Proyecto Segunda Entrega

Pontificia Universidad Javeriana

Materia:

Procesamiento de datos a gran escala

Docente:

John Corredor Franco

Presentado por:

Daniel Mauricio Ordoñez Saavedra

Laura Sofia Salamanca Barrera

Neyl Peñuela Bernate

Noviembre

2023

Tabla de Contenidos:

Introducción

La ciudad de Nueva York ha brindado una política de datos abiertos reflejado en bases de datos, lo que hace que la riqueza de datos públicos generados por varias agencias de la ciudad y otras organizaciones municipales esté disponible para uso público. Este proceso se llevó a cabo con el propósito de crear un plan de acción basado en el procesamiento de datos para mejorar algunos indicadores territoriales de interés para el equipo de gobierno. Los indicadores que más preocupan al equipo de gobierno son la cantidad de arrestos y la cantidad de accidentes viales.

Por lo tanto, se han proporcionado los siguientes conjuntos de datos con el fin de desarrollar un proyecto analítico que permita generar un plan de acción para mejorar estos indicadores. Una vez que se cuenta con una comprensión profunda del problema de negocio y los datos, es necesario completar la limpieza y transformación de los mismos para abordar las preguntas de negocio planteadas previamente. También se aplicarán modelos de aprendizaje automático para enriquecer el ejercicio.

Objetivos:

Dar respuesta a las preguntas planteadas mediante el entendimiento del negocio y de sus datos planteados.

Generar un ambiente de preentrenamiento óptimo para poder hacer uso de los algoritmos de Machine Learning seleccionados en este proyecto. Este debe contener normalizaciones sobre valores aptos para este proceso, tratamientos convenientes según cada procedimiento de los *data frame,* y a su vez, una correlación entre las variables objetivo seleccionadas para el entrenamiento de cada algoritmo.

DESARROLLO

En este apartado se encontrará el procesamiento de los datos de cada base de datos, algunos de estos serán sobre la limpieza de las bases tales como en el manejo de los Nulos a pesar de ser parte del proyecto anteriores fueron documentados con el fin de redireccionar a un proceso más completo y eficaz concorde a la evolución de este proyecto, en sí para preparar el entorno al entrenamiento. Estos tratamientos se verían reflejados tanto en las variables(columnas) y sus registros (filas).

Bases de datos y su procesamiento

En este apartado se encuentran los filtros y transformaciones a las que fueron sometidos las bases de datos, más específicamente ante los valores nulos.

Dataset de arrestos (_NYPD Arrest Data (Year to Date)):

Primeramente, se decidió renombrar las columnas para volver más representativos con respecto a la información que presentan:

Columnas antes de renombrar	Columnas después de renombrar
['ARREST_KEY', 'ARREST_DATE', 'PD_CD', 'PD_DESC', 'KY_CD', 'OFNS_DESC', 'LAW_CODE', 'LAW_CAT_CD', 'ARREST_BORO', 'ARREST_PRECINCT', 'JURISDICTION_CODE', 'AGE_GROUP', 'PERP_SEX', 'PERP_RACE', 'X_COORD_CD', 'Y_COORD_CD', 'Latitude', 'Longitude', 'New Georeferenced Column']	['ARREST_KEY', 'ARREST_DATE', 'CODIGO_ESPECIFICO_DE_DELITO', 'DESCRIPCIÓN_CODIGO_ESPECIFICO', 'CATEGORIA_SECUNDARIA_INTERNA', 'DESCRIPCIÓN_DEL_DELITO', 'LAW_CODE', 'NIVEL_DEL_DELITO', 'DISTRITO_DE_ARRESTO', 'COMISARIA_DE_ARRESTO', 'CODIGO_DE_JURISDICCIÓN_RESPONSABLE', 'AGE_GROUP', 'PERP_SEX', 'PERP_RACE', 'X_COORD_CD', 'Y_COORD_CD', 'Latitude', 'Longitude', 'New Georeferenced Column']

Manejo de nulos :

Observaciones del conteo de nulos:

Los nulos de "codigo_especifico_de_delito" son despreciables por el tamaño de los registros y lo que representa para la investigación, al ser un identificador secundario es posible obtener la misma calidad descriptiva tomando otro identificador como lo puede ser el codigo de ley.

