UNIVERSIDAD NACIONAL DE MOQUEGUA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

EVALUACION DE MODELOS PREDICTIVOS EMPLEANDO TÉCNICAS DE MACHINE LEARNING PARA EL PRONÓSTICO DE LA VIOLENCIA FISICA CONTRA LA MUJER EN LA PROVINCIA DE ILO, AÑO 2021

TESIS

PRESENTADO POR:

AMPARO LISBETH NUÑEZ GALINDO

PARA OPTAR EL TITULO PROFESIONAL DE:

INGENIERO DE SISTEMAS E INFORMÁTICA

MOQUEGUA – PERÚ

2024

UNAM
UNIVERSIDAD NACIONAL DE MOQUEGUA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

Folio **Nº 014**

ACTA DE SUSTENTACIÓN

I. Modalidad:

Tesis	: [X]
Trabajo de Suficiencia Profesional	: [ ]
Otro:	: [ ]

II. Documento de aprobación para la sustentación:

Resolución de Facultad Nº 220-2024-FACIA-UNAM

Fecha: 06 Noviembre 2024

III. El jurado ha dictaminado la (el) Tesis, TITULADO:

"Evaluación de Modelos Predictivos empleando Técnicas de Machine Learning para el Pronóstico de la Violencia Física contra la mujer en la Provincia de Ilo, Año 2021"

Bachiller: Amparo Lisbeth Nuñez Galindo

Asesor Principal: Vaneza Flores Gutierrez

Co-Asesor: —

IV. Se dictamina:

Como: Aprobado por Unanimidad la sustentación de manera:

Presencial [X] Virtual [ ]

Con nota: Quince (15) por lo que; Sí [X] No [ ]

Opta el Título Profesional de INGENIERO DE SISTEMAS E INFORMÁTICA.

Siendo las 17:22 horas, del 25 de Noviembre del año 2024 se procede con la firma en conformidad de lo dictaminado por el Jurado.

Presidente
Ender Tito Chura

Primer Miembro
Victor Yanz Mamani

Segundo Miembro
Juan Carlos Clares Pérez

Bachiller
Amparo Lisbeth Nuñez Galindo

DEDICATORIA

A Dios, por cada bendición recibida. A mi familia, por su amor y apoyo incondicional.

AGRADECIMIENTO

A mi familia, por su apoyo incondicional, por creer en mí incluso en los momentos difíciles. Gracias a mis padres, su amor, comprensión y sacrificio han sido el motor que me impulsó a alcanzar este logro.

A mis amigos y compañeros, por su compañía y motivación durante este proceso, y por estar siempre dispuestos a escuchar y alentar.

A mis profesores y asesora, en especial a la Dra. Vaneza Flores Gutierrez y a los miembros del jurado, por compartir sus conocimientos, por su paciencia y por contribuir a mi formación profesional.

A la Universidad Nacional de Moquegua, por brindarme la oportunidad de formarme académica y profesionalmente en una institución comprometida con la excelencia educativa.

Finalmente, a todos aquellos que, de alguna manera, aportaron a este logro, les estaré eternamente agradecido.

RESUMEN

En el presente trabajo de investigación se tiene como objetivo la evaluación de modelos predictivos basado en machine learning para pronosticar la violencia física contra la mujer en la Provincia de Ilo. Se tomó en consideración los registros en hojas HIS como instrumento de recojo de datos de los establecimientos de salud en toda la ciudad de Ilo del año 2021 y 2022. Dicha información permitió realizar la evaluación de modelos predictivos con diferentes algoritmos de Machine Learning, y de esta manera definir cuál de ellos tiene una mejor precisión en la predicción de la violencia física contra la mujer. Determinando que el algoritmo más eficiente en predicción es Random Forest con 0.9354 de F1 Score, seguido por Naive Bayes con 0.9340 de F1 Score.

Palabras clave: machine learning, predicción, violencia física contra la mujer, evaluación, F1 Score.

ABSTRACT

The objective of this research is to evaluate predictive models based on machine learning to predict physical violence against women in the province of Ilo. The records in HIS sheets were taken into consideration as an instrument for data collection from health facilities throughout the city of Ilo for the years 2021 and 2022. This information allowed the evaluation of predictive models with different Machine Learning algorithms, and thus define which of them has a better accuracy in predicting physical violence against women. It was determined that the most efficient algorithm in prediction is Random Forest with 0.9354 F1 Score, followed by Decision Trees with 0.9340 F1 Score.

Keywords: machine learning, prediction, physical violence against women, evaluation, F1 Score.

INDICE

RESUMEN 4

ABSTRACT 5

Introducción 14

I. Planteamiento del Problema 16

1.1 Descripción y Formulación del Problema 16

- Descripción de la realidad problemática 16
- Formulación del problema 19
- Interrogante General 19
- Interrogantes Específicos 19

1.2 Objetivos 19

- Objetivo General 19
- Objetivos Específicos 19

1.3 Justificación e Importancia 20

1.4 Hipótesis 20

- Hipótesis General 20
- Hipótesis Especificas 20

1.5 Variables 21

- Variable Independiente 21
- Variable Dependiente 21

II. Marco Teórico 23

2.1 Antecedentes 23

2.2 Bases Teóricas 25

A. Violencia contra la Mujer 25

B. Tipos de Violencia 26

C. Modelo Ecológico 26
D. Machine Learning 27
E. Técnicas de Machine Learning 27
- Decision Trees 27
- Naive Bayes 30
- K-Nearest Neighbors (KNN) 33
- Random Forest 35

F. Métricas de Evaluación del Modelo 37
G. Técnica SMOTE 38
H. Label Encoding 39
I. One-Hot Encoding 40
J. Validación Cruzada 40
2.3 Definición de Términos 41
III. Metodología 42
3.1 Tipo y Nivel de Investigación 42
3.2 Ámbito Temporal y Espacial 42
3.3 Población y Muestra 42
3.4 Instrumentos 42
3.5 Procedimientos 43
3.6 Análisis de Datos 44
IV. Resultados y Discusión 45
- Resultados 45

4.1 Comprensión del Negocio 45
4.2 Comprensión de los Datos 46
4.3 Preparación de los Datos 53

4.4 Modelado 59
4.5 Evaluación.....100
a. Evaluación de los Modelos 100
4.6 Despliegue.....101
• Discusión.....102
V. Conclusiones 103
1. Conclusión General.....104
2. Conclusiones Especificas.....104
VI. Recomendaciones.....107
VII. Referencias.....108
VIII. Anexos 112

INDICE DE TABLAS

Tabla 1 Descripción de data del Ítem R456 47

Tabla 2 Resumen Evaluación KNN 65

Tabla 3 Resumen de Evaluación RF 69

Tabla 4 Resumen de Evaluación NB.....73

Tabla 5 Resumen de Evaluación DT.....77

Tabla 6 Resumen de Evaluación KNN Optimizado.....82

Figura 7 Resumen RF Optimizado.....88

Tabla 8 Resumen de Evaluación NB Optimizado.....93

Tabla 9 Resumen de Evaluación DT Optimizado.....98

Tabla 10 Comparativa de los Resultados de las Técnicas Utilizadas.....100

Tabla 11 Comparación con Antecedentes.....102

INDICE DE FIGURAS

Figura 1 Árbol de Decision Simple....29

Figura 2 Diagrama Naive Bayes....32

Figura 3 Algoritmo K-Nearest Neighbors....34

Figura 4 Diagrama Random Forest....37

Figura 5 Técnica SMOTE....38

Figura 6 Técnica Label Encoding....39

Figura 7 Técnica One-Hot-Encoding....40

Figura 8 Datos no procesados del Ítem R456....46

Figura 9 Visualización de datos únicos de cada columna....52

Figura 10 Eliminación de variables según verificación de calidad de los datos....53

Figura 11 Identificación de variables redundantes....54

Figura 12 Muestra de cantidad en variable dependiente....56

Figura 13 Aplicación de Smote en datos de entrenamiento....58

Figura 14 Función de la Matriz de Confusión....60

Figura 15 Función de Curva ROC....60

Figura 16 Función de F1 Score....61

Figura 17 Función de MCC....61

Figura 18 Código del Algoritmo KNN....62

Figura 19 Predicción de Probabilidades KNN....62

Figura 20 Ejecutar Función de Métricas para KNN....63

Figura 21 Matriz de Confusión KNN....63

Figura 22 F1 Score KNN....64

Figura 23 Curva ROC KNN....64

Figura 24 Gráfico MCC de KNN....65

Figura 25 Código de Random Forest....66

Figura 26 Predicción de Probabilidades RF....66

Figura 27 Matriz de Confusión RF....67

Figura 28 Ejecución de Función de Métricas para RF....67

Figura 29 F1 Score RF....68

Figura 30 Curva ROC RF....68

Figura 31 Gráfico MCC de RF....69

Figura 32 Código del Algoritmo NB....70

Figura 33 Predicción de Probabilidades NB....70

Figura 34 Matriz de Confusión NB....71

Figura 35 Ejecución de las métricas para NB....71

Figura 36 F1 Score NB....72

Figura 37 Curva ROC NB....72

Figura 38 Gráfico MCC de NB....73

Figura 39 Código del Algoritmo DT....74

Figura 40 Predicción de Probabilidades DT....74

Figura 41 Matriz de Confusión DT....75

Figura 42 Ejecutar Función de Métricas DT....75

Figura 43 F1 Score DT....76

Figura 44 Curva ROC DT....76

Figura 45 Gráfico MCC DT....77

Figura 46 Ajuste de Hiperparametros KNN....78

Figura 47 Ejecución de Predicción KNN optimizado....79

Figura 48 Matriz de Confusión KNN Optimizado....80

Figura 49 Evaluación de KNN Optimizado....80

Figura 50 Curva ROC KNN Optimizado....80

Figura 51 Gráfico MCC de KNN Optimizado....81

Figura 52 F1 Score KNN Optimizado....81

Figura 53 Feature Importances KNN Optimizado....83

Figura 54 Ajuste de Hiperparametros RF Optimizado....84

Figura 55 Ejecución de Predicción RF Optimizado....85

Figura 56 Evaluación de RF Optimizado....86

Figura 57 Matriz de Confusión RF Optimizado....86

Figura 58 F1 Score RF Optimizado....87

Figura 59 Curva ROC RF Optimizado....87

Figura 60 Gráfico MCC de RF Optimizado....88

Figura 61 Feature Importances RF Optimizado....89

Figura 62 Ajuste de Hiperparametros NB....90

Figura 63 Ejecución de Predicción NB optimizado....91

Figura 64 Evaluación de NB Optimizado....91

Figura 65 Matriz de Confusión NB Optimizado....91

Figura 66 F1 Score NB Optimizado....92

Figura 67 Curva ROC NB Optimizado....92

Figura 68 Gráfico MCC NB Optimizado....93

Figura 69 Feature Importances NB Optimizado....94

Figura 70 Ajuste de Hiperparametros DT....95

Figura 71 Ejecución de Predicción DT optimizado....96

Figura 72 Evaluación de DT Optimizado 96

Figura 73 Matriz de Confusión DT Optimizado 96

Figura 74 F1 Score DT Optimizado 97

Figura 75 Curva ROC DT Optimizado 97

Figura 76 Grafica MCC de DT Optimizado 98

Figura 77 Feature Importances DT Optimizado 99

Figura 78 Feature Importances en base a Random Forest 101

Figura 79 Determinación de Feature Importances 112

Figura 80 Excel con la data de violencia física y psicológica 112

Figura 81 Documento de solicitud de Información 113

Introducción

La violencia contra la mujer es un fenómeno que ocurre desde la antigüedad, con todo el avance ya sea político, ambiental y tecnológico no se llega a erradicar en su totalidad, y como una ficha de dominó la violencia hacia la mujer desencadena una serie eventos desafortunados que no solo afecta a ella sino a las personas más cercanas en su entorno social y familiar. Por ello con el presente trabajo de investigación se plantea evaluar modelos predictivos basado en Machine Learning para pronosticar la violencia física contra la mujer en la Provincia de Ilo, y de esta manera contribuir de manera favorable en la reducción de la violencia contra la mujer futura.

La provincia de Ilo cuenta con la dirección de Red de Salud Ilo, el cual es un órgano desconcentrado de la Dirección Regional de Salud Moquegua, cuya responsabilidad es asegurar una adecuada prestación de servicios de salud a la población de su ámbito(Dirección Regional de Salud Moquegua, 2008). Para esta investigación se tomará en consideración las hojas HIS específicamente en referencia a los problemas relacionados con violencia de los establecimientos de salud situados en diferentes zonas centralizadas de la ciudad de Ilo registradas durante el año 2021 y 2022, entre estos datos se encuentra la modalidad de violencia física, psicológica y puntos claves para los factores identificados en la violencia contra la mujer. En otro ámbito, el estado de arte de la investigación enfocado en la disciplina de Machine Learning nos propone diversos algoritmos y métodos abarcados para la predicción de violencia contra la mujer en diferentes Países.

De esta manera, se realizó la evaluación de diferentes modelos de machine learning para subsidiar el problema planteado y asegurar resultados óptimos, las cuales serán detalladas y explicadas en los resultados y conclusiones.

El presente informe de investigación consta de cuatro partes principales los cuales son: Introducción, marco teórico, metodología, resultados y conclusiones.

I. Planteamiento del Problema

1.1 Descripción y Formulación del Problema

- **Descripción de la realidad problemática**

La violencia contra la mujer es un fenómeno global en diversos sectores sociales y afecta no solo a las mujeres sino también a las personas directamente relacionadas con la víctima como niños, familiares y amigos. Esto puede afectar la salud de una persona y causar daños graves a largo plazo a los niños que han crecido en familias disfuncionales como esta.

Las Naciones Unidas definen la violencia contra la mujer como cualquier acto de violencia de género que resulte en, o pueda resultar en, daño o sufrimiento físico, sexual o mental a la mujer, incluyendo amenazas de tales actos, coerción o privación arbitraria de libertad, ya sea en la vida pública o privada (World Health Organization, 2021).

En el Perú la violencia contra la mujer se ha mantenido en cifras alarmantes estableciendolo como un problema que se debe abordar urgentemente y ello se ve reflejado en el documento brindado por el comité estadistico interinstitucional de la criminalidad Rafael et al. (2021) conformado por diferentes entidades importantes para la recabacion de información referente al feminicidio y violencia contra la mujer a nivel nacional del año 2020 en el cual establece que , de un total de 97 926 victimas de violencia contra la mujer el 47,7% fueron por violencia psicologica y/o verbal, 39,1% por violencia fisica y 13,2% por violencia sexual. Esta cifra ha ido aumentando para el año 2021 en el cual según expresa el Programa Nacional AURORA (2021), 140 833 victimas de violencia contra la mujer el 45% fueron por violencia psicologica

y/o verbal, 39,9% por violencia fisica y 15,1% por violencia sexual. Con ello se puede resaltar un incremento del 30,47% entre el año 2020 y 2021 un porcentaje de aumento preocupante.

En los últimos años el Departamento de Moquegua ha reflejado un nivel muy alto respecto a la violencia contra la mujer en las tres formas principales mencionadas anteriormente como son violencia a psicológica y/o verbal, violencia física y sexual. El Programa Nacional AURORA (2020) ha reflejado que, de una población total de 174 863 en el Departamento de Moquegua, 762 mujeres han sido víctimas de violencia, en lo cual esta cifra ha ido en aumento para el año 2021 que se reportaron 1 350 casos de violencia contra la mujer en el Departamento de Moquegua (Programa Nacional AURORA, 2021).

Actualmente de lo que va de Enero a Marzo de 2022 ya se han sido atendidos 126 casos de violencia contra la mujer en la Provincia de Ilo, si bien los metodos establecidos frente a estos casos son de apoyo frente a la violencia ya ejercida es necesario contar con diferentes herramientas para la prevencion del mismo(Programa Nacional AURORA, 2022).

Se ha visto necesario enfocarse en una modalidad especifica en la violencia contra la mujer que se ha mantenido constante desde 2015-2020 pero ha ido aumentando en el año 2021 y es la violencia fisica. Según el informe principal ENDES 2020 del Instituto Nacional de Estadística e Informática (2021) la violencia fisica a tenido una baja minima de 10% quedando en 27,1% a nivel nacional en lo que respecta del año 2009 a 2020 que fue su ultimo censo demografico de salud familiar a comparacion de la violencia psicologica y/o verbal que bajo un 23% y considerandose en 50,1% a nivel nacional en el mismo periodo de tiempo, y esto ha tenido un cambio

considerable en el informe estadistico del Programa Nacional Aurora en el que resalta que la violencia psicologica se disminuyó actualmente en un 45% en comparacion a la violencia fisica que aumentó a un 39,9% durante el perido Enero-Diciembre de 2021(Programa Nacional AURORA, 2021).

Frente a todo ello se ha visto necesario realizar una evaluación de modelos predictivos empleando técnicas de machine learning para el pronóstico de la violencia física contra la mujer y así determinar la conducta futura frente a este fenómeno grave que afecta a diferentes Países. Esta investigación se llevará a cabo en la Provincia de Ilo considerando que en la actualidad no se cuenta con un modelo predictivo que contribuya a la prevención de la violencia física contra la mujer, y se analizará los registros de hojas HIS en los establecimientos de salud de la Provincia de Ilo, esto tomando en cuenta que el Instituto Nacional de Estadística e Informática (2021) refleja en la Encuesta Demográfica y de Salud Familiar 2020 que el 83,9% de mujeres víctimas de violencia recurren a una comisaría en busca de ayuda y como tal según la ley N° 30364 expresa que la entidad policial tiene la obligación de proceder de manera que garantice la protección y bienestar de la víctima con una evaluación médica, informe legal médico, protección, atención, denuncia y procedimientos legales(Peruano PODER LEGISLATIVO & Presidente La República, 2015). La atención medica es realizada en el establecimiento de salud más cercano a la comisaria donde se hace la denuncia por ello se ha considerado obtener los registros de estas entidades.

En base a ello se considera que con un modelo predictivo se puede brindar un aporte importante para el riesgo de violencia física futura.

- **Formulación del problema**

  - **Interrogante General**

    A partir de datos de entrada, ¿se podrá evaluar modelos predictivos basados en técnicas de machine learning para pronosticar la violencia física contra la mujer en la Provincia de Ilo?

  - **Interrogantes Específicos**

    1. ¿Cuáles son los factores más influyentes en los modelos predictivos que explica la violencia física contra la mujer en la Provincia de Ilo?
    2. Un modelo predictivo basado en técnicas de machine learning ¿podrá predecir la violencia física contra la mujer en la Provincia de Ilo?
    3. ¿Es posible precisar el modelo más eficaz a partir del desarrollo y aplicación comparativa del mismo para la predicción de la violencia física contra la mujer en la Provincia de Ilo?

1.2 Objetivos

- **Objetivo General**

  Evaluar modelos predictivos empleando técnicas de machine learning para pronosticar la violencia física contra la mujer en la Provincia de Ilo.

- **Objetivos Específicos**

  1. Determinar los factores más influyentes en los modelos predictivos que expliquen la violencia física contra la mujer en la Provincia de Ilo.
  2. Construir un modelo predictivo basado en técnicas de machine learning para la predicción de la violencia física contra la mujer en la Provincia de Ilo

3. Aplicar métricas de evaluación para precisar el modelo más óptimo que permita la predicción de la violencia física contra la mujer en la Provincia de Ilo.

1.3 Justificación e Importancia

La violencia contra la mujer es un delito de gravedad que da pie a la opresión en la cual se ve sometida la mujer, ello afecta en todos los aspectos de su vida y trae como consecuencia la falta de autonomía y baja productividad sin la capacidad de cuidar por sí mismas y/o de sus hijos, conllevando a mantener una pésima calidad de vida.

Con lo expuesto se considera de vital importancia realizar una evaluación y predicción de la violencia física contra la mujer en la provincia de Ilo, tomando en cuenta que no se cuenta actualmente con un modelo predictivo fiable que permita predecir el comportamiento violento futuro al que se podría encontrar expuesta una mujer y el cual se encuentra en una delgada brecha del feminicidio, evitar ello es fundamental para la seguridad de la mujer.

1.4 Hipótesis

- **Hipótesis General**

El modelo predictivo construido con técnicas de machine learning predice considerablemente la violencia física contra la mujer en la Provincia de Ilo.

- **Hipótesis Especificas**

1. La determinación de los factores más influyentes contribuye a predecir considerablemente la violencia física contra la mujer en la Provincia de Ilo.
2. Un modelo predictivo, basado en técnicas de machine learning, puede predecir considerablemente la violencia física contra la mujer en la Provincia de Ilo.

3. El resultado de la comparación de modelos predictivos, basados en técnicas de machine learning, determina el modelo más significativo para predecir la violencia física contra la mujer en la Provincia de Ilo.

1.5 Variables

- **Variable Independiente**

"Modelo predictivo empleando técnicas de Machine Learning"

Para el presente proyecto se emplea técnicas de Machine Learning como variables independientes esto debido a la capacidad para identificar patrones complejos en los datos y generar modelos predictivos basados en múltiples características, para este caso nos basamos en el modelo ecológico propuesto por (Heise, 1998) que permite establecer una relación de los factores en base a la modalidad de violencia en cada víctima.

Las técnicas utilizadas permitirán automatizar la detección de violencia física a través del análisis de diferentes variables, ya que estos pueden manejar datos tanto numéricos como categóricos, permitiendo la creación de modelos robustos y generalizables que maximicen la precisión en la predicción de eventos de violencia física.

- **Variable Dependiente**

"Violencia Física contra la mujer"

La variable dependiente seleccionada es la Violencia Física contra la mujer, la cual está categorizada como 'SI' o 'NO' en el conjunto de datos. Esta variable es crítica ya que representa el evento que el modelo intenta predecir. La clasificación binaria de

esta variable facilita la aplicación de algoritmos de clasificación supervisada. Dada la importancia social y la naturaleza sensible del problema, es vital que el modelo minimice tanto los falsos positivos como los falsos negativos, para no clasificar incorrectamente a mujeres que no han sufrido violencia, ni omitir casos reales y realice una predicción eficiente en la predicción de violencia física contra la mujer. Esto justifica el uso de métricas como la precisión, el recall, y el F1-score, entre otros en la evaluación del modelo.

II. Marco Teórico

2.1 Antecedentes

En el estudio realizado por Saboya et al. (2019) presentaron el articulo *"Predictive Model Based on Machine Learning for the Detection of Physically Mistreated Women in the Peruvian Scope"* Donde es motivo de estudio la elaboración de un modelo predictivo que permita el reconocimiento de la futura mujer físicamente maltratada, en el cual utilizaron tres modelos de aprendizaje supervisado, Random Forest, Clasificación con Arboles de Decision y Clasificación con Arboles adicionales, dichos modelos fueron utilizados para comprobar los resultados y seleccionar el de mejor rendimiento. En cuanto a los resultados se llegó a determinar que el mejor modelo fue Random Forest con un 64% en la métrica de F1 Score como segundo mejor modelo se considera Clasificación de Arboles adicionales con un F1 Score de 50% y por último el modelo de Clasificación de Arboles de decisión con un F1 Score de 48%.

El método que se utilizó en la investigación de Timaran-Pereira et al. (2017) en el artículo *"Aplicación de los árboles de decisión en la identificación de patrones de lesiones fatales por causa externa en el municipio de Pasto, Colombia"* presentan la detección de patrones delictivos con técnicas de minería de datos para el cual se construyó un modelo de clasificación basado en arboles de decisión utilizando el algoritmo J48 de la herramienta Weka, que permitió descubrir patrones de las muertes externas con una precisión del 82% lo cual fue considerado aceptable, confiable y eficiente para la clasificación de nuevos casos como homicidios entre otros.

Considerando el estudio de Amusa et al. (2020), en el artículo *"Predicting the Vulnerability of Women to Intimate Partner Violence in South Africa: Evidence from*

Tree-based Machine Learning *Techniques"* presentan la construcción de un modelo predictivo que clasifique de forma eficiente la probabilidad de que ocurra violencia contra la mujer. Tomando en cuenta los factores asociados, y la relación de agresor-victima, construyendo un modelo basado en arboles de decisión, Random forest, aumento de gradiente y modelo convencional LR para predecir la concurrencia de violencia a la mujer. Llegando a la conclusión de que los modelos con mejor desempeño en la predicción fueron Arboles de Decisión y Random Forest por sobre los demás modelos, y con ello tomando en consideración un modelo de clasificación interpretable para la identificación de factores, prefirieron Arboles de decisión sobre Random Forest.

La investigación realizada por Tamilarasi & Rani (2020) denominada "*Diagnosis of Crime Rate against Women using k-fold Cross Validation through Machine Learning*" en el cual realizaron una evaluación de la eficacia y aplicación de los algoritmos como KNN, arboles de decisión, Naive Bayes y arboles de clasificación y regresión para el análisis de datos de las numerosas denuncias de violencia contra la mujer, llegando a la conclusión de que entre la evaluación y comparación de la eficiencia de todos los algoritmos ya mencionados es KNN con un F1 Score de 30% el cual destaca por sobre los demás considerándolo el algoritmo más eficiente para la predicción del tipo de crimen que ocurre periódicamente en la India.

El estudio realizado por Al-Taleb & Saeed (2019) en el artículo "*A study on application of data mining techniques in crime forecasting*" presentaron modelos predictivos con los algoritmos de minería de datos tales como Support Vector Machine (SVM) y Naive Bayes para determinar los patrones delictivos y la categoría de delitos que es probable que ocurra, llegando a la conclusión que el clasificador Naive Bayes

supera ligeramente al algoritmo SVM en eficacia para la predicción de patrones delictivos con un 75% de F1 Score.

El estudio de Cumbicus-Pineda et al. (2021)titulado “Data Mining to Determine the Causes of Gender-Based Violence against Women in Ecuador” tiene como objetivo determinar las causas de violencia contra la mujer, en el cual utilizaron algoritmos como arboles de decisión (J48), redes neuronales, KNN y Random Forest en el cual obtuvieron como resultado que el algoritmos de KNN fue el mejor clasificador para los datos de forma global mientras Random Forest se tuvo una mejor precisión de clasificación.

La investigación realizada por Rodríguez-Rodríguez et al. (2020)denominado “Modeling and forecasting gender-based violence through machine learning techniques” en el cual desean pronosticar la violencia de genero para interferir su incidencia futura para ellos utilizaron algoritmos como Random Forest, KNN, Regresión lineal y Support Vector Machine (SVM) llegando a la conclusión de que el algoritmo Random Forest ofrece una mejor precisión del modelo en una aproximado de 6 meses de anticipación.

2.2 Bases Teóricas

A. Violencia contra la Mujer

World Health Organization (2021) define la violencia contra la mujer como “todo acto de violencia de género que resulte, o pueda tener como resultado un daño físico, sexual o psicológico para la mujer, inclusive las amenazas de tales actos, la coacción o la privación arbitraria de libertad, tanto si se producen en la vida pública como en la privada”.

B. Tipos de Violencia

- Violencia Sexual: Cualquier acto que agrede contra el cuerpo y/o sexualidad de la víctima y que por lo mismo atente contra su libertad, dignidad e integridad física (Olivares et al., 2011).
- Violencia Física: Agresiones físicas dirigidas al cuerpo de la víctima por parte del agresor causando daños temporales o permanentes (Olivares et al., 2011).
- Violencia Emocional/Psicológica: Todo tipo de acto que produzca desvalorización, sufrimiento o agresión psicológica, el cual no se limita a insultos o gritos sino a humillaciones hacia la víctima (Olivares et al., 2011).

