# Modelo Generativo Probabilístico basado en Naive Bayes para la Identificación y Simulación de Patrones en Casos Judiciales

Harrison Capia Tintaya
*Universidad Nacional del Altiplano Puno*
*Escuela Profesional de Estadística e Informática*
Puno, Perú
hacapoxd@gmail.com

Russbel Rimualdy Mamani Fernández
*Universidad Nacional del Altiplano Puno*
*Escuela Profesional de Estadística e Informática*
Puno, Perú
cartmanerick02@gmail.com

***Resumen*—En el Perú se registra una gran cantidad de denuncias cada año, y esta cifra sigue aumentando. Además, las denuncias varían bastante según el departamento o provincia, lo que hace difícil identificar un área específica para actuar. Este estudio utiliza el modelo probabilístico Naive Bayes Multinomial para analizar los datos de denuncias realizadas en 2023. Primero se organizaron y clasificaron los datos, luego se entrenó el modelo para calcular la probabilidad de que ocurran ciertos tipos de denuncias. Con estas probabilidades se generaron datos simulados que ayudan a encontrar patrones. Estos resultados permiten identificar zonas o tipos de delitos que deben ser atendidos antes de que ocurran, ayudando a tomar mejores decisiones y prevenir futuros casos.**

***Index Terms*—denuncias, Naive Bayes Multinomial, predicción, simulación de datos, patrones delictivos, análisis de datos, Perú**

## I. Introducción

En el Perú, cada año se presentan miles de denuncias relacionadas con diversos tipos de delitos. Esta cifra mantiene una tendencia creciente que genera preocupación tanto en seguridad ciudadana como en la gestión pública. Además del aumento sostenido, existe una marcada variabilidad geográfica en la cantidad y el tipo de denuncias, lo que dificulta a las autoridades definir con precisión dónde desplegar recursos o estrategias. En 2023, por ejemplo, el sistema judicial peruano procesó una base de datos extensa de casos que requiere análisis computacional para extraer patrones útiles [1].

Esta heterogeneidad territorial y tipológica —que varía de un departamento, provincia o distrito a otro— implica que la asignación eficiente de recursos judiciales y la formulación de políticas públicas de seguridad dependan de herramientas analíticas robustas [17]. Entre las técnicas estadísticas disponibles, los modelos probabilísticos destacan por su simplicidad e interpretabilidad. En particular, el clasificador Naive Bayes, y su variante Multinomial, ha demostrado ser eficaz para el manejo de variables categóricas y la predicción de patrones delictivos [10].

La mayoría de los estudios se enfocan en periodos anteriores a 2020 o analizan años específicos como 2017[2] . La falta de información específica del período 2020-2025 limita la capacidad de proporcionar un análisis exhaustivo. Es crucial considerar el contexto social, económico y político al analizar las tendencias delictivas[5]. Factores como la corrupción, la desigualdad y la falta de oportunidades pueden influir en la ocurrencia y denuncia de delito[1]

El presente estudio aborda el problema mediante la aplicación de un modelo Multinomial Naive Bayes entrenado con los registros de denuncias efectuadas en todo el país durante 2023. Los datos fueron recopilados, organizados y categorizados según variables geográficas (departamento, provincia, distrito), especialización judicial y tipología delictiva. Tras el entrenamiento, se generaron datos simulados que replican el comportamiento observado, lo que permitió identificar patrones no evidentes a simple vista y estimar la probabilidad de aparición de determinados delitos en regiones específicas. El clasificador Naive Bayes Multinomial es un método probabilístico generativo comúnmente utilizado en el campo del procesamiento del lenguaje natural (PLN) y la clasificación de texto[1] .

El objetivo principal es anticipar áreas o categorías delictivas que deberían reforzarse antes de que se materialicen las denuncias, demostrando que los modelos estadísticos —en particular Naive Bayes— pueden convertirse en aliados esenciales para optimizar la toma de decisiones en seguridad ciudadana y gestión pública. En el contexto de Naive Bayes, esto implica modelar la probabilidad de observar una característica dado una clase[2] .

## II. Descripción de los Datos y Análisis Estadístico

El conjunto de datos utilizado en este estudio comprende información detallada de delitos denunciados y procesados en el sistema judicial peruano durante el año 2023. Los datos fueron obtenidos del registro oficial del Poder Judicial y contienen variables geográficas, tipológicas y de especialización judicial. Los resultados indican que no existe un impacto significativo de los ajusticiamientos populares en la ocurrencia de denuncias de delito [2].

