

Documentación prueba técnica para científico de datos Axity

Caso de deserción estudiantil



7 de octubre de 2024

Fabian enrique Pedreros camargo

Contenido

[Alcance 2](#_Toc179162139)

[Metodología 2](#_Toc179162140)

[1. Entendimiento del negocio o problema. 4](#_Toc179162141)

[2. Aproximación analítica. 4](#_Toc179162142)

[3. Requerimientos de los datos. 4](#_Toc179162143)

[4. Recolección de los datos 4](#_Toc179162144)

[5. Limpieza y Entendimiento de los datos – EDA 5](#_Toc179162145)

[6. Preparación de los datos 18](#_Toc179162146)

[7. Modelado 18](#_Toc179162147)

[8. Implementación 20](#_Toc179162148)

[9. Feedback 20](#_Toc179162149)

[10. Deuda técnica 21](#_Toc179162150)

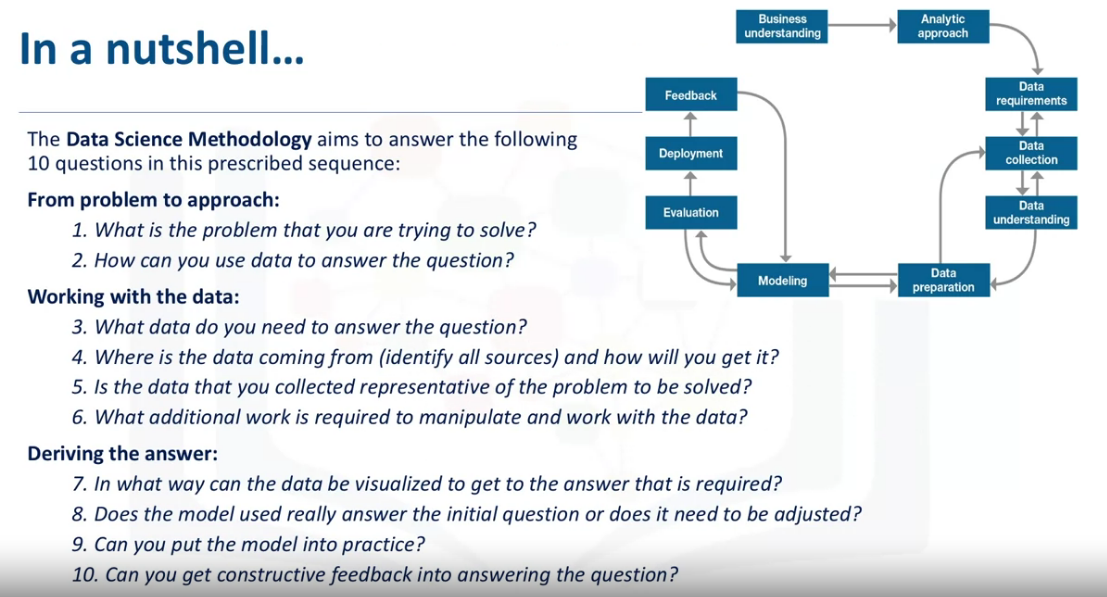
# Alcance

El presente documento tiene como alcance el servir como informe del proceso para dar una solución al problema planteado como prueba para el cargo de científico de datos para la compañía Axity, por lo tanto, en este se incluye un resumen del proceso realizado bajo la metodología aplicada por IBM para la solución de problemas relacionados con ciencia de datos, las decisiones tomadas en el desarrollo de las actividades, y finalmente las recomendaciones basadas en los resultados del modelamiento.

Se obvian los pasos para la implementación y retroalimentación dado el alcance de la prueba.

# Metodología

Se selecciona para el desarrollo de la prueba técnica la metodología para la ciencia de datos aplicada por IBM, que a su vez esta basada en CRISP – DM y busca ejecutar diez actividades principales de una manera cíclica, abarcando desde el entendimiento de la problemática hasta la retroalimentación por parte de los interesados (stakeholders) del proyecto. A continuación, se detallan cada uno de los pasos propuestos por la metodología y las preguntas a resolver.



Actividades enmarcadas en la metodología para ciencia de datos de IBM.