C. Modelo Ecológico

El modelo ecológico propuesto por Heise (1998) estudia factores que actúan en diferentes ámbitos: individual, familiar, comunicativo y sociocultural; este modelo consiste en la interacción que se realiza en factores de distinto nivel. El cual explica que no es solo un factor casual sino una interacción de diferentes factores en distintos niveles en la violencia a la mujer. Por ello se tomará el enfoque en los siguientes factores tomando como base el modelo ecológico:

- **Factores individuales:** Se enfoca al conjunto de hechos y situaciones que intervienen en el desarrollo de un individuo en un punto determinado de su vida.
- **Factores socioculturales:** Es un agregado de relaciones sociales entre individuos que comparten una misma cultura e interactúan entre sí, con el cual conforman una sociedad.

- **Factores familiares:** Hace relación a los primeros patrones en el comportamiento de una persona, la familia aporta considerablemente en la desarrollo emocional y actitudinal de una niña.
- **Factores comunicativos:** Se referencia en relación en el aislamiento social y familiar de las mujeres y en formar parte de grupos violentos.

D. Machine Learning

Kotsiantis (2007) argumenta que el aprendizaje supervisado consiste en encontrar algoritmos de razonamiento a partir de ejemplos externos para generar hipótesis que luego predicen casos futuros.

El método de aprendizaje supervisado construye modelos predictivos probando una gran cantidad de ejemplos de entrenamiento, cada ejemplo de entrenamiento tiene una etiqueta que indica su salida. Los algoritmos de aprendizaje supervisado incluyen regresión lineal, regresión logística, Random Forest, Support Vector Machine Neural Network Artificial, Decision Tree, k-Nearest Neighbors entre otros (Kotsiantis, 2007).

E. Técnicas de Machine Learning

- **Decision Trees**

Los árboles de decisión son populares porque tienen una representación sencilla de problemas con un número finito (y a ser posible reducido) de clases. Además, son modelos comprensibles y proposicionales, se puntualizan en que un árbol de decisión es una estructura en la cual cada nodo interno significa una prueba sobre uno o varios atributos, donde cada rama representa una salida de la prueba y los nodos hojas representan clases (Hernández et al., 2005).

Para cada nodo, el algoritmo evalúa cada característica y elige la que maximiza la reducción de impureza en los subconjuntos resultantes. Las métricas más comunes para evaluar esta impureza son:

- **Entropía (para ganancia de información)**: La entropía mide la impureza de un conjunto de datos. Para un conjunto S con c clases, la entropía se define como:

$$H(S)=-i=1\sum cpilog2(pi)$$

  Donde Pi es la proporción de ejemplos en la clase i.

- **Gini Impurity**: La impureza de Gini es otra medida de impureza. Para un conjunto S con c clases, se define como:

$$Gini(S)=1-i=1\sum cpi2$$

  Donde Pi es la proporción de ejemplos en la clase i.

- **Ganancia de Información (Information Gain)**: La ganancia de información es la reducción en la impureza (entropía) después de dividir el conjunto de datos. Si se divide un conjunto S en subconjuntos S1 S2 basados en una característica, la ganancia de información es:

$$Gain\ (S, A) = H(S)-i=1\sum k|S||Si|H(Si)$$

  Donde A es la característica utilizada para la división, y Si son los subconjuntos resultantes de la división.

- **Reducción de Gini**: Similar a la ganancia de información, la reducción de Gini mide cuánto reduce la impureza de Gini al realizar una división:

$$\Delta Gini\ (S, A) = Gini(S)-i=1\sum k|S||Si|Gini\ (Si)$$

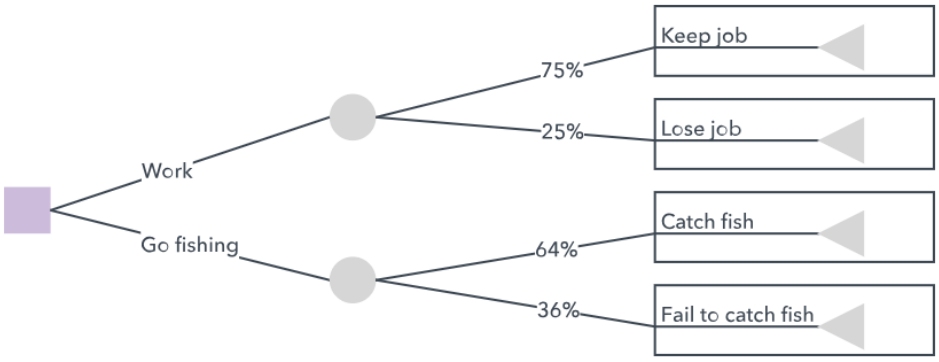
  Donde Si son los subconjuntos tras la división por la característica A.

- **Criterio de división**: En cada nodo, el algoritmo selecciona la característica A y el valor de corte v que maximiza la ganancia de información o la reducción de Gini.
- **Determinación de hojas**: Las hojas del árbol contienen las predicciones finales. Cuando un nodo no puede ser dividido de manera que aumente significativamente la ganancia de información (o reduzca la impureza de Gini), se detiene la partición, y el nodo se convierte en una hoja, asignando la clase más frecuente en ese nodo.

Este proceso se repite en cada nodo del árbol, seleccionando en cada paso la característica que maximiza la ganancia de información o la reducción de Gini, hasta que se cumplen las condiciones de parada(Quinlan, 1986).

Figura 1

Árbol de Decision Simple



Nota. Adaptado de (Lucidchart, 2024).

- **Naive Bayes**

Una de las características primordiales de los métodos bayesianos es el uso de distribuciones de probabilidad para cuantificar incertidumbre de los datos que se desea modelar. Estos métodos proporcionan una metodología práctica para la inferencia y predicción y, en última instancia, para tomar decisiones que involucran cantidades inciertas (Hernández et al., 2005).

El clasificador Naive Bayes según expresa Mccallum & Nigam, (1998) se basa en el Teorema de Bayes y asume que todas las características son independientes entre sí (condicionalmente independientes dado la clase). La fórmula general del teorema de Bayes es:

$$P(C|X) = \frac{P(X|C)\cdot P(C)}{P(X)}$$

donde:

P(C|X): Probabilidad de que la instancia X pertenezca a la clase C (posterior).

P(X|C): Probabilidad de observar X dado que la clase es C (verosimilitud).

P(C): Probabilidad previa de la clase C (prior).

P(X): Probabilidad total de X (constante para todas las clases, puede omitirse en clasificación).

El clasificador Naive Bayes toma decisiones sobre la clase más probable dado un conjunto de características X= {x1, x2, …, xn} Dado que asume independencia condicional entre las características, la verosimilitud P(X|C) se descompone en el producto de probabilidades individuales:

$$P(X|C) = P(x1|C) \cdot P(x2|C) \cdot \ldots \cdot P(xn|C)$$

Así, la fórmula para la probabilidad de que una instancia pertenezca a la clase CCC se convierte en:

$$P(C|X) \propto P(C) \cdot i = \prod_{i=1}^{n} P(xi \mid C)$$

Pasos en la predicción con Naive Bayes:

- **Calcular la probabilidad previa** P(C) para cada clase.
- **Calcular la verosimilitud** P(Xi|C) para cada característica xi dado cada clase C.
- **Calcular el producto** P(C)·∏P(Xi|C) para cada clase.
- **Seleccionar la clase** que maximice P(C|X).

Variantes de Naive Bayes

Gaussiano Naive Bayes: Se asume que las características siguen una distribución normal. La verosimilitud P(Xi|C) para cada característica continua se calcula como:

$$P(xi \mid C) = \frac{1}{\sqrt{2\pi\sigma^2 C}} \exp\left(-\frac{(xi - \mu C)^2}{2\sigma^2 C}\right)$$

Donde μC y σ2C son la media y varianza de la característica xi en la clase C.

Multinomial Naive Bayes: Utilizado para datos de conteo como texto. La verosimilitud es:

$$P(xi \mid C) = \frac{Ni, C}{\sum Ni', C}$$

Donde Ni, C es el número de veces que el atributo i aparece en la clase C, y el denominador es la suma de todos los atributos en la clase C.

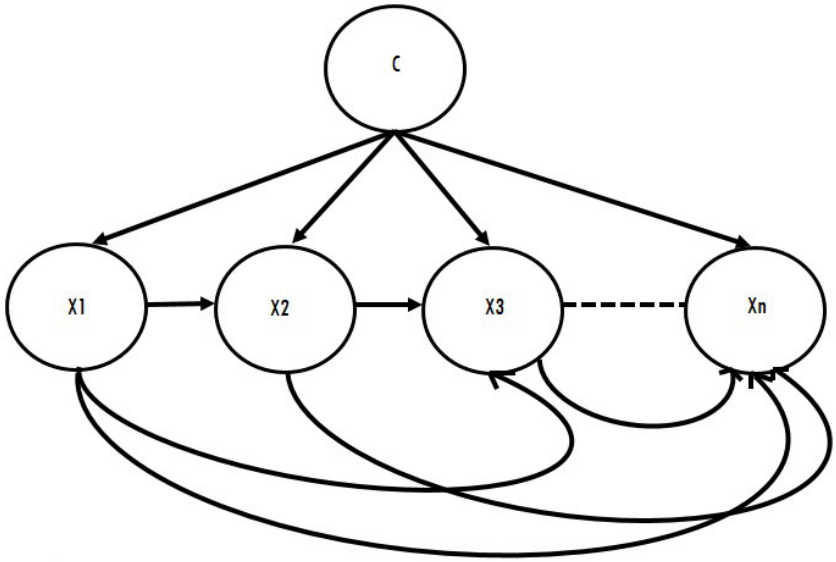
Bernoulli Naive Bayes: Asume que las características son binarias, por lo que la verosimilitud se calcula como:

$$\mathrm{P}(\mathrm{xi} \mid \mathrm{C}) = p_{\mathrm{C}}^{xi}.(1 - p\mathrm{C})^{1-xi}$$

Donde pC es la probabilidad de que la característica sea 1 en la clase C.

Figura 2

Diagrama Naive Bayes



Nota. Adaptado de (Mccallum & Nigam, 1998).

- **K-Nearest Neighbors (KNN)**

  El método de clasificación basado en k-Nearest Neighbors se basa en el hecho de que los atributos de los datos la entrada x es la misma que la especificada, su entorno pertenece al mismo que la clase más común de vecinos más cercano (Zapata-Tapasco et al., 2014).

  Paso a paso del algoritmo:

  - **Distancia entre puntos**: El KNN se basa en la noción de distancia para determinar los vecinos más cercanos. La métrica de distancia más común es la **distancia euclidiana.** Para dos puntos X= (x1, x2, …, xn) y Y= (y1, y2, …, yn), la distancia euclidiana está dada por:

$$\mathrm{d(X,Y)} = \sqrt{(x1 - y1)^2 + \ (x2 - y2)^2 + \cdots + \ (xn - yn)^2}$$

  - **Selección de los k vecinos más cercanos**: Dada una instancia XXX no etiquetada, el algoritmo busca los k puntos más cercanos en el conjunto de entrenamiento utilizando una de las métricas de distancia mencionadas. Los vecinos más cercanos son aquellos que minimizan d (X, Y).
  - **Asignación de clase** (para clasificación):

    El algoritmo asigna a X la clase mayoritaria entre sus k vecinos más cercanos. La predicción se realiza mediante votación de mayoría**:**

$$\hat{y} = \arg \underbrace{max}_{c} \sum_{i=1}^{k} 1(yi = c)$$

    donde 1(yi=c) es una función indicadora que toma el valor 1 si el vecino i pertenece a la clase c, y 0 en caso contrario.

- **Asignación de valor** (para regresión):

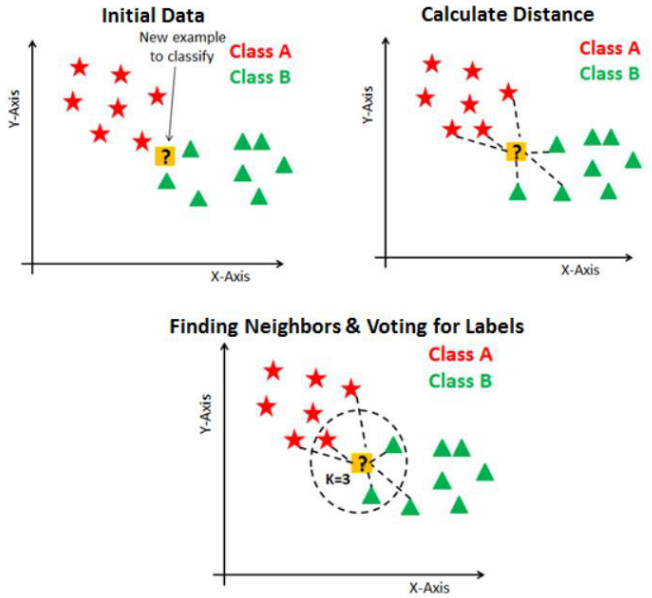
  En el caso de problemas de regresión, el valor predicho es el **promedio** de los valores de los k vecinos más cercanos:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} yi$$

- **Parámetro k**: El valor de k afecta significativamente el rendimiento de KNN. Un k pequeño puede resultar en un modelo ruidoso (alta varianza), mientras que un k grande puede generar predicciones más estables, pero menos sensibles a los detalles locales (alta tendencia).
- **Normalización de los datos**: Debido a que KNN utiliza distancias, es importante normalizar o escalar las características para que todas tengan el mismo rango, especialmente si tienen unidades diferentes o escalas muy distintas(Aha et al., 1991).

Figura 3

Algoritmo K-Nearest Neighbors



Nota. Adaptado de (Aha et al., 1991)

- **Random Forest**

Breiman (2001) define un bosque aleatorio como un clasificador formado por un conjunto de árboles clasificadores, es un método de aprendizaje supervisado que genera múltiples árboles de decisión sobre el conjunto de datos de entrenamiento: los resultados obtenidos se combinan para obtener un modelo más robusto que los resultados de cada árbol por separado.

La predicción final de un Random Forest depende del tipo de problema:

- **Clasificación**: La predicción final es la clase más votada por todos los árboles.

$$y\hat{}=\text{mode}(T1(x), T2(x), \ldots, TB(x))$$

donde y ^ es la predicción final, Ti(x) es la predicción del i-ésimo árbol, y B es el número total de árboles.

- **Regresión**: La predicción final es el promedio de las predicciones individuales de los árboles.

$$\hat{y} = \frac{1}{B}\sum_{i=1}^{B} Ti(x)$$

donde Ti(x) es la predicción del i-ésimo árbol para la instancia x.

Pasos clave del algoritmo Random Forest:

- **Muestreo por Bootstrap**: Se selecciona aleatoriamente, con reemplazo, un subconjunto de ejemplos de entrenamiento para cada árbol. Este proceso es conocido como bagging**.**
- **Selección aleatoria de características**: Para cada nodo del árbol de decisión, se selecciona un subconjunto aleatorio de características, y la división se realiza en función de la mejor característica de ese subconjunto. Esto introduce una mayor diversidad en los árboles y reduce la correlación entre ellos.

- **Construcción de árboles**: Cada árbol de decisión se construye utilizando el subconjunto de datos de entrenamiento y el subconjunto de características, siguiendo el algoritmo estándar de árboles de decisión.
- **Agregación de predicciones**:

  Para problemas de clasificación: se toma el voto mayoritario entre los árboles.

  Para problemas de regresión: se toma el promedio de las predicciones de todos los árboles.

  Fórmulas matemáticas clave:

  Entropía (para clasificación):

$$\mathrm{H(S)} = -\sum_{i=1}^{k} pi\ \log_2(pi)$$

  donde pi es la proporción de instancias de la clase i en el nodo S, y k es el número de clases.

  Índice de Gini (para clasificación):

$$\mathrm{Gini(S)} = 1 - \sum_{i=1}^{k} Pi^2$$

  donde pi es la proporción de la clase i en el nodo S.

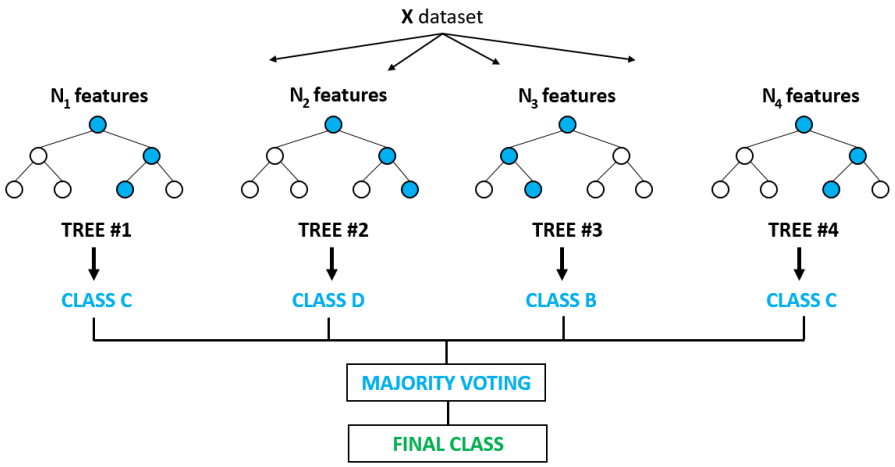
  Error cuadrático medio (para regresión):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(yi - \widehat{yi})^2$$

  donde yi es el valor real, $\widehat{yi}$ es la predicción del modelo, y n es el número de instancias(Rokach, 2010).

Figura 4

Diagrama Random Forest



Nota. Adaptado de (Breiman, 2001).

F. Métricas de Evaluación del Modelo

- **Matriz de confusión**

  Según Kotsiantis (2007) la matriz de confusión es una herramienta que ayuda a visualizar que tan bien se comporta un modelo esto determinado en 4 cuadrantes que informa en que se equivocó el modelo.

- **Curva ROC**

  Es una herramienta grafica que se utiliza para evaluar el rendimiento del modelo, la curva de ROC muestra cómo cambia el rendimiento del modelo en base al cambio de umbral de decisión(Kotsiantis, 2007).

- **F1 Score**

  Es una métrica que obtiene el promedio armonioso de la precisión y la sensibilidad, lo cual infiere que encuentra un equilibrio entre ambas métricas, dicha métrica da más importancia a los FP y FN (Kotsiantis, 2007).

- **MCC**

Coeficiente de correlación de Matthews es una métrica que evalúa la calidad de un modelo de clasificación binaria, es confiable en clases desbalanceadas ya que ofrece una medida más equilibrada del modelo (Kotsiantis, 2007).

G. Técnica SMOTE

Es una técnica utilizada para resolver el problema del desequilibrio de clases en conjuntos de datos clasificados. La idea principal de SMOTE es generar nuevas instancias sintéticas para algunas categorías en lugar de simplemente copiar instancias existentes.

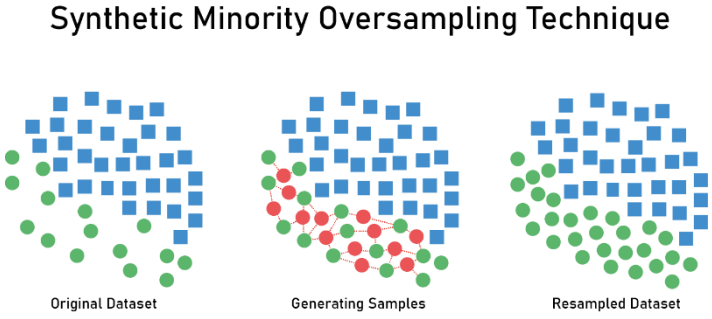
El proceso de SMOTE según lo detalla Chawla et al., (2002b) sigue unos pasos específicos como la selección aleatoria de un punto minoritario luego realiza la búsqueda de los K-vecinos más cercanos y por último una interpolación lineal, se muestra la formula general de SMOTE a continuación:

$$xsynt = xi + \lambda \times (xi' - xi),\ \lambda \in [0,1]$$

Esta fórmula garantiza que los puntos sintéticos estén distribuidos aleatoriamente en la vecindad de los puntos originales minoritarios, contribuyendo a un aumento balanceado de la clase minoritaria

Figura 5

Técnica SMOTE



Nota. Adaptado de (Chawla et al., 2002).

H. Label Encoding

Según se lo especificado por (Alakh, 2024) , es una técnica utilizada en el procesamiento de datos para convertir etiquetas categóricas en valores numéricos utilizados por algoritmos de aprendizaje automático, esto permite asignar a cada categoría un numero entero único, esto se puede describir como una función de mapeo según se expresa en la siguiente formula:

$$f:C \rightarrow Z$$

donde f(Ci)= i es el índice correspondiente a la categoría Ci y Z es el conjunto de números enteros.

Figura 6

Técnica Label Encoding

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29



Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

Nota. Adaptado de (Alakh, 2024).

I. One-Hot Encoding

Es una técnica de codificación que se utiliza para convertir datos categóricos en un formato que pueda ser interpretado por algoritmos de aprendizaje automático. A diferencia de la Label Encoding, que asigna números enteros a categorías, la codificación One-Hot crea columnas binarias separadas para cada categoría (Alakh, 2024).

One Hot Encoding crea una representación binaria para cada categoría, transformando un valor categórico Ci en un vector de longitud n (el número total de categorías). La representación se puede expresar como:

$$OHE\ (ci) = [0, 0, ...,1, ..., 0]$$

Figura 7

Técnica One-Hot-Encoding



Nota. Adaptado de (Alakh, 2024).

J. Validación Cruzada

En lugar de evaluar un modelo solo con una partición de datos el cual comprende de entrenamiento y prueba, la validación implementa múltiples particiones para proporcionar una estimación más robusta del rendimiento del modelo (Mohammed & Kora, 2023).

2.3 Definición de Términos

A. Modelo Predictivo

Es un modelo de datos, basado en estadísticas inferenciales, que se utiliza para predecir la respuesta a una promoción de marketing directo. El modelo predictivo utiliza estadísticas para predecir los resultados (Arimetrics, 2022).

B. Violencia

"El uso intencional de la fuerza o el poder físico, de hecho, o como amenaza, contra uno mismo, otra persona o un grupo o comunidad, que cause o tenga muchas probabilidades de causar lesiones, muerte, daños psicológicos, trastornos del desarrollo o privaciones" (World Health Organization, 2021).

C. ROC: Características Operativas Del Receptor.

D. SVM: Support Vector Machine.

E. KNN: k-Nearest Neighbors o k-Vecinos más cercanos.

F. RF: Random Forest

G. NB: Naive Bayes

H. Hoja HIS: Hojas clínicas utilizadas por los establecimientos de Salud del Perú

I. Item R456: Característica de Violencia Física y psicológica

J. P.S: Puesto de Salud**.**

K. C.S: Centro de Salud.

L. Feature Importances: Características importantes.

III. Metodología

3.1 Tipo y Nivel de Investigación

La presente investigación se considera del tipo aplicada ya que se utilizarán conocimientos de machine learning para el logro del objetivo de la investigación(Vargas Cordero Zoila Rosa, 2009), el nivel de investigación se considera explicativo por las hipótesis planteadas que se desean verificar empíricamente(Fidias G. Arias, 2012).

3.2 Ámbito Temporal y Espacial

La presente investigación es realizada a partir de los registros en hojas HIS entre el año 2021 y 2022 esto se delimita a los establecimientos de salud de la provincia de Ilo.

3.3 Población y Muestra

- **Población**

  Se toma en consideración a la población de un total de 1280 registros de violencia contra la mujer extraído de los registros de hojas HIS de los establecimientos de salud de la provincia de Ilo.

- **Muestra**

  La muestra utilizada está conformada por una muestra censal, por lo tanto, se tomará el total de la población referenciada en la sección anterior.

3.4 Instrumentos

- Registro de hojas HIS sobre violencia contra la mujer en la Provincia de Ilo, información que se recolectará en una base de datos.
- Análisis de información de archivo de la Base de Datos para analizar atributos asociados a la violencia contra la mujer en la Provincia de Ilo.

- Realización de los modelos predictivos bajo el lenguaje de Python.

3.5 Procedimientos

Los procedimientos de la presente investigación serán basados en la metodología CRISP-DM el cual según (Sngular, 2021) en el artículo de Azevedo y Santos llegaron a la conclusión de que dicha metodología es la más completa, ya que tomará en consideración el entorno de negocio de los resultados, así mismo se ha establecido que es la más utilizada el cual cuenta con 6 fases (Martinez-Plumed et al., 2021).

- Comprensión del Negocio
    - Establecer los objetivos de negocio.
    - Evaluar la situación actual.
- Comprensión de los Datos
    - Ejecutar procesos de captura de datos.
    - Realizar tareas de exploración de datos.
    - Gestionar la calidad de los datos.
- Preparación de los Datos
    - Realizar tareas de limpieza de datos.
    - Construir un juego de datos apto para ser usado.
- Modelado
    - Seleccionar las técnicas de modelado más adecuadas para nuestros datos y objetivos.
    - Fijar una estrategia de verificación de la calidad del modelo.
    - Construir un modelo a partir de la aplicación de las técnicas seleccionadas.

- Ajustar el modelo evaluando su fiabilidad y su impacto en los objetivos anteriormente establecidos.

- Evaluación
  - Evaluar el modelo o modelos generados hasta el momento.
  - Revisar todo el proceso que nos ha llevado hasta este punto.

- Despliegue
  - Diseñar un plan de despliegue de modelos y conocimiento sobre nuestra organización.
  - Realizar seguimiento y mantenimiento de la parte más operativa del despliegue.

3.6 Análisis de Datos

El análisis de los datos se realizará en base a la metodología, para lo cual se determinará las características con mayor influencia en la predicción de la violencia física contra la mujer, en base a ello se podrá obtener los resultados de cada modelo seleccionado previamente.

IV. Resultados y Discusión

- **Resultados**

Acto seguido se presenta los resultados de los procedimientos ejecutados en base a las fases de la metodología CRISP – DM.

4.1 Comprensión del Negocio

El objetivo de aplicación de Machine Learning es la evaluación de modelos predictivos para pronosticar la violencia física contra la mujer en la Provincia de Ilo, en base a ello se tomará los registros de las hojas HIS de los establecimientos de salud que pertenecen a la Red de Salud Ilo, esta entidad maneja un sistema de registro de las atenciones que se brinda a los pacientes en todos los establecimientos de salud de la provincia, la atención por la cual los pacientes acuden a estos establecimientos son distintas de acuerdo al malestar, para este caso nos vamos a enfocar únicamente en la violencia psicológica y física que está determinada por el Ítem R456 atendida en los establecimientos de salud de la Provincia de Ilo, los cuales son:

- P.S. LOS ANGELES
- P.S. EL ALGARROBAL
- P.S. VARADERO
- P.S. 18 DE MAYO
- P.S. PACOCHA
- C.S. PAMPA INALAMBRICA
- HOSPITAL ILO
- C.S. JHON F. KENNEDY

- C.S. ALTO ILO
- C.S. MIRAMAR
- C.S MENTAL COMUNITARIO ILO

Con los resultados de las predicciones se podrá ayudar en la determinación de la violencia física futura en la Provincia de Ilo

4.2 Comprensión de los Datos

Para la comprensión de los datos se debe realizar una serie de procesos para el análisis correcto de cada variable y como ello puede incurrir en el modelo a utilizar de machine learning.