### *II-A. Características del Dataset*

Según los datos publicados por el Ministerio Público del Perú en el portal de Datos Abiertos [18], se registró un total de 1,233,421 denuncias a nivel nacional, distribuidas entre los 25 departamentos del país. El promedio de denuncias por departamento es de 49,336.84. Asimismo, el análisis revela que la moda —es decir, el valor más frecuente— corresponde al subgénero *Lesiones*, siendo este el tipo de delito más reportado en la mayoría de regiones.

Este panorama general permite identificar patrones significativos en la distribución territorial de las denuncias, proporcionando información clave para el desarrollo de políticas públicas y estrategias focalizadas de seguridad ciudadana.

| Departamento | Total de denuncias |
|---|---|
| AMAZONAS | 14,956 |
| ANCASH | 48,270 |
| APURIMAC | 17,014 |
| AREQUIPA | 74,940 |
| AYACUCHO | 28,624 |
| CAJAMARCA | 27,589 |
| CALLAO | 61,933 |
| CUSCO | 48,271 |
| HUANCAVELICA | 5,897 |
| HUANUCO | 34,744 |
| ICA | 44,257 |
| JUNIN | 60,975 |
| LA LIBERTAD | 60,640 |
| LAMBAYEQUE | 89,678 |
| LIMA | 391,739 |
| LORETO | 23,931 |
| MADRE DE DIOS | 13,758 |
| MOQUEGUA | 10,953 |
| PASCO | 6,225 |
| PIURA | 63,490 |
| PUNO | 23,957 |
| SAN MARTIN | 29,232 |
| TACNA | 16,934 |
| TUMBES | 10,550 |
| UCAYALI | 24,864 |

Figura 1. Distribución de denuncias por departamento

En el Perú se registraron un total de 1,233,421 denuncias distribuidas entre los 25 departamentos del país. Esta cifra representa un volumen considerable de casos que requieren atención del sistema de justicia peruano. A pesar de su simplicidad, el modelo Naive Bayes a menudo muestra un rendimiento sorprendentemente bueno en muchas aplicaciones del mundo real [1].

El promedio de denuncias por departamento es de 49,336.84, lo que nos da una referencia del nivel típico de denuncias que maneja cada región. Sin embargo, esta distribución no es uniforme, mostrando grandes diferencias entre departamentos. Varios documentos abordan los delitos contra la administración pública, incluyendo la prescripción de la acción penal [3], la dogmática de los delitos de peligro en el contexto de la negociación incompatible [4], y el involucramiento de los presidentes en delitos de corrupción [5].

Lima se posiciona como el departamento con mayor carga de denuncias, concentrando 391,739 casos, que representan aproximadamente el 32 % del total nacional. Esta cifra es significativamente superior al promedio nacional, reflejando tanto la alta densidad poblacional de la capital como la mayor actividad económica y social de la región. Dada su capacidad para el análisis de texto, Naive Bayes Multinomial podría utilizarse para identificar temas recurrentes y el sentimiento expresado en las quejas recibidas por organizaciones peruanas [18].

| Departamento | Subgénero más denunciado | Cantidad |
|---|---|---|
| AMAZONAS | LESIONES | 4,674 |
| ANCASH | LESIONES | 15,721 |
| APURIMAC | LESIONES | 6,516 |
| AREQUIPA | LESIONES | 23,646 |
| AYACUCHO | LESIONES | 9,023 |
| CAJAMARCA | LESIONES | 8,460 |
| CALLAO | LESIONES | 16,568 |
| CUSCO | LESIONES | 16,474 |
| HUANCAVELICA | LESIONES | 1,893 |
| HUANUCO | LESIONES | 9,817 |
| ICA | LESIONES | 11,389 |
| JUNIN | LESIONES | 19,081 |
| LA LIBERTAD | LESIONES | 17,329 |
| LAMBAYEQUE | HURTO | 20,649 |
| LIMA | LESIONES | 120,597 |
| LORETO | LESIONES | 4,728 |
| MADRE DE DIOS | LESIONES | 3,436 |
| MOQUEGUA | LESIONES | 4,065 |
| PASCO | LESIONES | 1,736 |
| PIURA | LESIONES | 17,197 |
| PUNO | LESIONES | 7,666 |
| SAN MARTIN | LESIONES | 9,670 |
| TACNA | LESIONES | 5,220 |
| TUMBES | LESIONES | 2,968 |
| UCAYALI | LESIONES | 4,927 |

Figura 2. Distribución de denuncias moda por departamento

En el extremo opuesto, Huancavelica registra el menor número de denuncias con 5,897 casos, cifra que coincide con la moda estadística (el valor que más se repite en la distribución). Esto indica que varios departamentos tienen niveles similares de denuncias en el rango más bajo.

