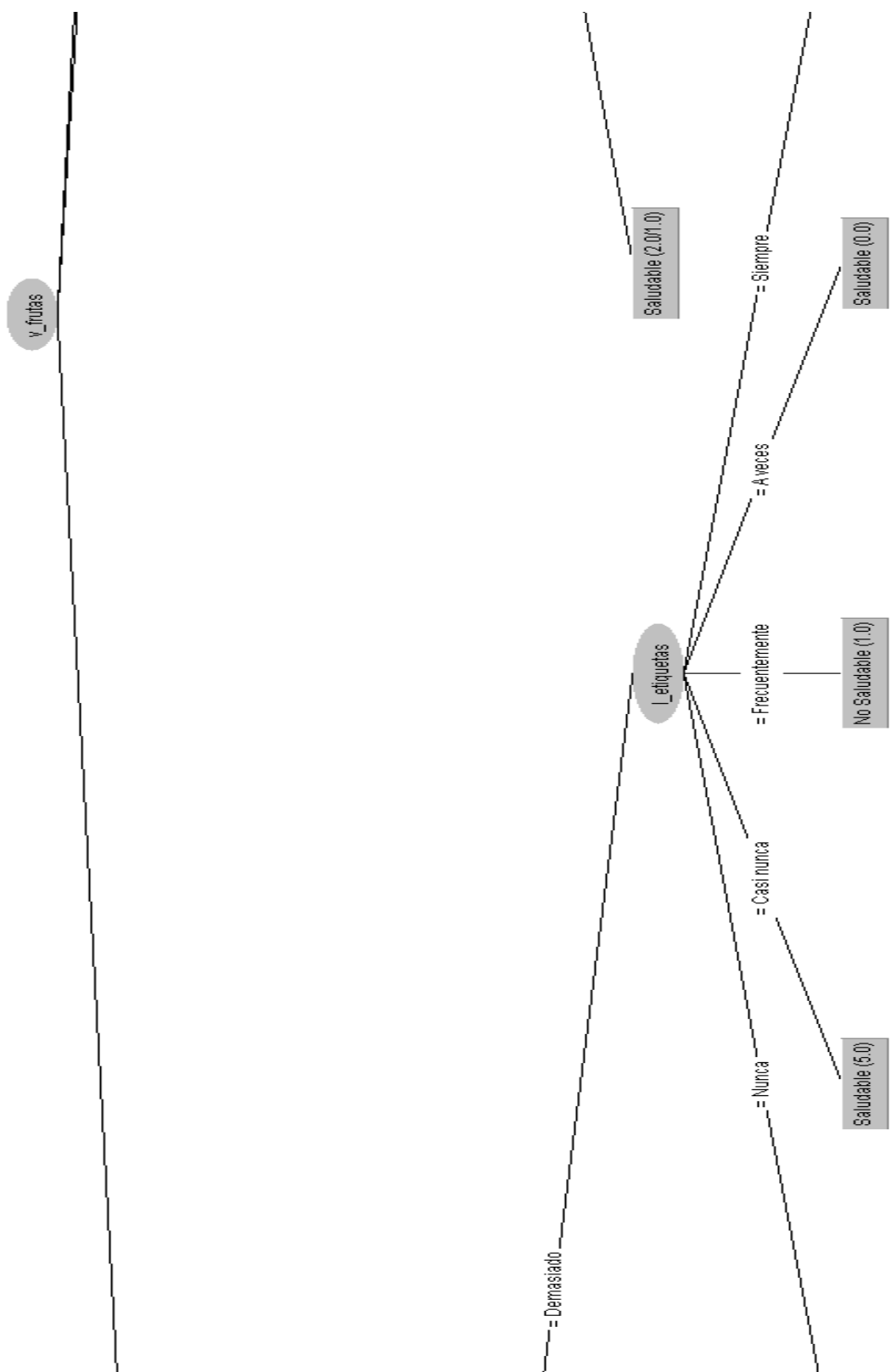
Fuente: Elaboración Propia.

Figura 26: Alimentación: Tercera parte de Árbol de decisión



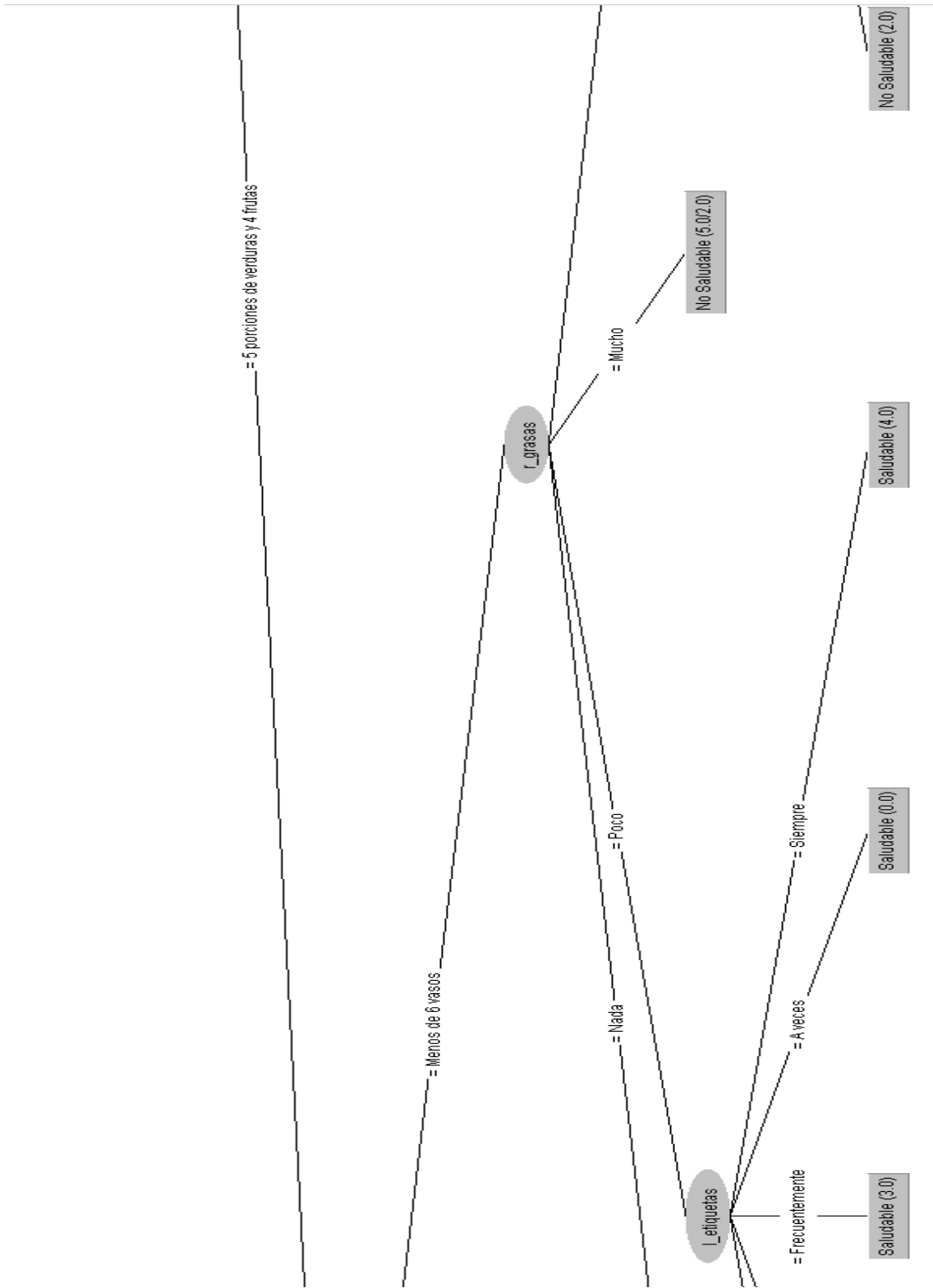
Fuente: Elaboración Propia.

Figura 27: Alimentación: Cuarta parte del Árbol de decisión



Fuente: Elaboración Propia.

Figura 28: Alimentación: Quinta parte del Árbol de decisión



Fuente: Elaboración Propia.

```

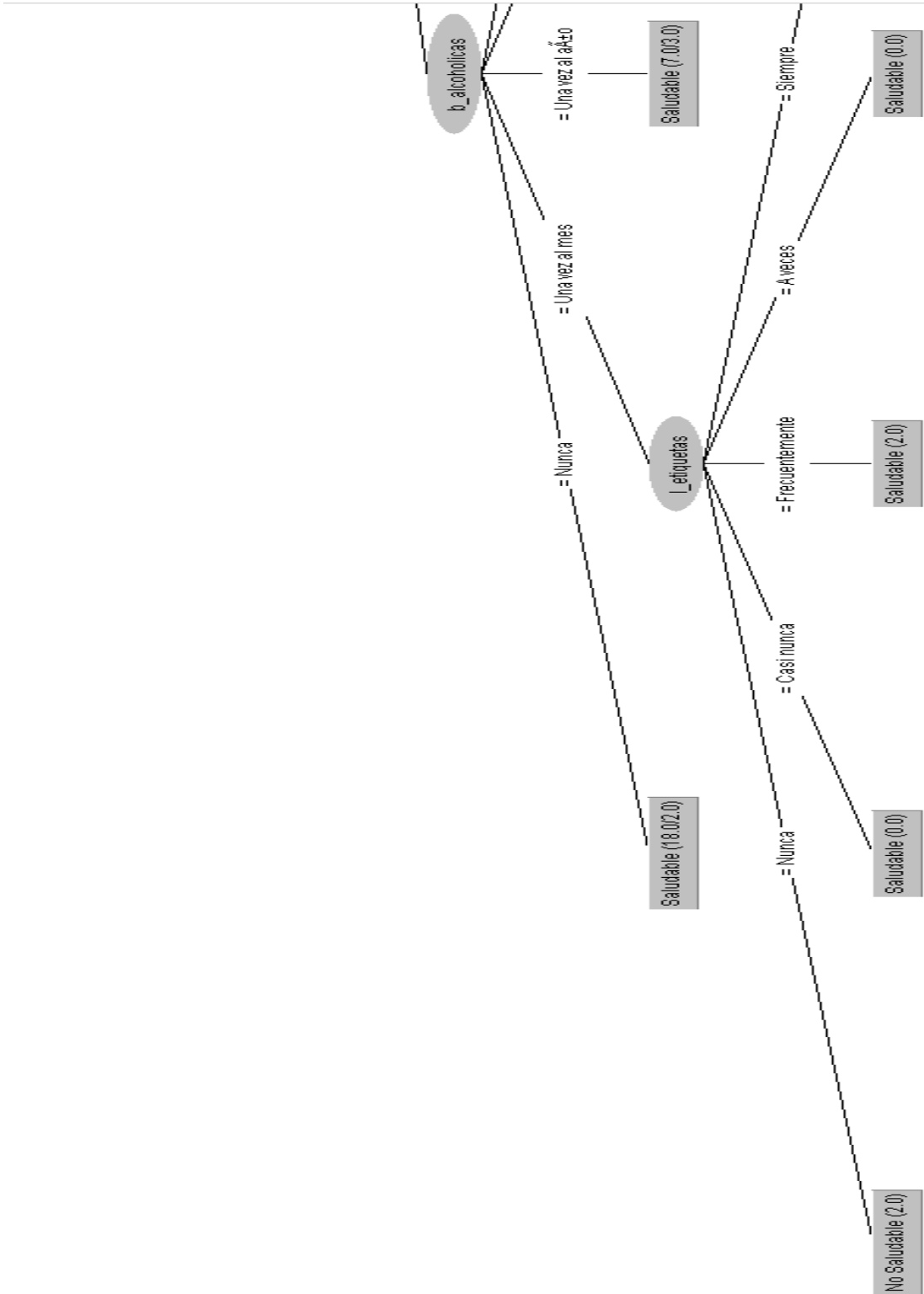
graph TD
    Root(( )) --- Agua((agua))
    Root --- Limpiador((limpiador))
    Root --- LimpiadorEcológico((limpiador ecológico))
    
    Agua --- P1[= Entre 6 a 8 vasos  
= 0.5]
    Agua --- P2[= Más de 8 vasos  
= 0.5]
    
    Limpiador --- P3[= Cada fin de semana  
= 0.5]
    Limpiador --- P4[= Menos de 12 veces al año  
= 0.5]
    
    LimpiadorEcológico --- P5[= Cada fin de semana  
= 0.5]
    LimpiadorEcológico --- P6[= Menos de 12 veces al año  
= 0.5]
    
    P1 --- R1[Saludable (62.07)]
    P2 --- R2[No Saludable (8.0)]
    
    P3 --- R3[Saludable (7.0)]
    P4 --- R4[No Saludable (4.0)]
    
    P5 --- R5[Saludable (7.0)]
    P6 --- R6[No Saludable (4.0)]
    
    R1 --- U1[= 62.07]
    R2 --- U2[= 8.0]
    
    R3 --- U3[= 7.0]
    R4 --- U4[= 4.0]
    
    R5 --- U5[= 7.0]
    R6 --- U6[= 4.0]
    
    U1 --- V1[= 62.07]
    U2 --- V2[= 8.0]
    
    U3 --- V3[= 7.0]
    U4 --- V4[= 4.0]
    
    U5 --- V5[= 7.0]
    U6 --- V6[= 4.0]
    
    V1 --- Final1[= 62.07]
    V2 --- Final2[= 8.0]
    
    V3 --- Final3[= 7.0]
    V4 --- Final4[= 4.0]
    
    V5 --- Final5[= 7.0]
    V6 --- Final6[= 4.0]
    
    Final1 --- FinalTotal[= 62.07]
    Final2 --- FinalTotal
    Final3 --- FinalTotal
    Final4 --- FinalTotal
    Final5 --- FinalTotal
    Final6 --- FinalTotal
  
```

El diagrama de árbol de decisión para la elección de un producto de limpieza se estructura de la siguiente manera:

- Nodo Raíz:** Se divide en tres ramas principales:
 - agua:** Conduce a un nodo de probabilidad con dos opciones:
 - = Entre 6 a 8 vasos (0.5):** Conduce a un nodo de resultado "Saludable (62.07)".
 - = Más de 8 vasos (0.5):** Conduce a un nodo de resultado "No Saludable (8.0)".
 - limpiador:** Conduce a un nodo de probabilidad con dos opciones:
 - = Cada fin de semana (0.5):** Conduce a un nodo de resultado "Saludable (7.0)".
 - = Menos de 12 veces al año (0.5):** Conduce a un nodo de resultado "No Saludable (4.0)".
 - limpiador ecológico:** Conduce a un nodo de probabilidad con dos opciones:
 - = Cada fin de semana (0.5):** Conduce a un nodo de resultado "Saludable (7.0)".
 - = Menos de 12 veces al año (0.5):** Conduce a un nodo de resultado "No Saludable (4.0)".
- Valores de Utilidad:** Los valores de utilidad se calculan para cada rama de probabilidad:
 - Para **agua**: $0.5 \times 62.07 + 0.5 \times 8.0 = 34.535$
 - Para **limpiador**: $0.5 \times 7.0 + 0.5 \times 4.0 = 5.5$
 - Para **limpiador ecológico**: $0.5 \times 7.0 + 0.5 \times 4.0 = 5.5$
- Elección Final:** El valor de utilidad para **agua** es el más alto (34.535), por lo que es la opción más saludable.

76

Figura 30: Alimentación: Séptima parte del Árbol de decisión



Fuente: Elaboración Propia.

En la mayoría de las flechas de lado izquierdo pertenecen a la respuesta “Saludable” y las de lado derecho a la respuesta de “No saludable”. El árbol de decisión encontró 41 reglas de decisión que predicen el estilo de vida que adoptan los estudiantes universitarios de la Universidad Nacional José María Arguedas. Gracias a estas reglas se puede conocer los hábitos de vida de los estudiantes que adoptan un estilo de vida saludable y no saludable.

