**Análisis de un modelo de predicción de gravedad de accidentes viales**

Yenny Betancur Torres, José Pinto Rodríguez, Gustavo Flórez Ortiz, Edgar Mauricio Flórez

*Universidad de Bogotá Jorge Tadeo Lozano; Bogotá D.C, Colombia*

# INTRODUCCIÓN

Actualmente la secretaría distrital de movilidad no dispone de un modelo de predicciones de los siniestros viales con el fin de catalogar rápida y correctamente la gravedad de un accidente para poder remitir a los recursos necesarios para atender adecuadamente el siniestro según sus características y posibles consecuencias generales. Dado lo anterior analizamos el comportamiento de los siniestros viales registrados en los informes policiales de accidentes de tránsito ocurridos en la ciudad de Bogotá durante el año 2018, proporcionados por la página de Datos Abiertos Colombia, con el fin de crear un modelo que logre predecir el tipo de gravedad a partir de las diferentes variables seleccionadas, utilizando las técnicas y conceptos de machine learning.

Es importante priorizar las medidas a tomar dado que existen unos tipos de gravedad en accidentes que tienen mayor impacto o que pueden afectar a los conductores y/o peatones, además del entorno familiar puesto que pueden presentarse con mayor frecuencia.

Dicho impacto ha sido fuente de análisis puramente estadístico como en el trabajo “Análisis estadístico sobre las influencias del entorno en los accidentes de tránsito en la ciudad de Bogotá” [1] en el cual se crea un modelo estadístico de regresión logística que da un acercamiento al número de víctimas fatales en un accidente de tránsito de acuerdo a la infraestructura vial y la clasificación de está. Sin embargo, en el trabajo “Modelo de predicción de gravedad de accidentes de tránsito: un análisis de los siniestros en Bogotá, Colombia” [2], se diseñó un modelo de regresión logística con datos de accidentes transcurridos entre 2007-2016 que permitió identificar las variables clave y se realizó la clasificación con la técnica k-prototypes la cual permite el uso de variables mixtas.

# METODOLOGIAS

Con el fin de abordar el problema y realizar el análisis de los datos se siguió la metodología CRISP-DM teniendo en cuenta una comparación entre los resultados obtenidos al aplicar un subconjunto de técnicas algorítmicas de clasificación mediante el uso de las librerías disponibles en Python en el entorno de ejecución Google Colab.

Teniendo en cuenta que nuestro caso de uso es un ejercicio puramente académico basado en datos reales, no todos los pasos de la metodología se aplicaron completamente, ya que no existe un cliente específico para ejecutar algunos de los pasos.

Diagrama

Descripción generada automáticamente

Figura 1 Pasos de la metodología CRISP - DM [3]

## Definición de necesidades del cliente

Se desea predecir la gravedad de los accidentes viales que suceden en la ciudad de Bogotá a partir de las variables de entrada proporcionadas por datos históricos del año 2018 extraídos de la página de Datos Abierto de Colombia.

A nivel funcional se espera tener un modelo usable que clasifique la gravedad de los accidentes teniendo en cuenta las variables ingresadas por el usuario con el fin de apoyar y poder remitir a los recursos necesarios para atender adecuadamente el siniestro según sus características y posibles consecuencias generales. Mientras que a nivel técnico se espera evaluar diversos modelos de clasificación con el fin de encontrar el modelo que más se ajuste a la necesidad funcional en términos de precisión aplicando analítica predictiva.

## Comprensión y exploración de los datos

En este caso de estudio no fue necesaria la recolección de datos adicionales ya que contamos con un dataset con una cantidad de variables y registros importantes que permitieron el análisis y construcción del modelo. Dicho dataset se encuentra en formato Excel con un total de 36953 registros y 50 columnas/variables

Sobre los datos obtenidos, se realizó inicialmente una exploración de las variables existentes y sus tipos de datos. En este primer acercamiento encontramos que la mayoría de los datos son categóricos, incluso algunos de estos tienen una representación a nivel de código, como se muestra en la siguiente tabla:

TABLA 1

TIPOS DE DATOS DE VARIABLES

| **CAMPO** | **TIPO DE DATOS** |
| --- | --- |
| idFormulario | int64 |
| Dia | object |
| Fecha | object |
| MES\_PROCESADO | object |
| Oficina | int64 |
| GravedadCod | int64 |
| GravedadNombre | object |
| ClaseCodigo | int64 |
| ClaseNombre | object |
| ChoqueCodigo | float64 |
| ChoqueNombre | object |
| ObjetoFijoCodigo | float64 |
| ObjetoFijoNombre | object |
| OtraClase | float64 |
| NombreOtraClase | object |
| Latitud | float64 |
| Longitud | float64 |
| Direccion | object |
| TipoVia1 | object |
| NumeroVia1 | object |
| LetraVia1 | object |
| CardinalVia1 | object |
| TipoVia2 | object |
| NumeroVia2 | float64 |
| LetraVia2 | object |
| CardinalVia2 | object |
| Complemento | float64 |
| Municipio | object |
| Localidad | object |
| FechaOcurrencia | object |
| HoraOcurrencia | object |
| HORA\_PROCESADA | int64 |
| Area | int64 |
| Sector | int64 |
| Zona | int64 |
| TipoDisenno | object |
| TipoTiempo | object |
| ZonaTransito | object |
| AreaTransito | object |
| CON\_BICICLETA | object |
| CON\_CARGA | object |
| CON\_EMBRIAGUEZ | object |
| CON\_HUECOS | object |
| CON\_MENORES | object |
| CON\_MOTO | object |
| CON\_PEATON | object |
| CON\_PERSONA\_MAYOR | object |
| CON\_RUTAS | object |
| CON\_TPI | object |
| CON\_VELOCIDAD | object |

En segundo lugar, se crearon tablas de frecuencia con el fin de realizar una exploración inicial de los datos, esto se realizó para cada variable teniendo en cuenta que son variables categóricas y esto permitiría un análisis de la distribución de los datos y validar si existen datos atípicos que puedan sesgar la información, a continuación, algunos ejemplos en las siguientes tablas:

TABLA 2

FRECUENCIA CLASE CODIGO

|  |  |  |
| --- | --- | --- |
| **ClaseCodigo** | **ClaseNombre** | **Cantidad** |
| 1 | Choque | 31601 |
| 2 | Atropello | 3759 |
| 3 | Volcamiento | 537 |
| 4 | Caída de ocupante | 920 |
| 5 | Incendio | 6 |
| 6 | Otro | 130 |
| Total |  | 36953 |

TABLA 3

FRECUENCIA LOCALIDAD

|  |  |
| --- | --- |
| **Localidad** | **Cantidad** |
| ANTONIO NARINO | 806 |
| BARRIOS UNIDOS | 2094 |
| BOSA | 1955 |
| CANDELARIA | 140 |
| CHAPINERO | 2146 |
| CIUDAD BOLIVAR | 1368 |
| ENGATIVA | 3925 |
| FONTIBON | 2986 |
| KENNEDY | 4005 |
| LOS MARTIRES | 1265 |
| PUENTE ARANDA | 2464 |
| RAFAEL URIBE URIBE | 800 |
| SAN CRISTOBAL | 1053 |
| SANTA FE | 1155 |
| SUBA | 3179 |
| SUMAPAZ | 2 |
| TEUSAQUILLO | 2013 |
| TUNJUELITO | 986 |
| USAQUEN | 3860 |
| USME | 751 |
| **Total** | **36953** |

TABLA 4

FECUENCIA DISEÑO

|  |  |
| --- | --- |
| **TipoDisenno** | **Cantidad** |
| Cicloruta | 65 |
| Glorieta | 270 |
| Interseccion | 5187 |
| Lote o predio | 282 |
| Paso a nivel | 23 |
| Paso elevado | 115 |
| Paso inferior | 22 |
| Ponton | 2 |
| Puente | 45 |
| Tramo de Via | 30913 |
| Tunel | 6 |
| Via peatonal | 23 |
| **Total** | **36953** |