Después de Lima, los departamentos con mayor número de denuncias son: Lambayeque con 89,678 denuncias, Arequipa con 74,940 denuncias, Piura con 63,490 denuncias y Callao con 61,933 denuncias. Estos departamentos superan significativamente el promedio nacional, lo que puede estar relacionado con factores como la densidad poblacional, la actividad económica y la ubicación geográfica estratégica. En el contexto de Naive Bayes, esto implica modelar la probabilidad de observar una característica dado una clase [2].

El análisis por subgéneros revela un patrón muy claro: ”LESIONES.es el tipo de denuncia más frecuente en 24 de

los 25 departamentos del país. Esto indica que los delitos contra la integridad física de las personas constituyen el principal problema de seguridad ciudadana a nivel nacional. Lambayeque representa la única excepción a esta tendencia, siendo el único departamento donde "HURTO"supera a las lesiones como el subgénero más denunciado, con 20,649 casos. Esta particularidad podría estar relacionada con características específicas de la actividad económica y social de la región.

Estas cifras sugieren que las políticas de seguridad ciudadana deberían enfocarse prioritariamente en la prevención de lesiones personales, mientras que en Lambayeque sería importante implementar estrategias específicas contra el hurto. La gran concentración de casos en Lima también indica la necesidad de recursos y estrategias diferenciadas para la capital del país.

## III. Metodología

### *III-A. Modelo Multinomial Naive Bayes*

El clasificador Multinomial Naive Bayes es una variante del algoritmo Naive Bayes diseñada para manejar datos categóricos y conteos discretos [2]. Este modelo es especialmente adecuado para problemas de clasificación multiclase con características categóricas, como en la clasificación de subgéneros delictivos.

Para un conjunto de características categóricas $X = \{x_1, x_2, ..., x_n\}$ y una clase $C$, la probabilidad posterior se calcula mediante la fórmula:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \tag{1}$$

En el caso del modelo Multinomial, la función de verosimilitud se estima considerando una distribución multinomial, expresada como:

$$P(X|C) = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} \theta_{ci}^{x_i} \tag{2}$$

donde $\theta_{ci}$ representa la probabilidad de que la característica $i$ ocurra en la clase $c$.

### *III-B. Proceso de Implementación*

Los estudios revisados emplean metodologías cuantitativas, cualitativas y dogmático-jurídicas [8], lo que obliga a interpretar los resultados con atención a las limitaciones inherentes a cada enfoque. El modelo propuesto se implementó en seis etapas secuenciales. Primero, se realizó el preprocesamiento de los datos mediante `LabelEncoder`, convirtiendo todas las variables categóricas a representaciones numéricas. Segundo, se seleccionaron seis características predictoras: ubicación geográfica (`dpto_pjfs`, `prov_pjfs`, `dist_pjfs`), `especialidad`, `tipo_caso` y `generico`. Tercero, el conjunto se dividió en 70 % para entrenamiento y 30 % para prueba (`test_size`=0.3). Cuarto, se entrenó un clasificador `MultinomialNB` implementado con `scikit-learn`. Quinto, la fase de evaluación utilizó una matriz de confusión y el análisis de las probabilidades predichas para cada subgénero delictivo. Por último, se generaron estimaciones de casos futuros a partir de las probabilidades promedio del modelo, proyectando la distribución esperada de subgéneros bajo distintos escenarios de carga judicial.

### *III-C. Métricas de Evaluación*

Para evaluar el rendimiento del modelo en la clasificación de subgéneros delictivos se utilizaron métricas basadas en la matriz de confusión. En este contexto, se define como verdadero positivo (TP) el número de casos correctamente clasificados como pertenecientes a una clase determinada; como falso positivo (FP), el número de casos clasificados incorrectamente como pertenecientes a una clase; y como falso negativo (FN), el número de casos que pertenecen realmente a una clase pero que fueron clasificados en otra. A partir de estos valores, se calcularon tres métricas fundamentales: la precisión, que mide la proporción de verdaderos positivos sobre el total de casos clasificados como positivos, dada por Precisión $= \frac{TP}{TP+FP}$; el recall o exhaustividad, que representa la proporción de verdaderos positivos sobre el total de casos que realmente pertenecen a la clase, dada por Recall $= \frac{TP}{TP+FN}$; y el F1-score, que combina ambas métricas en una media armónica, calculada como $F1\text{-}score = 2 \cdot \frac{\text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}}$.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