Los nulos de "categoría_secundaria_interna" son despreciables por el tamaño de los registros, al ser un identificador secundario es posible obtener la misma calidad descriptiva tomando otro identificador como lo puede ser "Law_code" que es una forma más general de identificar el delito.

Hay valores nulos que no se están teniendo en cuenta, dado que columnas como "Descripción_del_delito" ó "Descripción_codigo_especifico", cuentan con valores nulos que aparecen como '(null)'

Es importante eliminar los valores nulos de "NIVEL_DEL_DELITO" dado que al no poder contar con una forma de imputar los datos faltantes, puede que estos causen sesgos en el entrenamiento de modelos ML.

Después de este procedimiento se contempla los siguientes pasos para poder generar los análisis correspondientes:

Transformaciones:

- Transformación 1: Se completan los valores faltantes en la columna”Descripción del delito”, se reemplazará cada descripción general con las encontradas en la página web del senado de nueva york.
- Transfromación 2: Se normaliza el formato en el que se encuentran los valores nulos de la columna "PD_DESC", pasando los valores '(null)' a NULL.
- Transfromación 3: Se eliminan los valores nulos de la columna "NIVEL_DEL_DELITO".

En cuanto al manejo de duplicados este database no mostró necesidad de tratar puesto que no mostró registros idénticos.

Dataset de pobreza (NYCgov Poverty Measure Data)

En este dataframe los nombres de las columnas no eran intuitivos, este enmarcado por siglas que sin entendimiento del sistema en el que fue almacenada la información no se entendería a simple vista, por ende, se decidió renombrar estas columnas para mostrar en el análisis

Columnas antes de renombrar	Columnas después de renombrar
['SERIALNO', 'SPORDER', 'PWGTP', 'WGTP', 'AGEP', 'CIT', 'REL', 'SCH', 'SCHG', 'SCHL', 'SEX', 'ESR', 'LANX', 'ENG', 'MSP', 'MAR', 'WKW', 'WKHP', 'DIS', 'JWTR', 'NP', 'TEN', 'HHT', 'AgeCateg', 'Boro', 'CitizenStatus', 'EducAttain', 'EST_Childcare', 'EST_Commuting', 'EST_EITC', 'EST_FICAtax', 'EST_HEAP', 'EST_Housing', 'EST_IncomeTax', 'EST_MOOP', 'EST_Nutrition', 'EST_PovGap', 'EST_PovGapIndex', 'Ethnicity', 'FamType_PU', 'FTPTWork', 'INTP_adj', 'MRGP_adj', 'NYCgov_Income', 'NYCgov_Pov_Stat', 'NYCgov_REL', 'NYCgov_Threshold', 'Off_Pov_Stat', 'Off_Threshold', 'OI_adj', 'PA_adj', 'Povunit_ID', 'Povunit_Rel', 'PreTaxIncome_PU', 'RETP_adj', 'RNTP_adj', 'SEMP_adj', 'SSIP_adj', 'SSP_adj', 'TotalWorkHrs_PU', 'WAGP_adj']	'NumSerieHogar', 'NumOrdenPersonas', 'PesoPersona', 'Sueldo12Meses', 'CategoriaEdad', 'EstadoCiudadania', 'RelacionPrincipal', 'InscripcionEscolar', 'Grado', 'NivelEducativo', 'SEX', 'RegistroEstadoEmpleo', 'SegundoIdiomaHogar', 'HabilidadIngles', 'EstadoMarital', 'EstadoCivil', 'HorasTrabajadasSemana', 'PesoVivienda', 'RegistroDiscapacidad', 'MediTransporteTrabajo', 'NumPersonas', 'TenenciaVivienda', 'TipoHogar', 'AgeCateg', 'Boro', 'CitizenStatus', 'EducAttain', 'EST_Childcare', 'EST_Commuting', 'EST_EITC', 'EST_FICAtax', 'EST_HEAP', 'EST_Housing', 'EST_IncomeTax', 'EST_MOOP', 'EST_Nutrition', 'EST_PovGap', 'EST_PovGapIndex', 'Ethnicity', 'FamType_PU', 'FTPTWork', 'INTP_adj', 'MRGP_adj', 'NYCgov_Income', 'NYCgov_Pov_Stat', 'NYCgov_REL',

	'NYCgov_Threshold', 'Off_Pov_Stat', 'Off_Threshold', 'OI_adj', 'PA_adj', 'Povunit_ID', 'Povunit_Rel', 'PreTaxIncome_PU', 'RETP_adj', 'RNTP_adj', 'SEMP_adj', 'SSIP_adj', 'SSP_adj', 'TotalWorkHrs_PU', 'WAGP_adj']

Después de esto, al analizar el schema se encontró que todas las columnas contienen caracteres numéricos entre double y interger. Con ello puede representar una tarea de normalización en casos de tratamientos, tal como se va a exponer en continuación.