- **Recolección de los Datos Inicial:**

  Para la recolección de datos fue adquirida en un solo archivo de Excel el cual fue convertido a tipo “CSV” para la lectura en Python como se muestra en la Figura 8

  Figura 8

  Datos no procesados del Ítem R456

```
[ ] df= pd.read_csv(Data)
    # Descripción general del conjunto de datos
    df.head()
```

	Id_Cita	Anio	Mes	Dia	Fecha_Atencion	Lote	Num_Pag	Num_Reg	Id_Ups	Descripcion_Ups	...	Hemoglobina	Perimetro_Abdominal	Perimetro_Cefalico	Cant_Hijos
0	444752085	2021	1	6	2021-01-06 00:00:00	D14	2	1	303802	PSICOLOGIA	...	NaN	97.0	NaN	2
1	445057623	2021	1	2	2021-01-02 00:00:00	C10	2	1	303802	PSICOLOGIA	...	NaN	NaN	NaN	2
2	445059270	2021	1	2	2021-01-02 00:00:00	C10	2	2	303802	PSICOLOGIA	...	NaN	NaN	NaN	1
3	445154178	2021	1	7	2021-01-07 00:00:00	D15	2	1	303802	PSICOLOGIA	...	NaN	81.0	NaN	2
4	445154175	2021	1	7	2021-01-07 00:00:00	D15	2	1	303802	PSICOLOGIA	...	NaN	81.0	NaN	2

5 rows × 80 columns

- **Descripción de los Datos:**

  En esta sección se realizó la examinación de los datos respecto a sus características principales, tipo de datos, cantidad de datos y variables presentes. La descripción de cada variable se podrá apreciar en la tabla 1.

Tabla 1

Descripción de data del Ítem R456

Item	Variable	Cantidad	Tipo de dato	Descripción
0	Id_Cita	1280	int64	Código único de cita
1	Anio	1280	int64	Año de atención
2	Mes	1280	int64	Mes de atención
3	Dia	1280	int64	Dia de atención
4	Fecha_Atencion	1280	object	Fecha completa de atención
5	Lote	1280	object	Lote de la hoja HIS
6	Num_Pag	1280	int64	Numero de página de hoja HIS
7	Num_Reg	1280	int64	Numero de registro de hoja HIS
8	Id_Ups	1280	int64	Código único de unidad productora de servicios
9	Descripcion_Ups	1280	object	Nombre de la unidad productora de servicios
10	Descripcion_Sector	1280	object	Sector regional
11	Descripcion_Disa	1280	object	Región
12	Descripcion_Red	1280	object	Provincia
13	Descripcion_MicroRed	1280	object	Ciudad
14	Codigo_Unico	1280	int64	Código único de establecimiento
15	Nombre_Establecimiento	1280	object	Nombre
16	Abrev_Tipo_Doc_Paciente	1280	object	Tipo de documento

17	Numero_Documento_Paciente	788	object	Documento de paciente
18	Estado_Civil	1280	object	Estado civil
19	Apellido_Paterno_Paciente	0	float64	Apellido paterno
20	Apellido_Materno_Paciente	0	float64	Apellido materno
21	Nombres_Paciente	0	float64	Nombre de paciente
22	Fecha_Nacimiento_Paciente	1280	object	Fecha de nacimiento
23	Genero	1280	object	Genero
24	Id_Etnia	1280	int64	Código único de etnia
25	Descripcion_Etnia	1280	object	Etnia
26	Historia_Clinica	0	float64	Código de historia clínica
27	Ficha_Familiar	0	float64	Ficha familiar
28	Id_Financiador	1280	int64	Código único de financiador
29	Descripcion_Financiador	1280	object	Tipo de financiador
30	Zona_Sector	1280	object	Sector de establecimiento
31	Descripcion_Pais	1280	object	País de origen
32	Estado_Migrante	1280	object	Migrante
33	Nivel_Educativo	1280	object	Nivel educativo
34	Abrev_Tipo_Doc_Personal	0	float64	Tipo de documento de personal
35	Numero_Documento_Personal	0	float64	Documento de personal
36	Apellido_Paterno_Personal	0	float64	Apellido paterno de personal
37	Apellido_Materno_Personal	0	float64	Apellido materno de personal
38	Nombres_Personal	0	float64	Nombre de personal

39	Fecha_Nacimiento_Personal	0	float64	Fecha de nacimiento de personal
40	Id_Condicion	0	float64	Código único de condición
41	Descripcion_Condicion	0	float64	Nombre de condición
42	Id_Profesion	0	float64	Profesión de personal
43	Descripcion_Profesion	0	float64	Descripción de profesión de personal
44	Id_Colegio	0	float64	Código de colegio
45	Descripcion_Colegio	0	float64	Nombre de colegio
46	Numero_Colegiatura	0	float64	Nombre de colegiatura
47	Abrev_Tipo_Doc_Registrador	0	float64	Tipo de documento de registrador
48	Numero_Documento_Registrador	0	float64	Numero de documento de registrador
49	Apellido_Paterno_Registrador	0	float64	Apellido paterno de registrador
50	Apellido_Materno_Registrador	0	float64	Apellido materno de registrador
51	Nombres_Registrador	0	float64	Nombre de registrador
52	Fecha_Nacimiento_Registrador	0	float64	Fecha de nacimiento de registrador
53	Id_Condicion_Establecimiento	1280	object	Código de establecimiento
54	Id_Condicion_Servicio	1280	object	Código de servicio
55	Edad_Reg	1280	int64	Edad de paciente
56	Tipo_Edad	1280	object	Tipo de edad de paciente
57	Anio_Actual_Paciente	1280	int64	Año actual de paciente

58	Mes_Actual_Paciente	1280	int64	Mes actual de paciente
59	Dia_Actual_Paciente	1280	int64	Dia actual de paciente
60	Estado_Laboral	1280	object	Estado laboral de paciente
61	Id_Turno	1280	object	Turno de atención
62	Codigo_Item	1280	object	Código de tipo de atención
63	Descripcion_Item	1280	object	Tipo de atención al paciente
64	Violencia_Fisica	1280	object	Diagnóstico de violencia física
65	Tipo_Diagnostico	1280	object	Tipo de diagnostico
66	Exp_ViolenciaF_Infancia	1280	object	Experiencia de violencia física en infancia
67	Valor_Lab	71	object	Valor de laboratorio
68	Peso	389	float64	Peso de paciente
69	Talla	389	float64	Talla de paciente
70	Hemoglobina	26	float64	Hemoglobina de paciente
71	Perimetro_Abdominal	306	float64	Perímetro abdominal de paciente
72	Perimetro_Cefalico	22	float64	Perímetro cefálico de paciente
73	Cant_Hijos	1280	int64	Cantidad de hijos del paciente
74	Descripcion_Otra_Condicion	0	float64	Descripción de otra condición
75	Fecha_Ultima_Regla	11	object	Fecha de ultima regla de paciente

76	Fecha_Solicitud_Hb	0	float64	Fecha de solicitud de hemoglobina de paciente
77	Fecha_Resultado_Hb	0	float64	Fecha de resultado de hemoglobina de paciente
78	Fecha_Registro	1280	object	Fecha de registro de hoja HIS
79	Fecha_Modificacion	230	object	Fecha de modificación de hoja HIS

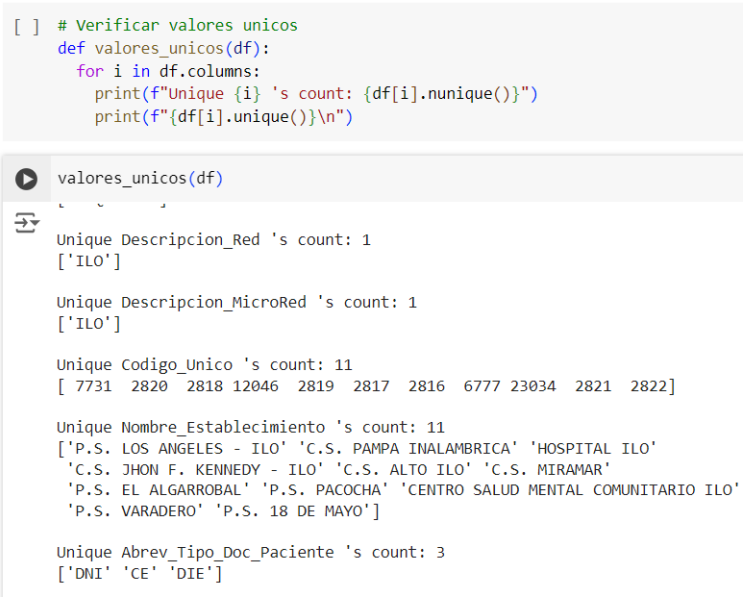
Nota. Data es propiedad de la Red de Salud Ilo.

- **Verificación de la Calidad de los Datos:**

En la verificación de la calidad de los datos se llegó a identificar datos faltantes, valores atípicos o datos inconsistentes de igual forma se identificó datos redundantes o con dependencia entre diferentes características una de otra, para ello se visualizó los datos únicos de cada columna con la función "Unique" como se muestra en la figura 9.

Figura 9

Visualización de datos únicos de cada columna



```
[ ]  # Verificar valores unicos
     def valores_unicos(df):
       for i in df.columns:
         print(f"Unique {i} 's count: {df[i].nunique()}")
         print(f"{df[i].unique()}\n")
```

```
valores_unicos(df)
```

```
Unique Descripcion_Red 's count: 1
['ILO']

Unique Descripcion_MicroRed 's count: 1
['ILO']

Unique Codigo_Unico 's count: 11
[ 7731  2820  2818 12046  2819  2817  2816  6777 23034  2821  2822]

Unique Nombre_Establecimiento 's count: 11
['P.S. LOS ANGELES - ILO' 'C.S. PAMPA INALAMBRICA' 'HOSPITAL ILO'
 'C.S. JHON F. KENNEDY - ILO' 'C.S. ALTO ILO' 'C.S. MIRAMAR'
 'P.S. EL ALGARROBAL' 'P.S. PACOCHA' 'CENTRO SALUD MENTAL COMUNITARIO ILO'
 'P.S. VARADERO' 'P.S. 18 DE MAYO']

Unique Abrev_Tipo_Doc_Paciente 's count: 3
['DNI' 'CE' 'DIE']
```

Nota. Información única de cada columna.

De igual forma, se realizó modificación a las variables y sus datos para una mejor visualización de la data, esto no alteró la data de forma brusca ya que solo se realizó cambio de espacios por guiones, reemplazo de abreviaciones por su nombre completo en variables categóricas para un mejor entendimiento de la data.

4.3 Preparación de los Datos

En este apartado se identificó datos faltantes, valores atípicos e inconsistencias, para ello se empezó descartando variables con datos de un valor único en todos los registros, variables de fechas y aquellos con pocos registros, se muestra código de proceso en la Figura 10.

Figura 10

Eliminación de variables según verificación de calidad de los datos

```
[ ]  # Borrar columnas con un valor unico, fechas y pocos registros

     df = df.drop(columns=['Descripcion_Sector',
     'Descripcion_Disa',
     'Descripcion_Red',
     'Descripcion_MicroRed',
     'Apellido_Paterno_Paciente',
     'Apellido_Materno_Paciente',
     'Nombres_Paciente',
     'Genero',
     'Historia_Clinica',
     'Ficha_Familiar',
     'Abrev_Tipo_Doc_Personal',
     'Numero_Documento_Personal',
     'Apellido_Paterno_Personal',
     'Apellido_Materno_Personal',
     'Nombres_Personal',
     'Fecha_Nacimiento_Personal',
     'Id_Condicion',
     'Descripcion_Condicion',
     'Id_Profesion',
     'Descripcion_Profesion',
     'Id_Colegio',
```

Nota. Se eliminó en total 44 variables inconsistentes.

A continuación, en esta sección se realizó la verificación de columnas con dependencia entre si lo cual podría afectar en el rendimiento del modelo. Según (Baru, 2019) la redundancia en los datos que están repitiendo la información puede provocar un sobreajuste en el modelo, se muestra en la figura 11 las variables redundantes que se eliminó.

Figura 11

Identificación de variables redundantes

```
[ ]  # Verificar columnas redundantes
     df = df.drop(columns=
     [
         'Id_Ups', # Redundante con la columna DESCRIPCION_UPS
         'Codigo_Unico', # Redundante con Nombre_Establecimiento y Zona_Sector
         'Nombre_Establecimiento', # Redundante con Zona_Sector
         'Abrev_Tipo_Doc_Paciente', # Redundante con Numero_Documento_Paciente
         'Id_Etnia', # Redundante con Descripcion_Etnia
         'Id_Financiador', # Redundante con Descripcion_Financiador
         'Codigo_Item', # Redundante con Descripcion_Item y Violencia_Fisica
         'Descripcion_Item', # Redundante con violencia_Fisica
         'Anio_Actual_Paciente', # Redundante con Edad_reg
         'Mes_Actual_Paciente',
         'Dia_Actual_Paciente'
     ])
```

Se inicio la transformación de datos realizando la creación de una nueva característica en base a la recurrencia de atención de un paciente por violencia psicológica o física esto a partir de la revisión de la columna de "Numero_Documento_Paciente" creando la nueva columna de "Numero Atenciones" con ello se elimina la columna de "Numero_Documento_Paciente" ya que al tener un numero único de paciente no es informativo para el modelo y brinda un riesgo de sobreajuste, esto al ser eliminado brinda mejores resultados al modelo.

De igual forma se ejecutó una técnica de preprocesamiento llamado Label Encoding para convertir variables con datos categóricos a numéricos para los casos que tienen datos como "SI" y "NO" se cambian por 1 y 0, en las variables categóricas que tienen más de dos características se aplica el método One-Hot Encoding el cual también ha permitido agrandar las columnas con las que se contaba, pasando de 22 columnas iniciales a 57 columnas después de aplicar dicho

método, se muestra la selección de variables que quedan en la data para su entrenamiento y prueba.

Item	Nombre Variable	Tipo de Dato	Cantidad de Datos
1	Anio	int64	832
2	Cant_Hijos	int64	832
3	Descripcion_Etnia_AIMARA	int64	832
4	Descripcion_Etnia_MESTIZO	int64	832
5	Descripcion_Etnia_OTRO	int64	832
6	Descripcion_Etnia_QUECHUAS	int64	832
7	Descripcion_Financiador_ESSALUD	int64	832
8	Descripcion_Financiador_EXONERADO	int64	832
9	Descripcion_Financiador_OTROS	int64	832
10	Descripcion_Financiador_S.I.S	int64	832
11	Descripcion_Financiador_USUARIO	int64	832
12	Descripcion_Pais_COLOMBIA	int64	832
13	Descripcion_Pais_PERU	int64	832
14	Descripcion_Pais_VENEZUELA	int64	832
15	Descripcion_Ups_ATENCION_EN_SALUD_FAMILIAR_Y_COMUNITARIA	int64	832
16	Descripcion_Ups_ENFERMERIA	int64	832
17	Descripcion_Ups_MEDICINA_GENERAL	int64	832
18	Descripcion_Ups_NUTRICION	int64	832
19	Descripcion_Ups_OBSTETRICIA	int64	832
20	Descripcion_Ups_PLANIFICACION_FAMILIAR	int64	832
21	Descripcion_Ups_PSICOLOGIA	int64	832
22	Descripcion_Ups_SERVICIOS_SOCIAL	int64	832
23	Edad_Reg	int64	832
24	Estado_Civil_CASADO	int64	832
25	Estado_Civil_DIVORCIADO	int64	832
26	Estado_Civil_SOLTERO	int64	832
27	Estado_Civil_VIUDO	int64	832
28	Estado_Laboral	int64	832
29	Estado_Migrante	int64	832
30	Exp_ViolenciaF_Infancia	int64	832
31	Id_Condicion_Establecimiento_CONTINUADOR	int64	832
32	Id_Condicion_Establecimiento_NUEVO	int64	832
33	Id_Condicion_Establecimiento_REINGRESO	int64	832
34	Id_Condicion_Servicio_CONTINUADOR	int64	832
35	Id_Condicion_Servicio_NUEVO	int64	832
36	Id_Condicion_Servicio_REINGRESO	int64	832
37	Id_Turno_MAÑANA	int64	832

38	Id_Turno_NOCHE	int64	832
39	Id_Turno_TARDE	int64	832
40	Nivel_Educativo_Educacion_Tecnica	int64	832
41	Nivel_Educativo_Educacion_Universitaria	int64	832
42	Nivel_Educativo_Estudios_Postgrado	int64	832
43	Nivel_Educativo_Nivel_Primaria	int64	832
44	Nivel_Educativo_Nivel_Secundaria	int64	832
45	Numero_Atenciones	int64	832
46	Tipo_Diagnostico_DEFINITIVO	int64	832
47	Tipo_Diagnostico_PRESUNTIVO	int64	832
48	Tipo_Diagnostico_REPETIDO	int64	832
49	Violencia_Fisica	int64	832
50	Zona_Sector_18_DE_MAYO	int64	832
51	Zona_Sector_ALTO_ILO	int64	832
52	Zona_Sector_EL_ALGARROBAL	int64	832
53	Zona_Sector_JHON_F._KENNEDY	int64	832
54	Zona_Sector_LOS_ANGELES	int64	832
55	Zona_Sector_MIRAMAR	int64	832
56	Zona_Sector_PACOCHA	int64	832
57	Zona_Sector_PAMPA_INALAMBRICA	int64	832
58	Zona_Sector_VARADERO	int64	832

Una vez realizada la transformación de las variables se identifica la cantidad tanto de pacientes que sufrieron violencia física como los pacientes que no sufrieron violencia física este último detectado por pacientes que solo han sufrido violencia psicológica, en la figura 12 se muestra la cantidad por cada tipo.

Figura 12

Muestra de cantidad en variable dependiente

```
[ ]  print(df_agrupado['Violencia_Fisica'].value_counts())

Violencia_Fisica
1    449
0    383
Name: count, dtype: int64
```

Nota. Cantidad total de cada clase.

Se realizó la división del conjunto de datos para entrenamiento y prueba, validación. De la data se dividió 80% para entrenamiento y 20% para prueba, esta división se mantuvo para todos los modelos utilizados.

Con el conjunto de datos de entrenamiento se puede enseñar al modelo de Machine Learning a relacionar e identificar patrones para con ello realizar una predicción en datos similares a los que se brinda en el entrenamiento, esto puede permitir que al momento de evaluar el rendimiento del modelo en los datos de prueba sea más preciso y brinde una predicción optima en valores reales fuera del conjunto de entrenamiento.

Según la figura 5 colocada anteriormente estaba presentado clases desbalanceadas para ello en el momento de realizar la división de entrenamiento y prueba se aplicó el método “Smote” para balancear las clases generando datos sintéticos para la clase minoritaria esto solamente para los datos de entrenamiento, al utilizar la técnica de “Smote” que está basado en Oversampling se busca balancear los datos de entrenamiento aumentando data sintética en la clase minoritaria esto con el fin de mantener toda la información posible considerando que nuestros datos en general no son de proporción mayor, al utilizar alguna técnica basado en undersampling se perdería información lo cual podría afectar en la capacidad del modelo para generalizar, normalmente se utilizaría en datos grandes para disminuir el tiempo de entrenamiento, en nuestro caso no se contiene datos grandes y se considera mas viable utilizar el método “Smote” basado en oversampling.

Con datos balanceados en cada clase se podrá mejorar la capacidad de generalización, mejorar las métricas de rendimiento y evitar el sobreajuste, en la

figura 13 se muestra el proceso de aplicación del método Smote en los datos de entrenamiento.

Figura 13

Aplicación de Smote en datos de entrenamiento

```
[ ] from sklearn.model_selection import train_test_split

    X1 = df_agrupado.drop('Violencia_Fisica', axis=1)
    X = X1.values
    y = df_agrupado['Violencia_Fisica'].values

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, stratify=y, random_state=42)
```

```
[ ] # Inicializar SMOTE
    smote = SMOTE(random_state=42)

    # Aplicar SMOTE al conjunto de entrenamiento
    X_train_smote, y_train_smote = smote.fit_resample(X_train, y_train)
```

```
[ ] print("Distribución de clases en y:", np.bincount(y))
    print("Distribución de clases en y_train:", np.bincount(y_train_smote))
    print("Distribución de clases en y_test:", np.bincount(y_test))
```

```
Distribución de clases en y: [383 449]
Distribución de clases en y_train: [359 359]
Distribución de clases en y_test: [77 90]
```

Nota. Aplicación del método Smote después de la división de entrenamiento y prueba.

4.4 Modelado

a. Selección de los Algoritmos de Machine Learning

Tras una revisión detallada de antecedentes referente al tema de investigación, se seleccionó 4 algoritmos que demostraron un desempeño superior en múltiples artículos académicos revisados.

Se seleccionaron los siguientes algoritmos:

- **K Nearest Neighbors:** Este algoritmo fue destacado en los artículos de Tamilarasi & Rani, 2020 y Cumbicus-Pineda et al., 2021 en los cuales mostró resultados más óptimos a comparación de los demás algoritmos.
- **Random Forest:** Según Saboya et al., 2019 y Rodríguez-Rodríguez et al., 2020, este algoritmo mostró ser particularmente más eficiente y preciso a comparación de los demás algoritmos utilizados en ambos artículos.
- **Naive Bayes:** Este algoritmo fue elegido a comparación de otros por su mayor precisión en los resultados obtenidos, como se evidenció en el trabajo de Al-Taleb & Saeed, 2019.
- **Decision Trees:** Finalmente este algoritmo fue considerado por su gran desempeño tal como se verifica en la investigación de Timaran-Pereira et al., 2017 y .Amusa et al., 2020.

La selección de estos 4 algoritmos se justifica por su desempeño destacado en las evaluaciones y su alineación con los objetivos específicos del presente trabajo de investigación.

b. Selección de Métricas de Evaluación

Las métricas de evaluación a utilizar son 4 ya que se busca evaluar el rendimiento de cada modelo con los conjuntos de prueba, estas fueron definidas en una función para ser llamadas en cada modelo a utilizar de igual forma con un gráfico representativo, las métricas a utilizar son las siguientes.

- **Matriz de Confusión**

Figura 14

Función de la Matriz de Confusión

```
[ ] # Función que muestra la matriz de confusión para un clasificador y el accuracy en entrenamiento y prueba
    def confusion_matrix_plot(X_train_smote, y_train_smote, X_test, y_test, classifier, y_pred, classifier_name):
        # Calcular la matriz de confusión sin normalizar
        cm = confusion_matrix(y_test, y_pred)

        # Crear una visualización de la matriz de confusión sin normalizar

        disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=["No Violencia_Fisica", "Violencia_Fisica"])
        disp.plot(cmap=plt.cm.Blues)
        plt.title(f'{classifier_name} - Confusion Matrix')
        plt.show()

        # Calcular la matriz de confusión normalizada
        cm_normalized = confusion_matrix(y_test, y_pred, normalize='true')

        # Crear una visualización de la matriz de confusión normalizada
        print("")
        disp_normalized = ConfusionMatrixDisplay(confusion_matrix=cm_normalized, display_labels=["No Violencia_Fisica", "Violencia_Fisica"])
        disp_normalized.plot(cmap=plt.cm.Blues)
        plt.title(f'{classifier_name} - Confusion Matrix (norm.)')
        plt.show()

        # Imprimir las métricas de precisión (accuracy)
        print(f'Accuracy Score Test: {accuracy_score(y_test, y_pred)}')
        print(f'Accuracy Score Train: {classifier.score(X_train_smote, y_train_smote)} (as comparison)')
        return print("")
```

Nota. Matriz de Confusión en conjunto con Accuracy

- **Curva ROC**

Figura 15

Función de Curva ROC

```
[ ]  # Función que muestra la curva ROC y su AUC
     def roc_curve_auc_score(X_test, y_test, y_pred_probabilities, classifier_name):

         y_pred_prob = y_pred_probabilities[:,1]
         fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)

         plt.plot([0, 1], [0, 1], 'k--')
         plt.plot(fpr, tpr, label=f'{classifier_name}')
         plt.xlabel('False Positive Rate')
         plt.ylabel('True Positive Rate')
         plt.title(f'{classifier_name} - ROC Curve')
         plt.show()

         return print(f'AUC Score (ROC): {roc_auc_score(y_test, y_pred_prob)}\n')
```

Nota. Curva ROC y su AUC

- **F1 Score**

Figura 16

Función de F1 Score

```
[ ]  # Muestra la curva precision-recall, el F1 score y AUC score
     def precision_recall_curve_and_scores(X_test, y_test, y_pred, y_pred_probabilities, classifier_name):

         y_pred_prob = y_pred_probabilities[:,1]
         precision, recall, thresholds = precision_recall_curve(y_test, y_pred_prob)

         plt.plot(recall, precision, label=f'{classifier_name}')
         plt.xlabel('Recall')
         plt.ylabel('Precision')
         plt.title(f'{classifier_name} - Precision-Recall Curve')
         plt.show()

         f1_score_result, auc_score_result = f1_score(y_test, y_pred), auc(recall, precision)

         return print(f'F1 Score: {f1_score_result} \nAUC Score (PR): {auc_score_result}\n')
```

Nota. F1 Score en conjunto con AUC Score

- **MCC**

Figura 17

Función de MCC

```
[ ]  # MCC
     def matthews_corrcoef_score(y_test, y_pred):
       y_true = y_test
       y_pred = y_pred
       mcc = matthews_corrcoef(y_true, y_pred)
       unique_labels = np.unique(y_true)
       true_counts = [sum(y_true == label) for label in unique_labels]
       pred_counts = [sum(y_pred == label) for label in unique_labels]

       plt.figure(figsize=(8, 6))
       plt.bar(unique_labels, true_counts, width=0.4, label='Etiquetas Verdaderas')
       plt.bar(unique_labels + 0.4, pred_counts, width=0.4, label='Predicciones')
       plt.xlabel('Clases')
       plt.ylabel('Cantidad')
       plt.title('Comparación de Etiquetas Verdaderas vs. Predicciones por Clase')
       plt.xticks(unique_labels)
       plt.legend()
       plt.grid(True)
       plt.show()
       print("Matthews Correlation Coefficient (MCC):")
       print(mcc)
```

Nota. MCC en representación con su grafica.

c. Set de Prueba

Los datos de entrenamiento y prueba fueron definidos en la fase de preparación de datos aplicando la técnica de Smote para crear data sintética en el

conjunto de entrenamiento sin alterar los datos de prueba lo cual se puede verificar en la figura 13.

d. Creación del Modelo

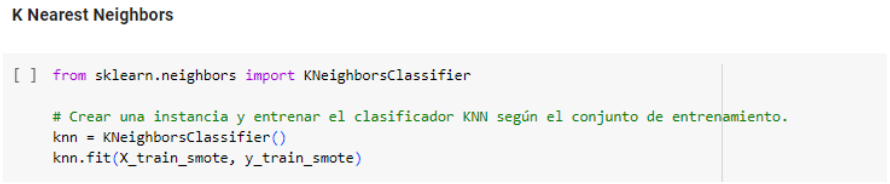
a. Algoritmo de K-Nearest Neighbors

i. Construcción del Modelo

En la construcción del modelo se entrenó los datos utilizando la librería Sklearn y los módulos para el algoritmo de K Nearest Neighbors como se puede observar en la figura 18.