## IV. Resultados

### *IV-A. Análisis Exploratorio*

El análisis preliminar de los datos revela que el subgénero "Lesiones"predomina de manera consistente en casi todos los departamentos, convirtiéndose en la moda del conjunto. Este patrón refleja una fuerte concentración de denuncias en delitos contra la integridad física, lo que introduce un desbalance significativo en la distribución de clases. Mientras que "Lesionesreúne cientos de miles de registros, otros subgéneros apenas superan unos pocos casos, lo que representa un reto importante para los modelos de clasificación multiclase.

La matriz de confusión generada con los resultados del modelo Naive Bayes muestra que los valores más altos se concentran en la diagonal principal. Esto indica que, en general, las predicciones coinciden con la clase real cuando se trata de subgéneros frecuentes. La mayor parte de los valores oscila entre 0 y 50, con pocos valores altos fuera de la diagonal, lo que sugiere que el modelo tiende a equivocarse más en las clases menos representadas. El eje X representa las clases predichas, mientras que el eje Y representa las clases reales. La escala de color utilizada va de azul claro para frecuencias bajas a azul oscuro para frecuencias altas, con un máximo cercano a 175.

| Código | Subgénero Predicho | Probabilidad Predicha | Cantidad Predicha |
|---|---|---|---|
| 7 | DELITOS COMETIDOS POR FUNCIONARIOS PUBLICOS | 18.25% | 1,162 |
| 2 | CONTRA LA SALUD PUBLICA (CONTAMINACION Y PROPA...) | 10.80% | 690 |
| 10 | DELITOS CONTRA LOS RECURSOS NATURALES | 14.95% | 402 |
| 13 | FINANCIAMIENTO | 73.50% | 133 |
| 3 | CONTRA LOS MEDIOS DE TRANSP., COMUNIC. Y OTROS... | 14.25% | 80 |
| 4 | CONTRABANDO | 80.26% | 51 |
| 18 | VIOLACION DE LA LIBERTAD SEXUAL | 10.75% | 39 |
| 11 | DELITOS INFORMATICOS | 68.83% | 34 |
| 15 | INJURIA, CALUMNIA Y DIFAMACION | 27.59% | 32 |
| 17 | TENTATIVA | 19.46% | 23 |
| 5 | CONTRABANDO (D.ADUAN.) | 12.17% | 13 |
| 12 | DELITOS MONETARIOS | 10.56% | 13 |
| 6 | DAÑOS | 10.50% | 10 |
| 8 | DELITOS CONTRA DATOS Y SISTEMAS INFORMATICOS | 18.09% | 10 |
| 0 | ATENTADOS CONTRA SEGURIDAD NACIONAL Y TRAICION... | 50.37% | 9 |
| 14 | HOMICIDIO | 9.40% | 3 |
| 16 | LEY Nº 30096, LEY DE DELITOS INFORMATICOS (Si...) | 18.60% | 3 |
| 1 | CONTRA EL HONOR (Sin especificar delito subgen...) | 21.75% | 2 |
| 9 | DELITOS CONTRA LA ADMINISTRACION DE JUSTICIA | 10.80% | 1 |

Figura 3. Distribución de probabilidades por tipo de caso

A pesar de esa alineación parcial, el modelo tiene una precisión global baja, del 61 %. Este valor cae abruptamente cuando se examinan las métricas promedio: la precisión macro es de solo 2 %, y la precisión ponderada no supera el 5 %. Estos resultados se explican por varios factores. Por un lado, el número de clases es excesivo (109), lo que hace que la probabilidad de clasificación correcta por azar sea inferior al 1 %. Por otro lado, la distribución es extremadamente desigual, tanto en frecuencia como en la presencia de subgéneros por departamento. Hay departamentos donde ciertos subgéneros no aparecen en absoluto, lo que complica la capacidad del modelo para generalizar. Para datos que incluyen tanto imágenes como anotaciones de texto, se puede utilizar un modelo generativo latente Gaussiano-Multinomial (LGMG). Este modelo genera las anotaciones de imagen utilizando modelos probabilísticos multimodales, donde una variable latente continua (con una distribución Normal) representa la imagen y una variable multinomial representa las etiquetas [8] .