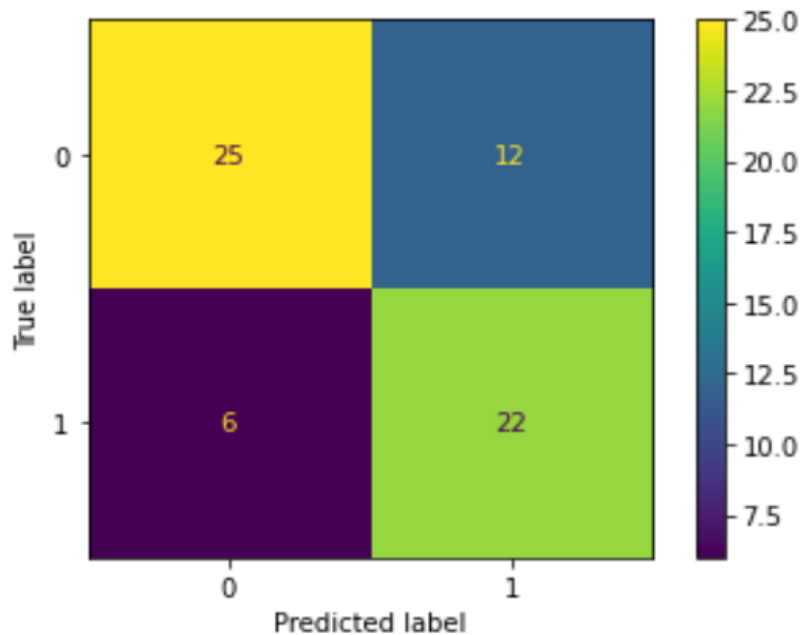
Tabla 7: Hábitos del estudiante que adopta un estilo de vida saludable

Reglas del modelo de árbol de decisión de los estudiantes que adoptan un estilo de vida saludable	Variable de salida (Diagnóstico)
Verduras y frutas = 5 porciones de verduras y 4 frutas Beber agua = Entre 6 a 8 vasos Bebidas alcohólicas = Nunca	Saludable
Verduras y frutas = 5 porciones de verduras y 4 frutas Beber agua = Entre 6 a 8 vasos Bebidas alcohólicas = Una vez al mes Lectura de etiquetas = Siempre	Saludable
Verduras y frutas = 4 porciones de verduras y 3 frutas Comer a deshoras = Falta de apoyo económico Lectura de etiquetas = Siempre	Saludable
Verduras y frutas = 4 porciones de verduras y 3 frutas Comer a deshoras = Tengo trabajar y estudiar Restricción de grasas = Demasiado	Saludable

Fuente: Elaboración Propia.

A continuación, en la figura 31 se muestra la matriz de confusión para los datos de alimentación con el algoritmo de Árboles de decisiones:

Figura 31. Alimentación: Matriz de confusión con Árboles de decisiones



Fuente: Elaboración propia

Conforme a la figura 31, los valores de verdaderos positivos que resultaron es de 25 datos, para los verdaderos negativos se hallaron 22 datos, los falsos positivos son 6 datos y los falsos negativos son 12 datos. Esto quiere decir, que al comparar los valores predichos con los valores verdaderos la mayoría de los datos predichos sí coinciden con los datos verdaderos, el porcentaje del error en esta matriz de confusión equivale al 27.69 %, lo cual es un porcentaje mínimo y menos del 50 %.

Seguidamente, con los resultados obtenidos en las matrices de confusión mostradas en las figuras 20, 23 y 31 se pudieron realizar las fórmulas de las métricas de evaluación propuestas, por eso en la tabla 8 se muestra los resultados de las métricas de evaluación halladas con las técnicas de Machine Learning Máquinas de Vectores de Soporte, K vecinos más cercanos y Árboles de decisiones:

Tabla 8. Métricas de Evaluación en la Alimentación

Alimentación: Algoritmos Supervisados

Métricas de evaluación	Máquinas de Vectores de Soporte	K vecinos más cercanos	Árboles de decisiones
Exactitud	83.08 %	75.38 %	73.85 %
Precisión	79.31 %	73.08 %	66.67 %
Tiempo total	0.0088	0.0175	0.0123

Fuente: Elaboración propia.

Se describe que la técnica que mayor resultado obtuvo es **Máquinas de Vectores de Soporte** con una exactitud de **83.08 %**, una precisión de **79.31 %**, marcó un tiempo total de **0.0088** segundos. Por tanto, la técnica de **Máquinas de Vectores de Soporte** es el más adecuado para la predicción de los estilos de vida en la Alimentación en los estudiantes universitarios.

Finalmente, se realizó la evaluación con validación cruzada con cinco pliegues que permitió dar un valor numérico, el cual indica que tan bueno es el modelo de Machine Learning. En código Python esto se realizó a través de una función llamada “cross_val_score”. Se muestra la tabla 9 de comparación de los resultados obtenidos sin validación cruzada y con validación cruzada con el conjunto de datos de Alimentación:

Tabla 9. Alimentación: Resultados con y sin validación cruzada

Valor de predicción	Sin Validación cruzada	Con Validación cruzada
	Máquinas de Vectores de Soporte	
puntaje	88.92 %	87.66 %
Algoritmo	K vecinos más cercanos	
puntaje	90.19 %	82.60 %
Algoritmo	Árboles de decisiones	
puntaje	99.68 %	80.38 %

Fuente: Elaboración propia.

Se puede visualizar que el **puntaje** con la validación cruzada es menor lo cual es más representativo, el puntaje aumenta la certeza de la predicción, entonces es importante que no se sobreestime la capacidad del modelo de aprendizaje automático. En este caso, el modelo para predecir los estilos de vida en la alimentación está basado en la técnica de Árboles de decisiones.

Por último, se acepta la hipótesis alterna indicando que, de acuerdo a los resultados presentados en la tabla 5, 8 y 9 el modelo para predecir los estilos de vida en la alimentación está basado en la técnica de Máquina de Vectores de Soporte.

5.3.2. Hipótesis Específica H2

H₀: El modelo para predecir los estilos de vida en la actividad física no está basado en la técnica de Árboles de decisiones.

H₁: El modelo para predecir los estilos de vida en la actividad física está basado en la técnica de Árboles de decisiones.

Para justificar la hipótesis específica 2, primeramente, una vez descargado los datos recolectados de la actividad física en formato Excel se procedió a cargar los datos a Google Colab como se demuestra.

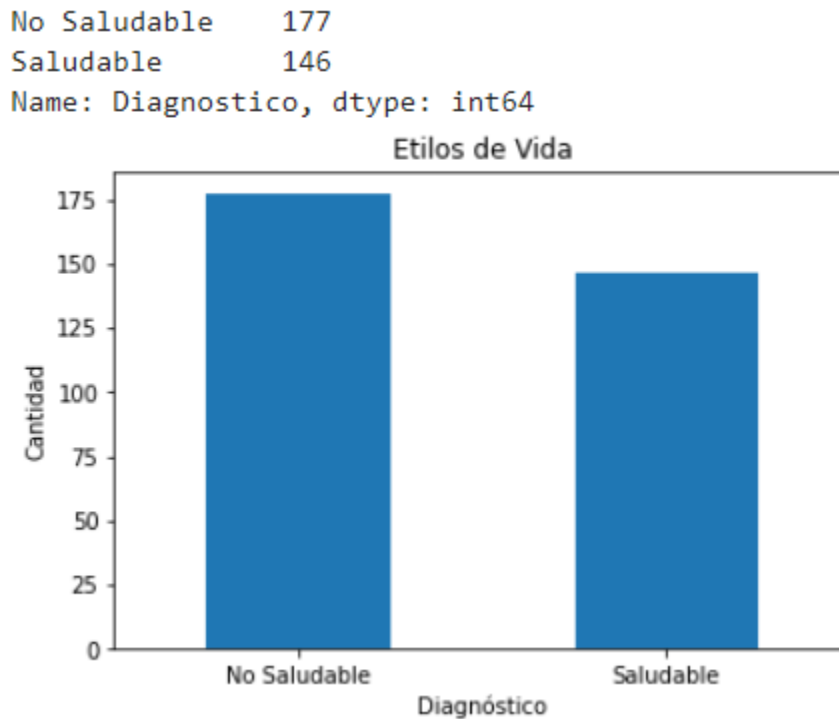
Figura 32. Variables de entrada (features) de la Actividad Física en Google Colab

edad	sexo	e_profesional	c_academico	trabaja	e_civil	a_moderadas	a_intensas	f_muscular	f_tendones	f_televisor	t_caminar	Diagnostico
Más de 21 años	Femenino	EPC	V	No	Soltero	2 h y 30 min a 5 h	1 h y 15 min a 2 h y 30 min	Menos de 2 días	Más 30 s	Más de 10 h	Más de 30 min	No Saludable
Más de 21 años	Masculino	EPIS	X	Sí	Soltero	Más de 5 h	Más de 2 h y 30 min	Más de 2 días	15 a 30 s	4 a 7 h	30 min	Saludable
Entre 18 a 21 años	Masculino	EPIS	IV	No	Soltero	Menos de 2 h y 30 min	Menos de 1 h y 15 min	Menos de 2 días	Menos de 15 s	Menos de 4 h	30 min	No Saludable
Menos de 18 años	Masculino	EPIAM	I	Sí	Soltero	Más de 5 h	1 h y 15 min a 2 h y 30 min	Más de 2 días	Más 30 s	2 h	Más de 30 min	Saludable
Entre 18 a 21 años	Masculino	EPEPI	IV	No	Soltero	Más de 5 h	Más de 2 h y 30 min	Más de 2 días	15 a 30 s	7 a 10 h	30 min	Saludable
Más de 21 años	Masculino	EPIAM	III	Sí	Soltero	Más de 5 h	Menos de 1 h y 15 min	Menos de 2 días	15 a 30 s	Más de 10 h	Más de 30 min	No Saludable
Más de 21 años	Masculino	EPEPI	IV	Sí	Soltero	2 h y 30 min a 5 h	Menos de 1 h y 15 min	Menos de 2 días	15 a 30 s	Menos de 4 h	Más de 30 min	Saludable
Entre 18 a 21 años	Femenino	EPEPI	IV	Sí	Soltero	Menos de 2 h y 30 min	Menos de 1 h y 15 min	Igual a 2 días	Menos de 15 s	7 a 10 h	Más de 30 min	No Saludable
Más de 21 años	Masculino	EPIA	VI	Sí	Soltero	2 h y 30 min a 5 h	1 h y 15 min a 2 h y 30 min	Más de 2 días	Más 30 s	4 a 7 h	Más de 30 min	Saludable
Entre 18 a 21 años	Masculino	EPC	II	No	Soltero	Menos de 2 h y 30 min	Menos de 1 h y 15 min	Menos de 2 días	Ninguno	Más de 10 h	30 min	No Saludable
Entre 18 a 21 años	Femenino	EPIA	I	Sí	Soltero	Menos de 2 h y 30 min	Menos de 1 h y 15 min	Más de 2 días	Menos de 15 s	4 a 7 h	Más de 30 min	No Saludable
Más de 21 años	Femenino	EPC	VI	Sí	Soltero	Menos de 2 h y 30 min	Menos de 1 h y 15 min	Menos de 2 días	Menos de 15 s	4 a 7 h	Más de 30 min	No Saludable
Entre 18 a 21 años	Femenino	EPEPI	IV	Sí	Soltero	Menos de 2 h y 30 min	Menos de 1 h y 15 min	Menos de 2 días	15 a 30 s	Más de 10 h	Más de 30 min	No Saludable
Entre 18 a 21 años	Masculino	EPIAM	I	Sí	Soltero	Más de 5 h	Más de 2 h y 30 min	Igual a 2 días	Más 30 s	Menos de 4 h	Más de 30 min	Saludable

Fuente: Elaboración propia

A continuación, se realizó una visualización de datos con el lenguaje de Python y la librería pandas acerca de las clases que existe en la variable de salida llamada “diagnóstico” para los datos de la actividad física:

Figura 33. Actividad Física: Visualización de datos de las clases de la variable de salida

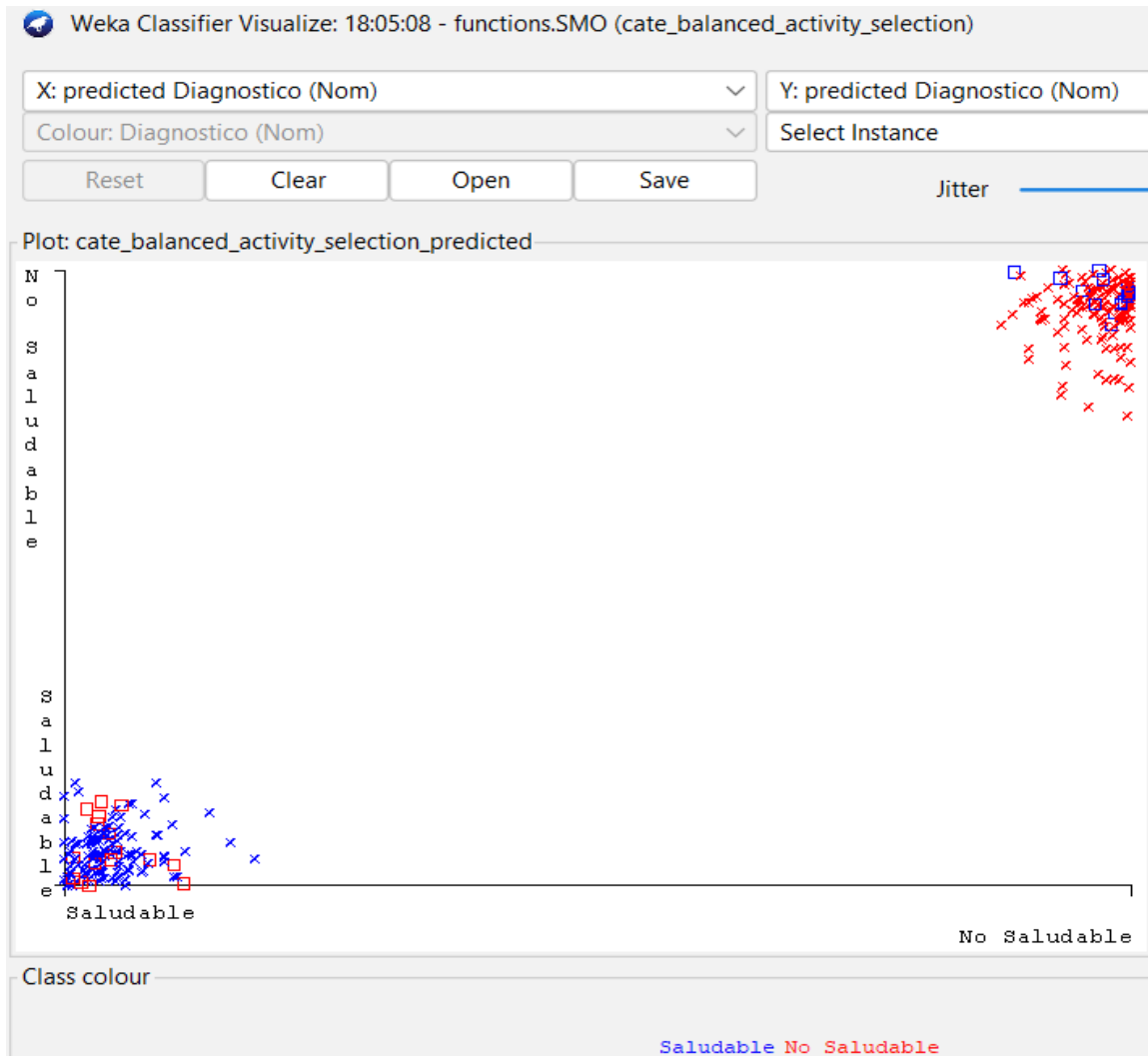


Fuente: Elaboración propia

Como se visualiza en la figura 33 se reportaron 177 datos que pertenecen a la clase de no saludable que pertenecen a aquellos estudiantes universitarios que adoptan un estilo de vida no saludable y 146 datos que pertenecen a la clase de saludable que pertenecen a aquellos estudiantes que adoptan un estilo de vida saludable.