Figura 18

Código del Algoritmo KNN



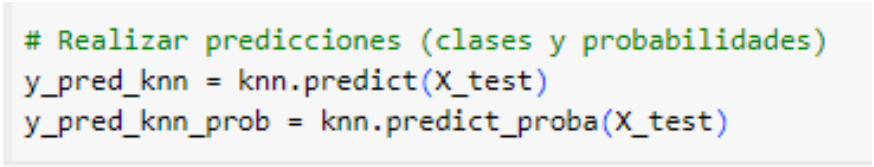
Nota. Se personalizó el algoritmo KNN

ii. Predicción del Algoritmo

Se realizó las predicciones del algoritmo con el modelo entrenado en el conjunto de prueba.

Figura 19

Predicción de Probabilidades KNN



iii. Evaluación del Modelo

Para la evaluación del modelo se llamó a las funciones de cada métrica a utilizar y así interpretar los resultados del modelo empleado.

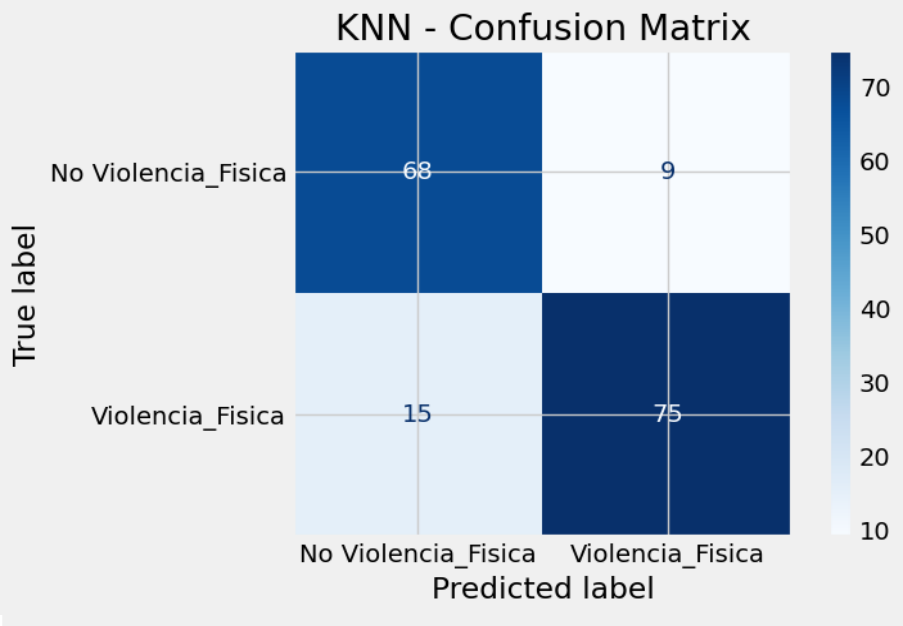
Figura 20

Ejecutar Función de Métricas para KNN

```
[ ] # Grafico de evaluaciones
    confusion_matrix_plot(X_train_smote, y_train_smote, X_test, y_test, knn, y_pred_knn, 'KNN')
    roc_curve_auc_score(X_test, y_test, y_pred_knn_prob, 'KNN')
    precision_recall_curve_and_scores(X_test, y_test, y_pred_knn, y_pred_knn_prob, 'KNN')
    matthews_corrcoef_score(y_test, y_pred_knn)
```

Figura 21

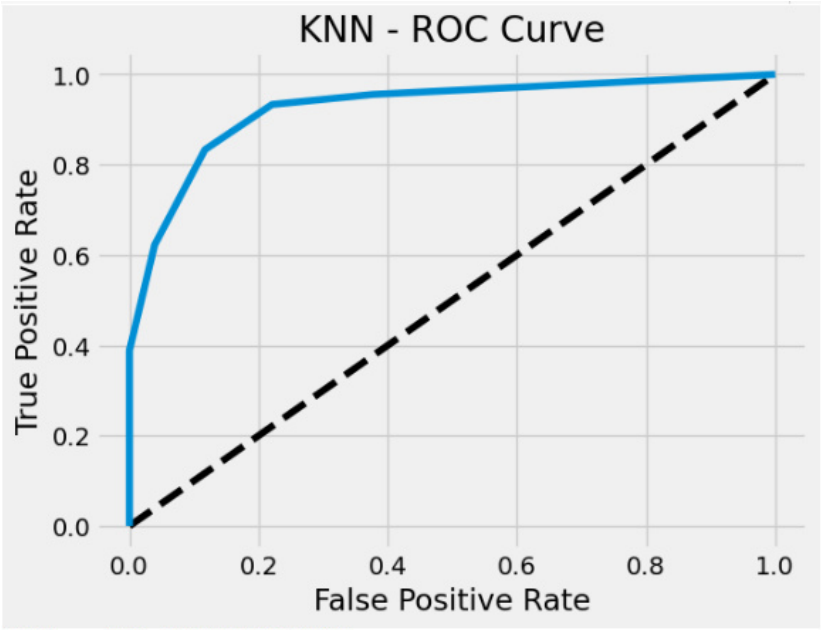
Matriz de Confusión KNN




```
Accuracy Score Test: 0.8562874251497006
Accuracy Score Train: 0.924791086350975 (as comparison)
```

Figura 23

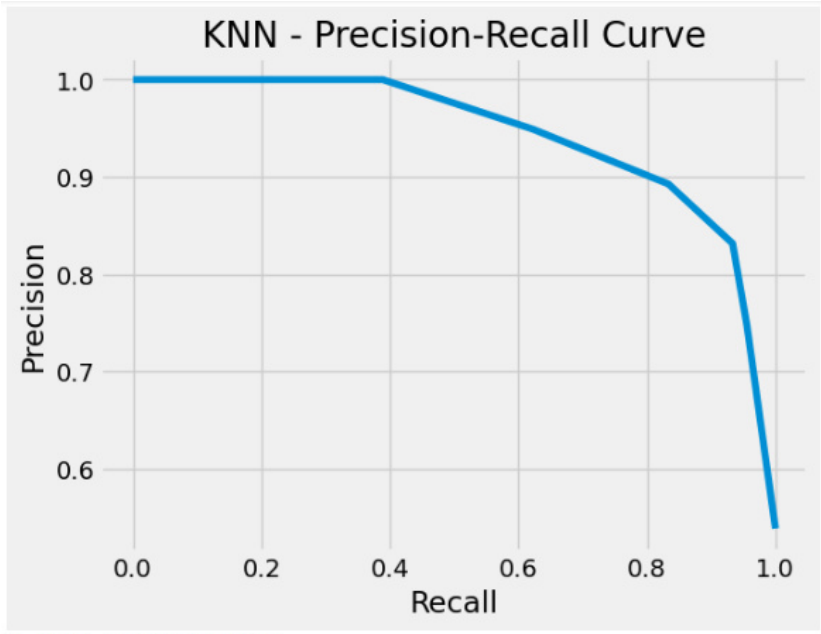
Curva ROC KNN



AUC Score (ROC): 0.9248917748917749

Figura 22

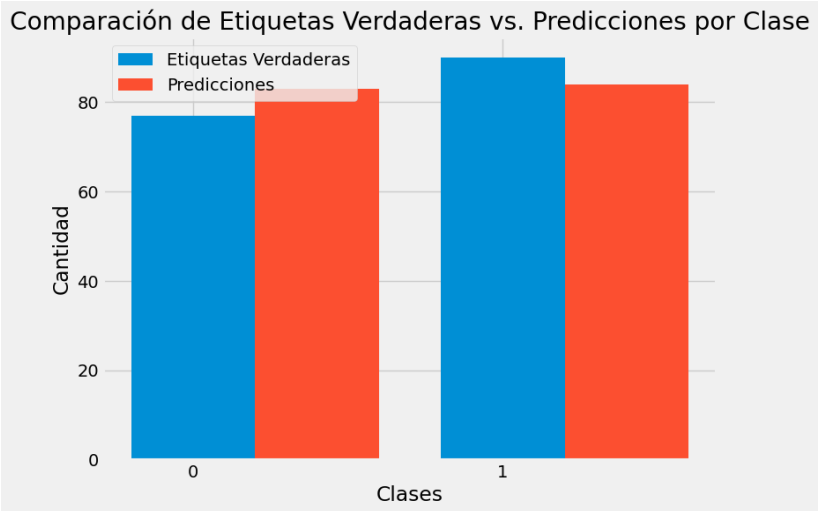
F1 Score KNN



F1 Score: 0.8620689655172413
AUC Score (PR): 0.9430958982827446

Figura 24

Gráfico MCC de KNN




```
Matthews Correlation Coefficient (MCC):
0.7142889740706924
```

Tabla 2

Resumen Evaluación KNN

Resumen	K Nearest Neighbors
Falso Positivos	9
Falso Negativos	15
Accuracy	0.8562874251497006
Curva ROC	0.9248917748917749
F1 Score	0.8620689655172413
MCC	0.7142889740706924

Según se verifica en la aplicación de las métricas de evaluación, se obtuvo en la matriz de confusión 15 falsos negativos y 9 falsos positivos, se realizará ajuste de Hiperparametros para bajar estos valores en la matriz de confusión, se observa que en las demás métricas se obtiene un valor óptimo

considerando los objetivos de la investigación de igual forma se realizará las verificaciones para mejorar el modelo y obtener mejores resultados.

b. Algoritmo Random Forest

i. Construcción del Modelo

Para la construcción del modelo se utilizó la Liberia de Sklearn y los módulos de Random Forest como se puede visualizar en la figura 25

Figura 25

Código de Random Forest

```
[ ]  from sklearn.ensemble import RandomForestClassifier

     # Crear una instancia y entrenar el modelo de RF según el conjunto de entrenamiento.
     rf = RandomForestClassifier()
     rf.fit(X_train_smote, y_train_smote)
```

Nota. Se personalizó algoritmo de RF

ii. Predicción del Algoritmo

Se realizó las predicciones del algoritmo con el modelo entrenado en el conjunto de prueba.

Figura 26

Predicción de Probabilidades RF

```
# Realizar predicciones (clases y probabilidades)
y_pred_rf = rf.predict(X_test)
y_pred_rf_prob = rf.predict_proba(X_test)
```

iii. Evaluación del Modelo

Para la evaluación del modelo se llamó a las funciones de cada métrica a utilizar y así interpretar los resultados del modelo empleado.

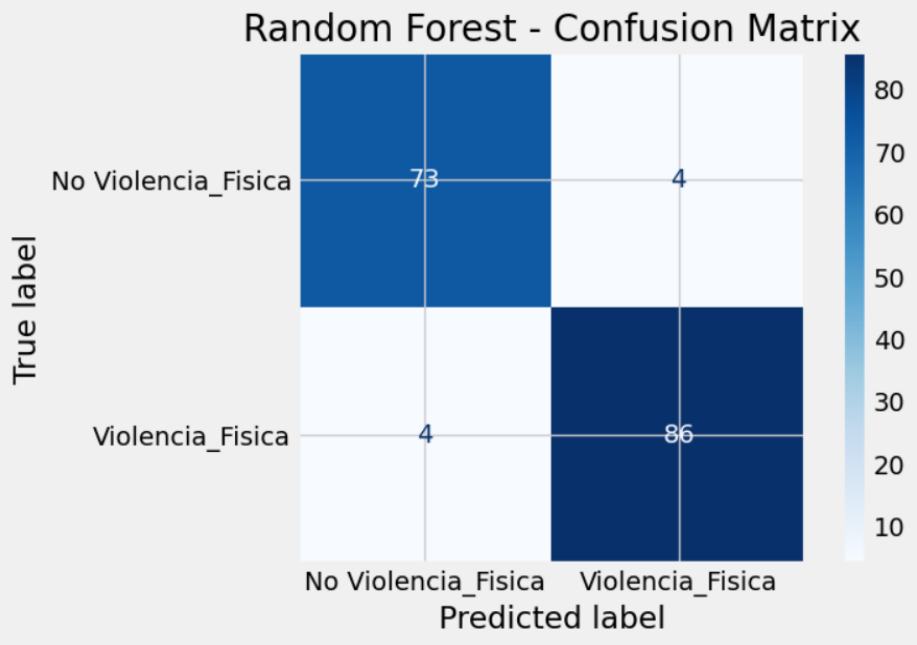
Figura 28

Ejecución de Función de Métricas para RF

```
# Grafico de evaluaciones
confusion_matrix_plot(X_train_smote, y_train_smote, X_test, y_test, rf, y_pred_rf, 'Random Forest')
roc_curve_auc_score(X_test, y_test, y_pred_rf_prob, 'Random Forest')
precision_recall_curve_and_scores(X_test, y_test, y_pred_rf, y_pred_rf_prob, 'Random Forest')
matthews_corrcoef_score(y_test, y_pred_rf)
```

Figura 27

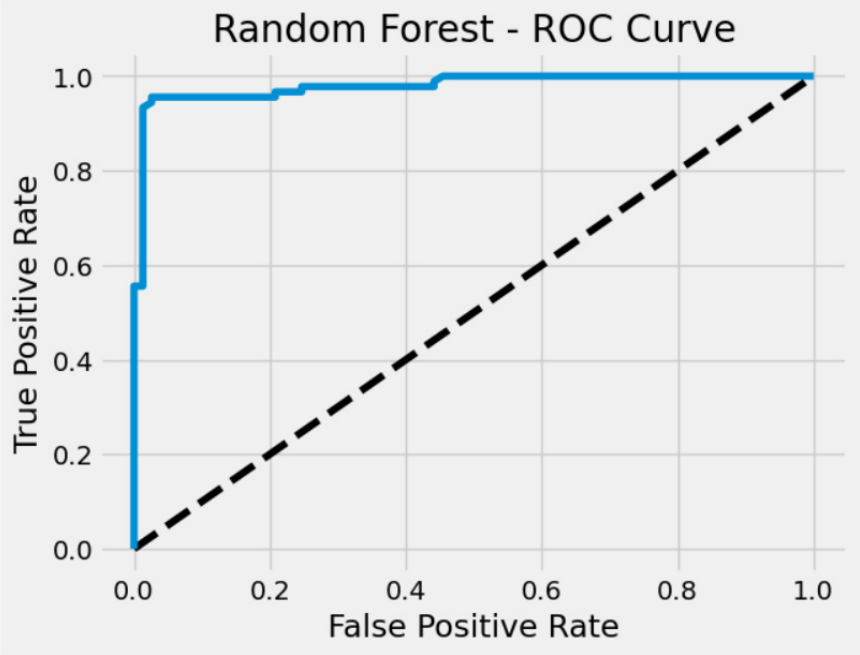
Matriz de Confusión RF




```
Accuracy Score Test: 0.9520958083832335
Accuracy Score Train: 0.9972144846796658 (as comparison)
```

Figura 30

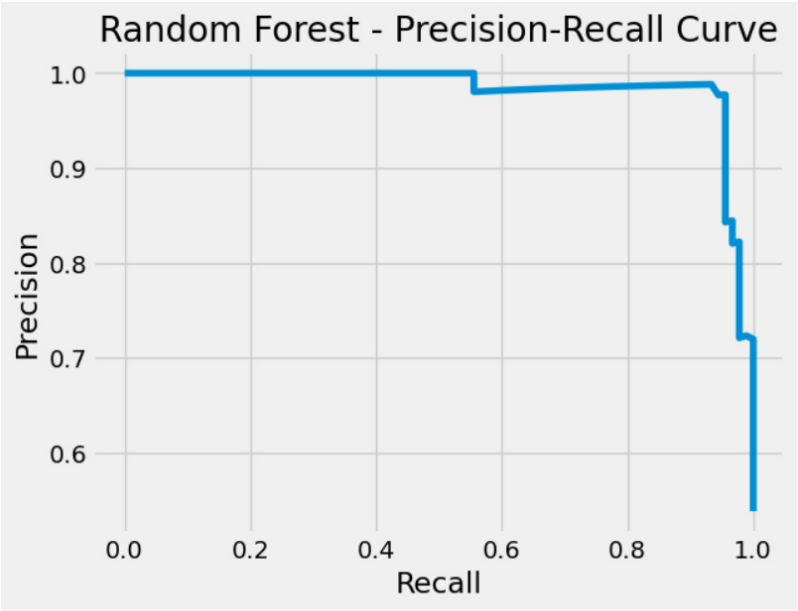
Curva ROC RF



AUC Score (ROC): 0.9796536796536797

Figura 29

F1 Score RF



F1 Score: 0.9555555555555556
AUC Score (PR): 0.9839815725483725

Figura 31

Gráfico MCC de RF

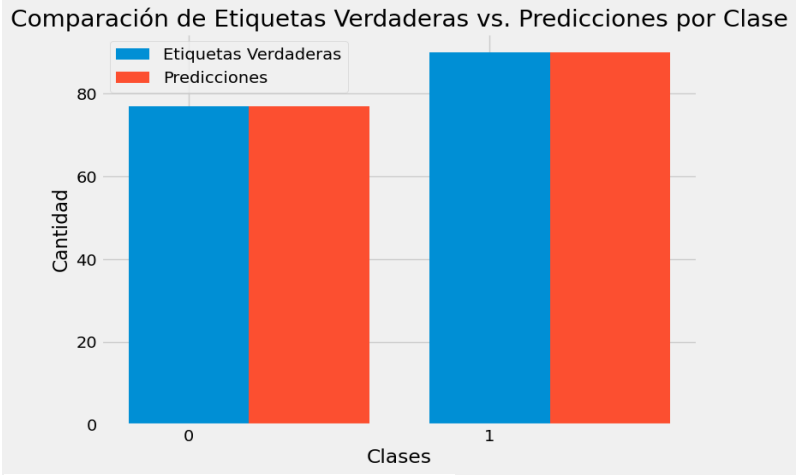


Tabla 3

Resumen de Evaluación RF

Resumen	Random Forest
Falso Positivos	4
Falso Negativos	4
Accuracy	0.9520958083832335
Curva ROC	0.9796536796536797
F1 Score	0.9555555555555556
MCC	0.9036075036075036

Se realiza el análisis de los resultados obtenidos, a pesar de tener más del 90% en los resultados de las métricas esto se evalúa para realizar unos ajustes del modelo y obtener mejores resultados en la matriz de confusión.

c. Algoritmo Naive Bayes

i. Construcción del Modelo

Como se especificó en los algoritmos anteriores se utilizará la librería de Sklearn y el módulo de Naive Bayes el cual se muestra en la figura 32.

Figura 32

Código del Algoritmo NB

```
[ ]  from sklearn.naive_bayes import BernoulliNB

     # Crear el modelo Bernoulli Naive Bayes
     nb = BernoulliNB()

     # Entrenar el modelo
     nb.fit(X_train_smote, y_train_smote)
```

Nota. Se personalizó algoritmo de NB

ii. Predicción del Algoritmo

Se realizó las predicciones del algoritmo con el modelo entrenado en el conjunto de prueba.

Figura 33

Predicción de Probabilidades NB

```
# Hacer predicciones
y_pred_nb = nb.predict(X_test)
y_pred_nb_prob = nb.predict_proba(X_test)
```

iii. Evaluación del Modelo

Para la evaluación del modelo se llamó a las funciones de cada métrica a utilizar y así interpretar los resultados del modelo empleado.

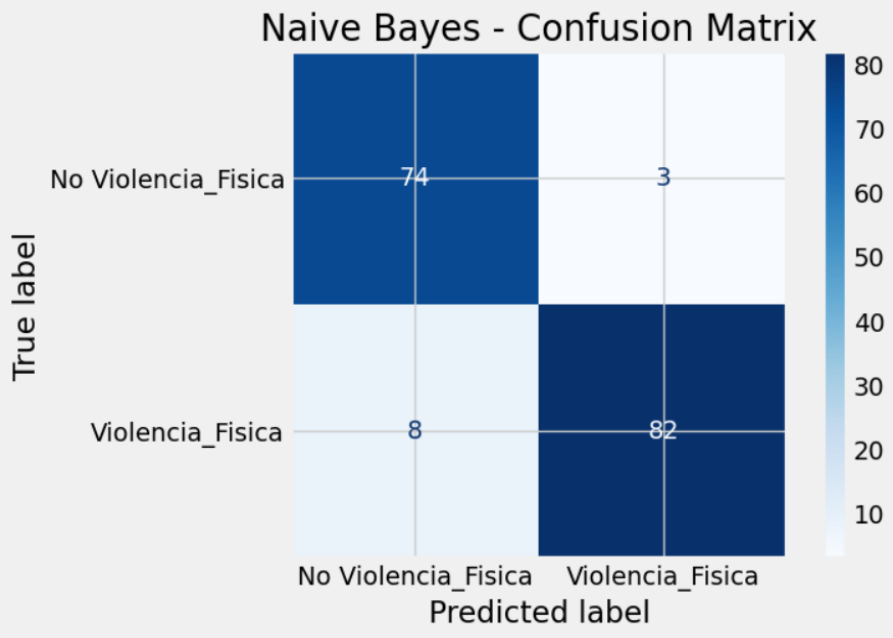
Figura 35

Ejecución de las métricas para NB

```
# Grafico de evaluaciones
confusion_matrix_plot(X_train_smote, y_train_smote, X_test, y_test, nb, y_pred_nb, 'Naive Bayes')
roc_curve_auc_score(X_test, y_test, y_pred_nb_prob, 'Naive Bayes')
precision_recall_curve_and_scores(X_test, y_test, y_pred_nb, y_pred_nb_prob, 'Naive Bayes')
matthews_corrcoef_score(y_test, y_pred_nb)
```

Figura 34

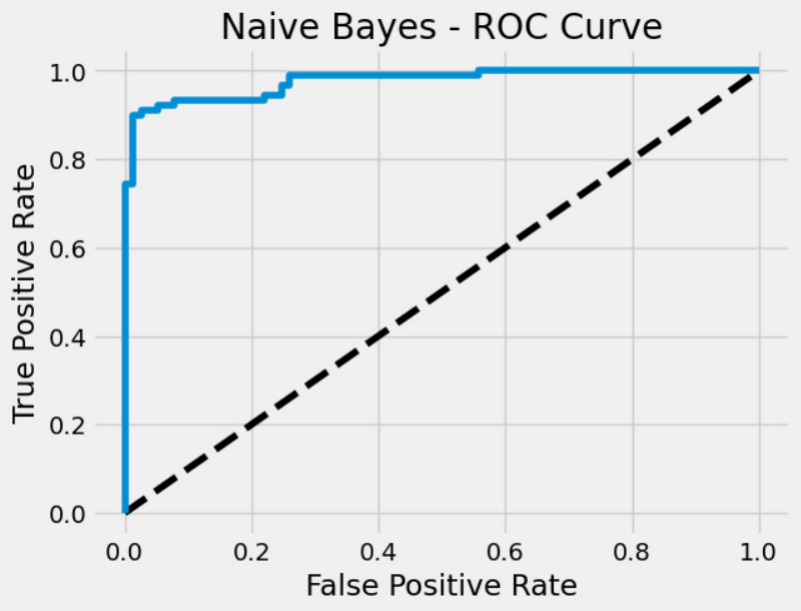
Matriz de Confusión NB




```
Accuracy Score Test: 0.9341317365269461
Accuracy Score Train: 0.9317548746518106 (as comparison)
```

Figura 37

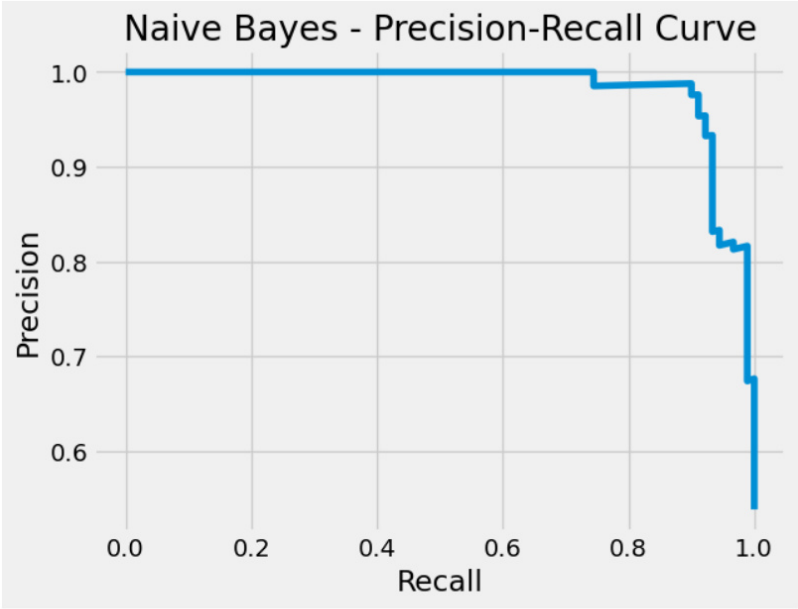
Curva ROC NB




```
AUC Score (ROC): 0.9763347763347763
```

Figura 36

F1 Score NB




```
F1 Score: 0.9371428571428572
AUC Score (PR): 0.9827912473296043
```

Figura 38

Gráfico MCC de NB

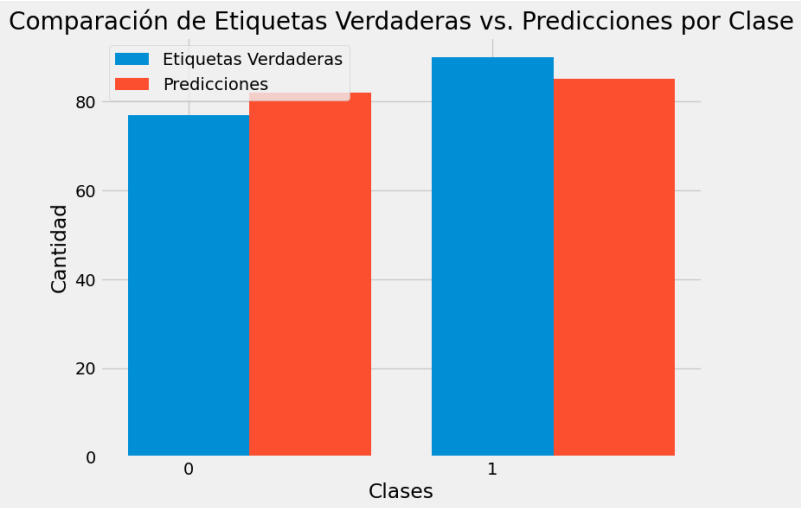


Tabla 4

Resumen de Evaluación NB

Resumen	Naive Bayes
Falso Positivos	3
Falso Negativos	8
Accuracy	0.9341317365269461
Curva ROC	0.9763347763347763
F1 Score	0.9371428571428572
MCC	0.8696438886007516

Se realiza el análisis de los resultados de las métricas ejecutadas para establecer ajustes en el modelo y obtener mejores resultados.

d. Algoritmo Decision Trees

i. Construcción del Modelo

Para la construcción del modelo se utilizó la librería Sklearn y el módulo para el algoritmo de Decision Trees como se puede observar en la figura 39

Figura 39

Código del Algoritmo DT

```
[ ]  from sklearn.tree import DecisionTreeClassifier

     # Crear y entrenar el modelo de Decision Trees
     dt = DecisionTreeClassifier(random_state=42)
     dt.fit(X_train_smote, y_train_smote)
```

Nota. Se personalizó el algoritmo DT

ii. Predicción del Algoritmo

Se realizó las predicciones del algoritmo con el modelo entrenado en el conjunto de prueba.