Manejo de nulos y transformaciones:

Las columnas que conteían nulos con su respectivo tratamiento:

- NivelEducativo: en este caso, el rango esta entre *(1,24)*, por lo que se toma la decisión de llenar los nulos con 0. Esto indicaría o que no hay registro o que no se tiene un nivel educativo.
- RegistroEstadoEmpleo: en este caso se llenará con 0. Por el diccionario de datos tenemos que se da principalmente porque las personas tienen menos de 16 años, o sea, no hacen parte de la fuerza laboral.
- SegundoIdiomaHogar: en este caso se llenará con 0. Por el diccionario de datos tenemos que se da principalmente porque las personas tienen menos de 5 años.
- HabilidadIngles: en este caso se llenará con 0. Por el diccionario de datos tenemos que se da principalmente porque las personas tienen menos de 5 años, por lo que es imposible clasificar a estas personas.
- EstadoMarital: en este caso se llenará con 0. Por el diccionario de datos tenemos que se da principalmente porque las personas tienen menos de 15 años, por lo que se consideran que no tienen o no clasifican en ningun estado marital.
- HorasTrabajadasSemana: en este caso se llenará con 0. Por el diccionario de datos tenemos que se da principalmente porque las personas tienen menos de 16 años, o sea, no hacen parte de la fuerza laboral.
- MediTransporteTrabajo: se llenará con 0. A pesar de que es una clasificación de 1 a 12, no se tiene información concreta de cada una de las categorias. Sin embargo, para no eliminar los registros, se decide llenarlos con 0.
- EducAttain: se llenará con 0, indicando que no tiene ningun logro, o no hay registro de ello.

  Esto se puede consolidar en que: todas las columnas con nulos se llenan de 0. En cada una, ese valor tiene un significado diferente.

Dataset de colisiones (_Motor Vehicle Collisions - Vehicles_):

En este database no se encontró necesario el cambio de nombre de columnas puesto a que estos representaban de manera correcta los datos que exponen.

Columnas de la base de datos
['UNIQUE_ID', 'COLLISION_ID', 'CRASH_DATE', 'CRASH_TIME', 'VEHICLE_ID', 'STATE_REGISTRATION', 'VEHICLE_TYPE', 'VEHICLE_MAKE', 'VEHICLE_MODEL', 'VEHICLE_YEAR', 'TRAVEL_DIRECTION', 'VEHICLE_OCCUPANTS', 'DRIVER_SEX', 'DRIVER_LICENSE_STATUS', 'DRIVER_LICENSE_JURISDICTION', 'PRE_CRASH', 'POINT_OF_IMPACT', 'VEHICLE_DAMAGE', 'VEHICLE_DAMAGE_1', 'VEHICLE_DAMAGE_2', 'VEHICLE_DAMAGE_3', 'PUBLIC_PROPERTY_DAMAGE', 'PUBLIC_PROPERTY_DAMAGE_TYPE', 'CONTRIBUTING_FACTOR_1', 'CONTRIBUTING_FACTOR_2']

Manejo de nulos y transformaciones:

Al seleccionar solo aquellos registros de siniestros viales que mostraban valores nulos en el registro del siniestro se acortó el dataframe significativamente, esto es necesario ya que son variables que se pueden establecer como variables objetivo. Después de esto, se muestra el tratado de los nulos de las demás columnas:

- VEHICLE_DAMAGE : Al mostrarse que si esta columna está con un valor "null", los valores de los daños de los otros vehiculos estaran a su vez en "null", se entendería que en este caso aunque sea un vehiculo debió estar involucrado, entonces se cambía a "No_Damage".
- VEHICLE_DAMAGE_1, VEHICLE_DAMAGE_2, VEHICLE_DAMAGE_3: Al saber que para haber un accidente se necesita solo de un vehiculo, los valores nulos se reemplazaron por "No_involved".
- PUBLIC_PROPERTY_DAMAGE_TYPE: Muestra que casi en su totalidad está denotada como valores "null", por ende se decide quitar la columna en su totalidad.
- VEHICLE_MAKE: Los nulos se presentan como valores que faltaron poner en el reporte, por ende se pone "Indefinite".
- VEHICLE_MODEL: Se decide quitar la columna en su totalidad ya que está en una gran mayoría retratado por nulos, y a su vez no aporta mucha información. La información sobre el vehiculo ya la aportan otras columnas tal como Vehicle_Make
- VEHICLE_YEAR: Este valor no se puede automatizar, por ende se cambia por un 0
- TRAVEL_DIRECTION: Los nulos en este caso se trabajan como direcciones y al haber nulos se cambia por "Unmarked", donde la dirección de la se provenia el accionante que produjo el accidente.
- PRE_CRASH: Los nulos se normalizan y se ponen como "Going straight ahead"
- POINT_OF_IMPACT: Los nulos se cambian por valores "Other"

- CONTRIBUTING_FACTOR_1, CONTRIBUTING_FACTOR_2: Los valores nulos se llevan a "Unspecified"

Dataset de educación (2016 - 2017 Health Educations)

Columnas
['School DBN', 'Community School District', 'City Council District', 'School Name', '# of students in grades 9-12', '# of students in grades 9-12 scheduled for at least one semester of health instruction', '%', '# of 16-17 June and August graduates', '# of 16-17 June and August graduates meeting high school health requirements', '% 1']

Revisando los registros, tenemos que en la mayoría de las columnas hay caracteres 's', incluso en columnas de tipo integer. Se remplazará ese carácter por 0.

De igual forma, vemos que la columna % y % 1 esta como string, siendo realmente un float. Por lo tanto, se le quitará el carácter '%' a cada registro y se procederá a castear la columna

Después de las anteriores transformaciones, algunas columnas se cambiaron a *string*. Por esta razón, se hace el casteo manual de las respectivas columnas

Además, se agregó una columna que contiene cada Borough en la que se encunetra cada escuela, esto gracias al código DBN de cada escuela

Manejo de nulos

Con respecto a este procedimiento al ser estos valores númericos, se cambiaron estos valores nulos por 0.

Preguntas de negocio y sus respuestas

Ante el análisis de las bases de datos se pudieron formalizar las preguntas que se iban a someter a estudio, y se brindó su respectiva respuesta.

- ¿Qué características (edad,raza,sexo) representan la mayor proporción de arrestos cometidos dentro del dataset ?
  Raza con mayor proporción de arrestos: BLACK Género con mayor proporción de arrestos: M Edad con mayor proporción de arrestos: 25-44

- ¿Qué nivel de infracción es el más recurrente?
  Cómo se puede ver; el nivel de delito más recurrente es "M" o "Misdemeanor", que corresponde a un nivel de delito medio.

- ¿Cuales son los Boroughs con mayor proporción de personas por hogar? ¿Hay alguna relación con el índice de pobreza?

Los boroughs con un mayor promedio de personas por cada hogar son:

- Queens
- Bronx
- Brooklyn
- Staten Island
- Manhattan

Por parte del promedio del indice de pobreza en cada barrio tenemos, en su respectivo orden de más pobreza a menos, los siguientes boroughs:

- Brooklyn
- Queens
- Bronx - Staten Island
- Manhattan

A partir de ello podemos decir que los dos boroughs con mayor promedio de personas por hogar son Queens y Bronx. Ademas, aunque los indices de pobreza no son los peores, si tienen un valor considerablemente alto. Por lo que se puede decir que el hecho de tener más gente por hogar indica una gran posibilidad de que la pobreza incremente.

- **¿Cuál es la relación entre la Etnicidad y los ingresos de las personas?**
  En este caso, vemos que las personas pertenecientes a Non-Hispanic White , son las que reciben mejores ingresos, a comparación de los otros grupos etnicos. De segundas le sigue la variedad de grupos étnicos diferentes a los considerados. En tercer y cuarto lugar están los Asiáticos no hispanicos y los Negros no hispánicos. Por último, los que menos ingresos reciben son los Hispánicos.