## Preparación de los datos.

Después de la verificación y análisis de los datos se decidieron realizar las siguientes acciones como parte de la preparación de los datos, teniendo en cuenta lo que aporta valor y lo que no a nuestro modelo:

1. Teniendo en cuenta la cantidad de registros por Clase de accidente, se decidió trabajar solo con accidentes tipo choque (31601 registros, más del 95% de la totalidad) para que el dataset no quede desbalanceado y lleguen a sesgar los datos

2. Se eliminarán las siguientes variables teniendo en cuenta que hacen referencia a otras variables categóricas del dataset:

● GravedadNombre = GravedadCod

● ClaseNombre = ClaseCodigo

● ChoqueNombre = ChoqueCodigo

● ObjetoFijoNombre = ObjetoFijoCodigo

● NombreOtraClase = OtraClase

● HoraOcurrencia = HORA\_PROCESADA

3. Se eliminan las siguientes variables ya que no aportan información al modelo:

● idFormulario: Identificador de cada accidente.

● Oficina: Valor único para todos los registros.

● Municipio: Valor único para todos los registros.

● Latitud: Valor único para todos los registros, excepto para un registro.

● Longitud: Valor único para todos los registros, excepto para un registro.

● FechaOcurrencia y fecha: No nos aporta al modelo ya que no trabajaremos con series de tiempo.

● ChoqueCodigo: Variable con datos en muy pocos registros, ya que decidimos trabajar sólo con choques.

● ObjetoFijoCodigo: Variable con datos en muy pocos registros, ya que decidimos trabajar sólo con choques.

● OtraClase: Valores nulos para todos los registros tipo choque

● ClaseCodigo: Valor único para todos los registros, ya que decidimos trabajar sólo con choques.

4. Dado que las siguientes variables hacen referencia o describen una ubicación dentro del distrito de Bogotá se tomó la decisión de eliminarlas y tomar de referencia la variable Localidad como forma de ubicación:

● Dirección:

● TipoVia1

● NumeroVia1

● LetraVia1

● CardinalVia1

● TipoVia2

● NumeroVia2

● LetraVia2

● CardinalVia2

● Complemento

● ZonaTransito

● AreaTransito

● Area

● Sector

● Zona

5. Identificando datos atípicos: Se identifican registros atípicos con un solo dato los cuales pueden llegar sesgar los datos de la variable, así que se eliminan los registros con menos de 11 repeticiones a excepción de niebla, como se muestra en la siguiente tabla:

TABLA 5

FRECUENCIA TIPO TIEMPO

|  |  |
| --- | --- |
| **TipoTiempo** | **Cantidad** |
| Normal | 30838 |
| Lluvia | 698 |
| Viento | 31 |
| Normal/Normal | 11 |
| Niebla | 11 |
| Granizo | 7 |
| Viento/Normal | 1 |
| Lluvia/Normal | 1 |
| Niebla/Normal | 1 |
| Viento/Lluvia | 1 |
| Normal/Viento | 1 |
| Total | 31601 |

6. Encodificacion de campos: Teniendo en cuenta que se eliminaron algunas variables categóricas y se dejaron las variables con sus respectivos códigos, a continuación, relacionamos el diccionario de datos de dichas variables:

● Gravedad Código

1=Con Muertos

2=Con Heridos

3=Solo Daños

● Choque Código

1=Vehiculo

2=Tren

3=Semoviente

4=Objeto Fijo

5=Otro

Si realizamos la validación de la cantidad de datos sobre nuestra variable objetivo, encontramos que presenta un desbalanceo respecto a un tipo de gravedad (Gravedad Tipo 1), por lo cual procedemos con la eliminación de los registros y será un análisis binomial:

TABLA 6

FRECUENCIA GRAVEDAD

|  |  |  |
| --- | --- | --- |
| **GravedadCod** | **GravedadNombre** | **Cantidad** |
| 1 | Con Muertos | 213 |
| 2 | Con Heridos | 7669 |
| 3 | Solo Danos | 23703 |