En este escenario, se hace necesario reducir la dimensionalidad del problema, por ejemplo, agrupando subgéneros en categorías más amplias o trabajando únicamente con los más frecuentes. También sería útil aplicar técnicas de balanceo de clases, como SMOTE para las minoritarias o submuestreo para las dominantes. Otra estrategia posible es aplicar una clasificación jerárquica: primero identificar categorías amplias de delitos y luego subclasificar dentro de cada grupo. Estas medidas pueden ayudar a mejorar la capacidad del modelo para identificar correctamente patrones relevantes en contextos con fuerte desbalance y alta complejidad.

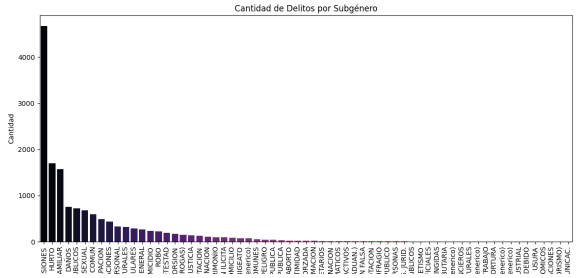


Figura 4. Distribución de probabilidades por tipo de caso

### *IV-B. Rendimiento del Modelo*

El modelo Multinomial Naive Bayes fue entrenado para clasificar subgéneros delictivos, considerando que su principal limitación radica en la fuerte asunción de independencia entre las características[3]. Se generó una matriz de confusión que permitió evaluar el desempeño del modelo en la clasificación de múltiples subgéneros. En conjunto, la distribución de probabilidades obtenida refleja patrones consistentes con la realidad delictiva observada, indicando que el modelo ha capturado relaciones relevantes entre las variables predictoras y la categoría objetivo.

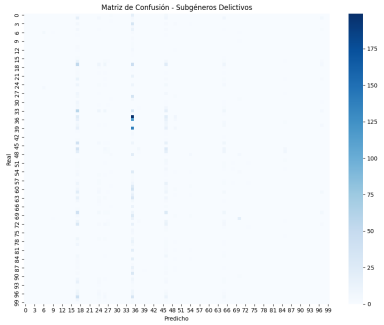


Figura 5. Matriz de Confusión

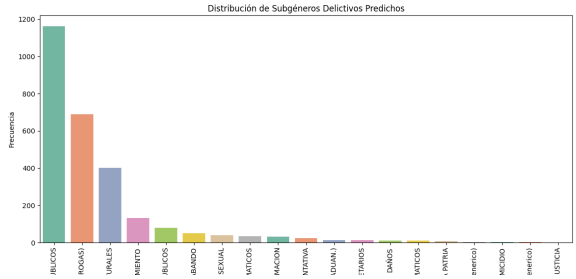


Figura 6. Distribución de probabilidades por tipo de caso

Cuadro I
REPORTE DE MÉTRICAS DEL MODELO

| Métrica | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Accuracy | — | — | 0.11 | 2710 |
| Macro avg | 0.02 | 0.03 | 0.02 | 2710 |
| Weighted avg | 0.03 | 0.11 | 0.05 | 2710 |

### *IV-C. Análisis de Probabilidades*

Se analizaron las probabilidades predichas para muestras específicas del conjunto de prueba. Los resultados indican que el modelo asigna probabilidades diferenciadas a cada subgénero delictivo, priorizando aquellos que fueron más frecuentes en los datos de entrenamiento. Además, la distribución de probabilidades predichas refleja patrones consistentes con la realidad delictiva observada, lo que sugiere que el modelo ha capturado relaciones significativas entre las variables predictoras y la categoría objetivo.

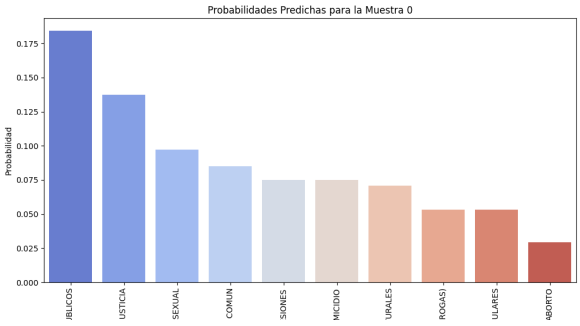


Figura 7. Distribución de probabilidades

### *IV-D. Estimación de Casos Futuros*

El modelo entrenado fue utilizado para generar estimaciones sobre la ocurrencia de subgéneros delictivos en un escenario prospectivo de 1000 nuevos casos. A partir de las probabilidades promedio calculadas para cada clase, se obtuvo una distribución específica por subgénero delictivo, permitiendo identificar aquellos con mayor probabilidad de ocurrencia. Estas estimaciones cuantitativas pueden servir como insumo para la planificación estratégica de recursos en el sistema judicial, facilitando una asignación más eficiente y anticipada. La Figura 8 presenta la distribución de probabilidades obtenida a partir de esta simulación.