Consecutivamente, se muestra el modelo con la técnica de Máquina de Vectores de Soporte realizada en el programa de Weka:

Figura 34. Actividad Física (SVM): Modelo con SVM en Weka



Fuente: Elaboración propia

Se puede visualizar que cada aspa de color azul son las predicciones correctas que dieron como “Saludable” y cada aspa de color rojo son las predicciones correctas que dieron como “No Saludable”, los cuadros azules que se encuentran en la parte de las predicciones “No Saludable”, son las predicciones que fallaron, o sea que predijeron como “Saludable” pero en realidad debieron de predecir como “No Saludable”, así mismo los cuadros rojos que se encuentran en la parte de lo “Saludable” son las predicciones que fallaron, o sea predijeron como “No saludable” pero en realidad debieron de predecir como “Saludable”.

Al hacer click en cada aspa se genera una interfaz como se demuestra:

Figura 35. Actividad Física (SVM): Modelo en Weka parte 2

```

Weka: Instance info

Plot : weka.classifiers.functions.SMO (cate_balanced_
Instance: 86
    a_moderadas : 2 h y 30 min a 5 h
    a_intensas : 1 h y 15 min a 2 h y 30 min
    f_muscular : Igual a 2 días-as
    f_tendones : 15 a 30 s
    prediction margin : 1.0
    predicted Diagnostico : Saludable
    Diagnostico : Saludable

```

Fuente: Elaboración propia

Se muestra el perfil de un estudiante que adopta hábitos para un estilo de vida saludable en la Actividad Física que la técnica de Máquina de Vectores de Soporte ha logrado predecir. Asimismo, se muestra más casos en la tabla 10:

Tabla 10. Actividad Física (SVM): Hábitos del estudiante que adopta un estilo de vida saludable

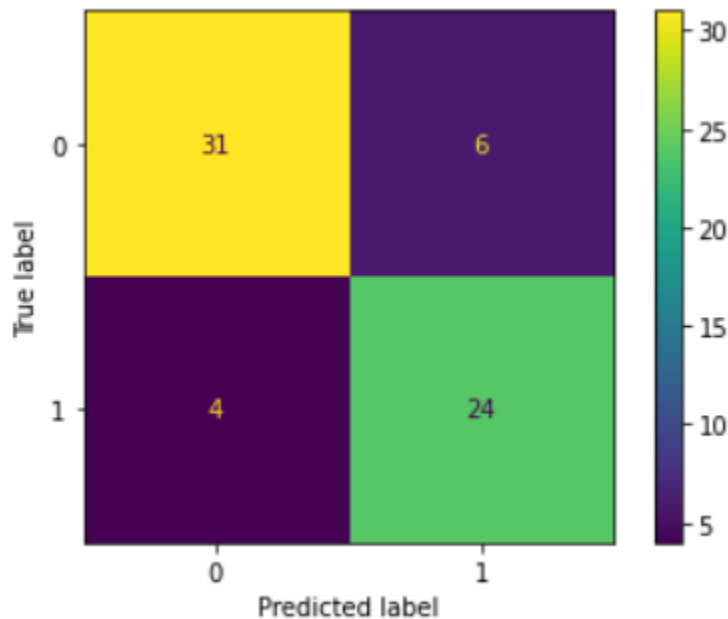
Reglas del modelo de Máquina de Vectores de Soporte de los estudiantes que adoptan un estilo de vida saludable	Variable de salida (Diagnóstico)
Actividades físicas aeróbicas moderadas = 2 h y 30 min a 5 h Actividades físicas aeróbicas intensas = 1 h y 15 min a 2 h y 30 min Actividades físicas de fortalecimiento muscular = Menos de 2 días Actividades físicas de flexibilidad en tendones = 15 a 30 s	Saludable
Actividades físicas aeróbicas moderadas = 2 h y 30 min a 5 h Actividades físicas aeróbicas intensas = 1 h y 15 min a 2 h y 30 min Actividades físicas de fortalecimiento muscular = Igual a 2 días	Saludable

Actividades físicas de flexibilidad en tendones = Más de 30 s	
Actividades físicas aeróbicas moderadas = Más de 5 h	Saludable
Actividades físicas aeróbicas intensas = Más de 2 h y 30 min	
Actividades físicas de fortalecimiento muscular = Más de 2 días	
Actividades físicas de flexibilidad en tendones = Más de 30 s	

Fuente: Elaboración propia

A continuación, en la figura 36 se muestra la matriz de confusión para los datos de actividad física con el algoritmo de Máquina de Vectores de Soporte:

Figura 36. Actividad Física: Matriz de confusión con SVM



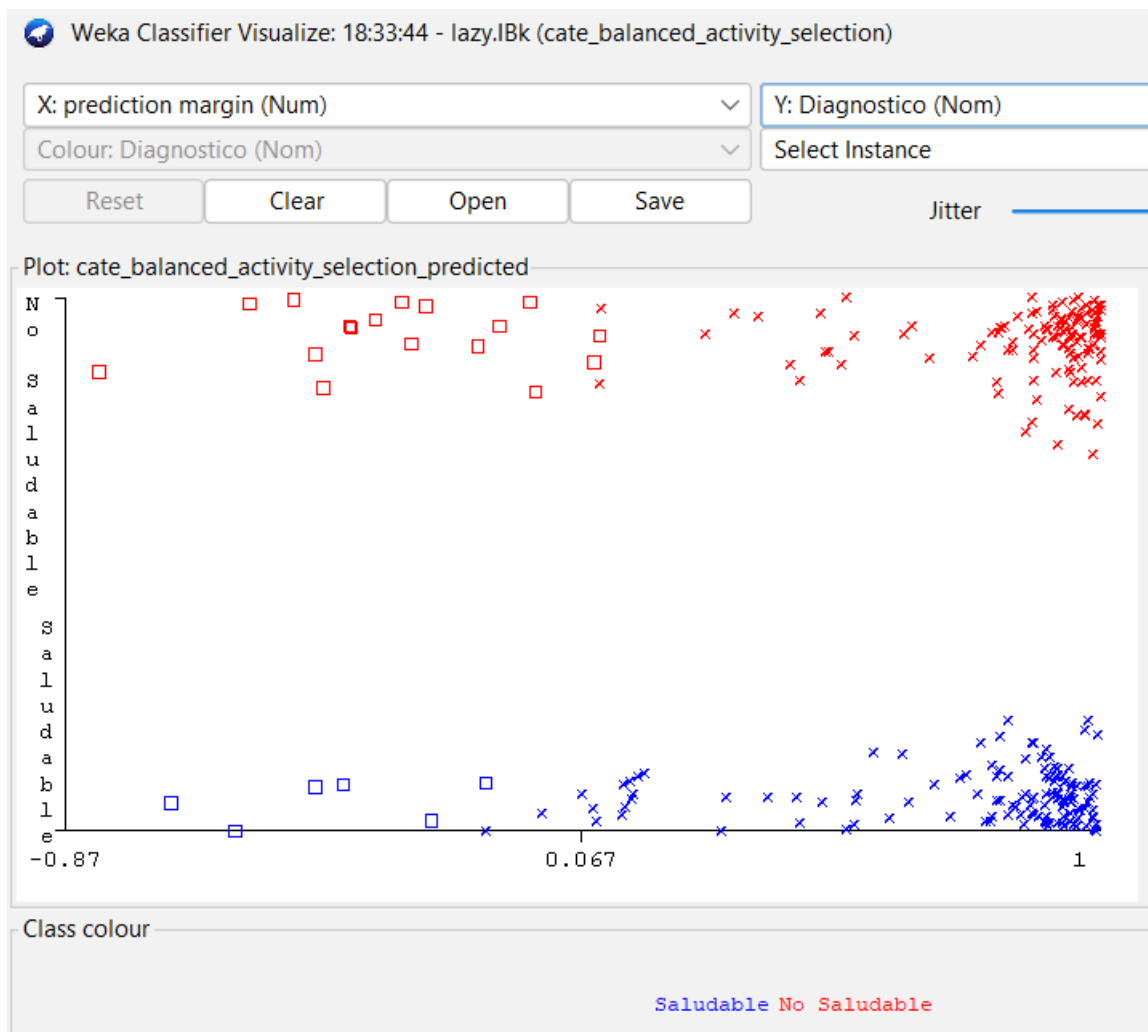
Fuente: Elaborado propia

De acuerdo a la figura 36, los valores de verdaderos positivos que resultaron es de 31 datos, para los verdaderos negativos se hallaron 24 datos, los falsos positivos son 4 datos y los falsos negativos son 6 datos. Esto quiere decir, que al comparar los valores predichos con los valores verdaderos la mayoría de los datos predichos sí coinciden con los datos verdaderos, el

porcentaje del error en esta matriz de confusión equivale al 15.38 %, lo cual es un porcentaje mínimo y menos del 50 %.

Seguidamente, se realizó el modelo con la técnica de K vecinos más cercanos con el programa de Weka:

Figura 37. Actividad Física (KNN): Modelo con KNN en Weka



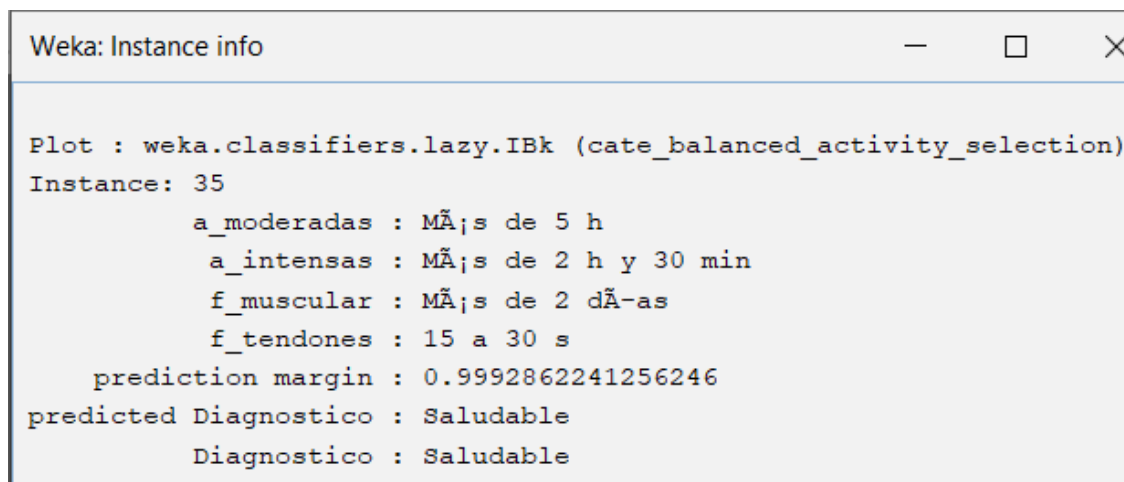
Fuente: Elaboración propia.

Se puede visualizar que cada aspa de color azul son las predicciones correctas que dieron como “Saludable” y cada aspa de color rojo son las predicciones correctas que dieron como “No Saludable”, los cuadros azules son las predicciones que fallaron, o sea que predijeron como “Saludable” pero en realidad debieron de predecir como “No Saludable“, así mismo los cuadros rojos son las predicciones que fallaron, o sea predijeron como “No saludable” pero en realidad

debieron de predecir como “Saludable”. Asimismo, se visualiza que cada clase busca estar junto a su vecino más cercano.

Al hacer click en cada aspa se genera una interfaz como se muestra en la figura 38:

Figura 38. Actividad Física (KNN): Modelo con KNN en Weka parte 2



```

Weka: Instance info

Plot : weka.classifiers.lazy.IBk (cate_balanced_activity_selection)
Instance: 35
    a_moderadas : MÃ;s de 5 h
    a_intensas : MÃ;s de 2 h y 30 min
    f_muscular : MÃ;s de 2 dÃ-as
    f_tendones : 15 a 30 s
    prediction margin : 0.9992862241256246
    predicted Diagnostico : Saludable
    Diagnostico : Saludable
  
```

Fuente: Elaboración propia

Se muestra el perfil de un estudiante que adopta hábitos para un estilo de vida saludable en la Actividad Física donde la técnica de K vecinos más cercanos ha logrado predecir.

Asimismo, se muestra más casos en la tabla 11:

Tabla 11. Actividad Física (KNN): Hábitos del estudiante que adopta un estilo de vida saludable

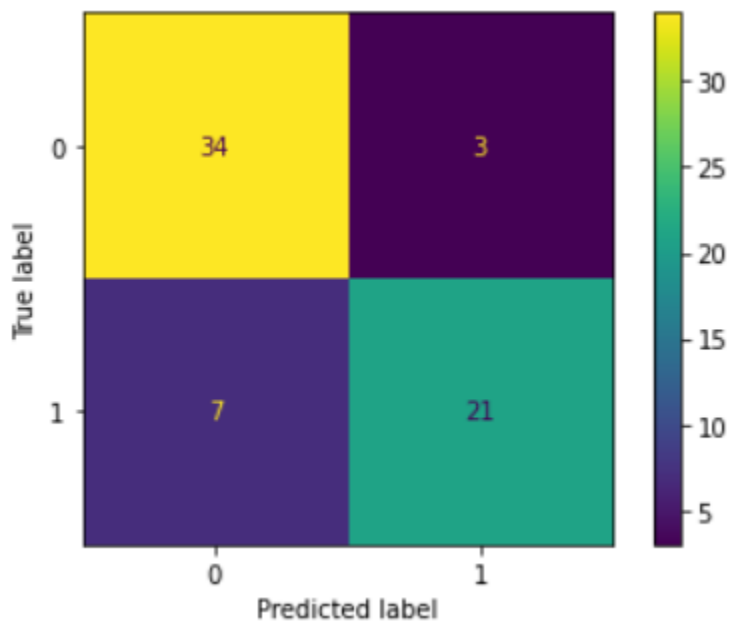
Reglas del modelo con K vecinos más cercanos para los estudiantes que adoptan un estilo de vida saludable	Variable de salida (Diagnóstico)
Actividades físicas aeróbicas moderadas = 2 h y 30 min a 5 h Actividades físicas aeróbicas intensas = 1 h y 15 min a 2 h y 30 min Actividades físicas de fortalecimiento muscular = Igual a 2 días Actividades físicas de flexibilidad en tendones = 15 a 30 s	Saludable
Actividades físicas aeróbicas moderadas = Más de 5 h	Saludable

Actividades físicas aeróbicas intensas = Más de 2 h y 30 min Actividades físicas de fortalecimiento muscular = Más de 2 días Actividades físicas de flexibilidad en tendones = Menos de 15 s	
Actividades físicas aeróbicas moderadas = 2 h y 30 min a 5 h Actividades físicas aeróbicas intensas = 1 h y 15 min a 2 h y 30 min Actividades físicas de fortalecimiento muscular = Igual a 2 días Actividades físicas de flexibilidad en tendones = Más de 30 s	Saludable

Fuente: Elaboración propia

A continuación, en la figura 39 se muestra la matriz de confusión para los datos de actividad física con el algoritmo de K vecinos más cercanos:

Figura 39. Actividad Física: Matriz de confusión con KNN

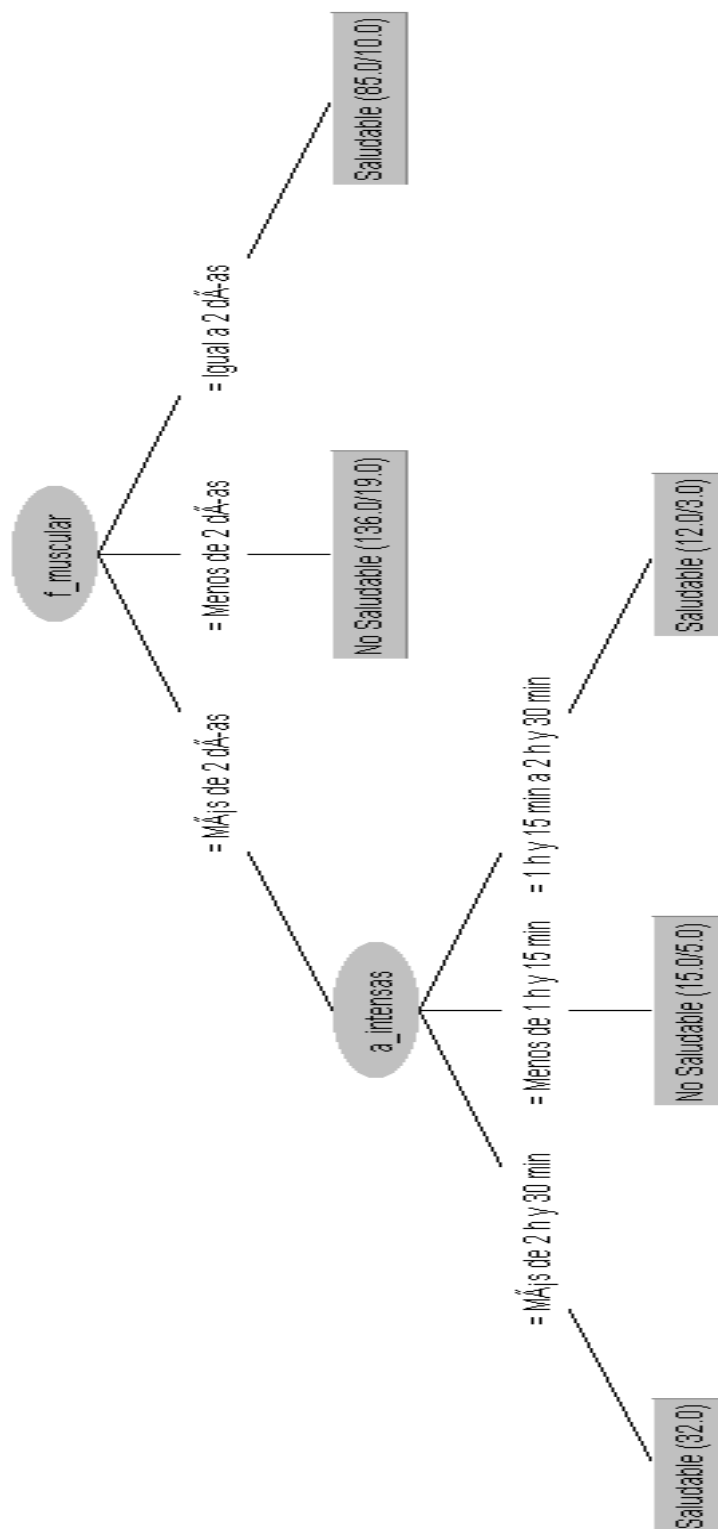


Fuente: Elaboración propia

De acuerdo a la figura 39, los valores de verdaderos positivos que resultaron es de 34 datos, para los verdaderos negativos se hallaron 20 datos, los falsos positivos son 7 datos y los falsos negativos son 3 datos. Esto quiere decir, que al comparar los valores predichos con los valores verdaderos la mayoría de los datos predichos sí coinciden con los datos verdaderos, el porcentaje del error en esta matriz de confusión equivale al 15.38 %, lo cual es un porcentaje mínimo y menos del 50 %.

Seguidamente, se muestra en la figura 40 la construcción del árbol de decisiones con los datos de Actividad Física, para analizar la información que contiene el árbol de decisión y como se utilizaría si un estudiante está adoptando un estilo de vida saludable o no. Se presenta el modelo del árbol de decisión:

Figura 40. Actividad Física: Modelo del Árbol de decisión



Fuente: Elaboración Propia.

Para la interpretación de la figura anterior del árbol de decisión, en la mayoría de las flechas de lado izquierdo pertenecen a la respuesta “Saludable” y las de lado derecho a la respuesta de “No saludable”. El árbol de decisión encontró 5 reglas de decisión que predicen el estilo de vida que adoptan los estudiantes universitarios de la Universidad Nacional José María Arguedas, gracias a estas reglas se puede conocer los hábitos de vida de los estudiantes que adoptan un estilo de vida saludable y no saludable en la Actividad Física.

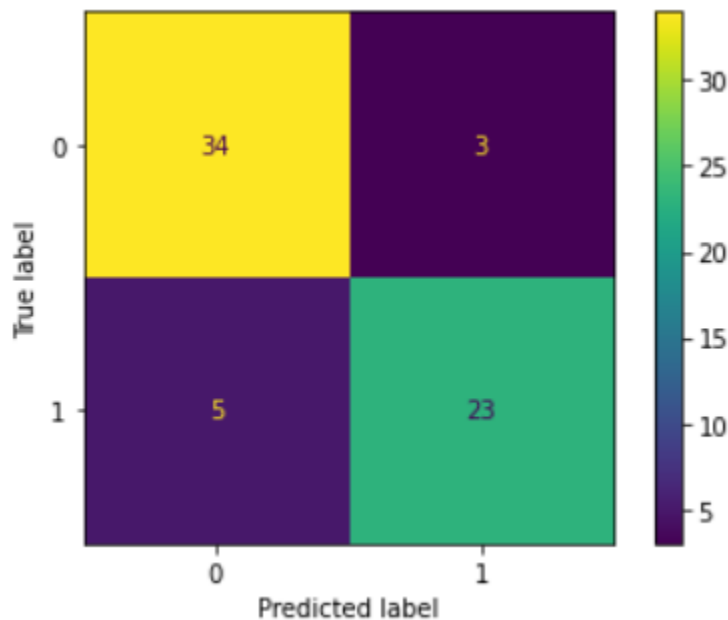
Tabla 12. Actividad Física: Hábitos del estudiante que adopta un estilo de vida saludable

Reglas del modelo del Árbol de decisión para los estudiantes que adoptan un estilo de vida saludable	Variable de salida (Diagnóstico)
Actividades físicas de fortalecimiento muscular = Más de 2 días Actividades físicas aeróbicas intensas = Más de 2 h y 30 min	Saludable
Actividades físicas de fortalecimiento muscular = Más de 2 días Actividades físicas aeróbicas intensas = 1 h y 15 min a 2 h y 30 min	Saludable
Actividades físicas de fortalecimiento muscular = Igual a 2 días	Saludable

Fuente: Elaboración Propia.

A continuación, en la figura 41 se muestra la matriz de confusión para los datos de actividad física con el algoritmo de Árboles de decisiones:

Figura 41. Actividad Física: Matriz de confusión con Árboles de decisiones



Fuente: Elaboración propia

De acuerdo a la figura 41, los valores de verdaderos positivos que resultaron es de 34 datos, para los verdaderos negativos se hallaron 23 datos, los falsos positivos son 5 datos y los falsos negativos son 3 datos. Esto quiere decir, que al comparar los valores predichos con los valores verdaderos la mayoría de los datos predichos sí coinciden con los datos verdaderos, el porcentaje del error en esta matriz de confusión equivale al 12.31 %, lo cual es un porcentaje mínimo y menos del 50 %.

Seguidamente, con los resultados obtenidos en las matrices de confusión mostradas en las figuras 36, 39 y 41 se pudieron realizar las fórmulas de las métricas de evaluación propuestas, por eso en la siguiente tabla se muestra los resultados de las métricas de evaluación halladas con las técnicas de Machine Learning Máquinas de Vectores de Soporte, K vecinos más cercanos y Árboles de decisiones con el lenguaje de programación de Python:

Tabla 13. Métricas de evaluación en la Actividad Física

Métricas de evaluación	Actividad Física: Algoritmos de Machine Learning		
	Máquinas de Vectores de Soporte	K vecinos más cercanos	Árboles de decisiones
Exactitud	84.62 %	84.62 %	87.69 %
Precisión	80.00 %	87.50 %	88.46 %
Tiempo total	0.0066	0.0318	0.0268

Fuente: Elaboración propia.

Para la interpretación de la tabla 5, se describe que la técnica que mayor resultado obtuvo es **Árboles de decisiones** con una exactitud de **87.69 %**, una precisión de **88.46 %**, y el algoritmo de **Máquina de Vectores de Soporte** marcó un tiempo total de **0.0066** segundos. Por lo tanto, la técnica de Árboles de decisiones es el más adecuado para la predicción de los estilos de vida en la actividad física en los estudiantes universitarios.

Finalmente, se realizó la evaluación con validación cruzada con cinco pliegues que permitió dar un valor numérico, el cual indica que tan bueno es el modelo de Machine Learning en código python esto se realizó a través de una función llamada “cross_val_score”. A continuación, se muestra la tabla 14 de comparación de los resultados obtenidos sin validación cruzada y con validación cruzada con el conjunto de datos de Actividad Física:

Tabla 14. Actividad Física: Resultados con y sin validación cruzada

Valor de predicción	Sin Validación cruzada	Con Validación cruzada
	Máquinas de Vectores de Soporte	
puntaje	87.50 %	87.14 %
Algoritmo	K vecinos más cercanos	
puntaje	89.29 %	86.79 %
Algoritmo	Arboles de decisiones	

puntaje	90.71 %	88.50 %
----------------	----------------	----------------

Fuente: Elaboración propia.

En la tabla 14 se puede visualizar que el **puntaje** con la validación cruzada es menor lo cual es más representativo, el puntaje aumenta la certeza de la predicción, entonces es importante que no se sobreestime la capacidad del modelo de aprendizaje automático. Por tanto, una vez más el modelo para predecir los estilos de vida en la actividad física está basado en la técnica de Árboles de decisiones.

Por último, se acepta la hipótesis alterna indicando que, de acuerdo a los resultados presentados en la tabla 12, 13 y 14, el modelo para predecir los estilos de vida en la actividad física está basado en la técnica de Árboles de decisiones.

5.3.3. Hipótesis Específica H3

H₀: El modelo para predecir los estilos de vida en el aspecto psicológico no está basado en la técnica de Árboles de decisiones.

H₁: El modelo para predecir los estilos de vida en el aspecto psicológico está basado en la técnica de Árboles de decisiones.

Para justificar la hipótesis específica 3, primeramente, una vez descargado los datos recolectados del aspecto psicológico en formato Excel se procedió a cargar los datos a Google Colab como se muestra en la figura 42.

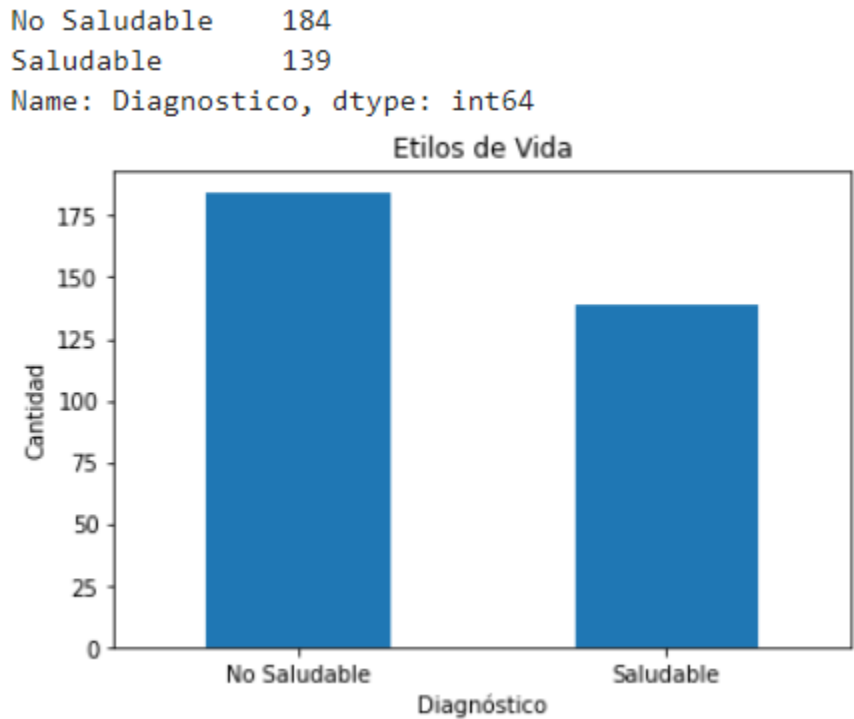
Figura 42. Variables de entrada (features) del aspecto Psicológico en Google Colab

edad	sexo	e_profesional	c_academico	trabaja	e_civil	personalidad	ansiedad	a_horas_clases	h_sueño	c_espacios	s_academica	Diagnostico
Entre 08 a 10 años	Masculino	EPAE	VIII	No	Soltero	Prepotencia	Trabajos grupales	Expresión de desvelado u ojeroso	Menos de 7 h	Bailar	Dificultad para concentrarte	No Saludable
Más de 10 años	Masculino	EPC	II	Sí	Soltero	Empatía	Trabajos finales	Expresión de deprimido	Más de 8 h	Jugar videojuegos	Dejas a última hora las tareas	No Saludable
Menos de 08 años	Femenino	EPEPI	II	Sí	Soltero	Intolerancia	Exposiciones finales	Fatiga o cansancio	Menos de 7 h	No me alcanza tiempo	Carga de horario	No Saludable
Más de 10 años	Masculino	EPC	V	No	Soltero	Intolerancia	Trabajos grupales	Desinterés por las clases	Menos de 7 h	Bailar	Carga de horario	Saludable
Entre 08 a 10 años	Masculino	EPAE	VIII	Sí	Soltero	Flexibilidad	Trabajos grupales	Expresión de deprimido	Más de 8 h	Jugar algún deporte	Dificultad para concentrarte	Saludable
Entre 08 a 10 años	Femenino	EPC	V	No	Soltero	Intolerancia	Trabajos finales	Desmotivado	Más de 8 h	Bailar	Dificultad para concentrarte	No Saludable
Entre 08 a 10 años	Femenino	EPIA	VI	Sí	Soltero	Empatía	Trabajos finales	Desmotivado	7 a 8 h	Jugar algún deporte	Carga de horario	Saludable
Entre 08 a 10 años	Femenino	EPEPI	VIII	No	Soltero	Empatía	Exposiciones finales	Desmotivado	Menos de 7 h	Ausencia de lugares para sentirse mejor	Carga de horario	Saludable

Fuente: Elaboración propia

A continuación, se realizó una visualización de datos con el lenguaje de Python y la librería pandas acerca de las clases que existe en la variable de salida llamada “diagnóstico” para los datos de los aspectos psicológico, como se muestra en la figura 43:

Figura 43. Aspecto psicológico: Visualización de datos de las clases de la variable de salida



Fuente: Elaboración propia

Como se visualiza en la figura 43 se reportaron 184 datos que pertenecen a la clase de no saludable que pertenecen a aquellos estudiantes universitarios que adoptan un estilo de vida no saludable y 139 datos que pertenecen a la clase de saludable que pertenecen a aquellos estudiantes que adoptan un estilo de vida saludable.

Consecutivamente, se muestra el modelo con la técnica de Máquina de Vectores de Soporte realizada en el programa de Weka:

Figura 44. Aspecto Psicológico (SVM): Modelo con SVM en Weka



Fuente: Elaboración propia

Se puede visualizar que cada aspa de color azul son las predicciones correctas que dieron como “Saludable” y cada aspa de color rojo son las predicciones correctas que dieron como “No Saludable”, los cuadros azules que se encuentran en la parte de las predicciones “No Saludable”, son las predicciones que fallaron, o sea que predijeron como “Saludable” pero en realidad debieron de predecir como “No Saludable”, así mismo los cuadros rojos que se encuentran en la parte de lo “Saludable” son las predicciones que fallaron, o sea predijeron como “No saludable” pero en realidad debieron de predecir como “Saludable”.