1. **Entendimiento del negocio o problema.** En esta etapa se debe entender el problema a resolver y las necesidades que tienen los stakeholders, la definci{on debe ser calara y no dejar dudas del producto final y la aplicación que se debe implementar. La pregunta por resolver es: ¿Cuál es el problema que se requiere resolver?
2. **Aproximación analítica.** Aquí se especifica cual es la solución a generar, esta depende del problema a resolver y el tipo de analítica a aplicar, pueden ser soluciones de tipo descriptivo (¿Qué paso?), diagnostico (¿Por qué paso?), predictivo (¿Va pasar?) o prescriptivo (¿Cómo se resuelve?). La pregunta por resolver es: ¿Cómo podemos usar los datos para resolver el problema?
3. **Requerimientos de datos.** En esta fase se debe especificar cuales son los datos utilizar, si estos ya existen o no y que proceso debe ser aplicado para su obtención. Esto se hace resolviendo preguntas relacionadas a los datos con; ¿Quién? ¿Cómo? ¿Cuándo? ¿Dónde? ¿Qué? ¿Por qué? La pregunta por resolver es: ¿Qué datos son requeridos para resolver el problema?
4. **Recolección de los datos.** Se deben identificar las fuentes de los datos y sus formatos, así como ejecutar la recolección de los datos o el levantamiento de los mismos. La pregunta por resolver es: ¿De dónde vienen los datos y cómo obtenerlos?
5. **Entendimiento de los datos (EDA Exploratory Data Analysis).** En este apartado se realiza la revisión de los datos, buscando establecer la completitud y calidad requerida, así como de entender cuál es el comportamiento de los datos a través del análisis exploratorio basado en estadística descriptiva. La pregunta por resolver es: ¿Los datos recolectados son representativos del problema a resolver?
6. **Preparación de los datos.** Aquí se dan las transformaciones requeridas para el uso de los datos como ingredientes de la aproximación analítica seleccionada. Se ejecutan tareas como la ingeniería de atributos. La pregunta por resolver es: ¿Qué tareas adicionales son requeridas para trabajar con los datos?
7. **Modelado.** Según el acercamiento seleccionado se entrena un modelo que permita resolver el problema planteado.
8. **Evaluación.** Se debe revisar con los stakeholders si el modelo seleccionado cumple con los requisitos, resuelve el problema y es aplicable. La pregunta por resolver es: ¿El modelo resuelve realmente el problema inicial o debe ser ajustado?
9. **Implementación.** En esta fase se desarrolla o implementa en un ambiente de pruebas o productivo el producto para que pueda ser usado por los stakeholders. La pregunta por resolver es: ¿Se puede poner el modelo en uso?
10. **Retroalimentación.** El modelo debe ser evaluado y se pueden dar mejoras o modificaciones al modelo planteado inicialmente, buscando que el problema sea resuelto según las métricas y definiciones de éxito. La pregunta por resolver es: ¿Se puede obtener retroalimentación constructiva para resolver el problema?

# Entendimiento del negocio o problema.

El problema planteado en la prueba es el de: “identificar la probabilidad de deserción estudiantil”, además de presentar un dashboard con los resultados obtenidos, de manera que usuarios no técnicos puedan usar los resultados del modelo y ahondar en los comportamientos de los datos y las predicciones hechas.

# Aproximación analítica.

La aproximación analítica es dada en las instrucciones de la prueba, esta obedece a una analítica predictiva, ya que se debe de entrenar un modelo supervisado para la clasificación de la deserción de los estudiantes, así como de la obtención de la probabilidad de deserción. Dadas las instrucciones y la aproximación, se debería de entrenar un modelo basado en una regresión logística, permitiendo la clasificación binaria (deserción si o no) y obtener la probabilidad de la clasificación dada, sin embargo, se indica la necesidad de usar tres distintos algoritmos, los seleccionados son Regresión logística, Lineal SVM classification y arboles de decisión.

Las métricas o medidas de desempeño para determinar cual de los modelos es el mejor serán: el recall, dado que desde la perspectiva del negocio se busca atender a la mayor totalidad de estudiantes que estén en riesgo de desertar.

# Requerimientos de los datos.

El set de datos inicial esta dado por los datos compartidos en la prueba, estos se encuentran en el archivo “BD\_2024.csv”, son datos estructurados almacenados en un formato tabular separado por punto y coma. El archivo contiene 542,123 registros y 73 columnas, los registros parecen obedecer a los datos registrados tras la finalización de un semestre para cada estudiante y materia cursada en un programa educativo. Los atributos obedecen a datos demográficos del estudiante, datos académicos, el programa, la materia, el puntaje obtenido en simulacros y puntajes del Saber Pro. Se requieren hacer transformaciones a los datos, dado que la granularidad parece encontrarse a nivel de estudiante y materia por programa, y el problema planteado indica una granularidad de estudiante y programa.

La columna “Estado” indica si el estudiante deserto o no del programa educativo, por lo que esta es utilizada como la etiqueta a determinar mediante el modelo de clasificación.

# Recolección de los datos

La recolección de los datos se simplifica dada la fuente de los mismos, por lo que la consulta al documento “BD\_2024.csv” se realiza de manera directa en el proceso de EDA.