Figura 40

Predicción de Probabilidades DT

```
# Realizar predicciones (clases y probabilidades) con el modelo entrenado en el conjunto de prueba.
y_pred_dt = dt.predict(X_test)
y_pred_dt_prob = dt.predict_proba(X_test)
```

iii. Evaluación del Modelo

Se realiza la ejecución de las métricas declaradas anteriormente, con ello se realiza la evaluación de los resultados obtenidos.

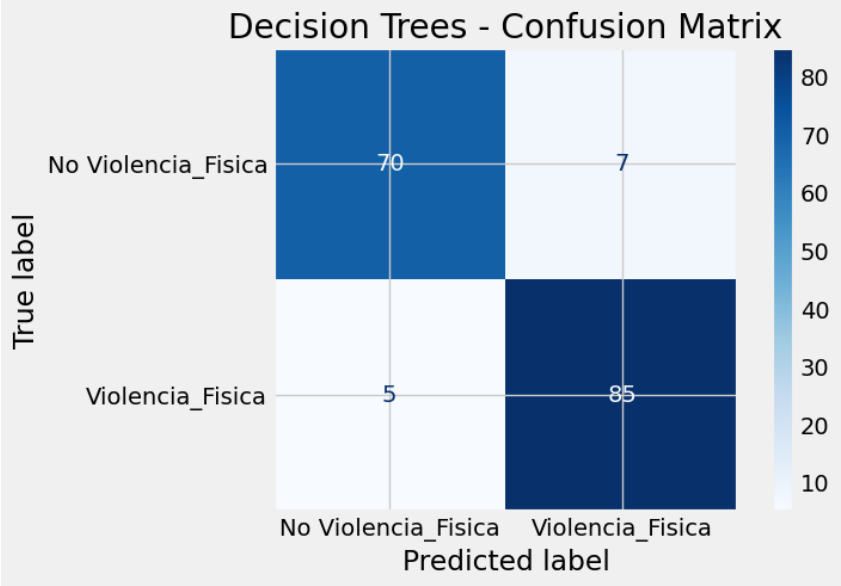
Figura 42

Ejecutar Función de Métricas DT

```
# Grafico de evaluaciones
confusion_matrix_plot(X_train_smote, y_train_smote, X_test, y_test, dt, y_pred_dt, 'Decision Trees')
roc_curve_auc_score(X_test, y_test, y_pred_dt_prob, 'Decision Trees')
precision_recall_curve_and_scores(X_test, y_test, y_pred_dt, y_pred_dt_prob, 'Decision Trees')
matthews_corrcoef_score(y_test, y_pred_dt)
```

Figura 41

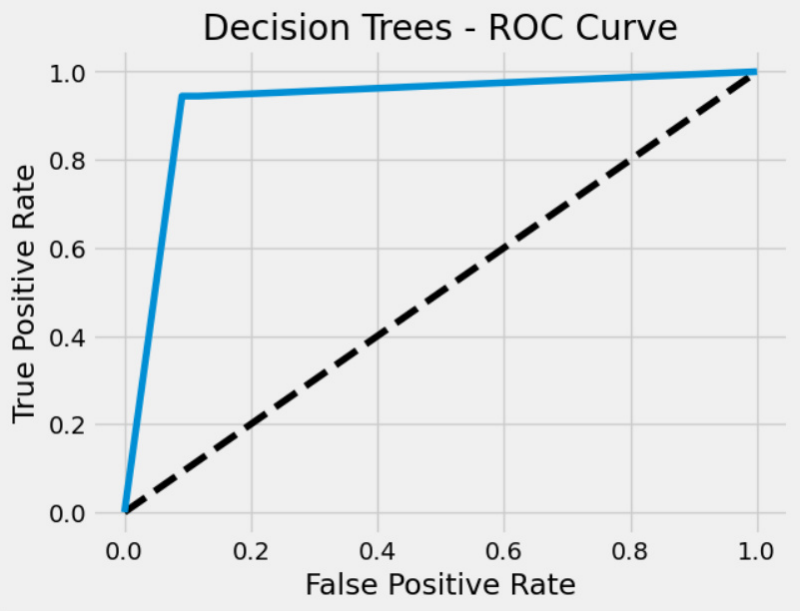
Matriz de Confusión DT




```
Accuracy Score Test: 0.9281437125748503
Accuracy Score Train: 0.9972144846796658 (as comparison)
```

Figura 44

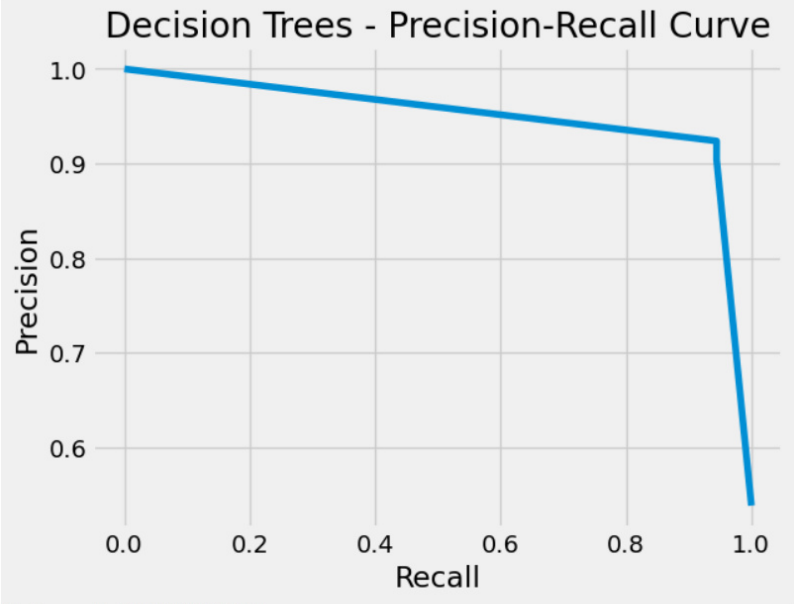
Curva ROC DT



AUC Score (ROC): 0.926046176046176

Figura 43

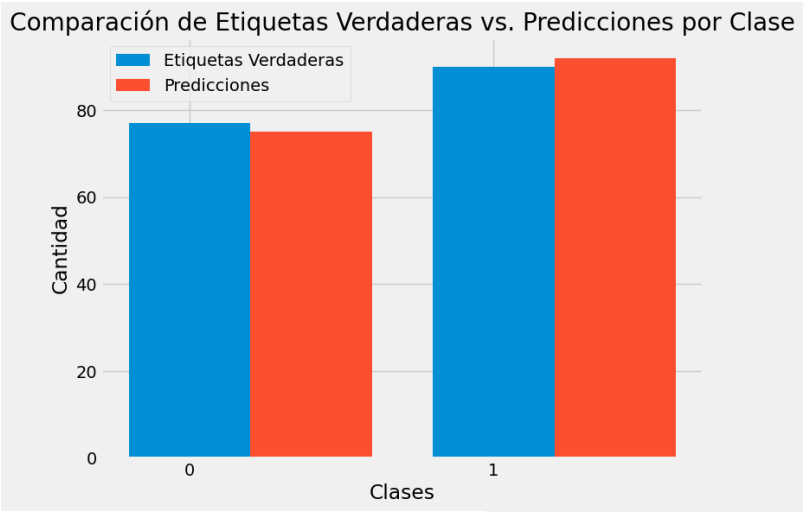
F1 Score DT



F1 Score: 0.9340659340659341
AUC Score (PR): 0.9486027559435553

Figura 45

Gráfico MCC DT




```
Matthews Correlation Coefficient (MCC):
0.8553888526855987
```

Tabla 5

Resumen de Evaluación DT

Resumen	Decision Trees
Falso Positivos	7
Falso Negativos	5
Accuracy	0.9281437125748503
Curva ROC	0.926046176046176
F1 Score	0.9340659340659341
MCC	0.8553888526855987

Se realiza el análisis de los resultados de las métricas ejecutadas para establecer ajustes en el modelo y obtener mejores resultados en la siguiente sección.

e. Ajustes de Parámetros (Hiperparametros)

En esta sección se realizará el ajuste de Hiperparametros a las técnicas de Machine Learning ya seleccionadas, como se verificó en la sección anterior se ejecutó cada modelo con la forma predeterminada del algoritmo, con los ajustes de Hiperparametros nos permitirá adecuar cantidades precisas para el entrenamiento del modelo, cada algoritmo cuenta con ajustes específicos que ayudarán a optimizar el rendimiento del modelo, adicionando a ello se utilizará el método de validación cruzada el cual nos permitirá obtener una estimación más precisa.

a. Algoritmo K-Nearest Neighbors Optimizado

i. Ajuste de Hiperparametros en el Modelo

Se realiza el ajuste de Hiperparametros para el algoritmo de K Nearest Neighbors.

Figura 46

Ajuste de Hiperparametros KNN

```
# Definir el pipeline
pipeline = Pipeline([
    ('scaler', StandardScaler()),  # Escalado de características
    ('knn', KNeighborsClassifier())
])

#Define parameter grid for GridSearch.
param_grid = {
    'knn__n_neighbors': np.arange(1, 30),
    'knn__weights': ['uniform', 'distance'],
    'knn__metric': ['euclidean', 'manhattan', 'minkowski']
}

# Configurar la validación cruzada estratificada
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Instanciar GridSearchCV
knn_cv = GridSearchCV(
    estimator=pipeline,
    param_grid=param_grid,
    cv=cv,
    n_jobs=-1,
    verbose=2
)

knn_cv.fit(X_train_smote, y_train_smote)
```

Los ajustes realizados para el modelo de KNN son los siguientes:

Pipeline: Es definido para escalar automáticamente las características y en base a ello aplicar el modelo de KNN en un solo paso, de esta forma los datos son procesados de forma óptima antes de hacer predicciones.

knn__n_neighbors: Se especifica el número de vecinos que el algoritmo va a considerar.

knn__weights: Es el peso que se asigna a los vecinos, que todos tengan el mismo peso y considerar a los más cercanos con más peso que los lejanos.

knn__metric: Define la métrica de distancia a utilizar para calcular la cercanía en cada punto.

ii. Predicción del Modelo

Se ejecuta la predicción del modelo entrenado a los datos de prueba.

Figura 47

Ejecución de Predicción KNN optimizado

```
# Predecir las probabilidades en el conjunto de prueba
y_pred_knn_tuned_prob = knn_cv.predict_proba(X_test)
threshold = 0.4 # Establecer un umbral más bajo para la clase positiva
y_pred_knn_tuned_threshold = (y_pred_knn_tuned_prob[:, 1] >= threshold).astype(int)

# Evaluar resultados
print('KNN best parameters:', knn_cv.best_params_)
```

```
Fitting 5 folds for each of 174 candidates, totalling 870 fits
KNN best parameters: {'knn__metric': 'euclidean', 'knn__n_neighbors': 11, 'knn__weights': 'distance'}
```

iii. Evaluación del Modelo

Se ejecutó las métricas predefinidas al modelo K Nearest Neighbors optimizado.

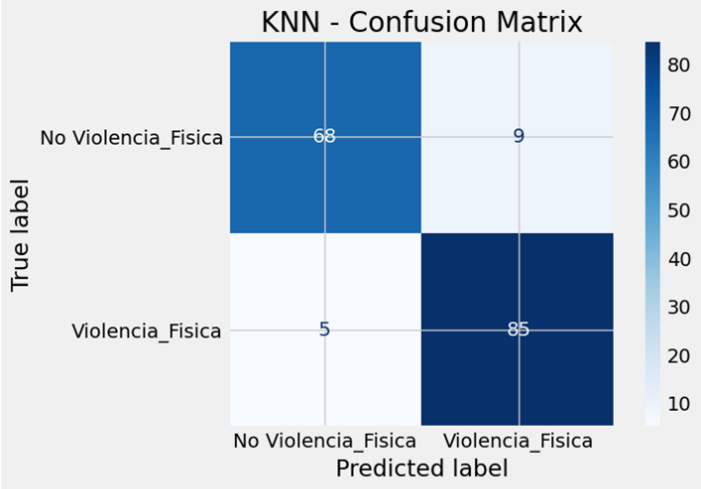
Figura 49

Evaluación de KNN Optimizado

```
# Grafico de evaluaciones
confusion_matrix_plot(X_train_smote, y_train_smote, X_test, y_test, knn_cv, y_pred_knn_tuned_threshold, 'KNN')
roc_curve_auc_score(X_test, y_test, y_pred_knn_tuned_prob, 'KNN')
precision_recall_curve_and_scores(X_test, y_test, y_pred_knn_tuned_threshold, y_pred_knn_tuned_threshold, 'KNN')
matthews_corrcoef_score(y_test, y_pred_knn_tuned_threshold)
```

Figura 48

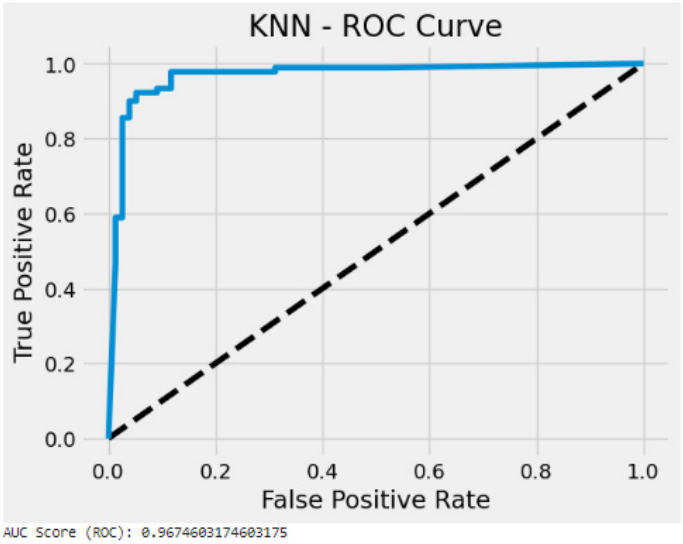
Matriz de Confusión KNN Optimizado




```
Accuracy Score Test: 0.9161676646706587
Accuracy Score Train: 0.9972144846796658 (as comparison)
```

Figura 50

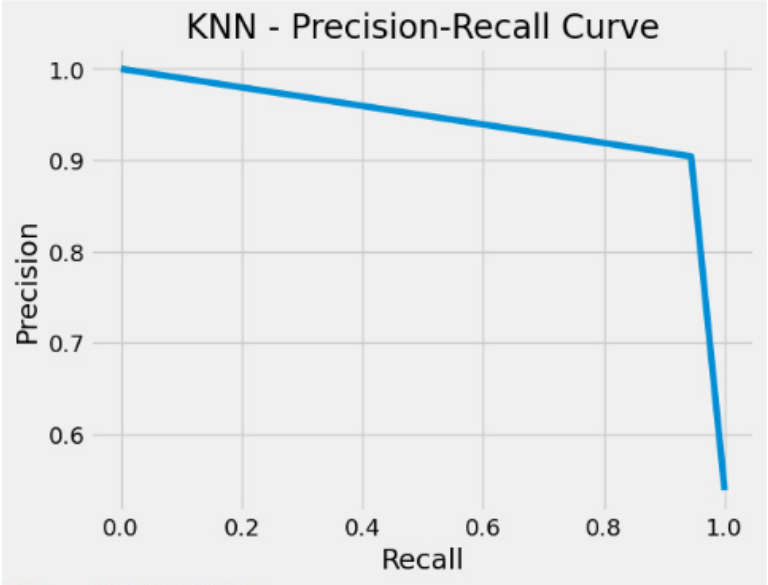
Curva ROC KNN Optimizado




```
AUC Score (ROC): 0.9674603174603175
```

Figura 52

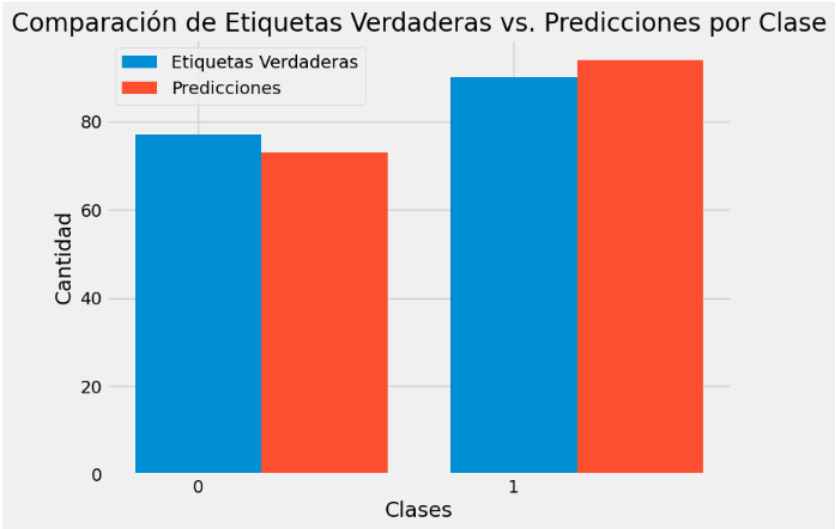
F1 Score KNN Optimizado



F1 Score: 0.9239130434782609
AUC Score (PR): 0.9393199416769298

Figura 51

Gráfico MCC de KNN Optimizado



Matthews Correlation Coefficient (MCC):
0.8316516396061548

Tabla 6

Resumen de Evaluación KNN Optimizado

Resumen	K Nearest Neighbors	K Nearest Neighbors Optimizado
Falso Positivo	9	9
Falso Negativo	15	5
Accuracy	0.8562874251497006	0.91616766467065
Curva ROC	0.9248917748917749	0.96746031746031
F1 Score	0.8620689655172413	0.92391304347826
MCC	0.7142889740706924	0.83165163960615

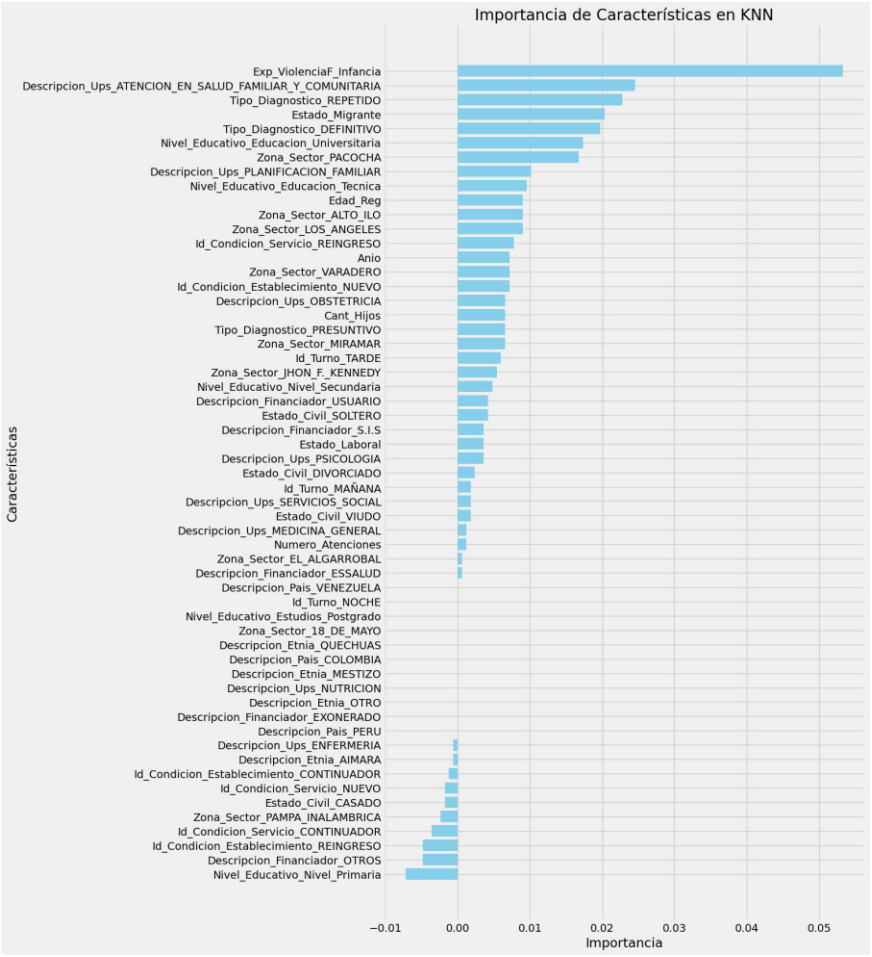
Los resultados obtenidos en el modelo de K Nearest Neighbors Optimizado presentan mayor precisión y exactitud que en la ejecución del modelo predeterminado, la matriz de confusión ha presentado ligeros cambios, para estos casos se considera que una menor cantidad de falso negativo para el objetivo planteado evitan tener consecuencias graves a comparación de los falsos positivos más que nada en la sección de salud como lo es la violencia física, para esta situación la no detección de eventos importantes puede ser mucho más peligroso(Mohammed & Kora, 2023).

iv. Interpretación del Modelo

Para la interpretación del modelo se mostrará las características más influyentes para KNN Optimizado en la figura 53

Figura 53

Feature Importances KNN Optimizado



Nota. Se verifica que la característica más influyente es Exp_ViolenciaF_Infancia.

b. Algoritmo Random Forest Optimizado

i. Ajuste de Hiperparametros en el Modelo

Se realiza el ajuste de Hiperparametros para el algoritmo de Random Forest.

Figura 54

Ajuste de Hiperparametros RF Optimizado

```
# Definir el modelo
rf = RandomForestClassifier(random_state=42)
# Definir la grid de hiperparámetros para RandomizedSearch
param_grid_rf = {
    'n_estimators': [100, 200, 300],
    'max_depth': [10, 20, 30],
    'min_samples_split': [10, 20, 30],
    'min_samples_leaf': [5, 7, 10],
    'max_features': ['sqrt', 'log2'],
    'class_weight': ['balanced', {0: 1, 1: 2}],
    'criterion': ['gini', 'entropy'],
    'bootstrap': [True, False]
}
# Configurar la validación cruzada estratificada
skf = StratifiedKFold(n_splits=5)

# Configurar RandomizedSearchCV
rf_random_grid = RandomizedSearchCV(
    estimator=rf,
    param_distributions=param_grid_rf,
    cv=skf,
    n_jobs=-1,
    verbose=2
)
```

Los ajustes realizados para el modelo de RF son los siguientes:

n_estimators: Numero de árboles en el bosque, para este caso se evalúa 3 opciones.

max_depth: Profundidad máxima de los árboles, para este caso limita los niveles que puede tener un árbol, esto ayuda para controlar la complejidad del modelo y evitar un sobreajuste, se evalúa 3 niveles.

min_samples_split: Número mínimo de muestras para dividir un nodo, se evalúa 3 divisiones mínimas.

min_samples_leaf: Número mínimo de muestras necesarias en una hoja, esto que los árboles se ramifiquen demasiado, para este caso se probó 3 casos mínimos de muestras.

max_features: Considera el número máximo de características que el modelo va a considerar para la división en cada nodo, para este caso se consideró “sqrt” y “log2” para el total de características.

class_weight: Controla el peso para cada clase en función de la prevalencia en este caso se consideró los dos tipos de ajustes para el balance y consideración de peso en variable.

criterion: Se utiliza para ajustar la calidad de una división en el nodo de un árbol.

Bootstrap: Con este ajuste va a controlar si se utiliza el muestreo con reemplazo para construir los árboles.

ii. Predicción del Modelo

Se ejecuta la predicción del modelo entrenado a los datos de prueba.

Figura 55

Ejecución de Predicción RF Optimizado

```
# Hacer predicciones
y_pred_rf_tuned_prob = rf_random_grid.predict_proba(X_test)

threshold = 0.4  # Ajustar el umbral para la clase positiva
y_pred_adjusted = (y_pred_rf_tuned_prob[:, 1] >= threshold).astype(int)

# Evaluar los resultados
print('Random Forest - Best Parameters: ', rf_random_grid.best_estimator_)
```

```
Fitting 5 folds for each of 10 candidates, totalling 50 fits
Random Forest - Best Parameters:  RandomForestClassifier(class_weight='balanced', criterion='entropy',
                       max_depth=10, min_samples_leaf=7, min_samples_split=30,
                       random_state=42)
```

iii. Evaluación del Modelo

Se ejecutó las métricas predefinidas al modelo Random Forest optimizado.