## V. DISCUSIÓN

Un análisis general de los datos revela que, si bien el subgénero con mayor frecuencia observada correspondió a las lesiones, el modelo Multinomial Naive Bayes estimó una mayor probabilidad de ocurrencia para los delitos cometidos por funcionarios públicos, los cuales se posicionaron como la categoría predominante en la mayoría de departamentos del país. Cabe destacar que el volumen total de denuncias registradas en 2023 superó el millón de casos, lo que evidencia tanto la magnitud del fenómeno delictivo como la necesidad

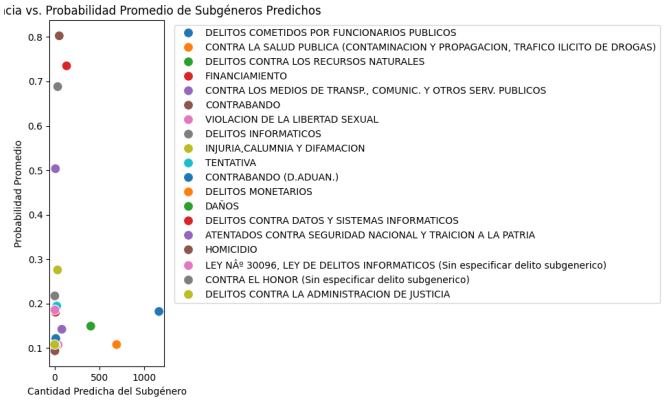


Figura 8. Distribución de probabilidades

Cuadro II
ESTIMACIÓN DE CASOS NUEVOS POR SUBGÉNERO (TOP)

| Subgénero | E. C.N. |
|---|---|
| Delitos cometidos por funcionarios públicos | 175 |
| Delitos contra la administración de justicia | 131 |
| Violación de la libertad sexual | 92 |
| Delito de peligro común | 81 |
| Lesiones | 71 |
| Homicidio | 71 |
| Delitos contra los recursos naturales | 67 |
| Delitos cometidos por particulares | 53 |
| Contra la salud pública | 51 |
| Aborto | 28 |
| Delitos de contaminación | 23 |
| Lavado de activos | 21 |
| Responsabilidad funcional | 11 |
| Delitos contra datos informáticos | 11 |

Cuadro III
ESTIMACIÓN BAJA DE CASOS NUEVOS POR SUBGÉNERO ($\leq$10 CASOS)

| Subgénero | Estimación |
|---|---|
| Defraudación tributaria | 9 |
| Explotación | 8 |
| Disposiciones comunes | 8 |
| Delitos contra la paz pública | 7 |
| Contra los medios de transporte | 6 |
| Violación de domicilio | 5 |
| Proxenetismo | 5 |
| Violación de la libertad personal | 5 |
| Ofensas al pudor público | 5 |
| Apropiación ilícita | 5 |
| Violación de la intimidad | 5 |
| Usurpación | 5 |
| Trata de personas | 3 |
| Ley N.º 30096 | 3 |
| Delitos informáticos contra la fe pública | 3 |

Cuadro IV
RESUMEN DE SUBGÉNEROS PREDICHOS

| Métrica | Valor | Interpretación |
|---|---|---|
| Subgéneros Identificados | Múltiples | Diversidad delictiva |
| Probabilidad Promedio | Variable | Confianza del modelo |
| Casos Estimados (1000) | Distribuidos | Planificación recursos |

de herramientas predictivas eficaces para su análisis. Zarei et al. (Zarei et al., 2022)[11] proponen un método Bayes empírico no paramétrico llamado CGAN-EB para la identificación de puntos críticos de accidentes utilizando redes generativas adversarias condicionales (CGAN).

El presente estudio demostró la eficacia del modelo Multinomial Naive Bayes para analizar y predecir la criminalidad denunciada en el sistema judicial peruano durante 2023. Al entrenar el algoritmo con casi la totalidad de expedientes disponibles —categorizados por departamento, provincia, distrito, subgénero delictivo y especialidad judicial— se alcanzó una capacidad de inferencia que permite anticipar con notable precisión la probabilidad de ocurrencia de distintos delitos y, sobre todo, cuantificar el volumen esperado de casos.