Al hacer click en cada aspa se genera una interfaz como se muestra en la figura 45:

Figura 45. Aspecto Psicológico (SVM): Modelo en Weka parte 2

```

Weka: Instance info

Plot : weka.classifiers.functions.SMO (cat
Instance: 72
      s_academica : Carga de horario
      c_espacios  : Jugar videojuegos
prediction margin : 1.0
predicted Diagnostico : Saludable
Diagnostico       : Saludable
  
```

Fuente: Elaboración propia

Se muestra el perfil de un estudiante que adopta hábitos para un estilo de vida saludable en el Aspecto Psicológico que la técnica de Máquina de Vectores de Soporte ha logrado predecir. Asimismo, se muestra más casos en la tabla 15:

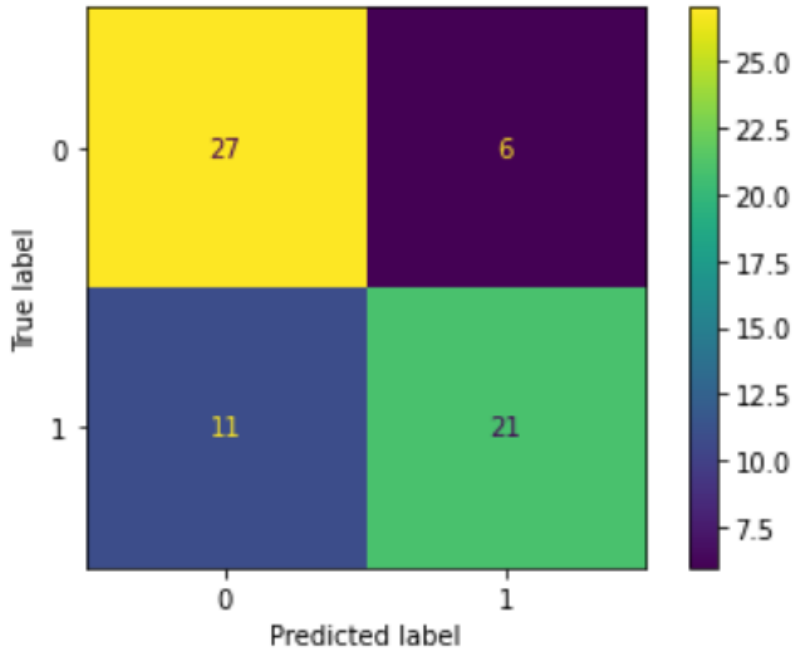
Tabla 15. Aspecto Psicológico (SVM): Hábitos del estudiante que adopta un estilo de vida saludable

Reglas del modelo de Máquina de Vectores de Soporte de los estudiantes que adoptan un estilo de vida saludable	Variable de salida (Diagnóstico)
Sobrecarga académica: Carga de horario Crear espacios: Ausencia de lugares para sentirse mejor	Saludable
Sobrecarga académica: Exigencia del curso Crear espacios: Jugar algún deporte	Saludable
Sobrecarga académica: Carga de horario Crear espacios: Jugar algún deporte	Saludable
Sobrecarga académica: Carga de horario Crear espacios: Bailar	Saludable

Fuente: Elaboración propia

A continuación, en la figura 46 se muestra la matriz de confusión para los datos del aspecto psicológico con el algoritmo de Máquina de Vectores de Soporte:

Figura 46. Aspecto Psicológico: Matriz de confusión con SVM

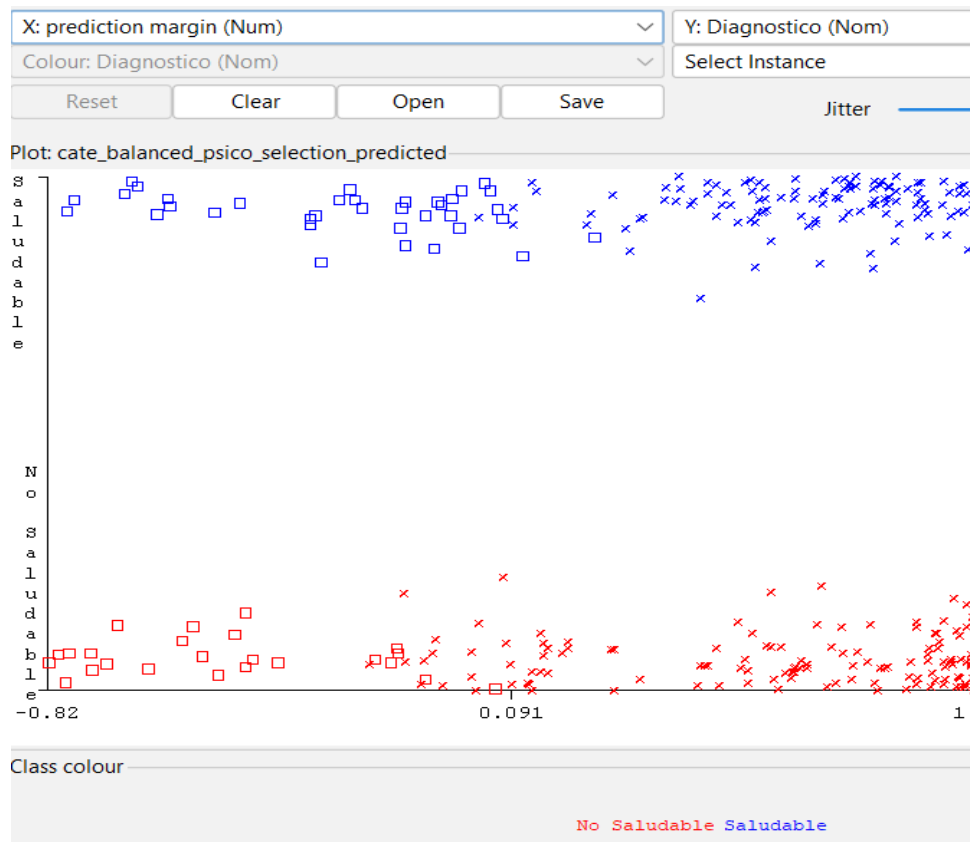


Fuente: Elaboración propia

De acuerdo a la figura 47 los valores de verdaderos positivos que resultaron es de 27 datos, para los verdaderos negativos se hallaron 21 datos, los falsos positivos son 11 datos y los falsos negativos son 6 datos. Esto quiere decir, que al comparar los valores predichos con los valores verdaderos la mayoría de los datos predichos sí coinciden con los datos verdaderos, el porcentaje del error en esta matriz de confusión equivale al 26.15 %, lo cual es un porcentaje mínimo y menos del 50 %.

Seguidamente, se realizó el modelo con la técnica de K vecinos más cercanos con el programa de Weka:

Figura 47. Aspecto Psicológico (KNN): Modelo con KNN en Weka



Fuente: Elaboración propia.

Se puede visualizar que cada aspa de color rojo son las predicciones correctas que dieron como “No Saludable” y cada aspa de color azul son las predicciones correctas que dieron como “Saludable”, los cuadros rojos son las predicciones que dieron como “Saludable” pero en realidad no lo son ósea que fallaron, así mismo los cuadros azules son las predicciones que dieron como “No Saludable” pero en realidad no lo son ósea que fallaron.

Al hacer click en cada aspa se genera una interfaz como se muestra en la figura 48:

Figura 48. Aspecto Psicológico (KNN): Modelo con KNN en Weka parte 2

```

Weka: Instance info
Plot : weka.classifiers.lazy.IBk (cate_balanced_psico_selection)
Instance: 282
    s_academica : Carga de horario
    c_espacios : Ausencia de lugares para sentirte mejor
    prediction margin : 0.6359807460890494
    predicted Diagnostico : Saludable
    Diagnostico : Saludable

```

Fuente: Elaboración propia.

Se muestra el perfil de un estudiante que adopta hábitos para un estilo de vida saludable en el aspecto Psicológico donde la técnica de K vecinos más cercanos ha logrado predecir. Asimismo, se muestra más casos en la tabla 16:

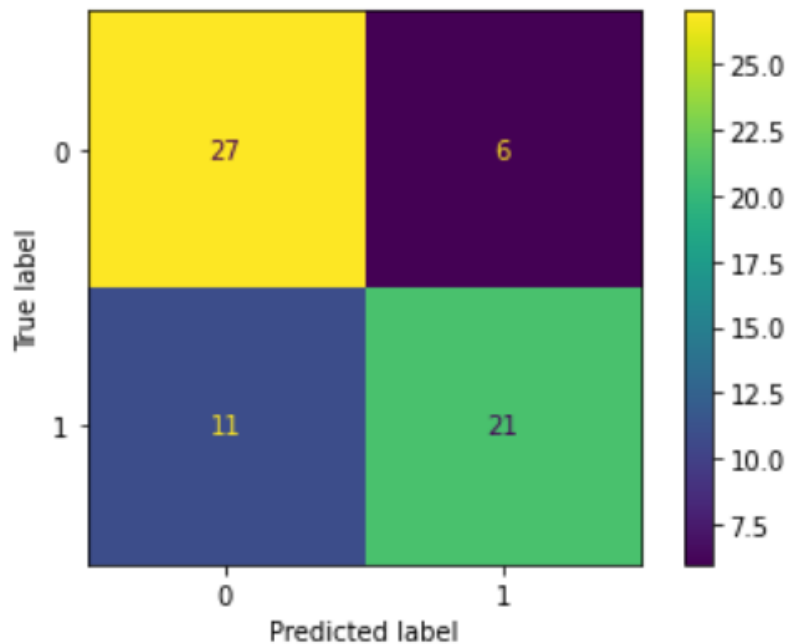
Tabla 16. Aspecto Psicológico (KNN): Hábitos del estudiante que adopta un estilo de vida saludable

Reglas del modelo de K vecinos más cercanos de los estudiantes que adoptan un estilo de vida saludable	Variable de salida (Diagnóstico)
Sobrecarga académica: Carga de horario Crear espacios: Jugar algún deporte	Saludable
Sobrecarga académica: Carga de horario Crear espacios: Bailar	Saludable
Sobrecarga académica: Exigencia del curso Crear espacios: Jugar algún deporte	Saludable
Sobrecarga académica: Dificultad para concentrarse Crear espacios: Jugar algún deporte	Saludable

Fuente: Elaboración propia.

A continuación, en la figura 49 se muestra la matriz de confusión para los datos del aspecto psicológico con el algoritmo de K vecinos más cercanos:

Figura 49. Aspecto Psicológico: Matriz de confusión con KNN

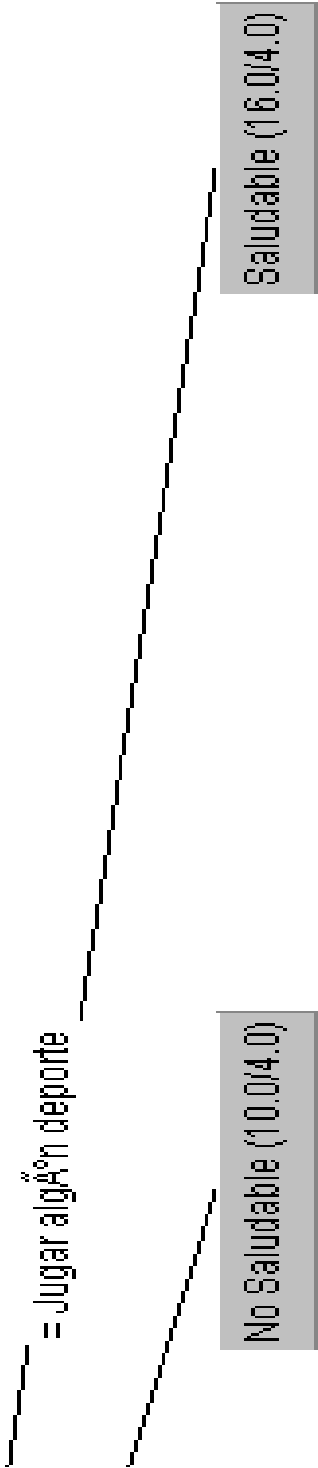


Fuente: Elaboración propia

De acuerdo a la figura 49 los valores de verdaderos positivos que resultaron es de 27 datos, para los verdaderos negativos se hallaron 21 datos, los falsos positivos son 11 datos y los falsos negativos son 6 datos. Esto quiere decir, que al comparar los valores predichos con los valores verdaderos la mayoría de los datos predichos sí coinciden con los datos verdaderos, el porcentaje del error en esta matriz de confusión equivale al 26.15 %, lo cual es un porcentaje mínimo y menos del 50 %.

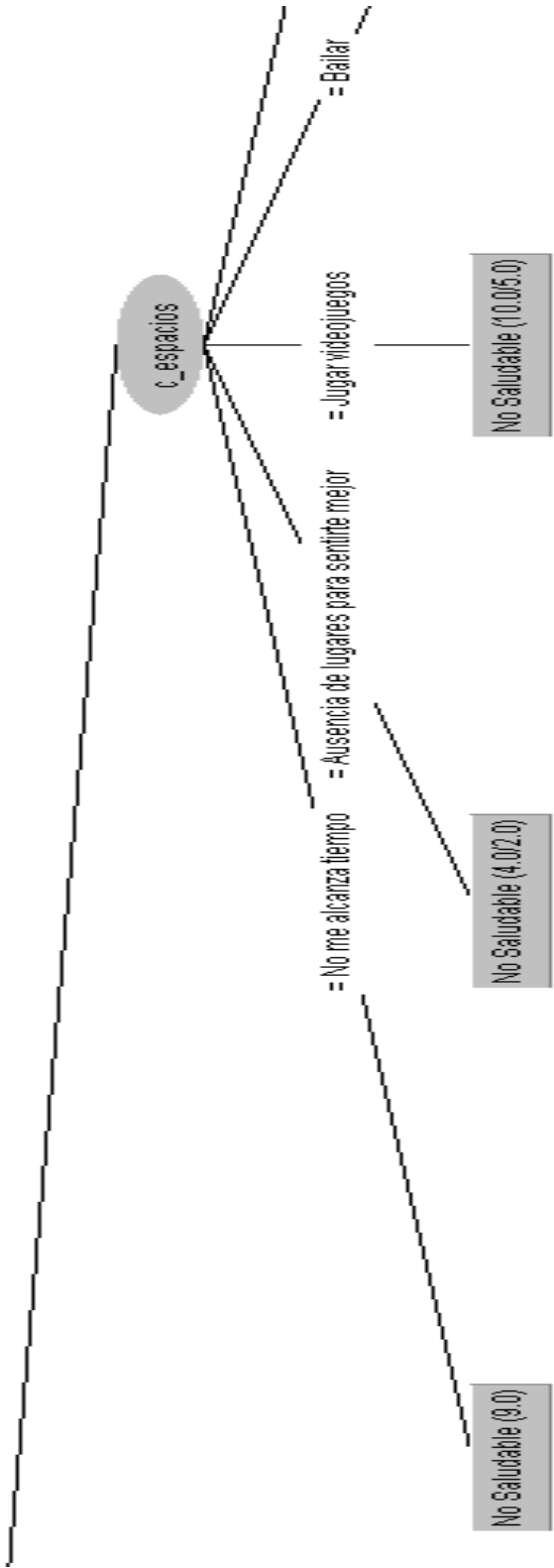
Seguidamente, se muestra la construcción del árbol de decisiones con los datos del aspecto Psicológico, para analizar la información que contiene el árbol de decisión y como se utilizaría si un estudiante está adoptando un estilo de vida saludable o no. Se presenta cinco páginas del modelo del árbol de decisiones para apreciarlo mejor:

Figura 50. Aspecto Psicológico: Primera parte del Árbol de decisión



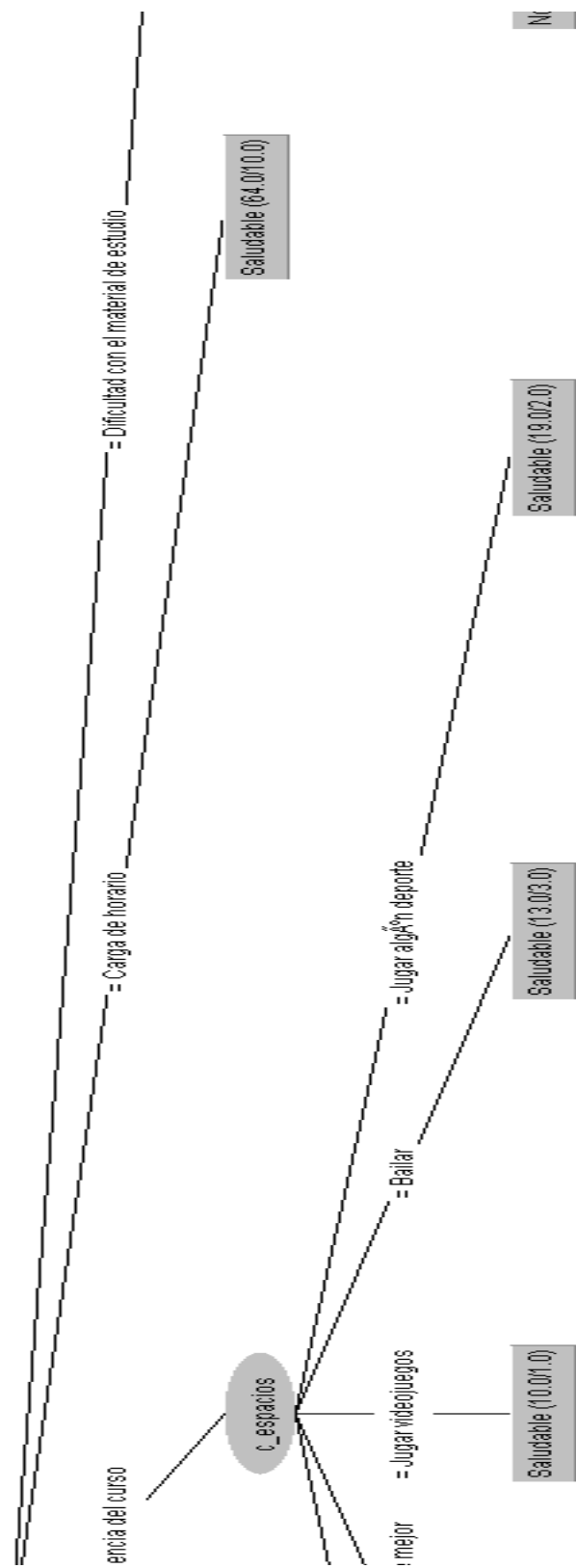
Fuente: Elaboración Propia.

Figura 51. Aspecto Psicológico: Segunda parte del Árbol de decisión



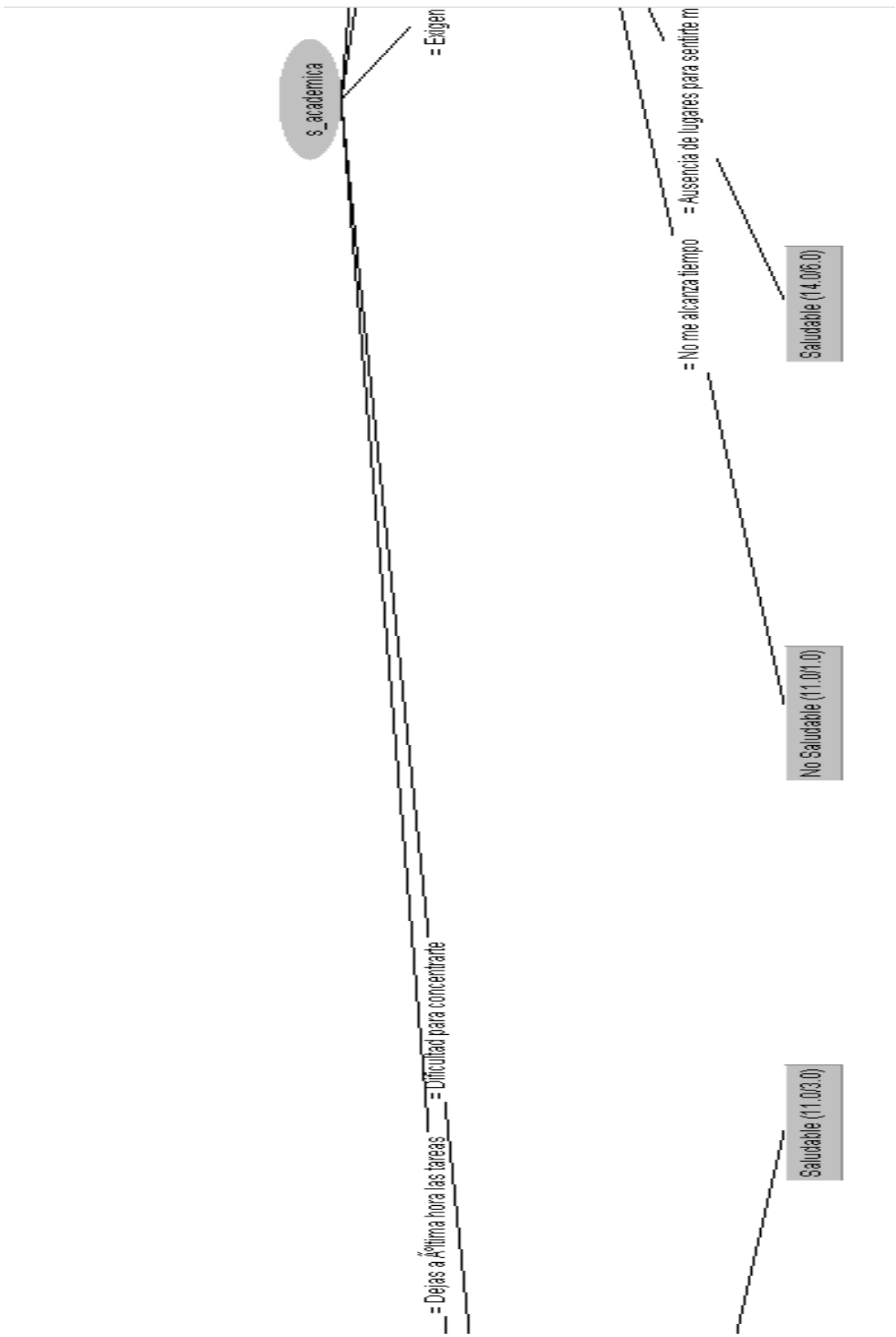
Fuente: Elaboración Propia.

Figura 52. Aspecto Psicológico: Tercera parte del Árbol de decisión



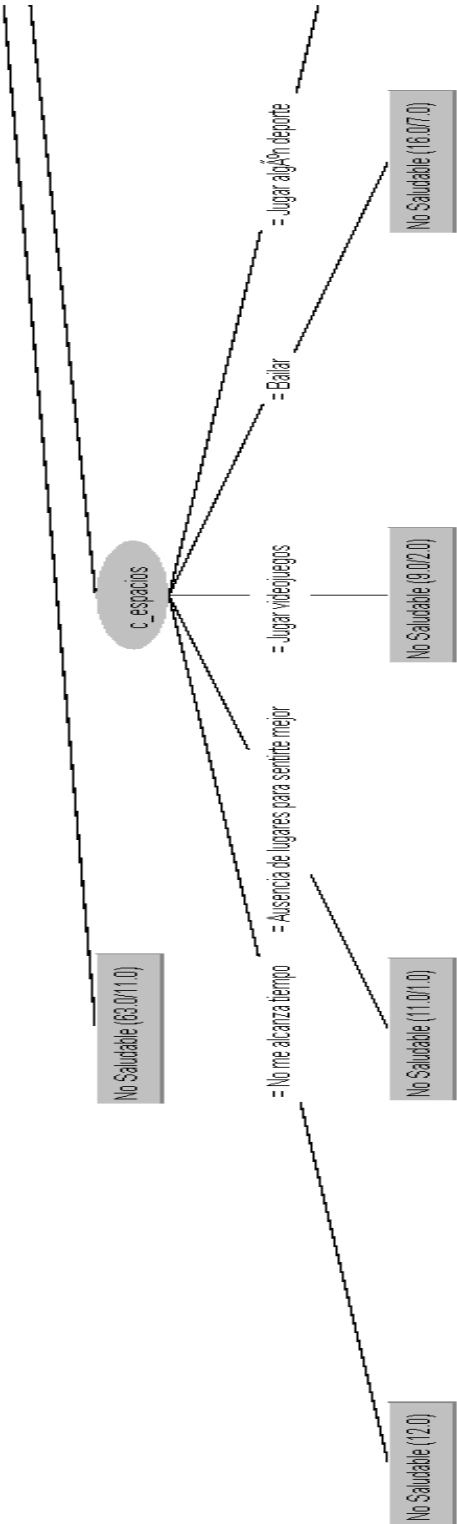
Fuente: Elaboración Propia.

Figura 53. Aspecto Psicológico: Cuarta parte del Árbol de decisión



Fuente: Elaboración Propia.

Figura 54. Aspecto Psicológico: Quinta parte del Árbol de decisión



Fuente: Elaboración Propia.

El árbol de decisión encontró 17 reglas de decisión que predicen el estilo de vida en el aspecto Psicológico que adoptan los estudiantes universitarios de la Universidad Nacional José María Arguedas, gracias a estas reglas se puede conocer los hábitos de vida de los estudiantes que adoptan un estilo de vida saludable y no saludable.

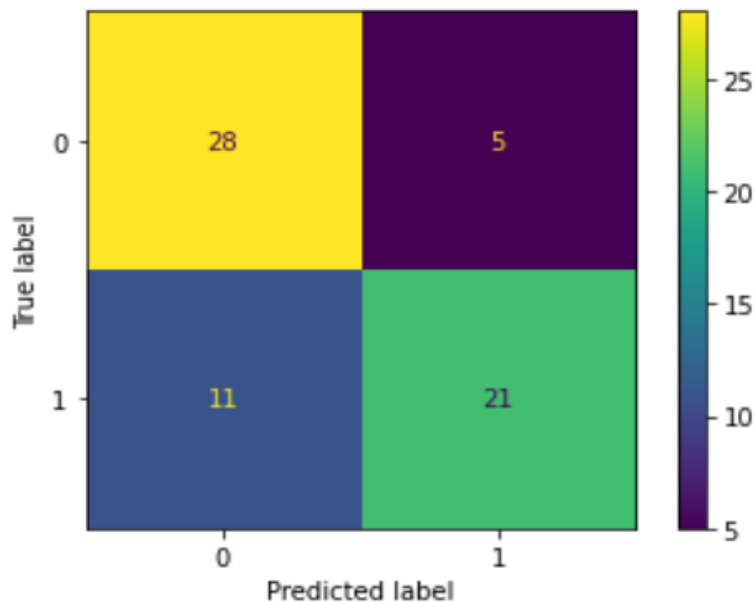
Tabla 17. Hábitos del estudiante que adopta un estilo de vida saludable

Reglas del modelo de Árbol de decisión de los estudiantes que adoptan un estilo de vida saludable	Variable de salida (Diagnóstico)
Sobrecarga académica: Dificultad para concentrarte Crear espacios: Jugar algún deporte	Saludable
Sobrecarga académica: Exigencia del curso Crear espacios: Ausencia de lugares para sentirte mejor	Saludable
Sobrecarga académica: Exigencia del curso Crear espacios: Jugar videojuegos	Saludable
Sobrecarga académica: Exigencia del curso Crear espacios: Bailar	Saludable

Fuente: Elaboración Propia.

A continuación, en la figura 55 se muestra la matriz de confusión para los datos del aspecto psicológico con el algoritmo de Árboles de decisiones:

Figura 55. Aspecto Psicológico: Matriz de confusión con Árboles de decisiones



Fuente: Elaboración propia.

Los valores de verdaderos positivos que resultaron es de 28 datos, para los verdaderos negativos se hallaron 21 datos, los falsos positivos son 11 datos y los falsos negativos son 5 datos. Esto quiere decir, que al comparar los valores predichos con los valores verdaderos la mayoría de los datos predichos sí coinciden con los datos verdaderos, el porcentaje del error en esta matriz de confusión equivale al 24.62 %, lo cual es un porcentaje mínimo y menos del 50 %.

Seguidamente, con los resultados obtenidos en las matrices de confusión mostradas en las figuras 46, 49 y 55 se pudieron realizar las fórmulas de las métricas de evaluación propuestas, por eso en la siguiente tabla se muestra los resultados de las métricas de evaluación halladas con las técnicas de Machine Learning Máquinas de Vectores de Soporte, K vecinos más cercanos y Árboles de decisiones:

Tabla 18. Métricas de evaluación en el aspecto Psicológico

Métricas de evaluación	Psicológico: Algoritmos de Supervisados		
	Máquinas de Vectores de Soporte	K Vecinos más cercanos	Árboles de Decisiones

Exactitud	73.85 %	73.85 %	75.38 %
Precisión	77.78 %	77.78 %	80.77 %
Tiempo total	0.0042	0.0093	0.0030

Fuente: Elaboración propia en software SPSS statistics

Para la interpretación de la tabla 18, se describe que el algoritmo que mayor resultado obtuvo es **Árboles de Decisiones** con una exactitud de **75.38 %**, una precisión de **80.77 %**, y un tiempo total de **0.0030** segundos. Por lo tanto, el algoritmo de Machine Learning es el más adecuado para la predicción de los estilos de vida en el aspecto psicológico en los estudiantes universitarios.

Finalmente, se realizó la evaluación con validación cruzada con cinco pliegues que permitió dar un valor numérico, el cual indica que tan bueno es el modelo de Machine Learning, en código python esto se realizó a través de una función llamada “cross_val_score”. A continuación, se muestra la tabla 19 de comparación de los resultados obtenidos sin validación cruzada y con validación cruzada con el conjunto de datos del aspecto Psicológico:

Tabla 19. Aspecto Psicológico: Resultados con y sin validación cruzada

Valor del puntaje	Sin Validación cruzada	Con Validación cruzada
	Máquinas de Vectores de Soporte	
Puntaje	79.14 %	77.51 %
Algoritmo	K vecinos más cercanos	
Puntaje	78.80 %	73.53 %
Algoritmo	Arboles de decisiones	
Puntaje	80.46 %	77.17 %

Fuente: Elaboración propia.

En la tabla 19 se puede visualizar que el **puntaje** con la validación cruzada es menor lo cual es más representativo, el puntaje aumenta la certeza de la predicción, entonces es importante

que no se sobreestime la capacidad del modelo de aprendizaje automático. Por tanto, una vez más el modelo para predecir los estilos de vida en el aspecto psicológico está basado en la técnica de Árboles de decisiones.

Por último, se acepta la hipótesis alterna indicando que, de acuerdo a los resultados presentados en la tabla 17, 18 y 19, el modelo para predecir los estilos de vida en el aspecto psicológico está basado en la técnica de Árboles de decisiones.

CAPÍTULO VI

6. DISCUSIÓN

A partir de los resultados encontrados, aceptamos la hipótesis alternativa específica que establece que el algoritmo de Machine Learning permite predecir los estilos de vida en el aspecto psicológico que adoptan los estudiantes universitarios en la Universidad Nacional José María Arguedas, para este caso el algoritmo que determinó fue Árboles de decisiones con una exactitud de 75.38 %. Estos resultados guardan relación con los resultados que sostiene en su estudio la autora Ovalle Toledo (2020) que el algoritmo de Árboles de decisiones predice la ansiedad con una exactitud de 80 % y que el algoritmo de Random Forest predice la depresión con una exactitud de 91 %, en los estudiantes universitarios en la Universidad de Chile, por lo tanto, los resultados en la exactitud de la tesis de la investigadora Ovalle Toledo (2020), fue alto debido a que los datos que utilizó para el entrenamiento no necesitaron balancearlos, ya que para la predicción de la ansiedad, la diferencia proporcional de las clases de la variable de salida asignadas para el entrenamiento fue mínima.