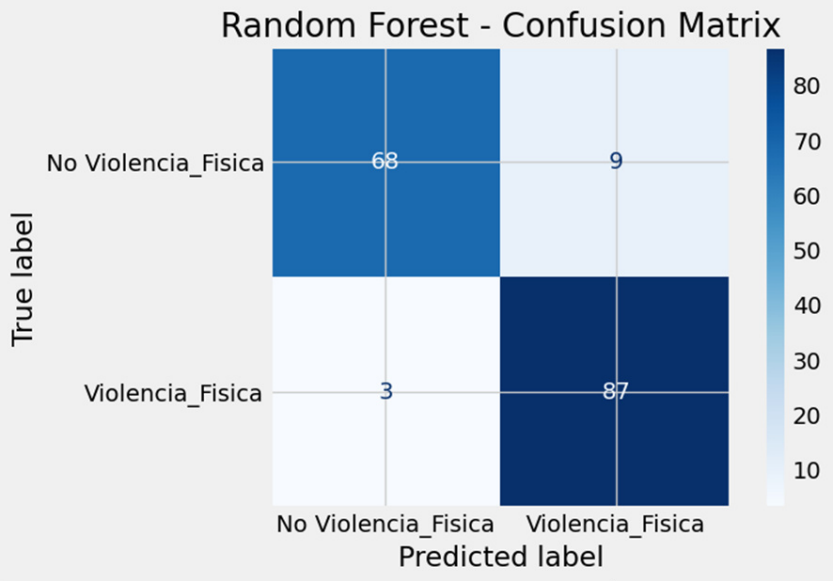
Figura 56

Evaluación de RF Optimizado

```
# Plot model evaluations.
confusion_matrix_plot(X_train_smote, y_train_smote, X_test, y_test, rf_random_grid, y_pred_adjusted, 'Random Forest')
roc_curve_auc_score(X_test, y_test, y_pred_rf_tuned_prob, 'Random Forest (tuned)')
precision_recall_curve_and_scores(X_test, y_test, y_pred_adjusted, y_pred_adjusted, 'Random Forest (tuned)')
matthews_corrcoef_score(y_test, y_pred_adjusted)
```

Figura 57

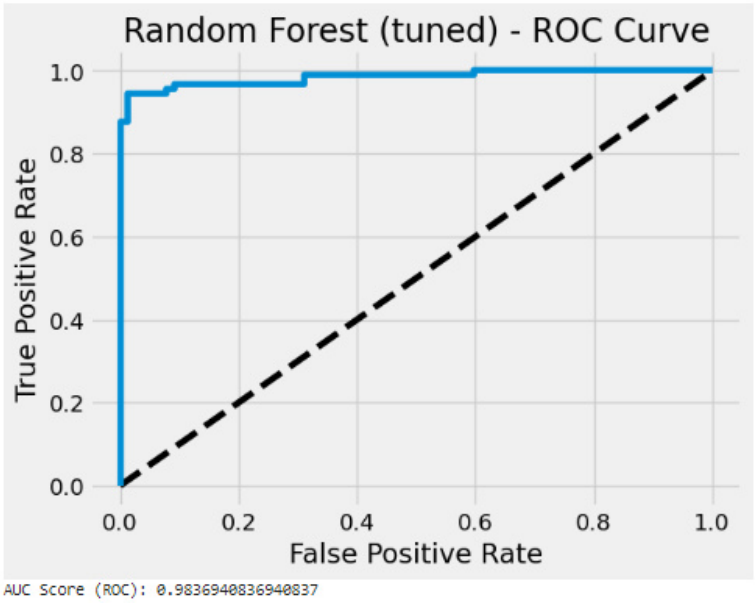
Matriz de Confusión RF Optimizado




```
Accuracy Score Test: 0.9281437125748503
Accuracy Score Train: 0.9540389972144847 (as comparison)
```

Figura 59

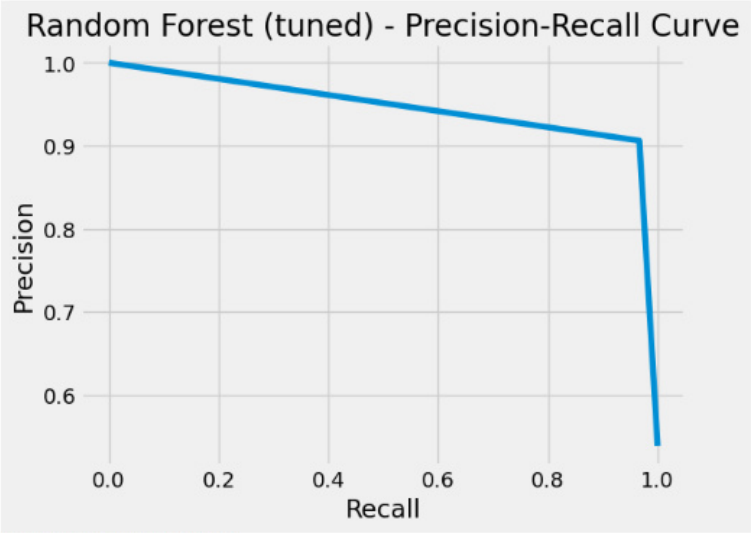
Curva ROC RF Optimizado



AUC Score (ROC): 0.9836940836940837

Figura 58

F1 Score RF Optimizado



F1 Score: 0.9354838709677419
AUC Score (PR): 0.9454403692614771

Figura 60

Gráfico MCC de RF Optimizado

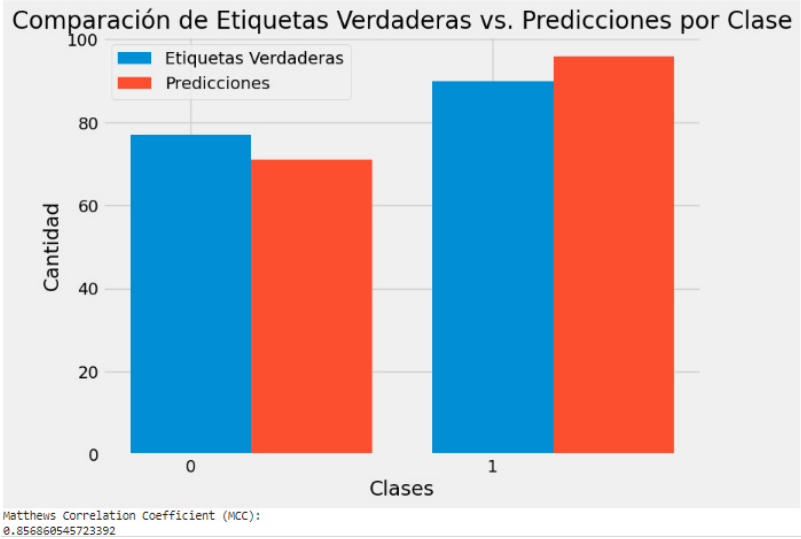


Figura 7

Resumen RF Optimizado

Resumen	Random Forest	Random Forest Optimizado
Falso Positivo	4	9
Falso Negativo	4	3
Accuracy	0.95209580838323	0.92814371257485
Curva ROC	0.98621933621933	0.98369408369408
F1 Score	0.95555555555555	0.93548387096774
MCC	0.90360750360750	0.8568605457233

Según los resultados obtenidos para el modelo de Random Forest Optimizados se puede llegar a verificar que en cada métrica se obtuvo resultados ligeramente por debajo del modelo sin optimizar esto realizado

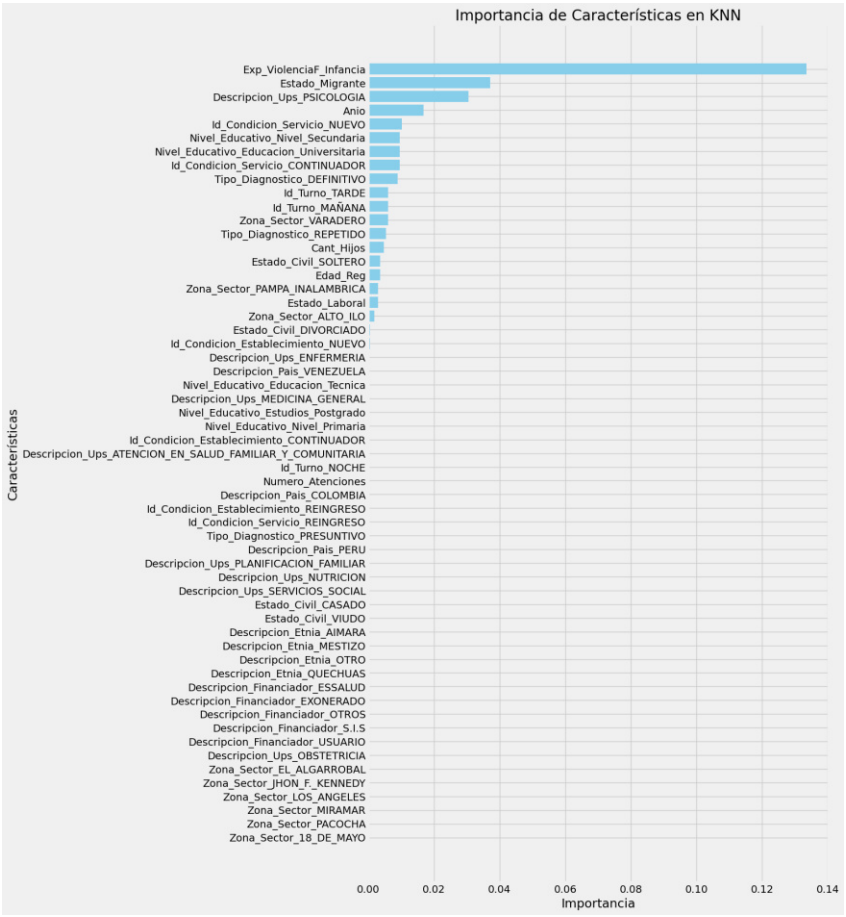
con el objetivo de disminuir los FN en la predicción del modelo ya que es importante considerar un modelo que logre capturar mas casos positivos en situaciones de riesgo.

iv. Interpretación del Modelo

Para la interpretación del modelo se mostrará las características más influyentes para RF Optimizado en la figura 61.

Figura 61

Feature Importances RF Optimizado



Nota. Se verifica que la característica más influyente es Exp_ViolenciaF_Infancia.

c. Algoritmo Naive Bayes Optimizado

i. Ajuste de Hiperparametros en el Modelo

Se realiza el ajuste de Hiperparametros para el algoritmo de Naive Bayes.

Figura 62

Ajuste de Hiperparametros NB

```
from sklearn.naive_bayes import BernoulliNB

# Ajuste de hiperparámetros con una cuadrícula ampliada
param_grid = {
    'alpha': [0.0001, 0.001],
    'fit_prior': [True, False]
}

# Instanciar el modelo Bernoulli Naive Bayes
nb = BernoulliNB()

nb_cv = GridSearchCV(estimator=nb, param_grid=param_grid, cv=5, n_jobs=-1, verbose=2)
nb_cv.fit(X_train_smote, y_train_smote)
```

Los ajustes realizados para el modelo de NB solo fueron dos y son los siguientes:

Alpha: Parámetro de suavización en este caso es un monto mínimo lo cual permite evitar problemas de probabilidades cero.

Fit_prior: Permite ajustar las probabilidades del conocimiento previo de las clases, para este modelo se considerará en 2 probabilidades.

ii. Predicción del Modelo

Se ejecuta la predicción del modelo entrenado a los datos de prueba.

Figura 63

Ejecución de Predicción NB optimizado

```
# Obtener las probabilidades
y_pred_nb_tuned_prob = nb_cv.predict_proba(X_test)

# Ajustar el umbral
threshold = 0.2
y_pred_adjusted = (y_pred_nb_tuned_prob[:, 1] >= threshold).astype(int)

# Imprimir los mejores parámetros y usar `y_pred_adjusted` para análisis adicionales.
print('Naive Bayes - Best Parameters: ', nb_cv.best_params_)
```

```
Fitting 5 folds for each of 4 candidates, totalling 20 fits
Naive Bayes - Best Parameters:  {'alpha': 0.0001, 'fit_prior': True}
```

iii. Evaluación del Modelo

Se ejecutó las métricas predefinidas al modelo Naive Bayes optimizado.

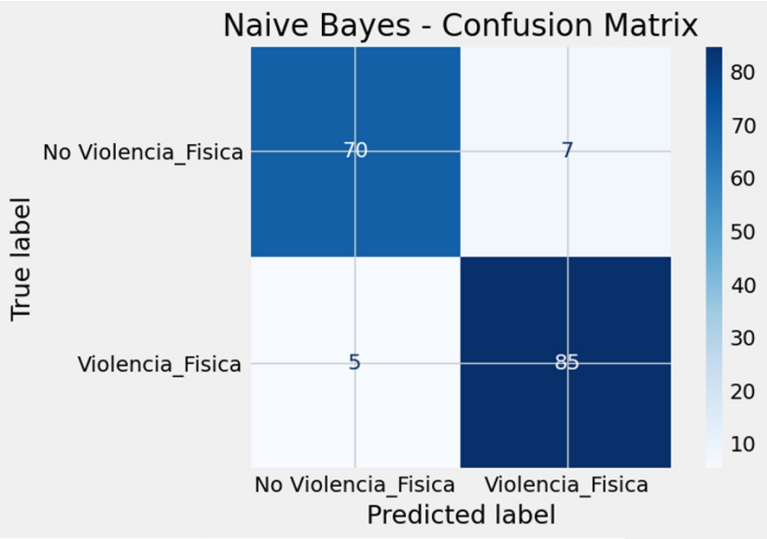
Figura 64

Evaluación de NB Optimizado

```
# Grafico de evaluaciones
confusion_matrix_plot(X_train_smote, y_train_smote, X_test, y_test, nb_cv, y_pred_adjusted, 'Naive Bayes')
roc_curve_auc_score(X_test, y_test, y_pred_nb_tuned_prob, 'Naive Bayes')
precision_recall_curve_and_scores(X_test, y_test, y_pred_adjusted, y_pred_adjusted, 'Naive Bayes')
```

Figura 65

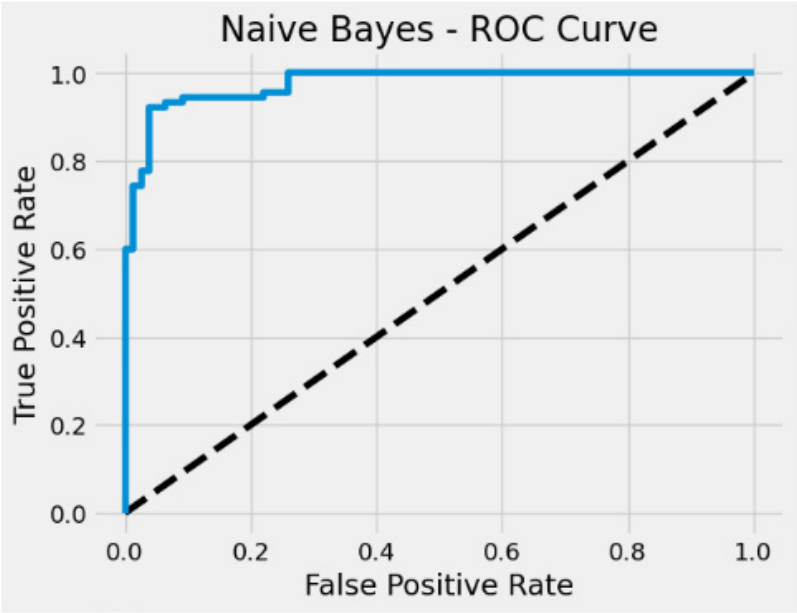
Matriz de Confusión NB Optimizado




```
Accuracy Score Test: 0.9281437125748503
Accuracy Score Train: 0.9345403899721448 (as comparison)
```

Figura 67

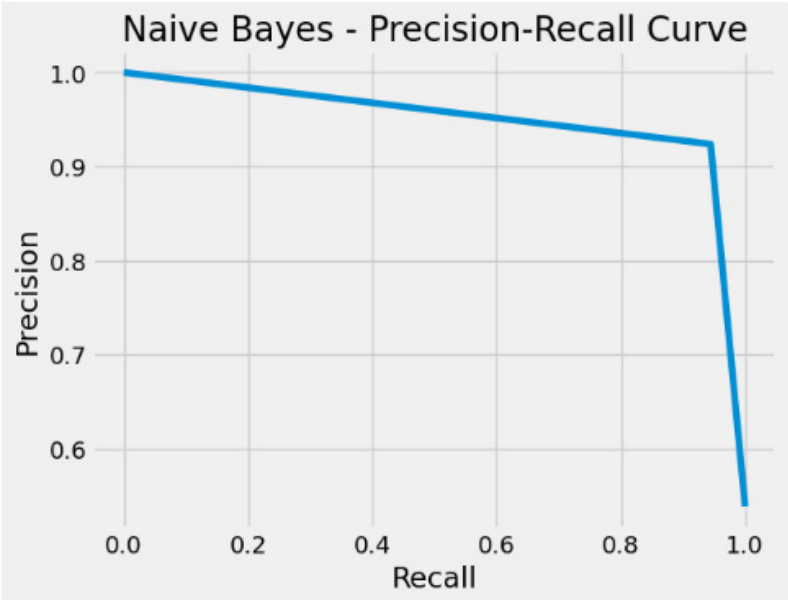
Curva ROC NB Optimizado



AUC Score (ROC): 0.9759018759018759

Figura 66

F1 Score NB Optimizado



F1 Score: 0.9340659340659341
AUC Score (PR): 0.9491488038415921

Figura 68

Gráfico MCC NB Optimizado

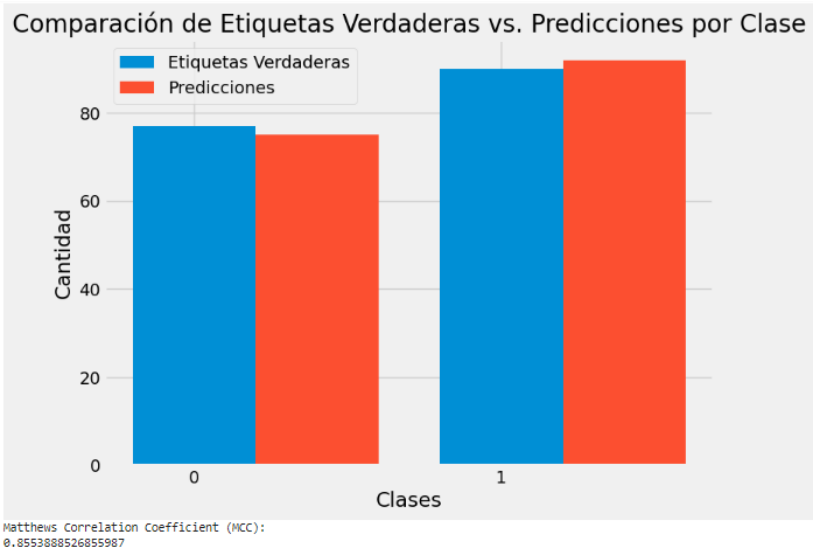


Tabla 8

Resumen de Evaluación NB Optimizado

Resumen	**Naive Bayes**	**Naive Bayes Optimizado**
Falso Positivo	3	7
Falso Negativo	8	5
Accuracy	0.93413173652694	0.92814371257485
Curva ROC	0.97633477633477	0.97590187590187
F1 Score	0.93714285714285	0.93406593406593
MCC	0.86964388860075	0.85538885268559

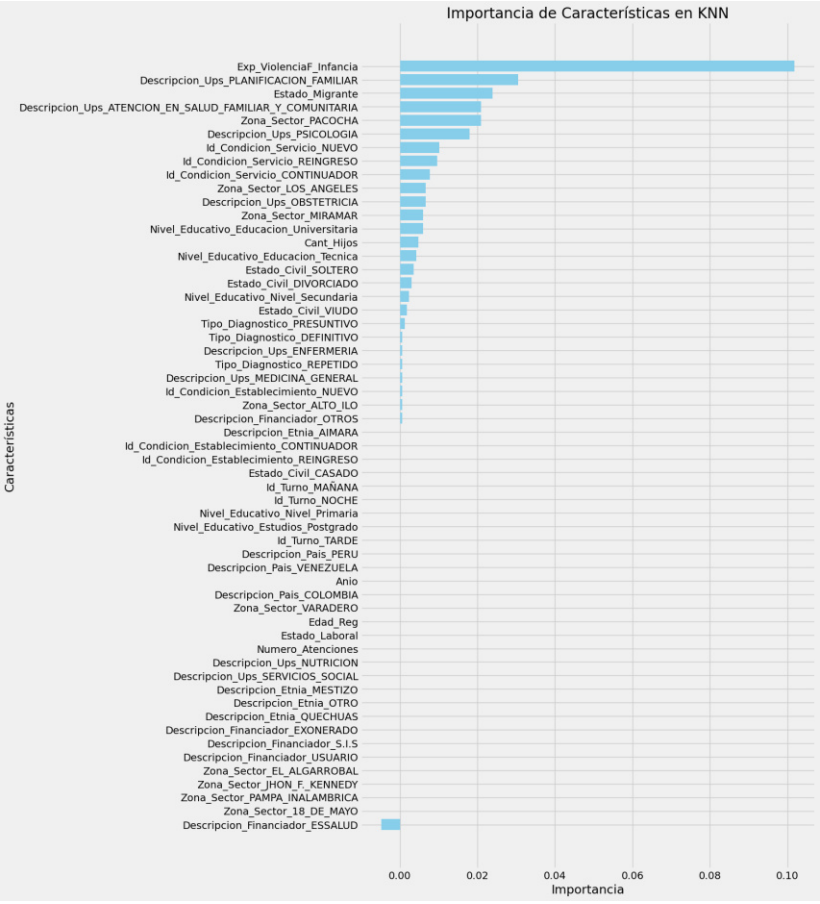
Los resultados obtenidos a partir del ajuste del modelo Naive Bayes se verifica que se tiene una ligera disminución en los resultados de cada métrica, sin embargo, la disminución de los FN es más considerable en el modelo optimizado lo cual ayuda en la determinación de más casos positivos.

iv. Interpretación del Modelo

Para la interpretación del modelo se mostrará las características más influyentes para NB Optimizado en la figura 69.

Figura 69

Feature Importances NB Optimizado



Nota. Se verifica que la característica más influyente es Exp_ViolenciaF_Infancia

d. Algoritmo Decision Trees Optimizado

i. Ajuste de Hiperparametros en el Modelo

Se realiza el ajuste de Hiperparametros para el algoritmo de Decision Trees.

Figura 70

Ajuste de Hiperparametros DT

```
from sklearn.tree import DecisionTreeClassifier

# Definir la cuadrícula de hiperparámetros
param_grid = {
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [5, 10, 20, 30],
    'min_samples_leaf': [2, 4, 6],
    'max_features': ['sqrt', 'log2', 0.9],
    'criterion': ['gini', 'entropy'],
    'splitter': ['best', 'random'],
    'max_leaf_nodes': [10, 20, 30]
}

# Instanciar el modelo de Decision Tree
dt = DecisionTreeClassifier(random_state=42)

# Configurar GridSearchCV
dt_cv = GridSearchCV(estimator=dt, param_grid=param_grid, cv=5, n_jobs=-1, verbose=2)

# Ajustar el modelo
dt_cv.fit(X_train_smote, y_train_smote)
```

Los ajustes realizados para el modelo de DT comparte similitud con Random Forest a excepción de Hiperparametros específicos para Decision Trees los cuales son:

max_leaf_nodes: Número máximo de hojas en el árbol.

splitter: Método para dividir los nodos.

ii. Predicción del Modelo

Se ejecuta la predicción del modelo entrenado a los datos de prueba.

Figura 71

Ejecución de Predicción DT optimizado

```
# Hacer predicciones de probabilidad
y_pred_dt_tuned_prob = dt_cv.predict_proba(X_test)

# Ajustar el umbral de clasificación
threshold = 0.4
y_pred_adjusted = (y_pred_dt_tuned_prob[:, 1] >= threshold).astype(int)

print('Decision Trees - Best Parameters: ', dt_cv.best_params_)
```

```
Fitting 5 folds for each of 1728 candidates, totalling 8640 fits
Decision Trees - Best Parameters:  {'criterion': 'entropy', 'max_depth': None,
```

iii. Evaluación del Modelo

Se ejecutó las métricas predefinidas al modelo Decision Trees optimizado.

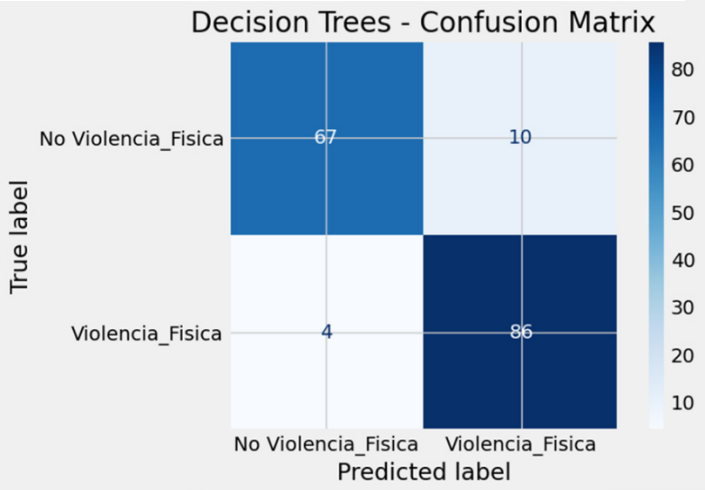
Figura 72

Evaluación de DT Optimizado

```
# Grafico de evaluaciones
confusion_matrix_plot(X_train_smote, y_train_smote, X_test, y_test, dt_cv, y_pred_adjusted, 'Decision Trees')
roc_curve_auc_score(X_test, y_test, y_pred_dt_tuned_prob, 'Decision Trees')
precision_recall_curve_and_scores(X_test, y_test, y_pred_adjusted, y_pred_adjusted, 'Decision Trees')
```

Figura 73

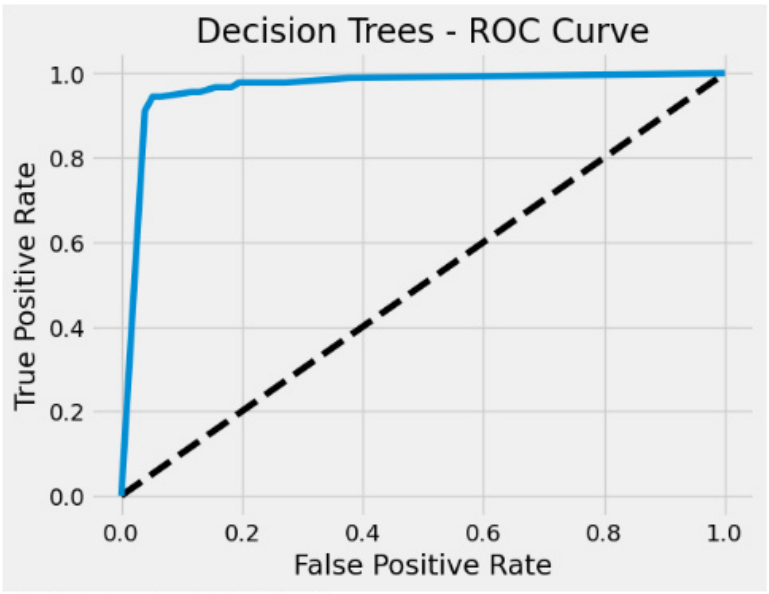
Matriz de Confusión DT Optimizado




```
Accuracy Score Test: 0.9161676646706587
Accuracy Score Train: 0.9526462395543176 (as comparison)
```

Figura 75

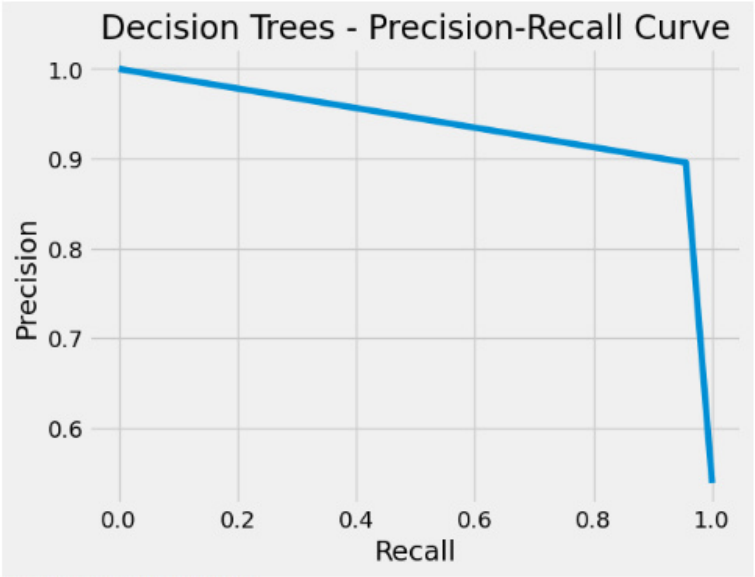
Curva ROC DT Optimizado



AUC Score (ROC): 0.9647907647907648

Figura 74

F1 Score DT Optimizado



F1 Score: 0.9247311827956989
AUC Score (PR): 0.9376704923486361

Figura 76

Grafica MCC de DT Optimizado

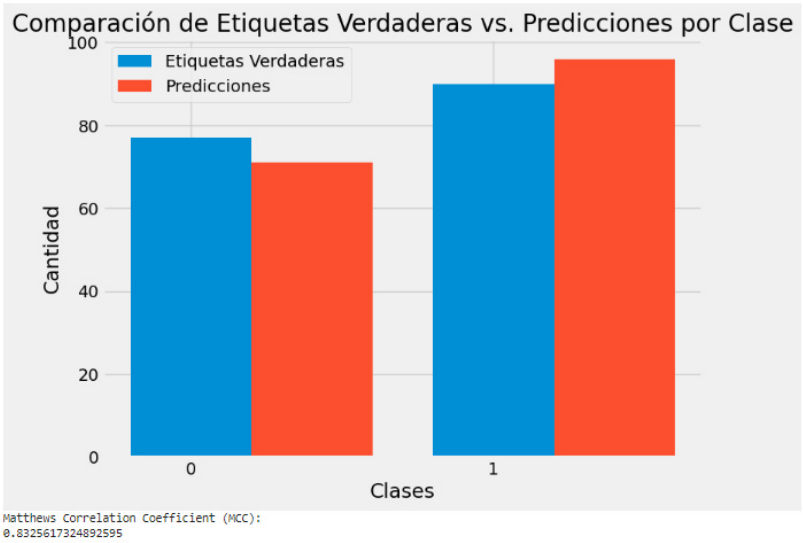


Tabla 9

Resumen de Evaluación DT Optimizado

Resumen	Decision Trees	Decision Trees Optimizado
Falso Positivo	7	10
Falso Negativo	5	4
Accuracy	0.92814371257485	0.91616766467065
Curva ROC	0.9260461760461	0.96479076479076
F1 Score	0.93406593406593	0.92473118279569
MCC	0.85538885268559	0.83256173248925

Según lo revisado en los resultados obtenidos del modelo Decision Trees optimizado, se han obtenido una Curva ROC más optima a

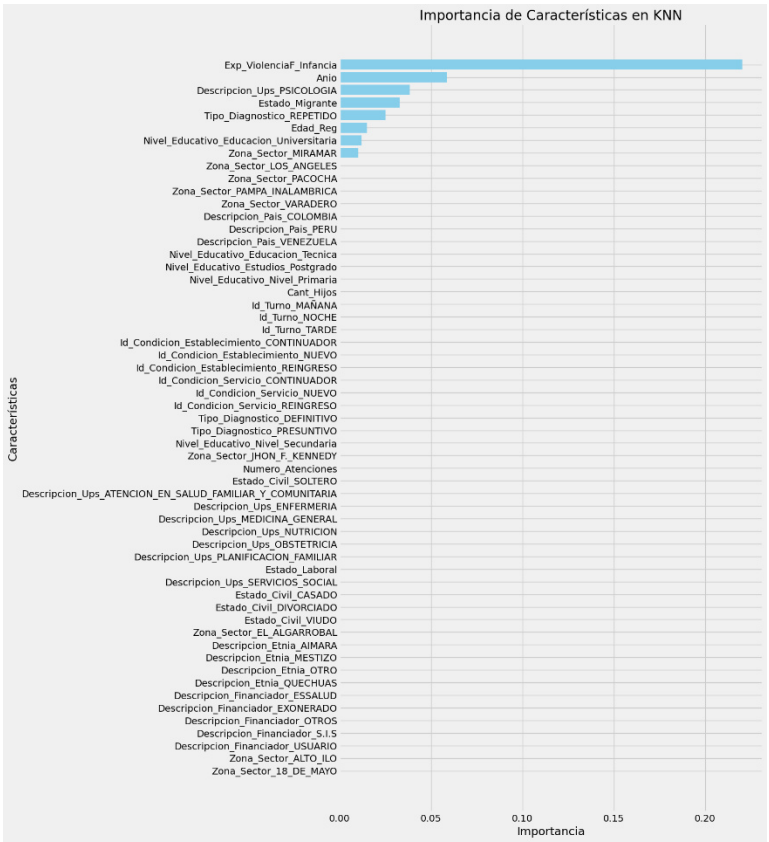
comparación del modelo sin ajuste, de igual forma la matriz de confusión tuvo una reducción considerable logrando disminuir los FN, se considerará el modelo optimizado como el resultado final para Decision Trees.

iv. Interpretación del Modelo

Para la interpretación del modelo se mostrará las características más influyentes para DT Optimizado en la figura 77.