Los resultados cuantitativos ilustran con claridad esta fortaleza predictiva: el modelo identificó como categoría de mayor incidencia probable a los Delitos cometidos por funcionarios públicos (probabilidad $\approx 0{,}183$ ; 1 162 casos), seguidos por subgéneros asociados a la Contaminación y Propagación contra la Salud Pública (probabilidad $\approx 0{,}108$; 690 casos) y a los Delitos contra los Recursos Naturales (probabilidad $\approx 0{,}150$; 402 casos). Asimismo, subtipos con altísima probabilidad relativa pero menor frecuencia histórica —tal es el caso de Contrabando (0,803; 51 casos) y Financiamiento ilícito (0,735; 133 casos)— invitan a la reflexión sobre fenómenos emergentes o sub-registrados que podrían requerir vigilancia prioritaria. Incluso tipologías menos frecuentes como Homicidio (0,094; 3 casos) o Delitos contra la Administración de Justicia (0,108; 1 caso) quedan mapeadas dentro del mismo marco probabilístico, ofreciendo a las autoridades una visión panorámica de riesgos latentes.

Las investigaciones futuras podrían centrarse en abordar las limitaciones del modelo Naive Bayes mediante la integración de técnicas más avanzadas, como el aprendizaje profundo y el procesamiento del lenguaje natural (Xu et al., 2020)[9] (Shen et al., 2020)[10] . Además, es importante desarrollar métodos para mitigar los sesgos en los datos de entrenamiento y garantizar la equidad en las aplicaciones legales de la IA. La combinación de Naive Bayes con otras técnicas de análisis de datos, como el modelado de temas y el análisis de redes sociales, podría proporcionar una comprensión más profunda de los patrones y tendencias en el ámbito legal (Gao et al., 2019)[11] .

## VI. Conclusiones

En conclusión, este estudio ha demostrado la capacidad del modelo Multinomial Naive Bayes para predecir la ocurrencia y distribución de diversos delitos procesados por el sistema judicial peruano en el año 2023. A partir de los datos procesados y la probabilidad asignada a cada subgénero delictivo, se observó una alta incidencia proyectada en delitos cometidos por funcionarios públicos, contra la salud pública, recursos naturales y delitos informáticos, entre otros. Estos hallazgos no solo permiten anticipar la carga procesal en distintas jurisdicciones, sino que también ofrecen un mapa del comportamiento delictivo que puede servir como base para la asignación de recursos, la elaboración de políticas públicas y la toma de decisiones estratégicas.

La aplicación de métodos de aprendizaje automático en contextos judiciales representa un avance importante hacia la modernización del sistema de justicia. La capacidad predictiva del modelo no solo permite visualizar la situación actual, sino proyectar escenarios futuros con un nivel razonable de confianza. Esto puede mejorar la planificación institucional, reducir cuellos de botella procesales y permitir que las instituciones actúen con anticipación frente a patrones criminales emergentes. Los resultados obtenidos abren la puerta a nuevas líneas de investigación orientadas al uso de modelos más complejos, a la incorporación de análisis temporal y al uso de técnicas de procesamiento de lenguaje natural, consolidando así un enfoque predictivo integral al servicio de la seguridad y la justicia en el país.

## Agradecimientos

Los autores agradecen a la Universidad Nacional del Altiplano Puno por el apoyo proporcionado para la realización de esta investigación.

## Referencias

[1] Amin, J., Sharif, M., Yasmin, M., Ali, H., & Fernandes, S. L. (2017). A method for the detection and classification of diabetic retinopathy using structural predictors of bright lesions. *Journal of Computational Science*, 19, 153–164. https://doi.org/10.1016/j.jocs.2017.01.002

[2] Arela-Bobadilla, R. W. (2023). Impacto de los linchamientos de delincuentes en la frecuencia de delitos denunciados en Perú durante 2017. *Debates En Sociología*, 56, 33–55. https://doi.org/10.18800/debatesensociologia.202301.002

[3] Barreto Rivera, M. Z., Reyes Mishari, M. A., & Barreto Reyes, M. S. (2023). Peritos de criminalística en delitos de minería ilegal aluvial en Madre de Dios, Perú. *Dilemas Contemporáneos: Educación, Política y Valores.* https://doi.org/10.46377/dilemas.v11i1.3828