  Acá vemos una relación marcada entre las razas y etnias junto con los ingresos recibidos por las personas. En Nueva York hay una marcada discriminación salarial hacia los Hispánicos, mientras que los demás grupos tienden a recibir mejores salarios.

- **¿Hay alguna relación entre el *borough* y el porcentaje de personas graduadas con los requisitos de fundamentales de salud en secundaria?**

Se puede observar con claridad que, casi el 100% de estudiantes que se graduaron cumplen con los requisitos de salud en secundaria, pues en la mayoría de boroughs el promedio está muy cercano al 100%. Sin embargo, el borough que menor proporción de personas que cumplen el requisito es Bronx, pues cada 2 personas de 100, NO cumplen aquel requisito.

- **¿Existe alguna relación entre el distrito del colegio y la proporción de estudiantes graduados de este en los años 2016-17?**

Se propone una ANOVA que responda a las siguientes hipotesis

H0: No hay diferencias significativas en la proporción de estudiantes graduados entre los diferentes distritos 'Boro'.

H1: Existen diferencias significativas en la proporción de estudiantes graduados entre al menos dos de los grupos de distritos 'Boro'.

Observaciones de los resultados

Al obtener un P-valor tan bajo (menor a 0.05), podemos concluir que la hipotesis nula (H0) se rechaza, mostrando que si hay una diferencia significativa entre las proporciones de estudiantes graduados para cada distrito.

Ambos resultados indican una diferencia significativa para la proporción de estudiantes graduados en el periodo determinado.

- **¿Cuáles son las características predominantes de los accidentes de tráfico registrados en cada año? (En función de factores como el punto de impacto, el tipo de vehículo más involucrado y la causa previa al accidente)**

- Se encuentra que en este proceso desde el año 2012 al 2021, el punto de impacto más recurrente fue el "Center Front End". Esta siendo una gran parte de los puntos de impactos en todos los accidentes.

- El tipo de vehículo encontrado de forma más recurrente en los accidentes fueron los Sedan, esto siendo recurrente entre los años 2012 y 2021

- Y marcando una gran totalidad de la acción previa al accidente más recurrente fue "going straight ahead", siendo este todos los años el valor dominante.

Ahora interpretando estas 3 características con una relación : es posible que los vehículos "Sedan" están presentando un problema en sus frenos al ser registrados los daños en la parte delantera del capo, y que a su vez estos se encontraban yendo hacía al frente.

- **¿Cuál es el promedio de personas involucradas en los accidentes por año?**

- Los años 2015 a 2018 muestran una relativa estabilidad en el promedio de pasajeros involucrados en accidentes, con valores cercanos a 1.3 en promedio, lo que demuestra que tiende a ser más usual los accidentes donde solo va el conductor.

Selección de técnicas de aprendizaje de máquina

Aplicación Modelo No supervisado K-Means

Durante el análisis del conjunto de datos de arrestos, surgió una pregunta fundamental que se abordó mediante técnicas de aprendizaje automático: ¿Existe una relación significativa entre la probabilidad de que un registro esté asociado con un tipo de infracción grave y la pertenencia del perpetrador a un grupo sociodemográfico minoritario (Black o White Hispanic, American Indian/Alaskan Native, Asian/Pacific Islander, Black)?

Este proceso implicó la preparación previa de los datos y la creación de un modelo de regresión lineal. Para llevar a cabo este análisis, se seleccionaron cuidadosamente las variables PERP_RACE, NIVEL_DEL_DELITO, Longitude, Latitude, X_COORD_CD e Y_COORD_CD.

La regresión reveló una conexión entre la raza de una persona y la probabilidad de que el registro esté relacionado con un delito grave; sin embargo, esta relación parece estar más inclinada hacia las personas blancas en comparación con otros grupos raciales.

Para abordar este análisis, se empleó el método K-Means, modelo ML no supervisado haciendo este procedimiento dos veces. El primero, con las variables de latitud y longitud como características clave. El proceso comenzó dividiendo los datos en conjuntos de entrenamiento y prueba. Se optó por dividir los datos en 5 clústeres, lo que corresponde a los 5 "boroughs" de Nueva York.