Teniendo en cuenta que tenemos varias variables cuya información está en formato texto y solo con dos valores diferentes, se encodificaron para facilitar su procesamiento. Para las siguientes variables 0 representa "No" y 1 representa "Si", como se muestra en la siguiente tabla:

TABLA 7

FRECUENCIA DE VARIABLES BINOMIALES

|  |  |  |  |
| --- | --- | --- | --- |
| **Valor** | **0** | **1** | **Total** |
| CON\_BICICLETA | 29378 | 2207 | 31585 |
| CON\_CARGA | 26521 | 5064 | 31585 |
| CON\_EMBRIAGUEZ | 31183 | 402 | 31585 |
| CON\_HUECOS | 31460 | 125 | 31585 |
| CON\_MENORES | 30719 | 866 | 31585 |
| CON\_MOTO | 24263 | 7322 | 31585 |
| CON\_PEATON | 31300 | 285 | 31585 |
| CON\_PERSONA\_MAYOR | 26028 | 5557 | 31585 |
| CON\_RUTAS | 31566 | 19 | 31585 |
| CON\_TPI | 24811 | 6774 | 31585 |
| CON\_VELOCIDAD | 31332 | 253 | 31585 |

Teniendo en cuenta la cantidad de datos y la carga de procesamiento que representa, también se decidió encodificar las siguientes variables:

● Localidad: Se ajustaron los datos de la siguiente forma:

0: 'ANTONIO NARINO',

1: 'BARRIOS UNIDOS',

2: 'BOSA',

3: 'CANDELARIA',

4: 'CHAPINERO',

5: 'CIUDAD BOLIVAR',

6: 'ENGATIVA',

7: 'FONTIBON',

8: 'KENNEDY',

9: 'LOS MARTIRES',

10: 'PUENTE ARANDA',

11: 'RAFAEL URIBE URIBE',

12: 'SAN CRISTOBAL',

13: 'SANTA FE',

14: 'SUBA',

15: 'SUMAPAZ',

16: 'TEUSAQUILLO',

17: 'TUNJUELITO',

18: 'USAQUEN',

19: 'USME'

● TipoDisenno: Se ajustaron los datos de la siguiente forma:

0: 'Cicloruta',

1: 'Glorieta',

2: 'Interseccion',

3: 'Lote o predio',

4: 'Paso a nivel',

5: 'Paso elevado',

6: 'Paso inferior',

7: 'Ponton',

8: 'Puente',

9: 'Tramo de Via',

10: 'Tunel',

11: 'Via peatonal'

● TipoTiempo

0: 'Granizo',

1: 'Lluvia',

2: 'Niebla',

3: 'Normal',

4: 'Viento'

● Dia

'LUNES': 1,

'MARTES': 2,

'MIERCOLES': 3,

'JUEVES': 4,

'VIERNES': 5,

'SABADO': 6,

'DOMINGO':7

● MES\_PROCESADO:

'ENERO': 1,

'FEBRERO': 2,

'MARZO': 3,

'ABRIL': 4,

'MAYO': 5,

'JUNIO': 6,

'JULIO':7,

'AGOSTO': 8,

'SEPTIEMBRE': 9,

'OCTUBRE': 10,

'NOVIEMBRE': 11,

'DICIEMBRE': 12

● HORA\_PROCESADA: Se asignaron rangos de horas para interpretar mejor los datos:

1 (Mañana) =horas de 6 a 13

2 (Tarde) =horas de 14 a 21

3 (Noche) =horas de 22 a 24 y 0 a 5

Finalmente, al terminar el proceso de encodificación y eliminación de las variables, se procede a convertir los campos a categóricos, quedando el listado final con los tipos de datos así:

TABLA 8

TIPOS DE DATOS AJUSTADOS

|  |  |
| --- | --- |
| **CAMPO** | **TIPO DE DATOS** |
| Dia | category |
| MES\_PROCESADO | category |
| GravedadCod | category |
| ChoqueCodigo | category |
| Localidad | category |
| HORA\_PROCESADA | category |
| TipoDisenno | category |
| TipoTiempo | category |
| CON\_BICICLETA | int64 |
| CON\_CARGA | int64 |
| CON\_EMBRIAGUEZ | int64 |
| CON\_HUECOS | int64 |
| CON\_MENORES | int64 |
| CON\_MOTO | int64 |
| CON\_PEATON | int64 |
| CON\_PERSONA\_MAYOR | int64 |
| CON\_RUTAS | int64 |
| CON\_TPI | int64 |
| CON\_VELOCIDAD | int64 |

7. Análisis de correlaciones

Dentro del análisis de correlaciones encontramos que existe poca relación entre las variables, lo cual nos indica que la mayoría de las variables son importantes para el modelo. Dado lo anterior tomamos 'CON\_EMBRIAGUEZ', 'CON\_VELOCIDAD', 'CON\_BICICLETA', 'CON\_CARGA', 'CON\_MOTO', 'CON\_PERSONA\_MAYOR' ya que presentan poca correlación y consideramos que son características importantes a evaluar teniendo en cuenta nuestra variable objetivo.

**Escala de tiempo

Descripción generada automáticamente**

Figura 2 Matriz de correlación entre variables

## Modelo seleccionado

Teniendo en cuenta el análisis previo de datos (Análisis de datos atípicos y faltantes, correlaciones e independencia entre variables), obtuvimos que la mayoría de las variables son categorías y binarias (las cuales manejaremos de manera numérica). Dado lo anterior tomamos la decisión de usar las variables 'Dia', 'Localidad', 'MES\_PROCESADO','HORA\_PROCESADA', 'TipoDisenno', 'CON\_EMBRIAGUEZ', 'CON\_VELOCIDAD', 'CON\_BICICLETA', 'CON\_CARGA', 'CON\_MOTO', 'CON\_PERSONA\_MAYOR'. Las cuales consideramos que son importantes y pueden ayudarnos a explicar mejor el modelo

Se toma como variable objetivo GravedadCod y se establece como tamaño de los conjuntos de prueba y entrenamiento el 30% del total de datos.

Se validaron los mejores modelos para clasificación para realizar el entrenamiento y validar su comportamiento, por lo cual ingresarán en nuestra validación los siguientes:

● XGBoost

● DecisionTreeClassifier

● Naive\_bayes

● LinearDiscriminantAnalysis

● RandomForestClassifier

Teniendo en cuenta la cantidad de variables del modelo, se definió la profundidad de la mayoría de estos en 12 y se procede a entrenarlos por primera vez con los siguientes hiperparámetros como se muestra en la siguiente tabla:

TABLA 9

HIPERPARAMETROS DE MODELOS SELECCIONADOS

| **MODELO** | **HIPERPARAMETRO** | **VALOR** |
| --- | --- | --- |
| XGBoost | base\_score | 0.5 |
| booster | gbtree |
| colsample\_bylevel | 1 |
| colsample\_bynode | 1 |
| colsample\_bytree | 1 |
| gamma | 0 |
| learning\_rate | 0.1 |
| max\_delta\_step | 0 |
| max\_depth | 12 |
| min\_child\_weight | 1 |
| missing | None |
| n\_estimators | 100 |
| n\_jobs | 1 |
| nthread | None |
| objective | binary:logistic |
| random\_state | 0 |
| reg\_alpha | 0 |
| reg\_lambda | 1 |
| scale\_pos\_weight | 1 |
| seed | None |
| silent | None |
| subsample | 1 |
| verbosity | 1 |
| DecisionTreeClassifier | ccp\_alpha | 0.0 |
| class\_weight | None |
| criterion | gini |
| max\_depth | 12 |
| max\_features | None |
| max\_leaf\_nodes | None |
| min\_impurity\_decrease | 0.0 |
| min\_samples\_leaf | 1 |
| min\_samples\_split | 2 |
| min\_weight\_fraction\_leaf | 0.0 |
| random\_state | None |
| splitter | best |
| Naive\_bayes | priors | None |
| var\_smoothing | 1e-09 |
| LinearDiscriminantAnalysis | covariance\_estimator | None |
| n\_components | 1 |
| priors | None |
| shrinkage | None |
| solver | svd |
| store\_covariance | False |
| tol | 0.0001 |
| RandomForestClassifier | bootstrap | True |
| ccp\_alpha | 0.0 |
| class\_weight | None |
| criterion | gini |
| max\_depth | 12 |
| max\_features | auto |
| max\_leaf\_nodes | None |
| max\_samples | None |
| min\_impurity\_decrease | 0.0 |
| min\_samples\_leaf | 1 |
| min\_samples\_split | 2 |
| min\_weight\_fraction\_leaf | 0.0 |
| n\_estimators | 100 |
| n\_jobs | None |
| oob\_score | False |
| random\_state | None |
| verbose | 0 |
| warm\_start | False |