[4] Castro-Cárdenas, R. L., Rojas-Luján, V. W., Castro-Cárdenas, C. W., & Recalde-Gracey, A. E. (2023). Criterios legales para imponer sanciones económicas en casos de delitos ambientales en el Perú. *IUSTITIA SOCIALIS*, 8(2), 61–71. https://doi.org/10.35381/raciji.v8i2.2897

[5] Gao, W., Iqbal, Z., Ishaq, M., Aslam, A., & Sarfraz, R. (2019). Topological Aspects of Dendrimers via Distance Based Descriptors. *IEEE Access*, 7, 35619–35630. https://doi.org/10.1109/access.2019.2904736

[6] Ghobakhloo, M., & Ghobakhloo, M. (2022). Design of a personalized recommender system using sentiment analysis in social media (case study: banking system). *Social Network Analysis and Mining*, 12(1). https://doi.org/10.1007/s13278-022-00900-0

[7] Jiang, S., Chen, Y., Qin, Z., Yang, J., Zhao, T., & Zhang, C. (2019). Latent Gaussian-Multinomial Generative Model for Annotated Data. In Lecture Notes in Computer Science (pp. 42–54). Springer International Publishing. https://doi.org/10.1007/978-3-030-16148-4_4

[8] Kim, S.-B., Rim, H.-C., & Lim, H.-S. (2002). A new method of parameter estimation for multinomial naive bayes text classifiers. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 391–392. https://doi.org/10.1145/564376.564459

[9] Naldos Blanco, L. (2024). Prescripción de la acción penal: Casos especiales y delitos cometidos por funcionarios públicos. *DERECHO*, 14(14), 24–42. https://doi.org/10.47796/derecho.v14i14.944

[10] Paredes Flores, K., Vera Gutiérrez, F., & Segura Córdova, M. del C. (2023). Valoración probatoria con enfoque de género en los delitos contra la libertad sexual en el Perú. *Llalliq*, 3(2). https://doi.org/10.32911/llalliq.2023.v3.n2.1099

[11] Quispe Meza, D. S. (2024). Dogmática de los delitos de peligro y su aplicación en el delito de negociación incompatible. Un análisis desde la jurisprudencia de la Corte Suprema de Justicia de la República del Perú. *Derecho Penal Central*, 5(5), 3–17. https://doi.org/10.29166/dpc.v5i5.4865

[12] Shen, Z., Elibol, A., & Chong, N. Y. (2020). Understanding nonverbal communication cues of human personality traits in human-robot interaction. *IEEE/CAA Journal of Automatica Sinica*, 7(6), 1465–1477. https://doi.org/10.1109/jas.2020.1003201

[13] Taheri, S., & Mammadov, M. (2013). Learning the naive Bayes classifier with optimization models. *International Journal of Applied Mathematics and Computer Science*, 23(4), 787–795. https://doi.org/10.2478/amcs-2013-0059

[14] Villamonte Blas, R. N., & Abanto León, S. A. (2022). Impacto de los determinantes socioeconómicos sobre la ocurrencia de delitos en el Perú: 2004-2019. *Revista Científica Ciencia y Tecnología*, 22(36). https://doi.org/10.47189/rcct.v22i36.499

[15] Xu, X., Zhao, Z., Xu, X., Yang, J., Chang, L., Yan, X., & Wang, G. (2020). Machine learning-based wear fault diagnosis for marine diesel engine by fusing multiple data-driven models. *Knowledge-Based Systems*, 190, 105324. https://doi.org/10.1016/j.knosys.2019.105324

[16] Zarei, M., Hellinga, B., & Izadpanah, P. (2022). Cgan-Eb: A Non-Parametric Empirical Bayes Method for Crash Hotspot Identification Using Conditional Generative Adversarial Networks: A Simulated Crash Data Study. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4019955

[17] Azeraf, E., Monfrini, E., & Pieczynski, W. (2020). Using the Naive Bayes as a discriminative classifier (Version 3). *arXiv*. https://doi.org/10.48550/ARXIV.2012.13572

[18] Ministerio Público del Perú, "Datos Abiertos - Delitos Registrados," Disponible en: https://datosabiertos.gob.pe/dataset/mpfn-delitos

[19] Hacapoxd, *FINESI_Estadistica_Computacional*, GitHub repository, Disponible en: https://github.com/Hacapoxd/FINESI_Estadistica_Computacional/tree/main/homeworks/paper_pattern_recognition