Con base en estas predicciones, se extrajeron las coordenadas correspondientes y se llevaron a cabo las visualizaciones pertinentes, lo que nos proporcionó lo siguiente: el modelo resultante exhibe un coeficiente silhouette de 0.642. Esto indica que la clusterización se ajusta de manera efectiva para predecir la forma que deberían tener las categorías.

El scatterplot, por su parte, nos brinda una representación gráfica más clara de la disposición de los datos. Como se puede apreciar, la forma del gráfico se asemeja a la geografía de Nueva York, con categorías que corresponden a los diferentes distritos de la ciudad.

Luego, se llevó a cabo un proceso similar utilizando las variables 'Y_COORD_CD' y 'X_COORD_CD' como características en el algoritmo K-Means. El resultado de este proceso arrojó un modelo con un coeficiente silhouette de 0.630, lo que indica que la clusterización se ajusta de manera efectiva para predecir la forma que deberían tener las categorías.

Dentro de las agrupaciones resultantes, se observó que el distrito del Bronx presenta la mayor proporción de arrestos, con un porcentaje del 0.307; Esta información proporciona una comprensión valiosa de la distribución de arrestos en los diferentes distritos de la ciudad.

Aplicación Modelo supervisado KNN

Este método se aplicó en el dataframe de pobreza: Para esto, en primer lugar, se eliminaron aquellos valores que son únicos para cada registro u hogar (*NumSerieHogar*), pues no tiene sentido realizar el analisis del modelo sobre esta variable. De igual forma, se eliminarán las columnas que no tienen un significado para nosotros y nuestro analisis. Con ello, se formó

matriz de correlación para poder observar cuales variables están más relacionadas entre sí con respecto a las demás. Y esto, demuestra que la variable de interes es NYCgov_Pov_Stat, indica el estado de pobreza de la persona {1: en pobreza, 2: no en pobreza}.

Si observamos la matriz de correlación, esta variable tiene variables muy cercanas a cero, lo que a primera vista, indicaria una relación baja. Sin embargo, hay que recordar que la correlación mide la relación lineal entre dos variables. Esto nos lleva a seguir con la misma variable de interes (NYCgov_Pov_Stat), suponiendo que, con las otras variables, tiene otro tipo de relación.

Luego, al sacar el rango de cada variable se observó una variabilidad significativa en el orden de magnitud de las variables (rango, diferencia entre el maximo y el mínimo). Esto podría advertir un deterioro de rendimiento de cualquier algoritmo de machine learning, pues al construir el modelo puede existir un sesgo entre las variables que altere el algoritmo. Por ello, se podría señalar que la desviación estándar en muchos casos es bastante alta, lo que indica que los datos están bastante dispersos. Y, por otro lado, hay variables que, por lo contrario, tienen muy poca dispersión.

Al realizar la estandarización el orden de magnitud en algunas variables se mostró bastante alto, lo que puede afectar los modelos de Machine Learning que se basen en distancias euclideanas. De igual forma, los modelos de Machine Learning, de manera general, se van a ver distorsionados con las variables que tengan un rango bastante alto.

Justificación del modelo

En este caso, nuestra variable de interes es si una persona vive en la pobreza o no (*1: pobreza, 2: no pobreza*). De ahí se concluye que se puede buscar un algoritmo de clasificación (o sea, un algoritmo **supervisado**). De forma especifica se obtiene que el KNN nos puede ayudar a predecir si una nueva observación (una persona en este caso) se encuentra en la categoría de pobreza o no, en función de las variables explicativas disponibles del dataset.

Tabla comparativa de rendimiento:

Modelo Supervisado KNN	**Modelo no Supervisado K-Means**
Mediante la matríz de confusión se obtuvieron los siguientes resultados: (Debemos recordar que {1: personas en pobreza, 1: personas no en pobreza}.)	Con la matriz de confusión no es suficiente para concluir sobre el modelo, necesitamos metricas de rendimiento del modelo que son dadas por el reporte generado a traves de classification_report(Y_test,Y_pred).
1. **Verdaderos positivos:** Las personas que predijo como pobres y si son pobres fueron 433.	**Precisión** • Para la clase 1, la precisión es de 0.64 • Para la clase 2, la precisión es de 0.88 • Su *macro-avg* es de 0.76