Con el fin de contrastar el desempeño de cada modelo, se realizó la validación cruzada para ver cual modelo presenta la mejor precisión. Elegimos esta medición ya que queremos que el modelo clasifique correctamente la gravedad de los accidentes. A continuación, los resultados por modelo:

TABLA 10

RESULTADOS PRIMERA MATRIZ DE VALIDACION CRURZADA

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MODELO** | **RESULTADOS** | | | | | **MEDIA** | **DESV.** |
| XGBClassifier (max\_depth=12) | 0,89305761 | 0,89710231 | 0,89323843 | 0,89710231 | 0,8966967 | 0,89543947 | 0,001877685 |
| DecisionTreeClassifier (max\_depth=12) | 0,89237138 | 0,90368357 | 0,89769277 | 0,91044323 | 0,90386978 | 0,901612147 | 0,006133988 |
| GaussianNB() | 0,85009437 | 0,85309973 | 0,85077048 | 0,85010707 | 0,85422343 | 0,851659015 | 0,00169104 |
| LinearDiscriminantAnalysis (n\_components=1) | 0,93087121 | 0,92983008 | 0,93801917 | 0,93595434 | 0.93380463 | 0,933695887 | 0,003057007 |
| RandomForestClassifier (max\_depth=12) | 0,91797718 | 0,91972831 | 0,92547552 | 0,92592593 | 0,92327044 | 0,922475476 | 0,003140682 |

Realizando la validación de los modelos, se encuentran sobreajustados y se identifica aún un desbalanceo en los datos:

TABLA 11

FRECUENCIA DE GRAVEDAD

|  |  |  |
| --- | --- | --- |
| **GravedadCod** | **GravedadNombre** | **Cantidad** |
| 2 | Con Heridos | 7669 |
| 3 | Solo Danos | 23703 |

Por lo anterior se procede a balancear los datos utilizando over\_sampling, con lo cual se aumenta de forma aleatoria (RandomOverSampler) la cantidad de datos del grupo con menor cantidad de datos y entrenamos nuevamente los modelos con los datos balanceados.

A continuación, los resultados obtenidos por cada modelo en la validación cruzada, en los cuales encontramos que los datos mejoraron, sin embargo, pueden continuar sobreajustados, de lo cual obtuvimos los siguientes resultados:

TABLA 12

RESULTADOS SEGUNDA MATRIZ DE VALIDACION CRURZADA

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MODELO** | **RESULTADOS** | | | | | **MEDIA** | **DESV.** |
| XGBClassifier (max\_depth=12) | 0,874 | 0,879 | 0,888 | 0,894 | 0,893 | 0,886 | 0,008 |
| DecisionTreeClassifier (max\_depth=12) | 0,854 | 0,848 | 0,857 | 0,859 | 0,866 | 0,857 | 0,006 |
| GaussianNB() | 0,798 | 0,808 | 0,800 | 0,811 | 0,808 | 0,805 | 0,005 |
| LinearDiscriminantAnalysis (n\_components=1) | 0,821 | 0,827 | 0,827 | 0,830 | 0,827 | 0,826 | 0,002 |
| RandomForestClassifier (max\_depth=15) | 0,863 | 0,862 | 0,874 | 0,874 | 0,876 | 0,870 | 0,005 |