Figura 77

Feature Importances DT Optimizado



Nota. Se verifica que la característica más influyente es Exp_ViolenciaF_Infancia.

4.5 Evaluación

a. Evaluación de los Modelos

Una vez realizado los modelos con los algoritmos de K Nearest Neighbors, Random Forest, Naive Bayes, Decision Trees. Se podrá determinar el modelo más optimo que cumple con los objetivos planteados de la presente investigación, para ello se detallará los resultados de los modelos seleccionados en la tabla 10.

Tabla 10

Comparativa de los Resultados de las Técnicas Utilizadas

Resumen	FP	FN	Accuracy	Curva ROC	F1 Score	MCC
K Nearest Neighbors Optimizado	**9**	**5**	**0.91616**	**0.96746**	**0.92391**	**0.83165**
Random Forest Optimizado	**9**	**3**	**0.92814**	**0.98369**	**0.93548**	**0.85686**
Naive Bayes Optimizado	**7**	**5**	**0.92814**	**0.97590**	**0.93406**	**0.85538**
Decision Trees Optimizado	**10**	**4**	**0.91616**	**0.96479**	**0.92473**	**0.83256**

Nota. En esta tabla se puede visualizar la validación de cada modelo en base a las métricas utilizadas en el cual se determina que el algoritmo de Random Forest es el más optimo.

Figura 78

Feature Importances en base a Random Forest

Feature	KNN	Random Forest	Naive Bayes	Decision Trees
Exp_ViolenciaF_Infancia	0.053293	0.133533	0.101796	0.220359
Estado_Migrante	0.020359	0.037126	0.023952	0.032934
Descripcion_Ups_PSICOLOGIA	0.003593	0.030539	0.017964	0.038323
Anio	0.007186	0.016766	0.000000	0.058683
Id_Condicion_Servicio_NUEVO	-0.001796	0.010180	0.010180	0.000000
Id_Condicion_Servicio_CONTINUADOR	-0.003593	0.009581	0.007784	0.000000
Nivel_Educativo_Nivel_Secundaria	0.004790	0.009581	0.002395	0.000000
Nivel_Educativo_Educacion_Universitaria	0.017365	0.009581	0.005988	0.011976
Tipo_Diagnostico_DEFINITIVO	0.019760	0.008982	0.000599	0.000000
Id_Turno_TARDE	0.005988	0.005988	0.000000	0.000000
Id_Turno_MAÑANA	0.001796	0.005988	0.000000	0.000000
Zona_Sector_VARADERO	0.007186	0.005988	0.000000	0.000000
Tipo_Diagnostico_REPETIDO	0.022754	0.005389	0.000599	0.025150
Cant_Hijos	0.006587	0.004790	0.004790	0.000000
Estado_Civil_SOLTERO	0.004192	0.003593	0.003593	0.000000
Edad_Reg	0.008982	0.003593	0.000000	0.014970
Zona_Sector_PAMPA_INALAMBRICA	-0.002395	0.002994	0.000000	0.000000
Estado_Laboral	0.003593	0.002994	0.000000	0.000000
Zona_Sector_ALTO_ILO	0.008982	0.001796	0.000599	0.000000
Estado_Civil_DIVORCIADO	0.002395	0.000599	0.002994	0.000000
Id_Condicion_Establecimiento_NUEVO	0.007186	0.000599	0.000599	0.000000

Nota. Se identifica las características importantes en base al modelo más optimo determinado el cual es Random Forest**.**

4.6 Despliegue

El despliegue no es una fase a considerar para cumplir con los objetivos planteados, ya que se tiene como propósito evaluar los modelos predictivos empleando técnicas de Machine Learning para pronosticar la violencia física contra la mujer en la provincia de Ilo, sin embargo, se puede evaluar como un paso a tomar en cuenta para futuras investigaciones o implementaciones en el cual consideren poner en práctica las conclusiones obtenidas de la presente investigación.

- **Discusión**

Se realiza la verificación de los resultados alcanzados en comparación con los resultados obtenidos en otros modelos de nuestro antecedente, En cada uno han determinado el modelo más óptimo para su investigación.

Tabla 11

Comparación con Antecedentes

Item	Autor	Modelo Optimo	F1 Score
1	Saboya et al.	Random Forest	0.6420
2	Timaran-Pereira et al.	Decision Trees	0.8235
3	Amusa et al.	Decision Trees	0.465
4	Tamilarasi & Rani	K Nearest Neighbors	0.3003
5	Al-Taleb & Saeed	Naive Bayes	0.7536
7	Cumbicus-Pineda et al.	K Nearest Neighbors	0.4578
8	Rodríguez-Rodríguez et al.	Random Forest	0.6389
9	Investigación Propuesta	Random Forest	0.9354

El modelo propuesto en este proyecto, basado en Random Forest, logró un F1 Score de 0.9354, lo cual es significativamente superior al rendimiento de los demás estudios presentados. Este resultado destaca que el modelo desarrollado es altamente eficaz en la predicción de la violencia física contra la mujer, superando ampliamente los valores observados en trabajos como los de Saboya et al., (2019) (0.6420) y Rodríguez-Rodríguez et al., (2020) (0.6389), quienes también utilizaron Random Forest. Esto sugiere que, aunque Random Forest es un modelo robusto en diferentes contextos, los

ajustes realizados en los Hiperparámetros y la preparación de los datos en esta investigación han mejorado su rendimiento.

Otros modelos como Decision Trees utilizados por Timaran-Pereira et al., (2017) (0.8235) y Amusa et al., (2020) (0.465), y Naive Bayes aplicado por Al-Taleb & Saeed, (2019) (0.7536), muestran un buen desempeño, pero no alcanzan los niveles de precisión del modelo más optimo del presente proyecto. De igual forma con el modelo de K Nearest Neighbors aplicado por Tamilarasi & Rani, (2020) (0.3003) y Cumbicus-Pineda et al., (2021) (0.4578) se puede identificar valores considerablemente bajos a comparación de los demás modelos, lo cual determina que el modelo más optimo seleccionado por la presente investigación demuestra ser más efectivo, la diferencia en el rendimiento pueden atribuirse a la correcta selección y ajuste del algoritmo, así como a una adecuada preparación y balanceo de los datos, incluyendo la aplicación de técnicas como SMOTE para manejar el desbalance de clases.

V. Conclusiones

A lo largo del presente trabajo de investigación, se plantearon tanto un objetivo general como tres objetivos específicos relacionados con la predicción de la violencia física contra la mujer en la provincia de Ilo, mediante el uso de técnica de machine learning. A continuación, se exponen las conclusiones en base al cumplimiento de dichos objetivos.

1. Conclusión General

Se logró evaluar modelos predictivos empleando técnicas de machine learning para pronosticar considerablemente la violencia física contra la mujer en la Provincia de Ilo.

A través de la evaluación de varios algoritmos de clasificación, como Random Forest, K Nearest Neighbors, Naive Bayes y Decision Trees, se lograron identificar modelos capaces de predecir de manera efectiva este fenómeno. Los resultados obtenidos, especialmente con el modelo de Random Forest, muestran que es posible generar predicciones con altos niveles de precisión y con un F1 Score destacado, lo que justifica el uso de técnica de machine learning para este tipo de problemas sociales.

2. Conclusiones Especificas

2.1 Se consiguió determinar los factores más influyentes en los modelos predictivos que contribuyó considerablemente en la predicción de la violencia física contra la mujer en la Provincia de Ilo.

Para esta primera conclusión especifica se utilizó el modelo de Random Forest para analizar la importancia de las variables en la predicción.

El análisis de Feature Importance permitió identificar las variables que tienen mayor impacto en la predicción de la violencia física, entre los factores más influyentes destacan:

- Exp_ViolenciaF_Infancia: Esta variable fue la más significativa, con un valor de 0.133533, lo que sugiere que es uno de los principales factores que predicen la violencia física en la adultez.
- Estado_Migrante: Con una importancia de 0.037126, esta variable indica que la condición de migrante es un factor relevante para predecir la violencia física contra la mujer.
- Descripción_Ups_PSICOLOGIA: Con un valor de 0.035976, esta variable, relacionada con el acceso o necesidad de servicios psicológicos, también demostró ser un factor determinante en la ocurrencia de violencia física.
- Otros factores como Id_Condicion_Servicio_NUEVO, Nivel_Educativo_Nivel_Secundaria, y Estado_Civil_DIVORCIADO también aportaron información valiosa para entender mejor las dinámicas detrás de este fenómeno.

Este análisis permitió no solo construir un modelo predictivo más preciso, sino también identificar claramente las características claves que explican la violencia física contra la mujer en la Provincia de Ilo, cumpliendo así con el primer objetivo específico de la investigación.

2.2 Se llevó a cabo la construcción de un modelo predictivo basado en técnicas de machine learning que predijo considerablemente la violencia física contra la mujer en la provincia de Ilo.

En base a esta segunda conclusión especifica se construyeron varios modelos, siendo el modelo basado en Random Forest el que mostró el mejor rendimiento. El modelo fue entrenado utilizando un conjunto de datos que fue previamente procesado y balanceado mediante distintas técnicas. Además, se aplicaron técnicas de ajuste de Hiperparámetros para optimizar el rendimiento del modelo.

2.3 El resultado de la comparación de modelos predictivos determinó el modelo más significativo para predecir la violencia física contra la mujer en la provincia de Ilo.

Esta última conclusión especifica se logró mediante el uso de diversas métricas de evaluación, como el F1 Score, Accuracy, Curva ROC y MCC. Estas métricas fueron aplicadas a todos los modelos desarrollados, lo que permitió identificar que el modelo Random Forest es el más adecuado, logrando un F1 Score de 0.9354, el cual supera ampliamente a otros modelos evaluados. Este resultado asegura que el modelo propuesto es capaz de realizar predicciones con alta precisión.

VI. Recomendaciones

Las recomendaciones que se puede realizar a partir de los resultados obtenidos en la investigación son las siguientes:

Realizar más ajustes iterativos en los Hiperparametros de cada modelo para mejorar la predicción en base a la matriz de confusión.

A partir de las características más influyentes se podría agregar más características relevantes ya sea añadiendo nuevas características o la creación de características en base a las existentes, de igual forma se podría aplicar técnicas de selección de características más avanzadas, esto con el fin de optimizar el modelo.

Con la determinación del modelo más optimo en esta investigación se podría profundizar más complementándolo con otros algoritmos para lograr un modelo ensamblado que permita aprovechar la utilidad de los algoritmos y crear un modelo más robusto.

Por último, se recomienda en un futuro realizar la implementación del modelo en un entorno de producción para que pueda predecir en tiempo real, esto permitirá que se pueda monitorear su rendimiento y realizar los ajustes con datos nuevos.

VII. Referencias

Aha, D. W., Kibler, D., Albert, M. K., & Quinian, J. R. (1991). *Instance-Based Learning Algorithms* (Vol. 6).

Alakh. (2024). *One Hot Encoding vs. Label Encoding in Machine Learning*. AnalyticsVidhya.

Al-Taleb, N., & Saeed, S. (2019). A study on application of data mining techniques in crime forecasting. *ICIC Express Letters, Part B: Applications*, *10*(8), 657–663. https://doi.org/10.24507/icicelb.10.08.657

Amusa, L. B., Bengesai, A. V., & Khan, H. T. A. (2020). Predicting the Vulnerability of Women to Intimate Partner Violence in South Africa: Evidence from Tree-based Machine Learning Techniques. *Journal of Interpersonal Violence*. https://doi.org/10.1177/0886260520960110

Arimetrics. (2022). *Qué es Modelo predictivo - Definición, significado y ejemplos*. https://www.arimetrics.com/glosario-digital/modelo-predictivo

Baru, Chaitanya. (2019). *2019 IEEE International Conference on Big Data : proceedings : Dec 9 - Dec 12, 2019, Los Angeles, CA, USA*. IEEE.

Breiman, L. (2001). *Random Forests* (Vol. 45).

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In *Journal of Artificial Intelligence Research* (Vol. 16).

Cumbicus-Pineda, O. M., Abad-Eras, T. E., & Neyra-Romero, L. A. (2021). Data Mining to Determine the Causes of Gender-Based Violence against Women in Ecuador. *2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM)*, 1–6. https://doi.org/10.1109/ETCM53643.2021.9590664

Direccion Regional de Salud Moquegua. (2008). *ROF RED ILO*.

Fidias G. Arias. (2012). *EL PROYECTO DE INVESTIGACIÓN 6a EDICIÓN*. https://www.researchgate.net/publication/301894369

Heise, L. L. (1998). Violence against women: An integrated, ecological framework. *Violence Against Woman*, *4*(3), 262–290. https://doi.org/10.1177/1077801298004003002

Hernández, J., Ma, O., Ramírez, J., Cèsar, Q., & Ramírez, F. (2005). *Introducción a la Minería de Datos*.

Instituto Nacional de Estadística e Informática. (2021). *INFORME_PRINCIPAL_ENDES_2020*.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. In *Informatica* (Vol. 31).

Lucidchart. (2024). *Símbolos de los árboles de decisión*. Https://Www.Lucidchart.Com/Pages/Es/Que-Es-Un-Diagrama-de-Arbol-de-Decision.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Mccallum, A., & Nigam, K. (1998). *A Comparison of Event Models for Naive Bayes Text Classification*. www.aaai.org

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. In *Journal of King Saud University - Computer and*

Information Sciences (Vol. 35, Issue 2, pp. 757–774). King Saud bin Abdulaziz University. https://doi.org/10.1016/j.jksuci.2023.01.014

Olivares, E., Teresa, F., & Romero, I. (2011). *Modelo ecológico para una vida libre de violencia de género*.

Peruano PODER LEGISLATIVO, E., & Presidente La República, E. DE. (2015). *LEY No 30364 LEY PARA PREVENIR, SANCIONAR Y ERRADICAR LA VIOLENCIA CONTRA LAS MUJERES Y LOS INTEGRANTES DEL GRUPO FAMILIAR*.

Programa Nacional AURORA. (2020). *Resumen Regional Moquegua 2020*.

Programa Nacional AURORA. (2021a). *PROGRAMA NACIONAL PARA LA PREVENCIÓN Y ERRADICACIÓN DE LA VIOLENCIA CONTRA LAS MUJERES E INTEGRANTES DEL GRUPO FAMILIAR-AURORA*.

Programa Nacional AURORA. (2021b). *Resumen Regional Moquegua 2021*.

Programa Nacional AURORA. (2022). *Resumen Regional Moquegua 2022*.

Quinlan, J. R. (1986). Induction of Decision Trees. In *Machine Learning* (Vol. 1).

Rafael, D., Bonett, C., Aguilar, A. S., Espino, C. G., Arias Chumpitaz, A., Dilcia, P., Carrión, D., Tapia, H., Catherine, E. E., Cervantes, R., Peña, H., Juan, A., Coaguila, R., Alexis, F., Rojas, H., Manuel, F., Buleje, M., De Mapas, E., Huertas, L., … Khang, M. (2021). *Perú: Feminicidio y Violencia contra la Mujer, 2015 - 2020*.

Rodríguez-Rodríguez, I., Rodríguez, J.-V., Pardo-Quiles, D.-J., Heras-González, P., & Chatzigiannakis, I. (2020). Modeling and forecasting gender-based violence through machine learning techniques. *Applied Sciences (Switzerland)*, *10*(22), 1–22. https://doi.org/10.3390/app10228244

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1–2), 1–39. https://doi.org/10.1007/s10462-009-9124-7

Saboya, N., Angel Sullon, A., & Loaiza, O. L. (2019). Predictive model based on machine learning for the detection of physically mistreated women in the Peruvian scope. *ACM International Conference Proceeding Series*, 18–23. https://doi.org/10.1145/3369114.3369143

Sngular. (2021). *CRISP-DM: La metodología para poner orden en los proyectos - Sngular*. https://www.sngular.com/es/data-science-crisp-dm-metodologia/

Tamilarasi, P., & Rani, R. U. (2020). Diagnosis of Crime Rate against Women using k-fold Cross Validation through Machine Learning. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 1034–1038. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000193

Timaran-Pereira, R., Calderón-Romero, A., & Hidalgo-Troya, A. (2017). Aplicación de los árboles de decisión en la identificación de patrones de lesiones fatales por causa externa en el municipio de Pasto, Colombia. *Universidad y Salud*, *19*(3), 388. https://doi.org/10.22267/rus.171903.101

Vargas Cordero Zoila Rosa. (2009). *Investigacion aplicada*.

World Health Organization. (2021). *Violence against women*. https://www.who.int/news-room/fact-sheets/detail/violence-against-women

Zapata-Tapasco, A., Pérez-Londoño, S., & Mora-Flórez. Juan. (2014). *A fault location method applied in power distribution systems based on k-NN classifiers parameterized using genetic algorithms and the reactance estimation*.

VIII. Anexos

Figura 79

Determinación de Feature Importances

```
# Caracteristicas importantes

feature_names = X1.columns

# Calcular la importancia por permutación
result_dt = permutation_importance(dt_cv, X_test, y_test, n_repeats=10, random_state=42)

# Crear un DataFrame para almacenar los resultados
importances_df = pd.DataFrame({
    'Feature': feature_names,
    'Importance': result_dt.importances_mean
})

# Ordenar las características por importancia
importances_df = importances_df.sort_values(by='Importance', ascending=False)

# Crear gráfico de barras de la importancia de características
plt.figure(figsize=(10, 20))  # Tamaño del gráfico
plt.tight_layout()
plt.barh(importances_df['Feature'], importances_df['Importance'], color='skyblue')
plt.xlabel('Importancia')
plt.ylabel('Características')
plt.title('Importancia de Características en KNN')
plt.gca().invert_yaxis()  # Invertir el eje Y para que la característica más importante esté arriba
plt.show()
```

Figura 80

Excel con la data de violencia fisica y psicológica

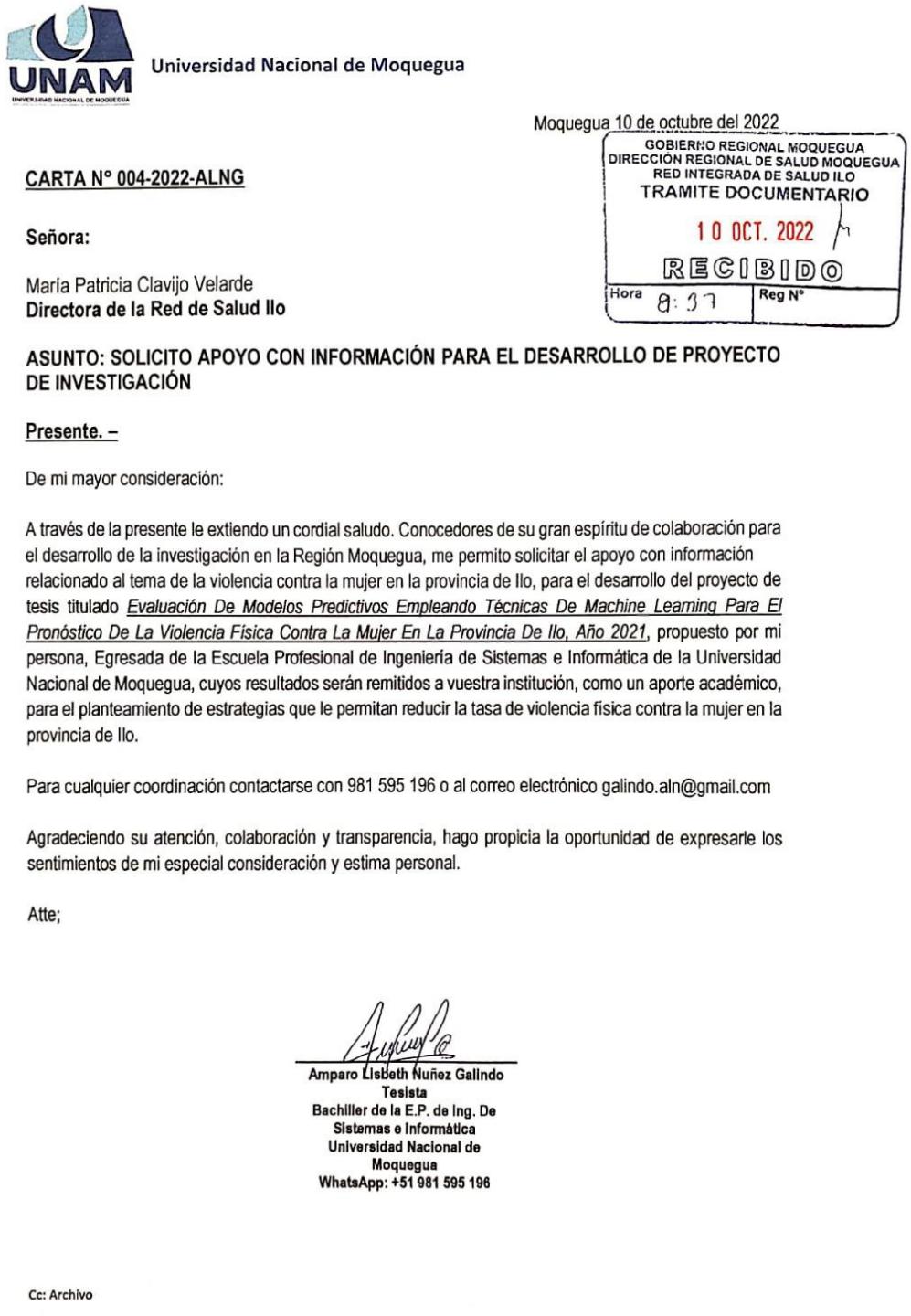
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Id_Cita	Anio	Mes	Dia	cha_Atenc	Lote	Num_Pag	Num_Reg	Id_Ups	cripcion_I	ripcion_S	cripcion_I	cripcion_	pcion_Mi	odigo_Uni	_Establecipo_Doc	locumento	stado_Civ_	
444752085	2021	1	6	########	D14	2	1	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		SOLTERO
445057623	2021	1	2	########	C10	2	1	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI		CASADO
445059270	2021	1	2	########	C10	2	2	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI		SOLTERO
445154178	2021	1	7	########	D15	2	1	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		SOLTERO
445154175	2021	1	7	########	D15	2	1	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI	76824409	CASADO
445157094	2021	1	7	########	D15	3	4	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		SOLTERO
445157091	2021	1	7	########	D15	3	4	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI	48416297	SOLTERO
445448985	2021	1	6	########	C10	9	4	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI		SOLTERO
445493712	2021	1	8	########	C10	16	1	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI		SOLTERO
445493716	2021	1	8	########	C10	16	1	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI	30646589	VIUDO
445493869	2021	1	8	########	C10	16	2	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI		SOLTERO
445493862	2021	1	8	########	C10	16	2	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI	78949553	SOLTERO
445670543	2021	1	9	########	D17	2	1	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		SOLTERO
445670617	2021	1	9	########	D17	2	2	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		SOLTERO
445670890	2021	1	9	########	D17	2	4	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		SOLTERO
445671575	2021	1	11	2021-01-11	D18	1	3	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		SOLTERO
446158006	2021	1	13	########	D20	4	7	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		SOLTERO
446596540	2021	1	14	########	D21	1	2	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		SOLTERO
446599682	2021	1	15	########	D22	1	1	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		SOLTERO
447118223	2021	1	12	########	C13	5	2	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI		SOLTERO
447118244	2021	1	12	########	C13	5	3	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI		SOLTERO
447123621	2021	1	15	########	C13	13	3	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI		SOLTERO
447123623	2021	1	15	########	C13	13	3	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI	14646743	SOLTERO
447124100	2021	1	15	########	C13	14	2	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI		SOLTERO
447345619	2021	1	18	########	D24	2	3	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		SOLTERO
447345737	2021	1	18	########	D24	2	4	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		DIVORCIAD
447952044	2021	1	20	########	VAT	7	22	303203	OBSTETRI	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI	78168045	VIUDO
448403256	2021	1	22	########	D28	1	1	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI		SOLTERO
448403257	2021	1	22	########	D28	1	1	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000007731	P.S. LOS A	DNI	61084397	SOLTERO
448504335	2021	1	18	########	C13	20	1	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI		SOLTERO
448504510	2021	1	18	########	C13	20	2	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI		CASADO
448504615	2021	1	18	########	C13	20	3	303802	PSICOLOG	GOBIERNO	MOQUEGU	ILO	ILO	000002820	C.S. PAMP	DNI		SOLTERO

Nota. Data de Red de Salud Ilo

- **Evidencia del Tramite Realizado para la obtención de Datos.**

Figura 81

Documento de solicitud de Información



Universidad Nacional de Moquegua

Moquegua 10 de octubre del 2022

GOBIERNO REGIONAL MOQUEGUA
DIRECCION REGIONAL DE SALUD MOQUEGUA
RED INTEGRADA DE SALUD ILO
TRAMITE DOCUMENTARIO
10 OCT. 2022
RECIBIDO
Hora 8:37 | Reg N°

CARTA N° 004-2022-ALNG

Señora:

María Patricia Clavijo Velarde
Directora de la Red de Salud Ilo

ASUNTO: SOLICITO APOYO CON INFORMACIÓN PARA EL DESARROLLO DE PROYECTO DE INVESTIGACIÓN

Presente. –

De mi mayor consideración:

A través de la presente le extiendo un cordial saludo. Conocedores de su gran espíritu de colaboración para el desarrollo de la investigación en la Región Moquegua, me permito solicitar el apoyo con información relacionado al tema de la violencia contra la mujer en la provincia de Ilo, para el desarrollo del proyecto de tesis titulado *Evaluación De Modelos Predictivos Empleando Técnicas De Machine Learning Para El Pronóstico De La Violencia Física Contra La Mujer En La Provincia De Ilo, Año 2021*, propuesto por mi persona, Egresada de la Escuela Profesional de Ingeniería de Sistemas e Informática de la Universidad Nacional de Moquegua, cuyos resultados serán remitidos a vuestra institución, como un aporte académico, para el planteamiento de estrategias que le permitan reducir la tasa de violencia física contra la mujer en la provincia de Ilo.