2. **Verdaderos negativos:** Las personas que predijo como no pobres y no son pobres fueron 4798. 3. **Falsos negativos:** Las personas que predijo como no pobres pero en realidad son pobres fueron 656. 4. **Falsos positivos:** Las personas que predijo como pobres pero en realidad son no pobres fueron 240 .	• Su *weighted-avg* es de 0.84 **Recall** • Para la clase 1, el recall es de 0.40 • Para la clase 2, el recall es de 0.95 • Su *macro-avg* es de 0.67 • Su *weighted-avg* es de 0.85 **F1 Score** • Para la clase 1, el f1 - score es de 0.49 • Para la clase 2, el f1 - score es de 0.91 • Su *macro-avg* es de 0.70 • Su *weighted-avg* es de 0.84
Exactitud: (Verdaderos Positivos+ Verdaderos negativos)/total de observaciones. **Exactitud:** 0.854	**Exactitud:** 0.85

Tabla1: Tabla de comparación eficiencia entre los algoritmos KNN y K-Means

Ambos modelos muestran un rendimiento generalmente sólido, con métricas de Precision y Recall que indican una buena capacidad para clasificar ambas clases. Cada parámetro cumple su función y estas predicciones a ponerlas a prueba cumplieron su objetivo.

KNN	Primer Modelo	Segundo Modelo	Tercer Modelo
Matriz de Confusión	• **Verdaderos positivos:** Las personas que predijo como pobres y si son pobres fueron 1926. • **Verdaderos negativos:** Las personas que predijo como no pobres y no son pobres fueron 10735. • **Falsos negativos:** Las personas que predijo como no pobres pero en	• **Verdaderos positivos:** Las personas que predijo como pobres y si son pobres fueron 1923. • **Verdaderos negativos:** Las personas que predijo como no pobres y no son pobres fueron 10742. • **Falsos negativos:** Las personas que predijo como no pobres pero en	• **Verdaderos positivos:** Las personas que predijo como pobres y si son pobres fueron 1973. • **Verdaderos negativos:** Las personas que predijo como no pobres y no son pobres fueron 414. • **Falsos negativos:** Las personas que predijo como no pobres pero en

	realidad son pobres fueron 461. • **Falsos positivos:** Las personas que predijo como pobres pero en realidad son no pobres fueron 354 .	realidad son pobres fueron 464. • **Falsos positivos:** Las personas que predijo como pobres pero en realidad son no pobres fueron 347 .	realidad son pobres fueron 344. • **Falsos positivos:** Las personas que predijo como pobres pero en realidad son no pobres fueron 10745 .
Reporte arrojado	**Precisión** • Para la clase 1, la precisión es de 0.84 • Para la clase 2, la precisión es de 0.96 • Su *macro-avg* es de 0.90 • Su *weighted-avg* es de 0.94 **Recall** • Para la clase 1, el recall es de 0.81 • Para la clase 2, el recall es de 0.97 • Su *macro-avg* es de 0.89 • Su *weighted-avg* es de 0.94 **F1 Score** • Para la clase 1, el f1 - score es de 0.83 • Para la clase 2, el f1 - score es de 0.96 • Su *macro-avg* es de 0.89 • Su *weighted-avg* es de 0.94	**Precisión** • Para la clase 1, la precisión es de 0.85 • Para la clase 2, la precisión es de 0.96 • Su *macro-avg* es de 0.90 • Su *weighted-avg* es de 0.94 **Recall** • Para la clase 1, el recall es de 0.81 • Para la clase 2, el recall es de 0.97 • Su *macro-avg* es de 0.89 • Su *weighted-avg* es de 0.94 **F1 Score** • Para la clase 1, el f1 - score es de 0.83 • Para la clase 2, el f1 - score es de 0.96 • Su *macro-avg* es de 0.89 • Su *weighted-avg* es de 0.94	• Para la clase 1, la precisión es de 0.85 • Para la clase 2, la precisión es de 0.96 • Su *macro-avg* es de 0.91 • Su *weighted-avg* es de 0.94 **Recall** • Para la clase 1, el recall es de 0.83 • Para la clase 2, el recall es de 0.97 • Su *macro-avg* es de 0.90 • Su *weighted-avg* es de 0.94 **F1 Score** • Para la clase 1, el f1 - score es de 0.84 • Para la clase 2, el f1 - score es de 0.97 • Su *macro-avg* es de 0.90 • Su *weighted-avg* es de 0.94

Exactitud	• Es de 0.94%.	• Es de 0.94%.	• Es de 0.94%.