# Limpieza y Entendimiento de los datos – EDA

La limpieza de los datos se efectúa en el archivo llamado “LimpiezaDatos.ipynb”, mientras que la exploración de los datos se realiza en el archivo “EDA.ipynb” esto para facilitar la visualización del proceso y los hallazgos realizados. Para la ejecución de estas actividades se hace uso de estadística descriptiva y visualización de datos, así como del lenguaje de programación Python y frameworks que facilitan el procesamiento de datos estructurados. A continuación, se comparten algunos de los hallazgos más relevantes o aspectos sobre los cuales hay alguna duda.

A continuación, se señalan algunas de las decisiones tomadas respecto a la limpieza de los datos:

- Se muestran atributos con valores nulos y otros que en la revisión general no tienen, pero debe ser más específica, dado que pueden haber valores nulos tratados como valores.

- La columna llamada "2" se debe eliminar del dataset dado que no contiene valores, estando totalmente vacía.

- La columna con nombre "1" se elimina dado que contiene tan solo 137912 valores y estos parecen obedecer a una concatenación de distintas columnas para los últimos registros.

- "Municipio Nacimiento", "Nombre Adicional(MAT. Equiv.)", "Municipio Colegio", se deciden eliminar dada la gran cantidad de valores categóricos y que obedecen a ubicaciones muy específicas con pocas ocurrencias

- "ID Departamento", "ID Municipio Nacimiento", "SNIES", "Plan de Estudios ID.", "Plan de Estudios Cód", "Código Materia", "Código AC ICFES" se eliminan dado que son identificadores

- "Nivel" se elimina dado que tiene un único valor.

- Se decide eliminar los puntajes de simulacro, "Año Presentación Simulacro Saber Pro" y los de puntaje del Saber Pro, dado que estos se desarrollan en instancias muy adelantadas de la carrera, siendo incluso requisitos para el grado, por lo que pueden representar un bias y leaking de datos.

- Se deben modificar los tipos de datos para las siguientes columnas, pasándolas de int a object:

- ID

- Año Académico

- Periodo Académico

- SNIES

- Plan de Estudios ID.

- Tipo Plan

- Se deben modificar los tipos de datos para las siguientes columnas, pasándolas de float a int:

- Semestre Ubicación

- Numero crédito de la materia

- Se deben modificar los tipos de datos para las siguientes columnas, pasándolas de object a int:

- Nota Normal --> Presenta valores de texto "APRO" y "REPR" que deben ser tratados si se quiere hacer uso de una nota promedio, son 39528 registros.

- Promedio Ponderado de la materia

- Puntaje Simulacro Razonamiento Cuantitativo --> Revisar valores con texto

- Puntaje Simulacro Lectura Crítica --> Revisar valores con texto

- Puntaje Simulacro Inglés --> Revisar valores con texto

- Puntaje Simulacro Competencias Ciudadanas --> Revisar valores con texto

- Puntaje Simulacro Puntaje Global --> Revisar valores con texto

- PUNT\_LECTURA\_CRITICA --> Revisar valores con texto

- PUNT\_MATEMATICAS --> Revisar valores con texto

- PUNT\_C\_NATURALES --> Revisar valores con texto

- PUNT\_SOCIALES\_CIUDADANAS --> Revisar valores con texto

- PUNT\_INGLES --> Revisar valores con texto

- PUNT\_GLOBAL --> Revisar valores con texto

- PROMEDIO\_1er\_SEMESTRE --> Revisar valores con texto

- PROMEDIO\_2do\_SEMESTRE --> Revisar valores con texto

- Se deben modificar los tipos de datos para las siguientes columnas, pasándolas a datetime

- Fecha de Nacimiento

- Fecha de Ingreso

Se observan múltiples ocurrencias de un mismo ID para distintos SNIES, esto puede ocurrir por el paso o el curso de distintos programas por una misma persona. Por lo tanto, la granularidad de los datos se debe llevar a estudiante y SNIES, esto lo da la columna ID\_SNIES, lo que tiene lógica desde el negocio, dado que un estudiante puede tener distintos comportamientos y cursar varios programas.