Para cualquier coordinación contactarse con 981 595 196 o al correo electrónico galindo.aln@gmail.com

Agradeciendo su atención, colaboración y transparencia, hago propicia la oportunidad de expresarle los sentimientos de mi especial consideración y estima personal.

Atte;

Amparo Lisbeth Nuñez Galindo
Tesista
Bachiller de la E.P. de Ing. De Sistemas e Informática
Universidad Nacional de Moquegua
WhatsApp: +51 981 595 196

Cc: Archivo

- **Código Fuente de los Algoritmos de Machine Learning**
    - **K-Nearest Neighbors Optimizado**

```
from sklearn.model_selection import GridSearchCV, StratifiedKFold
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.metrics import make_scorer, f1_score, recall_score,
precision_score
from sklearn.inspection import permutation_importance

# Definir el pipeline
pipeline = Pipeline([
    ('scaler', StandardScaler()),  # Escalado de características
    ('knn', KNeighborsClassifier())
])

# Define parameter grid for GridSearch
param_grid = {
    'knn__n_neighbors': np.arange(1, 30),
    'knn__weights': ['uniform', 'distance'],
    'knn__metric': ['euclidean', 'manhattan', 'minkowski']
}
# Configurar la validación cruzada estratificada
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Instanciar GridSearchCV
knn_cv = GridSearchCV(
    estimator=pipeline,
    param_grid=param_grid,
    cv=cv,
    n_jobs=-1,
    verbose=2
)
# Ajustar el modelo con los datos de entrenamiento
knn_cv.fit(X_train_smote, y_train_smote)

# Predecir las probabilidades en el conjunto de prueba
y_pred_knn_tuned_prob = knn_cv.predict_proba(X_test)
threshold = 0.4 # Establecer un umbral más bajo para la clase
positiva
y_pred_knn_tuned_threshold = (y_pred_knn_tuned_prob[:, 1] >=
threshold).astype(int)
# Evaluar resultados
print('KNN best parameters:', knn_cv.best_params_)
```

- **Random Forest Optimizado**

```
from sklearn.model_selection import RandomizedSearchCV
from sklearn.ensemble import RandomForestClassifier

# Definir el modelo
rf = RandomForestClassifier(random_state=42)
# Definir la grid de hiperparámetros para RandomizedSearch
param_grid_rf = {
    'n_estimators': [100, 200, 300],
    'max_depth': [10, 20, 30],
    'min_samples_split': [10, 20, 30],
    'min_samples_leaf': [5, 7, 10],
    'max_features': ['sqrt', 'log2'],
    'class_weight': ['balanced', {0: 1, 1: 2}],
    'criterion': ['gini', 'entropy'],
    'bootstrap': [True, False]
}
# Configurar la validación cruzada estratificada
skf = StratifiedKFold(n_splits=5)

# Configurar RandomizedSearchCV
rf_random_grid = RandomizedSearchCV(
    estimator=rf,
    param_distributions=param_grid_rf,
    cv=skf,
    n_jobs=-1,
    verbose=2
)

# Entrenar el modelo
rf_random_grid.fit(X_train_smote, y_train_smote)

# Hacer predicciones
y_pred_rf_tuned_prob = rf_random_grid.predict_proba(X_test)

threshold = 0.4  # Ajustar el umbral para la clase positiva
y_pred_adjusted = (y_pred_rf_tuned_prob[:, 1] >=
threshold).astype(int)

# Evaluar los resultados
print('Random Forest - Best Parameters: ',
rf_random_grid.best_estimator_)
```

- **Naive Bayes Optimizado**

```
from sklearn.naive_bayes import BernoulliNB

# Ajuste de hiperparámetros con una cuadrícula ampliada
param_grid = {
    'alpha': [0.0001, 0.001],
    'fit_prior': [True, False]
}

# Instanciar el modelo Bernoulli Naive Bayes
nb = BernoulliNB()

nb_cv = GridSearchCV(estimator=nb, param_grid=param_grid, cv=5,
n_jobs=-1, verbose=2)
nb_cv.fit(X_train_smote, y_train_smote)

# Obtener las probabilidades
y_pred_nb_tuned_prob = nb_cv.predict_proba(X_test)

# Ajustar el umbral
threshold = 0.2
y_pred_adjusted = (y_pred_nb_tuned_prob[:, 1] >=
threshold).astype(int)

# Imprimir los mejores parámetros y usar `y_pred_adjusted` para
análisis adicionales.
print('Naive Bayes - Best Parameters: ', nb_cv.best_params_)
```

- **Decision Trees Optimizado**

```
from sklearn.tree import DecisionTreeClassifier

# Ajuste de hiperparámetros
param_grid = {
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [5, 10, 20, 30],
    'min_samples_leaf': [2, 4, 6],
    'max_features': ['sqrt', 'log2', 0.9],
    'criterion': ['gini', 'entropy'],
    'splitter': ['best', 'random'],
    'max_leaf_nodes': [10, 20, 30]
}

# Instanciar el modelo con class_weight ajustado
dt = DecisionTreeClassifier(random_state=42,
class_weight='balanced')

# GridSearchCV para encontrar los mejores hiperparámetros
dt_cv = GridSearchCV(estimator=dt, param_grid=param_grid, cv=5,
n_jobs=-1, verbose=2)
dt_cv.fit(X_train_smote, y_train_smote)

# Hacer predicciones de probabilidad
y_pred_dt_tuned_prob = dt_cv.predict_proba(X_test)

# Ajustar el umbral de clasificación
threshold = 0.4
y_pred_adjusted = (y_pred_dt_tuned_prob[:, 1] >=
threshold).astype(int)

print('Decision Trees - Best Parameters: ', dt_cv.best_params_)
```

"UNIVERSIDAD NACIONAL DE MOQUEGUA"

VICERRECTORADO DE INVESTIGACIÓN

FORMATO DE AUTORIZACIÓN DE PUBLICACIÓN EN EL REPOSITORIO INSTITUCIONAL UNAM

Formulario de autorización para la publicación electrónica en la página web del Repositorio Institucional de la Universidad Nacional de Moquegua, de conformidad con la Ley N° 30035 Ley que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto y Resolución de Comisión Organizadora N° 041-2022-UNAM que aprueba el Reglamento que regula el Repositorio Institucional Digital de la UNAM.

1. DATOS DEL AUTOR

Nombres y Apellidos : Amparo Lisbeth Nuñez Galindo.....
DNI/Carnet de Extranjería/Pasaporte N° : 72240905.....
Correo electrónico : galindo.aln@gmail.com.....
Teléfono/Celular : 981595196.....

2. GRADO, TITULO O ESPECIALIZACIÓN (Marque con una X)

Bachiller	Título Profesional	Magister	Doctor	Segunda Especialidad
	X			

3. TIPO DE PUBLICACIÓN (Marque con una X)

a) Articulo ( )
b) Tesis de pregrado (X)
c) Tesis de Maestría ( )
d) Tesis de Doctorado ( )
e) Monografía ( )
f) Reporte de caso científico o hallazgos ( )
g) Libro ( )
h) Parte de un libro ( )
i) Revisión ( )
j) Conferencia ( )
k) Presentación ( )
l) Artículo preliminar ( )
m) Articulo antes de ser publicado ( )
n) Artículo de prensa ( )
o) Documentos internos ( )
p) Imagen ( )
q) Documentos técnicos ( )
r) Ponencias presentadas en eventos científicos (Jornada, Congreso, Simposio Conferencias, etc.) ( )
s) Informes de suficiencia profesional ( )
t) Patentes producidas por académicos/as ( )

4. TIPO DE TRABAJO DE INVESTIGACIÓN (Marque con una X)

a) Tesis **(X)**
b) Trabajo de Investigación **( )**
c) Trabajo de suficiencia profesional **( )**
d) Trabajo académico **( )**

5. DATOS DE LA PUBLICACION

a) Título: EVALUACION DE MODELOS PREDICTIVOS EMPLEANDO TÉCNICAS DE MACHINE LEARNING PARA EL PRONÓSTICO DE LA VIOLENCIA FISICA CONTRA LA MUJER EN LA PROVINCIA DE ILO, AÑO 2021
b) Campo del conocimiento OCDE (Indicar el código de clasificación OCDE de la obra: https://catalogos.concytec.gob.pe/vocabulario/ocde_ford.html): 2.00.00
c) Fecha de publicación: ..... diciembre de 2024.....

6. DATOS DEL ASESOR:

a) Nombres y Apellidos : Vaneza Flores Gutierrez .....
b) DNI/Pasaporte/Cédula : 40060249.....
c) ORCID : 0000-0001-7534-5810.....

FIRMA : .....

DOCUMENTO DE IDENTIDAD : 72240905.....
NOMBRES Y APELLIDOS : Amparo Lisbeth Nuñez Galindo.....

ANEXO VIII

DECLARACIÓN JURADA DE ORIGINALIDAD DE TRABAJO DE INVESTIGACIÓN / TESIS / TRABAJO DE SUFICIENCIA PROFESIONAL

Yo, Amparo Lisbeth Nuñez Galindo,
bachiller de la Escuela Profesional de:
Ingeniería de Sistemas e Informática,
de la Universidad Nacional de Moquegua, identificado(a) con DNI N°... 72240905, autor(a) del Trabajo de Investigación / Tesis / Trabajo de suficiencia profesional denominado:

EVALUACION DE MODELOS PREDICTIVOS EMPLEANDO TÉCNICAS DE MACHINE LEARNING PARA EL PRONOSTICO DE LA VIOLENCIA FISICA CONTRA LA MUJER EN LA PROVINCIA DE ILO, AÑO 2021

Declaro bajo juramento que:

1. El trabajo de investigación es auténtico, siendo resultado de mi trabajo personal, que no se ha copiado, que no se ha utilizado ideas, formulaciones, citas integrales e ilustraciones diversas, sacadas de cualquier tesis, obra, artículo, memoria, etc., (en versión digital o impresa), sin mencionar de forma clara y exacta su origen o autor, tanto en el cuerpo del texto, figuras, cuadros, tablas u otros que tengan derechos de autor. Por tanto, el trabajo de investigación no ha sido plagiado total ni parcialmente.
2. El trabajo de investigación no ha sido autoplagiada, es decir, no ha sido publicada ni presentada anteriormente para obtener algún grado académico previo o título profesional.
3. Los datos presentados en los resultados son reales, no han sido falseados, ni duplicados, ni copiados y por tanto los resultados que se presentan en el trabajo de investigación constituirán aporte a la realidad investigada.

Soy consciente de que, el hecho de no respetar los derechos de autor y hacer plagio, es objeto de sanciones universitarias y/o legales, por lo que asumo cualquier responsabilidad que pudiera derivarse de irregularidades en el trabajo de investigación, así como de los derechos sobre la obra presentada.

Moquegua, 5 de diciembre de 2024



Nombres y Apellidos: Amparo Lisbeth Nuñez Galindo
DNI: 72240905

"UNIVERSIDAD NACIONAL DE MOQUEGUA"

VICERRECTORADO DE INVESTIGACIÓN

LICENCIAS

Licencia Estándar

Bajo los siguientes términos, autorizo el depósito de mi trabajo de investigación en el Repositorio Institucional Digital de la UNAM.

Con la autorización de depósito da mi trabajo de investigación, otorgo a la Universidad Nacional de Moquegua, una Licencia No Exclusiva para reproducir. distribuir, comunicar al público, transformar (únicamente mediante su traducción a otros idiomas) y poner a disposición del público mi trabajo de investigación (incluido el resumen), en formato físico o digital, en cualquier medio, conocido o por conocerse, a través de los diversos servicios provistos por la Universidad, creados o por crearse, tales como el Repositorio Institucional Digital de la UNAM, Colección de Tesis, entre otros, en el Perú y en el extranjero, por el tiempo y veces que considere necesarias, y libre de remuneraciones.

En virtud de dicha Licencia, la UNIVERSIDAD NACIONAL DE MOQUEGUA, podrá reproducir mi trabajo de investigación en cualquier tipo de soporte y en más de un ejemplar, sin modificar su contenido, solo con los propósitos de seguridad, respaldo y preservación.

Declaro que el trabajo de investigación es una creación de mi autoría y exclusiva titularidad, o coautoría con titularidad compartida, y me encuentro facultado a conceder la presente licencia y, asimismo, garantizo que dicho trabajo de investigación no infringe derechos de autor de terceras personas.

La UNIVERSIDAD NACIONAL DE MOQUEGUA consignara el nombre del/los autor/es del trabajo de investigación, y no le hará ninguna modificación más de la permitida en la presente licencia.

Autorizo el depósito (marque con una X).

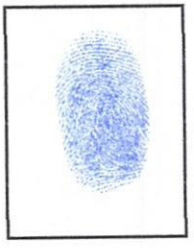
(**X**) **SI autorizo que se deposite inmediatamente.**

FIRMA : .....

DOCUMENTO DE IDENTIDAD : 72240905.....

NOMBRES Y APELLIDOS : Amparo Lisbeth Nuñez Galindo.....

FECHA : 05 de diciembre de 2024.....



AMPARO NUÑEZ

4-12-2024 VERSION_FINAL_TESIS_AMPARO.pdf

PROYECTOS ERICK SALDAÑA

My Files

Universidad Nacional De La Amazonia Peruana

Detalles del documento

Identificador de la entrega
trn:oid:::20208:412558131

Fecha de entrega
4 dic 2024, 10:24 a.m. GMT-5

Fecha de descarga
4 dic 2024, 10:36 a.m. GMT-5

Nombre de archivo
4-12-2024 VERSION_FINAL_TESIS_AMPARO.pdf

Tamaño de archivo
3.9 MB

117 Páginas

15,202 Palabras

92,604 Caracteres

turnitin Página 1 of 128 - Portada Identificador de la entrega trn:oid:::20208:412558131

17% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Coincidencias menores (menos de 10 palabras)

Exclusiones

- N.º de coincidencias excluidas

Fuentes principales

13% Fuentes de Internet
3% Publicaciones
13% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

DECENIO DE LA IGUALDAD DE OPORTUNIDADES PARA MUJERES Y HOMBRES
"AÑO DE LA RECUPERACIÓN Y CONSOLIDACIÓN DE LA ECONOMÍA PERUANA"

RESOLUCIÓN DE CONSEJO UNIVERSITARIO Nº 029-2025-CU-UNAM

Moquegua, 30 de enero de 2025

VISTOS:

El Acuerdo de Sesión Ordinaria de Consejo Universitario de fecha 30 de enero de 2025; Proveído N° 286-2025 de fecha 28 de enero de 2025 emitido por el Rectorado; Informe N° 011-2025-UNAM-R/SG de fecha de recepción 27 de enero de 2025 emitido por la Secretaría General; Carta N° 002-2025-UNAM/SEGE/GYT/SVR de fecha de recepción 23 de enero de 2025 emitido por la Especialista de Grados y Títulos; Sobre el otorgamiento del Título Profesional de Ingeniero de Sistemas e Informática, y;

CONSIDERANDO:

Que, la Constitución Política del Perú en su artículo 18º prevé que cada Universidad es autónoma en su régimen normativo, de gobierno, académico, administrativo y económico, y el artículo 8º de la Ley Nº 30220, Ley Universitaria, establece que la referida autonomía inherente a las Universidades, se ejerce de conformidad con la Constitución y las Leyes de la República e implica las organizaciones de su sistema académico, económico y administrativo, así como la administración de sus bienes y rentas, elaborar su presupuesto y aplicar sus fondos con la responsabilidad que impone la Ley;

Que, de conformidad con la Ley N° 28520 de fecha 23/05/2005, se creó la Universidad Nacional de Moquegua, cuyo artículo 1º dispone: "créase la Universidad Nacional de Moquegua como persona jurídica de derecho público interno, con sede en la ciudad de Moquegua, Provincia de Mariscal Nieto, Departamento de Moquegua". Asimismo, a través de la Resolución del Consejo Directivo N° 103-2017-SUNEDU/CD, el Consejo Directivo de la SUNEDU, le otorgó la Licencia Institucional a la Universidad Nacional de Moquegua; la misma que fue modificada a través de la Resolución del Consejo Directivo N° 135-2022-SUNEDU/CD;

Que, la Universidad Nacional de Moquegua, es una institución que brinda educación superior integral de calidad mediante una gestión académica y administrativa, centrada en los derechos humanos, con pertinencia social de las escuelas profesionales y posgrado, investigación científica y el ejercicio responsable del liderazgo universitario, vinculado al desarrollo de la región, el país y el mundo;

Que, en virtud a ello, mediante Resolución de Consejo Universitario Nº 001, 002 y 003-2025-CU-UNAM, de fecha 03 de enero de 2025, se oficializa la elección del Dr. Hugo Ruben Marca Maquera, como Rector, al Dr. Alejandro Manuel Ecos Espino como Vicerrector Académico y al Dr. Jhony Mayta Hancco como Vicerrector de Investigación de la Universidad Nacional de Moquegua;

Que, el artículo 58° de la Ley Universitaria, Ley Nº 30220, concordado con el artículo 14º del Estatuto de la Universidad Nacional de Moquegua, aprobado con Resolución de Comisión Organizadora Nº 578-2021-UNAM, establecen que el **Consejo Universitario es el máximo órgano de gestión, dirección y ejecución académica y administrativa de la universidad;**

Que, el artículo 44° de la Ley Nº 30220, Ley Universitaria, señala que las universidades otorgan los grados académicos de Bachiller, Maestro, Doctor y los Títulos Profesionales que correspondan, a nombre de la Nación (...);

Domicilio Fiscal: Prolongación Av., Ancash S/N - Moquegua
RUC: 20449347448 - Teléfono: (053) 670168
Página Web: www.unam.edu.pe

Que, el numeral 45.2 del artículo 45° de la Ley Nº 30220, Ley Universitaria, precisa **los requisitos mínimos para la obtención del título profesional, como son: el grado de Bachiller y la aprobación de una tesis** o trabajo de suficiencia profesional. Asimismo, establece que el título profesional solo se puede obtener en la universidad en la cual se haya obtenido el grado de bachiller;

Que, mediante Resolución de Comisión Organizadora Nº 1153-2021-UNAM, se aprobó el Reglamento de Grados y Títulos de la Universidad Nacional de Moquegua, el mismo que fue modificado a través de las Resoluciones de Comisión Organizadora N° 410; 453 y 890-2022, N° 880-2023-UNAM y N° 081, 484-2024-UNAM el cual establece los requisitos y norma los procedimientos para la obtención de los grados académicos y títulos profesionales conferidos por esta Casa Superior de Estudios;

Que, mediante la Carta N° 002-2025-UNAM/SEGE/GYT/SVR, el Especialista del Área de Grados y Títulos, remite a Secretaria General, la relación de tres (3) bachilleres de la Escuela Profesional de Ingeniería de Sistemas e Informática, quienes se encuentran aptos para la expedición del Título Profesional de Ingeniero de Sistemas e Informática, al haber cumplido satisfactoriamente con los requisitos legales y reglamentarios[1], exigidos en el Reglamento de Grados y Títulos de la Universidad Nacional de Moquegua, conforme lo señala la Dirección de Actividades y Servicios Académicos y la Dirección de la Escuela Profesional de Ingeniería de Sistemas e Informática, razón por la cual la Secretaría General incluye el expediente como punto de agenda para ser tratado en Sesión de Consejo Universitario de fecha 30 de enero de 2025;

Que, se ha establecido como atribución del Consejo Universitario de la UNAM, la facultad de conferir los grados académicos y los títulos profesionales (...), por lo que en mérito al pedido del Secretario General y debatido el caso en Sesión Ordinaria de fecha 30 de enero de 2025, se acordó por unanimidad **conferir a nombre de la Nación el Título Profesional de Ingeniero de Sistemas e Informática a tres (3) bachilleres de la Escuela Profesional de Ingeniero de Sistemas e Informática de la Facultad de Ingeniería y Arquitectura de la Universidad Nacional de Moquegua**;

[1] Artículo 7°. - Procedimiento y Trámite

Para obtener el Grado Académico, el egresado deberá iniciar el trámite presentando en su Escuela Profesional:

a) Solicitud dirigida al Director de la Escuela Profesional o quien haga sus veces, indicando la fecha y numero de comprobante de pago.
b) Dos (02) fotografías de frente, tamaño pasaporte, a color, con traje formal y fondo blanco: garantizando de esta forma la seguridad ciudadana.

La Dirección de Actividades y Servicios Académicos es la responsable de completar el expediente del egresado junto con las Escuela Profesional, en un plazo máximo de 7 días hábiles para su trámite correspondiente, con los siguientes documentos:

a) Constancia de Prácticas Preprofesionales expedido por el Director de la Escuela Profesional respectiva.
b) Certificado de estudios del idioma inglés o lengua nativa, nivel intermedio, expedido por el Centro de Idiomas de la Universidad Nacional de Moquegua o rendir el examen de suficiencia correspondiente en el Centro de Idiomas de la Universidad Nacional de Moquegua en caso de haber estudiado en otra institución.
c) Certificado de Estudios originales, debiendo acreditar el número de créditos exigidos por su Escuela Profesional.
d) Constancia de haber realizado Actividades Cocurriculares, expedida por la Dirección de Bienestar Universitario y la Dirección Proyección Social y Extensión Universitaria.
e) Constancia de "no adeudo" de Bienes a la Universidad, bienes a los Laboratorios Básicos y de Especialidad, material Bibliográfico; dichos documentos deberán ser refrendados por los responsables a cargo.
f) Copia simple de la Constancia de Egresado.
g) Copia simple de la Constancia de Inicio de Matricula.

La Dirección de Actividades y Servicios Académicos, previa verificación, deriva el expediente a la Oficina de Secretaría General para su trámite correspondiente.

Domicilio Fiscal: Prolongación Av., Ancash S/N - Moquegua
RUC: 20449347448 - Teléfono: (053) 670168
Página Web: www.unam.edu.pe

De conformidad con el artículo 59° de la Ley Universitaria, Ley Nº 30220, concordado con el artículo 16° del Estatuto de la Universidad Nacional de Moquegua, y estando a lo acordado en la I Sesión Ordinaria de fecha 30 de enero de 2025 de Consejo Universitario;

SE RESUELVE:

ARTÍCULO PRIMERO: CONFERIR, a nombre de la Nación el **TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS E INFORMÁTICA,** por la modalidad de Sustentación de Tesis, a los Bachilleres de la Escuela Profesional de Ingeniería de Sistemas de la Facultad de Ingeniería y Arquitecta de la Universidad Nacional de Moquegua, que se consigna en el listado siguiente:

1. **DIEGO HEDER PERCCA MAMANI**

   TESIS : "RECONOCIMIENTO DE EXPRESIONES FACIALES Y LA POSTURA DE LA CABEZA USANDO MODELOS DE DEEP LEARNING", cuyo Asesor fue el Dr. Víctor Yana Mamani.

2. **AMPARO LISBETH NUÑEZ GALINDO**

   TESIS : "EVALUACIÓN DE MODELOS PREDICTIVOS EMPLEANDO TÉCNICAS DE MACHINE LEARNIG PARA EL PRONÓSTICO DE LA VIOLENCIA FÍSICA CONTRA LA MUJER EN LA PROVINCIA DE ILO, AÑO 2021", cuyo Asesor fue el Dra. Vaneza Flores Gutierrez.

3. **GABRIELA KARINA VELASQUEZ QUISPE**

   TESIS : "RECONOCIMIENTO DE EXPRESIONES FACIALES Y LA POSTURA DE LA CABEZA USANDO MODELOS DE DEEP LEARNING", cuyo Asesor fue el Dr. Víctor Yana Mamani.

ARTÍCULO SEGUNDO: AUTORIZAR, a la Secretaría General de la Universidad Nacional de Moquegua, expedir y registrar los Diplomas correspondientes para su otorgamiento.

ARTÍCULO TERCERO: ENCARGAR al Área de Grados y Títulos de la Universidad Nacional de Moquegua, para que realice las acciones pertinentes conducentes, relacionadas con la inscripción de los grados y títulos en el Registro de Grados y Títulos de la Superintendencia Nacional de Educación Superior Universitaria - SUNEDU.

ARTÍCULO CUARTO: NOTIFICAR, de la presente Resolución a todas las instancias y dependencias administrativas y académicas de la Universidad Nacional de Moquegua, para su conocimiento, cumplimiento y demás fines pertinentes.

ARTÍCULO QUINTO: ENCARGAR, a la Oficina de Tecnologías de la Información, la publicación de la presente Resolución en el portal de transparencia de la Universidad Nacional de Moquegua.

REGÍSTRESE, COMUNÍQUESE, CÚMPLASE Y ARCHÍVESE.

DR. HUGO RUBEN MARCA MAQUERA
Rector
Universidad Nacional de Moquegua

ABG. GUILLERMO S. KUONG CORNEJO
Secretario General
Universidad Nacional de Moquegua

Domicilio Fiscal: Prolongación Av., Ancash S/N - Moquegua
RUC: 20449347448 - Teléfono: (053) 670168
Página Web: www.unam.edu.pe