Tabla2: Tabla de Comparación entre 3 modelos del algorítmo KNN

Se puede observar de estos 3 modelos que:

- Todos los modelos muestran un rendimiento destacado en la clase 2, con F1 Scores y otras métricas altas.
- Aunque el conjunto de datos está desequilibrado, los modelos logran buenos resultados en ambas clases.
- La métrica F1 Score proporciona una evaluación equilibrada entre precisión y recall, destacando la capacidad de los modelos para manejar ambas clases.
- Se recomienda un monitoreo continuo del rendimiento y, si es posible, explorar estrategias para abordar el desbalance en el conjunto de datos.

Conclusiones (Parciales)

Al realizar una comprensión de las respuestas de las preguntas de negocio, se puede comprender que el estado de Nueva York debería poner atención a la relación que hay entre las minorías, especialmente las comunidades Negras, que se encuentran en todos sus distritos, dándole mayor importancia al distrito Bronx, realizando un plan de acción que pueda implantar más en la cultura la educación. Esto podría guiar al proceso de investigación de organizaciones estatales que pueden no estar cumpliendo su labor.

Por otro lado, en cuanto a los modelos realizados al analizar su rendimiento encontramos que: En cuanto al modelo no supervisado aplicado al conjunto de datos de arrestos, se logra una agrupación precisa en los distritos definidos previamente. Ambas pruebas muestran rendimientos similares, sugiriendo que, para lograr mejoras significativas, sería necesario agregar una cantidad considerable de datos al conjunto. Aunque el modelo supervisado puede tener un mejor ajuste, parece priorizar la certeza sobre la cantidad de agrupaciones utilizadas en los datos. Por otro lado, el modelo con categorías agrupadas facilita la respuesta a preguntas futuras sobre la pertenencia de los delitos a nivel geográfico y la inclusión de otros indicadores propuestos como variables respuesta. La conclusión es que, para procesos mejor categorizados, se prefiere utilizar el formato de ubicación alternativo para lograr una mejor aproximación a los distritos.

En el contexto supervisado, el modelo exhibe un rendimiento generalmente sólido, alcanzando una precisión del 85%. Sin embargo, al analizar detenidamente cada clase, se revela un desequilibrio en el conjunto de datos, particularmente en las observaciones y predicciones de personas pobres. La clase de no pobres (clase 2) se clasifica de manera confiable, con un alto recall y precisión. En contraste, la clase de personas

pobres (clase 1) muestra un recall más bajo, indicando que el modelo tiene dificultades para detectar esta clase, aunque es altamente preciso cuando lo hace. El F1 Score también refleja esta discrepancia, siendo más alto para la clase 2 (0.91) en comparación con la clase 1 (0.49).

De forma general vemos que los 3 modelos se comportan de buena forma. En los 3 casos, las métricas son similares y cambian de forma mínima. La clase 2 es la que tiene mejor rendimiento a comparación de la clase 1. Es decir, el modelo predice mejor a las personas en no pobreza que a las que si están en pobreza. Sin embargo, la clase 1 no tiene malas métricas.

El mejor modelo de los 3 es el 3. En este caso, el k=3 exige al modelo ser más estricto. Se pensaba que, al ser más estricto, podría haber más datos mal clasificados. Sin embargo, se comportó de manera similar que los otros dos modelos. Esto nos lleva a concluir que las clases están lo suficientemente separadas para que el modelo pueda tener esa precisión. Además, si revisamos la precisión en las celdas anteriores, los decimales nos llevan a reforzar esta conclusión.

n general podriamos decir que es un buen modelo, pues con una exactitud del 85% en las predicciones nos podemos dar por bien servidos.

Referencias

Apache Spark. (2023, September 9). PySpark Overview . Spark. Retrieved 02 November. 2023, from https://spark.apache.org/docs/latest/api/python/index.html.

Fernández, R. (2023, January 24). Población de los EE. UU. por raza y origen hispano 2020-2060 | Statista. Es. Retrieved 02 November. 2023, from https://es.statista.com/estadisticas/600570/porcentaje-de-poblacion-de-estados-unidos--2060-por-raza-y-origen-hispano/.