Con el objetivo de predecir los estilos de vida que adoptan los estudiantes universitarios con el modelo predictivo basado en técnicas de Machine Learning en estudiantes de la Universidad Nacional José María Arguedas, los resultados que reflejaron por cada aspecto estudiado sus puntajes son altos, siendo en la Alimentación, se obtuvo una precisión de 79.31 % con el algoritmo de Máquina de Vectores de Soporte, en la Actividad Física se obtuvo una precisión de 88.46 %, con el algoritmo de Árboles de decisiones y en el aspecto psicológico se obtuvo una precisión de 80.77 % con el algoritmo de Árboles de decisiones; esto quiere decir que las técnicas de Machine Learning logran predecir los estilos de vida en los estudiantes universitarios como saludable o no saludable. Estos resultados guardan relación con lo referido por autor Colula (2018) quien determinó la elección del algoritmo de aprendizaje automático adecuado para clasificar las señales EEG del estrés como: estrés alto y estrés bajo, en el cual obtuvo una precisión de 70,26 % con el algoritmo de Random Forest. Analizando estos resultados podemos ver que la precisión alta en las métricas de evaluación que dieron al utilizar las técnicas de Machine Learning es la misma en estudiantes universitarios de la ciudad de Andahuaylas en el país de Perú y en la ciudad de Puebla del país de México. Por otro lado, los resultados altos que dieron en la precisión en esta investigación a comparación del trabajo de

investigación del autor Colula Medel (2018), a causa de que no utilizó la etapa de balanceo de datos, ya que esto influyó en los resultados de las métricas de evaluación.

De acuerdo a la investigación escrito por los autores José Velezmoro y Edwin Bazán (2020), en su investigación que tiene como objetivo: determinar el mejor algoritmo de Inteligencia Artificial para la implementación de un sistema web que prediga la depresión en adolescentes de la ciudad de Cajamarca, los resultados reflejaron que el algoritmo de Árbol de Decisión J48, ha sido el algoritmo ganador con un ratio de valores de precisión 91.70 %, entre los algoritmos de BayesNet, Multilayer Perceptron, Clasificador ForestPA y Clasificador Naive Bayes, estos resultados guardan relación con este trabajo de investigación en las que el algoritmo Árboles de decisiones tiene la mayor capacidad de predicción en el aspecto Psicológico en los estudiantes universitarios con una precisión de 80.77 %.

De acuerdo a la investigación escrito por el autor Hernán Vilca Masco (2019), en su investigación de tesis para los resultados que obtuvo en su tesis de investigación en las métricas de evaluación demuestran que el algoritmo supervisado de Árboles de decisión es el clasificador adecuado para la predicción del nivel de estrés en estudiantes universitarios con una tasa de precisión de 97 %, estos resultados guardan relación con este trabajo de investigación en las que el algoritmo de Árboles de decisiones fue el más adecuado para la predicción de los estilos de vida en el aspecto psicológico el cual presenta una tasa de precisión de 80.77 %.

CONCLUSIONES

- Se logró predecir los estilos de vida que adoptan los estudiantes universitarios con el modelo predictivo basado en técnicas de Machine Learning.
- Se logró establecer un modelo para predecir los estilos de vida en la alimentación de los estudiantes universitarios usando Máquina de Vectores de Soporte, con una exactitud de 83.08 %, lo cual es un resultado satisfactorio mayor al 50 %.
- Se logró establecer un modelo para predecir los estilos de vida en la actividad física de los estudiantes universitarios usando Árboles de decisiones, con una exactitud de 87.69 %, lo cual es un resultado satisfactorio mayor al 50 %.
- Se logró establecer un modelo para predecir los estilos de vida en el aspecto psicológico de los estudiantes universitarios usando Árboles de decisiones con una exactitud de 87.69 %, lo cual es un resultado satisfactorio mayor al 50 %.

RECOMENDACIONES

- Se recomienda que se realicen cada una de las etapas de Machine Learning basado en la metodología de KDD para que el modelo predictivo basado en técnicas de Machine Learning permita predecir los estilos de vida de los estudiantes universitarios.
- Se recomienda realizar un buen análisis, verificación, normalización, categorización, una buena exploración de datos hasta obtener una calidad de datos para que el algoritmo supervisado de Máquina de Vectores de Soporte prediga los estilos de vida en la alimentación de los estudiantes universitarios.
- Se recomienda realizar uno de los métodos del balanceo de datos dependiendo al conjunto de datos que se presente para que el algoritmo supervisado de Árboles de decisiones prediga los estilos de vida en la actividad física de los estudiantes universitarios.
- Se recomienda realizar el método de selección de características para que el algoritmo supervisado de Árboles de decisiones prediga con más eficiencia los estilos de vida en el aspecto psicológico de los estudiantes universitarios.

REFERENCIAS

- Azcona, Á. C. (2013). *Manual de Nutrición y Dietética* (Primera ed.). Madrid, España: Universidad Complutense. Recuperado el 2021 de 11 de 20, de <https://eprints.ucm.es/id/eprint/22755/1/Manual-nutricion-dietetica-CARBAJAL.pdf>
- Barradas, M. (2014). *Depresión en Estudiantes Universitarios: Una realidad indeseable* (Primera ed.). Bloomington, Indiana, Estados Unidos: Palibrio. Recuperado el 2021 de 12 de 10, de <https://es.scribd.com/read/523930310/Depresion-En-Estudiantes-Universitarios-Una-Realidad-Indeseable>
- Bezares, V., Cruz, R., Acosta, M., & Ávila, M. (2020). *Experiencias de investigación en estilo de vida saludable* (Primera ed.). Chiapas, Mexico: Editorial Universidad de Ciencias y Artes de Chiapas. Recuperado el 06 de 10 de 2021, de <https://www.maimonides.edu/descargas/Experiencias-de-investigacion-en-estilo-de-vida-saludable-2020-con-ISBN.pdf>
- Bobadilla, J. (2020). *Machine Learning y Deep Learning: Usando Python, Scikit y Keras* (1 ed.). Bogotá: Editorial Ra-ma. Obtenido de <https://books.google.com.pe/books?id=iAAyEAAQBAJ&pg=PA64&dq=algoritmo+de+machine+learning+KNN&hl=es&sa=X&ved=2ahUKEwik1fbuzNnyAhUWEbkGHbieCbgQ6AEwAXoECAgQAg#v=onepage&q=algoritmo%20de%20machine%20learning%20KNN&f=false>
- Cabanyes, J., & Monge, M. Á. (2017). *La salud mental y sus cuidados* (Cuarta ed.). Pamplona, España: Ediciones Universidad de Navarra, S.A. (EUNSA). Recuperado el 23 de 11 de 2021, de http://www.biblioteca.cij.gob.mx/Archivos/Materiales_de_consulta/Drogas_de_Abuso/Articulos/LibroSaludMentalCuidados.pdf
- Colula, I. (2018). *Método de Clasificación para detección de estrés empleado electroencefalograma*. Puebla, México: Universidad Autónoma de Puebla. Recuperado el 03 de 08 de 2022, de <https://repositorioinstitucional.buap.mx/handle/20.500.12371/7257>

- Dark, S. (2019). *Aprendizaje Automático La guía para Principiantes para comprender el Aprendizaje Automático* (1 ed.). (K. Gill, Ed.) España: Editorial Karanvir Gill. Obtenido de <https://es.b-ok.lat/book/11343378/11dca3>
- Gázquez, J. J., Molero, M. d., Pérez-Fuente, M. d., Martos, Á., Barragán, A. B., & Simón, M. d. (2017). Salud, Alimentación y Sexualidad en el ciclo vital. En J. J. Gázquez, M. d. Molero, M. d. Pérez-Fuente, Á. Martos, A. B. Barragán, & M. d. Simón, *Salud, Alimentación y Sexualidad en el ciclo vital* (pág. 544). Madrid, España: ASUNIVEP. Recuperado el 26 de 08 de 2022, de <https://www.pdfdrive.com/publicaci%C3%B3n-salud-alimentaci%C3%B3n-y-sexualidad-en-el-ciclo-vital-volumen-ii-e58607661.html>
- Gonzáles, L. (2019). *Machine Learning con Python - Aprendizaje Supervizado* (1 ed.). Caracas, Venezuela: Editorial Santillana. Obtenido de https://escuela.aprendeia.com/cursos/machine-learning-con-python-aprendizaje-supervisado/#tab-course-section__overview
- González, L. (2022). *aprendedeIA*. Recuperado el 16 de 07 de 2022, de aprendeIA: <https://aprendeia.com/conjunto-de-datos-desbalanceado/>
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (Tercera ed.). Waltham, USA: ELSEIVER. Recuperado el 25 de 09 de 2022, de <http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf>
- Hernández, R., Fernández, C., & Bapstista, P. (2014). *Metodología de la Investigación* (Sexta ed.). México DF, México: INTERAMERICA EDITORES. Recuperado el 06 de 10 de 2021, de <https://academia.utp.edu.co/grupobasicoclinicayaplicadas/files/2013/06/Metodología-de-la-Investigación.pdf>
- Jiménez, L. (2015). *Lo que dice la ciencia sobre dietas, alimentación y salud*. Ciudad de México, México: Plataforma Editorial. Recuperado el 30 de 08 de 2022, de <https://www.plataformaeditorial.com/libro/4892-lo-que-dice-la-ciencia-sobre-dietas-alimentacion-y-salud>

- Lara, J. (2014). *Minería de datos*. Madrid, España: UDIMA. Recuperado el 25 de 09 de 2022, de https://tienda.cef.udima.es/libros/mineria_datos.html
- López, E., & Maza, R. (2020). *Predicción de Enfermedades Cardiovasculares mediante Algoritmo de Inteligencia Artificial*. Tesis Bachiller, Málaga. Obtenido de <https://riuma.uma.es/xmlui/handle/10630/20458>
- Madaria, Z. (01 de 09 de 2018). *undación Española del Corazón*. Recuperado el 2022 de 08 de 28, de undación Española del Corazón: <https://fundaciondelcorazon.com/ejercicio/conceptos-generales/3151-la-piramide-de-la-actividad-fisica.html>
- Márquez, S., & Garatachea, N. (2013). *Actividad Física y Salud* (1 ed.). Madrid, España: Ediciones Díaz de Santos. Recuperado el 11 de 10 de 2021, de <https://www.pedagogicomadrededios.edu.pe/wp-content/uploads/2020/10/ACTIVIDAD-FISICA-Y-SALUD.pdf>
- Martínez, C. (2012). *Estadística y muestreo*. Bogotá: Ecoe. Recuperado el 03 de 07 de 2022, de <https://www.pdfdrive.com/estad%C3%ADstica-y-muestreo-de-ciro-mart%C3%ADnez-b-e42648572.html>
- Moiso, A., Mestorino, M. d., & Óscar, O. (2007). *Fundamentos de Salud Pública* (Primera ed.). La Plata, Argentina: Editorial de la Universidad Nacional de la Plata Edición. Recuperado el 06 de 10 de 2021, de <http://sedici.unlp.edu.ar/handle/10915/29128>
- Monsalve Torra, A. E. (2017). *Sistema de Ayuda a la Decisión Clínica en Enfermedades de Diagnóstico Complejo*. Alicante: Universidad de Alicante. Recuperado el 27 de 09 de 2021, de https://rua.ua.es/dspace/bitstream/10045/65334/1/tesis_monsalve_torra.pdf
- NHLBI. (13 de febrero de 2013). *National Heart, Lung, and Blood Institute*. Recuperado el 28 de 08 de 2022, de National Heart, Lung, and Blood Institute: <https://www.nhlbi.nih.gov/health/educational/wecan/espanol/tiempopantalla.htm#:~:text=Los%20expertos%20de%20la%20salud,relacionado%20con%20alguna%20tarea%20escolar.>

- Ñaupas, H., Palacios, J., Valdivia, M., & Romero, H. (2018). *Metodología de la investigación Cuantitativa - Cualitativa y Redacción de la Tesis* (Quinta ed.). Bogotá, Colombia: Ediciones de la U. Recuperado el 2021 de 10 de 06, de <https://corladancash.com/wp-content/uploads/2020/01/Metodologia-de-la-inv-cuanti-y-cuali-Humberto-Naupas-Paitan.pdf>
- OMS. (30 de 03 de 2018). Recuperado el 11 de 10 de 2021, de OMS: <https://www.who.int/es/news-room/fact-sheets/detail/mental-health-strengthening-our-response>
- OMS. (26 de 11 de 2020). Recuperado el 11 de 10 de 2021, de OMS: <https://www.who.int/es/news-room/fact-sheets/detail/physical-activity>
- OMS. (13 de 04 de 2021). *Organización Mundial de la Salud*. Recuperado el 2022 de 08 de 14, de Organización Mundial de la Salud: <https://www.who.int/es/news-room/fact-sheets/detail/noncommunicable-diseases>
- Ovalle, J. V. (2020). *Estudio de factibilidad de un modelo de clasificación de estudiantes universitarios según el estado crítico de estrés universitarios según el estado crítico de estrés, ansiedad y sintomatología depresiva*. Santiago de Chile: Universidad de Chile. Recuperado el 29 de 07 de 2022, de <https://repositorio.uchile.cl/handle/2250/176543>
- Python Software Foundation. (19 de Julio de 2021). *La Biblioteca Estándar de Python*. Obtenido de La Biblioteca Estándar de Python: <https://docs.python.org/3/library/timeit.html>
- Python Software Foundation. (31 de Mayo de 2022). *La Biblioteca Estándar de Python*. Recuperado el 01 de Junio de 2022, de La Biblioteca Estándar de Python: <https://docs.python.org/es/3/library/time.html#>
- Quispe, A., Calla, K., Yangali, J., Rodríguez, J., & Pumacayo, I. (2019). *Estadística no paramétrica aplicada a la investigación científica con software* (Primera ed.). Bogotá, Colombia: Editorial Eidec. Recuperado el 14 de 11 de 2021, de <https://editorialeidec.com/producto/estadistica-no-parametrica-aplicada-a-la-investigacion-cientifica-con-software-spss-minitab-y-excel/>

- Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning* (2 ed.). Birmingham, Reino Unido: Editorial Packt Publishing . Obtenido de https://books.google.com.pe/books?id=_plGDwAAQBAJ&printsec=frontcover&dq=Python+Machine+Learning+Second+Edition+Machine+Learning+and+Deep+Learning+with+Python,+scikit-learn,+and+TensorFlow&hl=es&sa=X&redir_esc=y#v=onepage&q=Python%20Machine%20Learning%20S
- RPAN. (11 de 02 de 2019). *Red Peruana de Alimentación y Nutrición*. Recuperado el 26 de 08 de 2022, de Red Peruana de Alimentación y Nutrición: <https://www.rpan.org/single-post/2019/02/24/la-pir%C3%A1mide-nutricional-educaci%C3%B3n-en-nutrici%C3%B3n>
- Ruiz Chércoles, E., & Cenarro Guerrero, T. (2016). *La importancia del etiquetado*. Madrid: LUA. Recuperado el 27 de 08 de 2022, de https://www.aepap.org/sites/default/files/4t2.11_la_importancia_del_etiquetado.pdf
- Russell, R. (2018). *Machine Learning Guía Detallada para Implementar Algoritmos de Machine Learning con Python* (1 ed.). España: CreateSpace Independent Publishing Platform. Obtenido de <https://www.amazon.com/-/es/Rudolph-Russell/dp/1720933685>
- Sebastián, R. (2015). *Python Machine Learning*. Mumbai: Packt Publishing Ltd. Recuperado el 10 de 12 de 2021, de <https://www.perlego.com/book/4113/python-machine-learning-pdf>
- Suca, C., Córdova, A., Condori, A., Cayra, J., & Sulla, J. (2018). *Comparacion de Algoritmos de Clasificación para la Predicción de Casos de Obesidad Infantil*. Artículo de Investigación, Arequipa.
- Torres-Carrión, P., González-Escarabay, J., López-Guerra, V., & Vaca, S. (2020). *Aprendizaje automático aplicado al análisis del consumo de alcohol y su relación con el estrés percibido*. Loja, Ecuador: Universidad Técnica Particular de Loja. Recuperado el 29 de 07 de 2022, de <https://www.proquest.com/openview/5d92effec869242e028dc6573748bfaa/1?pq-origsite=gscholar&cbl=1006393>