Teniendo en cuenta los datos obtenidos en la validación cruzada elegimos dos modelos para evaluar (GaussianNB y LinearDiscriminantAnalysis) ya que son los que menos sobreajuste presentan:

TABLA 13

PRIMERA EVALUACIÓN DE MODELOS SELECCIONADOS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MODELO** | **VALOR** | **precision** | **recall** | **f1-score** | **support** |
| GaussianNB() | Gravedad Tipo 3 | 0,7 | 0,81 | 0,75 | 2266 |
| Gravedad Tipo 2 | 0,94 | 0,89 | 0,91 | 7146 |
|  | | | | |
| accuracy |  |  | 0,87 | 9412 |
| macro avg | 0,82 | 0,85 | 0,83 | 9412 |
| weighted avg | 0,88 | 0,87 | 0,87 | 9412 |
| LinearDiscriminantAnalysis (n\_components=1) | Gravedad Tipo 3 | 0,7 | 0,81 | 0,75 | 2266 |
| Gravedad Tipo 2 | 0,94 | 0,89 | 0,91 | 7146 |
|  | | | | |
| accuracy |  |  | 0,87 | 9412 |
| macro avg | 0,82 | 0,85 | 0,83 | 9412 |
| weighted avg | 0,88 | 0,87 | 0,87 | 9412 |

Realizamos la evaluación del modelo y encontramos que aún es posible un sobreajuste. Dado lo anterior se procede a optimizar y/o mejorar los hiperparámetros iniciales de los modelos. En este caso bajamos la profundidad de los árboles para reducir la cantidad de validaciones y bajamos la tasa de aprendizaje a 0.001.

TABLA 14

RESULTADOS SEGUNDA MATRIZ DE VALIDACION CRURZADA

| **MODELO** | **RESULTADOS** | | | | | **MEDIA** | **DESV.** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| XGBClassifier (max\_depth=7, learning\_rate=0.001) | 0.832 | 0.836 | 0.832 | 0.840 | 0.835 | 0.835 | 0.002 |
| DecisionTreeClassifier (max\_depth=7) | 0,832 | 0,838 | 0,833 | 0,840 | 0,836 | 0,836 | 0,003 |
| GaussianNB() | 0,798 | 0,808 | 0,800 | 0,811 | 0,808 | 0,805 | 0,005 |
| LinearDiscriminantAnalysis (n\_components=1) | 0,821 | 0,827 | 0,827 | 0,830 | 0,827 | 0,826 | 0,002 |
| RandomForestClassifier (max\_depth=7) | 0,824 | 0,830 | 0,830 | 0,833 | 0,831 | 0,830 | 0,002 |

# resultado

Dada la última validación cruzada con los datos balanceados y con los modelos optimizados, encontramos que los modelos que menor sobreajuste presentan son GaussianNB y LinearDiscriminantAnalysis. Para elegir entre estos dos mostramos classification\_report para comparar los datos predichos con los datos de prueba:

TABLA 15

SEGUNDA EVALUACIÓN DE MODELOS SELECCIONADOS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MODELO** | **VALOR** | **precision** | **recall** | **f1-score** | **support** |
| GaussianNB() | Gravedad Tipo 3 | 0,7 | 0,81 | 0,75 | 2266 |
| Gravedad Tipo 2 | 0,94 | 0,89 | 0,91 | 7146 |
|  | | | | |
| accuracy |  |  | 0,87 | 9412 |
| macro avg | 0,82 | 0,85 | 0,83 | 9412 |
| weighted avg | 0,88 | 0,87 | 0,87 | 9412 |
| LinearDiscriminantAnalysis (n\_components=1) | Gravedad Tipo 3 | 0,7 | 0,81 | 0,75 | 2266 |
| Gravedad Tipo 2 | 0,94 | 0,89 | 0,91 | 7146 |
|  | | | | |
| accuracy |  |  | 0,87 | 9412 |
| macro avg | 0,82 | 0,85 | 0,83 | 9412 |
| weighted avg | 0,88 | 0,87 | 0,87 | 9412 |

Los modelos tienen mediciones muy similares, pero GaussianNB tiende a requerir más recursos máquina, por tanto, elegimos LinearDiscriminantAnalysis como el modelo de clasificación que mejor se ajusta.