- Uriarte, Y., & Vargas, A. (2018). *Estilos de vida de los estudiantes de Ciencias de la Salud de la Universidad Norbert Wiener*. Lima: Universidad Privada Norbert Wiener. Recuperado el 2021 de 09 de 27, de <http://repositorio.uwiener.edu.pe/handle/123456789/2390#:~:text=El%20tipo%20de%20muestreo%20fue,un%20estilo%20de%20vida%20Saludable>.
- Velezmoro, J. N., & Bazán, E. M. (2020). *Sistema web basado en algoritmos de inteligencia artificial para la predicción de la depresión en adolescentes de la ciudad de Cajamarca*. Cajamarca, Perú: Universidad Privada del Norte. Recuperado el 29 de 07 de 2022, de <https://repositorio.upn.edu.pe/handle/11537/24798#:~:text=La%20investigaci%C3%B3n%20titulada%20%E2%80%9CSistema%20web,de%20evaluar%20a%20la%20totalidad>
- Vilca, H. (2019). *Predicción del nivel de estrés en estudiantes universitarios utilizando técnicas de machine learning*. Puno: Universidad Peruana Unión. Recuperado el 05 de 03 de 2022, de <https://repositorio.upeu.edu.pe/handle/20.500.12840/4096>

ANEXOS

Anexo 01. Recopilación de datos

Para la recopilación de datos en las 6 escuelas profesionales se utilizó la técnica de encuesta, como se muestra a continuación:



**UNIVERSIDAD NACIONAL JOSÉ MARÍA
ARGUEDAS
FACULTAD DE INGENIERÍA
ESCUELA PROFESIONAL DE INGENIERÍA DE
SISTEMAS**



**CUESTIONARIO PARA ANALIZAR LOS ESTILOS DE VIDA DE LOS
ESTUDIANTES DE LA UNAJMA CON TÉCNICAS DE MACHINE LEARNING.
RESOLUCIÓN N° 379-2021-CFI-UNAJMA**

A continuación, se presenta un cuestionario sobre las actitudes que adoptan los estudiantes universitarios acerca de los estilos de vida, el cual es ANÓNIMO y confidencial, por favor responda con toda SINCERIDAD.

CUESTIONARIO

DATOS GENERALES

Rango de edad:

Menos de 18 años

Entre 18 a 21 años

Más de 21 años

Sexo:

Femenino

Masculino

Escuela Profesional: Elige

☐

EPIS.

☐

EPIA.

☐

EPIAM.

☐

EPAE.

☐

EPC.

☐

EPEPI.

Ciclo académico: Elige

☐

I.

☐

II.

☐

III.

☐

IV.

☐

V.

☐

VI.

☐

VII.

☐

VIII.

☐

IX.

☐

X.

Trabaja:

Sí

No

Estado Civil:

- a) Soltero
- b) Casado
- c) Conviviente
- d) Otro

ALIMENTACIÓN

- 1. ¿Cuántas porciones de verduras y frutas consumes diariamente?**
 - 5 porciones de verduras y 4 frutas
 - 4 porciones de verduras y 3 frutas
 - 3 porciones de verduras y 2 frutas
- 2. ¿Sueles restringir el consumo de grasas entre mantequillas, mayonesa, aceites, etc.?**
 - Nada
 - Poco
 - Mucho
 - Demasiado
- 3. ¿Lees las etiquetas de las comidas empaquetadas para identificar los nutrientes?**
 - Nunca
 - Casi nunca
 - A veces
 - Frecuentemente
 - Siempre
- 4. ¿Con que frecuencia consumes bebidas alcohólicas?**
 - Nunca
 - Una vez al año
 - Menos de 12 veces al año
 - Una vez al mes
 - Cada fin de semana
- 5. ¿Cuántos vasos de agua tomas al día?**
 - Menos de 6 vasos
 - Entre 6 a 8 vasos
 - Más de 8 vasos
- 6. ¿Por qué razones sueles comer a deshoras o dejas de comer el desayuno, el almuerzo o la cena?**
 - Presentación de proyectos Finales
 - Exámenes Parciales
 - Tengo que trabajar y estudiar
 - Vivo lejos a la universidad
 - Falta de apoyo económico
- 7. ¿Cuántos dulces entre pasteles, helados, gaseosas, snacks, galletas consumes a la semana?**
 - 0 dulces
 - Menos de 3 dulces
 - 4 a 7 dulces
 - Más de 7 dulces

8. ¿Cómo se encuentra con su peso?

- Debajo de su peso normal
- Peso normal
- Sobrepeso
- Obesidad

ACTIVIDAD FÍSICA

9. ¿Cuántas horas a la semana realizas actividades físicas aeróbicas moderadas?, como: caminar rápidamente, natación, bicicleta, correr, fútbol, vóley.

- Menos de 2 h y 30 min
- 2 h y 30 min a 5 h
- Más de 5 h

10. ¿Cuántas horas a la semana realizas actividades físicas aeróbicas intensas?, como: cucullas, planchas, saltos de tijera, abdominales, ejercicios con el abdomen o cintura, bailes.

- Menos de 1 h y 15 min
- 1 h y 15 min a 2 h y 30 min
- Más de 2 h y 30 min

11. ¿Cuántos días a la semana realizas actividades de fortalecimiento muscular?, como: alzar pesos, estiramientos, trabajos de albañilería o agricultura, alzar bolsas pesadas.

- Menos de 2 días
- Igual a 2 días
- Más de 2 días

13. ¿Cuánto tiempo le toma "en segundos" para realizar actividades físicas de flexibilidad de los músculos y tendones al día?

- Ninguno
- Menos de 15 s
- 15 a 30 s
- Más 30 s

14. ¿Cuántas horas de tu tiempo pasas al día frente al televisor, laptop, celular o videojuegos?

- 2 h
- Menos de 4 h
- 4 a 7 h
- 7 a 10 h
- Más de 10 h

15. ¿Cuánto tiempo caminas diariamente?

- Menos de 30 min
- 30 min
- Más de 30 min

ASPECTO PSICOLÓGICO

16. Cuando trabajas en grupos, uno de tus compañeros no cumplió con lo asignado. ¿Cómo actúas ante esta situación?

Intolerancia
Flexibilidad
Prepotencia
Empatía

17. ¿En qué casos te sientes con ansiedad o nerviosismo?

Exposiciones finales
Trabajos finales
Exámenes
Trabajos grupales

18. Generalmente, ¿Cuál es la actitud que mantienen durante las horas de clases en la universidad?

Expresión de deprimido
Desmotivado
Desinterés por las clases
Expresión de desvelado u ojeroso
Fatiga o cansancio

19. ¿Cuántas horas diarias concilias el sueño?

Menos de 7 h
7 a 8 h
Más de 8 h

20. ¿De qué manera creas espacios para sentirte mejor en la universidad?

Bailar
Ausencia de lugares para sentirte mejor
Jugar algún deporte
No me alcanza tiempo
Jugar videojuegos

21. ¿Por qué razones sueles tener sobrecarga académica?

Dejas a última hora las tareas
Dificultad con el material de estudio
Exigencia del curso
Carga de horario
Dificultad para concentrarte

Se recolectaron 323 encuestas, de acuerdo a la encuesta se tienen 27 variables de entrada incluida la variable de salida.

Anexo 02. Generación del Modelo Predictivo

Imágenes del código del entrenamiento con cada algoritmo supervisado y los valores predictivos.

Figura 1. Modelo predictivo con SVM

```
#Entrenando con SVM
from sklearn.svm import SVC
SupportVectorMachine=SVC()
SupportVectorMachine.fit(x_train,y_train)

SVC()

[135] x_test = x_test[['v_frutas','r_grasas','l_etiquetas','b_alcoholicas','agua','comer_deshoras']]

#Calculando las predicciones con el modelo entrenado y
#con los valores de prueba de x
import time #Módulo
start_time = time.time()
predictions_svm = SupportVectorMachine.predict(x_test)
total_time = time.time() - start_time |
print(predictions_svm)

[1 0 0 1 1 1 1 0 1 0 1 1 0 1 1 0 0 1 0 0 1 1 1 0 0 0 0 0 0 0 1 1 1 0 1 1 0
 1 0 1 0 1 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 1 1 0]
```

Figura 3. Modelo predictivo con KNN

```
[140] from sklearn.neighbors import KNeighborsClassifier
      knn = KNeighborsClassifier(n_neighbors=3)
      knn.fit(x_train,y_train)

      KNeighborsClassifier(n_neighbors=3)

[141] start_time = time.time()
      predictions_knn = knn.predict(x_test)
      total_time = time.time() - start_time
      print(predictions_knn)

[1 0 0 1 1 1 1 0 1 0 1 1 0 1 1 0 0 1 0 0 1 1 1 0 0 0 0 0 0 0 1 1 0 1 1 1 1
 1 0 0 0 1 0 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0 1 0 1 0 0 1 0]
```

Figura 4. Modelo predictivo con Árboles de decisiones

```
from sklearn import tree
tr = tree.DecisionTreeClassifier(max_depth=10)
tr.fit(x_train,y_train)

DecisionTreeClassifier(max_depth=10)

start_time = time.time()
predictions_tr = tr.predict(x_test)
total_time = time.time() - start_time
print(predictions_knn)

[1 0 0 1 1 1 1 0 1 0 1 1 0 1 1 0 0 1 0 0 1 1 1 0 0 0 0 0 0 0 1 1 0 1 1 1 1
 1 0 0 0 1 0 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0 1 0 1 0 0 1 0]
```


Anexo 03. Interpretación y Evaluación

Métricas de evaluación calculados entre los valores predichos y los valores de prueba de la variable “y” en la Alimentación:

Figura 7. Métricas de evaluación con el algoritmo supervisado de SVM

```
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score

#Muestra las estadísticas entre los valores reales(y_test) y las predicciones
print("\nAccuracy: ",accuracy_score(y_test, predictions_svm))
print("\nPrecision: ",precision_score(y_test, predictions_svm))
print("\nLlevó {} segundos en total.".format(total_time))

Accuracy:    0.8307692307692308

Precision:    0.7931034482758621

Llevó 0.008830070495605469 segundos en total.
```

Figura 8. Métricas de evaluación con el algoritmo supervisado de KNN

```
#Muestra las estadísticas entre los valores reales(y_test) y las predicciones
print("Accuracy: ",accuracy_score(y_test, predictions_knn))
print("\nPrecision: ",precision_score(y_test, predictions_knn))
print("\nLlevó {} segundos en total.".format(total_time))

Accuracy:    0.7538461538461538

Precision:    0.7307692307692307

Llevó 0.017535924911499023 segundos en total.
```

Figura 9. Métricas de evaluación con el algoritmo supervisado de Árboles de decisiones

```
#Muestra las estadísticas entre los valores reales(y_test) y las predicciones
print("\nAccuracy: ",accuracy_score(y_test, predictions_tr))
print("\nPrecision: ",precision_score(y_test, predictions_tr))
print("\nLlevó {} segundos en total.".format(total_time))

Accuracy:    0.7384615384615385

Precision:    0.6666666666666666

Llevó 0.012343645095825195 segundos en total.
```

Figura 10. Prueba de validación cruzada con el algoritmo de SVM

```
[183] x = df_seleccion.drop(['Diagnostico'], axis=1)
      y = df_seleccion['Diagnostico']

from sklearn.svm import SVC
SVC = SVC(gamma='auto', C=2.0, kernel='linear')
from sklearn.model_selection import cross_val_score

SVC.fit(x,y)
print(SVC.score(x,y))
print(cross_val_score(SVC, x,y, cv=5).mean())

0.8734177215189873
0.8672619047619048
```

Figura 11. Prueba de validación cruzada con el algoritmo de KNN

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(x,y)

print(knn.score(x,y))
print(cross_val_score(knn, x,y, cv=5).mean())

0.9177215189873418
0.7976190476190477
```

Figura 12. Prueba de validación cruzada con el algoritmo de Árboles de decisiones

```
[186] from sklearn import tree
      tr = tree.DecisionTreeClassifier(max_depth=10)
      tr.fit(x,y)

      print(tr.score(x,y))
      print(cross_val_score(tr, x,y, cv=5).mean())

0.990506329113924
0.8513888888888889
```

Anexo 04. Participación de estudiantes universitarios en la encuesta

Estudiantes universitarios de las 6 escuelas profesionales que participaron en la encuesta.

Estudiantes de EPIS

Figura 10. Estudiantes de EPIS caminando



Estudiantes de EPIA

Figura 11. Estudiantes de EPIA



Estudiantes de EPIAM

Figura 12. Estudiantes de EPIAM caminando



Estudiantes de EPAE

Figura 13. Estudiantes de EPAE descansando



Estudiantes de EPC

Figura 14. Estudiantes de EPC en sus aulas



Estudiantes de EPEPI

Figura 15. Estudiantes de EPEPI



Anexo 05. Matriz de Consistencia

Tabla 1. Matriz de Consistencia

PROBLEMA	OBJETIVO	HIPÓTESIS	VARIABLE DE ENTRADA	METODOLOGÍA	POBLACIÓN
Problema General: ¿De qué manera se puede predecir los estilos de vida que adoptan los estudiantes universitarios?	Objetivo General: Predecir los estilos de vida que adoptan los estudiantes universitarios con el modelo predictivo basado en técnicas de Machine Learning.	Hipótesis General: El modelo predictivo basado en técnicas de Machine Learning permite predecir los estilos de vida que adoptan los estudiantes universitarios.	Alimentación: - Verduras y frutas - Restricción de Grasas - Lectura de Información nutricional - Bebidas alcohólicas - Beber agua - Horario de comida - Dulces, golosinas y azúcar	Tipo de Investigación: El tipo de Investigación es predictivo Nivel de investigación: Nivel de investigación descriptivo. Donde: Para el desarrollo de las etapas se basó en la metodología KDD (Knowledge Discovery in Databases).	Población de estudio: La Población estuvo compuesta por el total de 2031 estudiantes universitarios. Muestra de estudio: Conformado por 323 registros de los estudiantes universitarios.
Problema Específicos: ¿Cuál es el modelo para predecir los estilos de vida en la alimentación de los estudiantes universitarios? ¿Cuál es el modelo para predecir los estilos de vida en la actividad física de los estudiantes universitarios? ¿Cuál es el modelo para predecir los estilos de vida en el aspecto psicológico de los estudiantes universitarios?	Objetivo Específicos: Establecer un modelo para predecir los estilos de vida en la alimentación de los estudiantes universitarios usando Máquina de Vectores de Soporte. Establecer un modelo para predecir los estilos de vida en la actividad física de los estudiantes universitarios usando Árboles de decisiones. Establecer un modelo para predecir los estilos de vida en el aspecto psicológico de los estudiantes universitarios usando Árboles de decisiones.	Hipótesis Específicas: El modelo para predecir los estilos de vida en la alimentación está basado en la técnica de Máquina de Vectores de Soporte. El modelo para predecir los estilos de vida en la actividad física está basado en la técnica de Árboles de decisiones. El modelo para predecir los estilos de vida en el aspecto psicológico está basado en la técnica de Árboles de decisiones.	Actividad Física: - Aeróbicas moderadas - Aeróbicas intensas - Fortalecimiento muscular - Flexibilidad - Televisión, laptop y videojuegos - Caminar Psicológico: - Personalidad - Manejo de estrés		

-Depresión

-Sueño

- Crear espacios

-Sobrecarga académica

Variable Dependiente:

Estilos de vida (Diagnóstico):

-Saludable

-No saludable