Para validar el poder de predicción del modelo vamos a probarlo con los datos de prueba, con la medición de la precisión macro y la matriz de confusión:

LinearDiscriminantAnalysis(n\_components=1)

● precision\_score: 0.8187601098127401

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Figura 3 Matriz de confusión

Teniendo en cuenta los resultados de los modelos, encontramos que el modelo LinearDiscriminantAnalysis es el que mejor precisión tiene respecto a la clasificación de los dos tipos de gravedad, obteniendo una precisión de 0.8187.

Es importante tener en cuenta que el modelo puede estar un poco sobreajustado dada la dispersión de los datos, sin embargo, logra ajustarse con lo requerido.

Adicionalmente con el fin de realizar un análisis diferente realizamos la comparación de los datos generados por el modelo seleccionado por la plataforma VERTEX (Herramienta de Auto Machine Learning de Google Cloud Plataform). En el cual también visualizamos un posible sobreajuste y se seleccionó como mejor modelo Boosted Tree (Árboles impulsados).

Tabla

Descripción generada automáticamente

Figura 4 Evaluación del modelo generado por AutoML

Imagen que contiene Gráfico

Descripción generada automáticamente

Figura 5 Matriz de confusión del modelo generado por AutoML

# DISCUSION

Es importante la evaluación y construcción del modelo ya que nos permitirá identificar la gravedad de los accidentes viales en la ciudad de Bogotá, lo que servirá como insumo para enviar la atención y ayuda necesaria acorde a la gravedad, lo cual representará una minimización en daños materiales y en la afectación de la salud y/o vida de los involucrados. Por esta razón tomamos como evaluador la precisión teniendo en cuenta lo importante que es clasificar de manera correcta estos eventos. Adicionalmente garantizamos que la información estará mejor controlada y clasificada para que las personas que lo tomen de insumo puedan tomar mejores decisiones y/o crear planes de acción.

Una forma de implementar y usar nuestro modelo podría ser una aplicación que capture todos los campos/variables por las entidades reguladoras como la secretaría de movilidad distrital o unidades de emergencia que requieran validar y clasificar la gravedad de los siniestros viales. Es importante realizar un control y evaluación periódica del modelo teniendo en cuenta la evolución y desarrollo de la ciudad y el comportamiento en la movilidad.

# CONCLUSIONES

- El conocimiento del dataset permite entender las variables más importantes que pueden apoyar el desarrollo del modelo, pero también las que realmente no aportan o simplemente generan ruido dentro de los datos.

- Es importante la exploración de los datos ya que podemos encontrar muchos registros/datos atípicos que logran sesgar la información.

- La encodificacion de los datos permite un mejor control y para los modelos permite un mejor procesamiento.

- El balanceo de datos permite mejorar la distribución y clasificación de los datos.

- La optimización y mejoramiento de los hiperparametros logran ajustar los modelos y reducen posibles sobre ajustes.

# REFERENCIA

[1] Lorena Fernanda Sacristán Celis, “Análisis estadístico sobre las influencias del entorno en los accidentes de tránsito en la ciudad de Bogotá,” 2016. Accessed: May 13, 2022. [Online]. Available: https://repositorio.uniandes.edu.co/bitstream/handle/1992/18957/u729492.pdf?sequence=1

[2] J. S. Edwin Urbano, “Modelo de predicción de gravedad de accidentes de tránsito: un análisis de los siniestros en Bogotá, Colombia,” Bogotá, 2018. Accessed: May 13, 2022. [Online]. Available: https://vicisev.institutoivia.org/wp-content/uploads/2018/11/edwin-urbano-CISEV-Articulo\_gravedad\_accidentes-VRFNL.pdf

[3] NICK HOTZ, “What is the Data Science Process? - Data Science Process Alliance.” https://www.datascience-pm.com/data-science-process/ (accessed May 13, 2022).