Hay valores en Departamento Nacimiento que son agrupados en uno mismo:

- Atlántico: Atlantico, Atlántico, ATLÁNTICO

- Bogotá D.C.: Bogota D.C., Bogotá D.C., BOGOTA DC

- Bolívar: Bolivar, Bolívar

- Boyacá: Boyaca, BOYACA, Boyacá

- Cesar: Cesar, CESAR

- Córdoba: Córdoba, CÓRDOBA

- Guainía: Guainía, Guanía

- Huila: Huila, HUILA

- Magdalena: Magdalena, MAGDALENA

- Meta: Meta, META

- Norte de Santander: Norte de Santander, Norte De Santander

- Quindío: Quindio, Quindío

- Valle del Cauca: Valle del Cauca, Valle Del Cauca

- Archipiélago de San Andrés, Providencia y Santa Catalina: "Archipiélago de San Andrés, Providencia y Santa Catalina", San Andrés

Se agrupan los valores en un nuevo atributo para el Estado del estudiante buscando reflejar la deserción de la siguiente manera:

- Si --> Desertor

- No --> Activo, Graduado

Debido a la elevada presencia de valores nulos se resuelve no hacer uso de algunos atributos como:

- Tipo de Colegio                  31.398639

- Recibió Apoyo Financiero         36.300375

- Recibió Apoyo Académico          36.300375

- Recibió Otros Apoyos             36.300375

- FAMI\_ESTRATOVIVIENDA             87.015370

- FAMI\_PERSONASHOGAR               86.977135

- FAMI\_CUARTOSHOGAR                86.984782

- FAMI\_EDUCACIONPADRE              87.007723

- FAMI\_EDUCACIONMADRE              87.015370

- FAMI\_TRABAJOLABORPADRE           91.588285

- FAMI\_TRABAJOLABORMADRE           91.592108

- FAMI\_TIENEINTERNET               86.996253

- FAMI\_TIENECOMPUTADOR             86.988606

- FAMI\_TIENEAUTOMOVIL              86.988606

- FAMI\_TIENEMOTOCICLETA            91.599755

- FAMI\_NUMLIBROS                   87.328898

- ESTU\_HORASSEMANATRABAJA          98.971477

- ESTU\_TIPOREMUNERACION            91.618873

- COLE\_AREA\_UBICACION              86.766843

- Semestre AC                      80.767760

- Semestre Coincidencia            86.751549

Se decide la eliminación de registros con valores nulos en los siguientes atributos:

- Sexo                           17.618720

- Estado Civil                     10.476409

- Etnia                            19.148123

- Discapacidad                     19.148123

- Departamento Nacimiento          21.140170

- PROMEDIO\_1er\_SEMESTRE             0.019118

- PROMEDIO\_2do\_SEMESTRE             8.331422

- Fecha de Nacimiento corregida    19.985471

- Fecha de Ingreso corregida       19.962530

- Desertor                          2.733807

- Semestre Ingreso                  0.814407

De los datos de las materias , periodos y semestre se pueden extraer los siguientes atributos a nivel de ID\_SNIES, para ser agragados a los datos que ya se tienen.

- Cantidad de materias vistas

- Cantidad de cursos libres reprobados

- Cantidad de cursos libres aprobados

- Promedio de las materias

- Cantidad de materias perdidas con nota menor a 3

- Cantidad de creditos inscritos

- Semestre Ubicación

Se ejecuta análisis descriptivo univariable y multivariable, aquí se destacan algunos de los hallazgos y decisiones tomadas:

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Hay una mayor cantidad de registros cuyo sexo es masculino, por lo que debe validarse si es u comportamiento normal de los datos o hay un bias en la recolección.

Gráfico, Gráfico de barras

Descripción generada automáticamente

El 99.1% de los registros son Solteros, de manera que puede que este sea un atributo a no ser utilizado.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Se debe consultar a que se refiere con el atributo Tipo Plan, pero hay una mayor cantidad de registros que pertenecen a el tipo 2.

Gráfico

Descripción generada automáticamente

El set de datos contiene un 14.7% de registros que son etiquetados como desertores.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Al ser datos históricos se tiene una gran cantidad de registros de los últimos semestres, por lo que se espera una elevada cantidad en los semestres iguales al 10 o el primero (al tratarse de pregrado), pero llama la atención que hay un elevada cantidad en semestres intermedios, debe validarse si hay alguna relación con la deserción.

Gráfico, Histograma

Descripción generada automáticamente

Se observa una distribución con un sesgo hacia la derecha, es decir que tiene una cola izquierda. Lo que indica que los estudiantes tienden a ver entre 50 y 60 materias, esto hace lógica dado que tenemos datos históricos con estudiantes ya graduados.

Gráfico, Histograma

Descripción generada automáticamente

El histograma con el promedio de las notas de las materias muestra como la distribución cuenta con una media cercana a 4, mientras que existen pocos registros con menos de 3. Esto nos lleva a pensar que son más los estudiantes que se gradúan o por lo menos lo pueden hacer según las notas obtenidas, que los que no logran graduarse. Además se observa una variación baja, dada la densidad del gráfico.

Gráfico

Descripción generada automáticamente

El gráfico nos indica que existe una mayor cantidad de estudiantes de sexo Femenino que desertan, en especial si lo comparamos con el 12.5% de hombres que lo hace.

Puede ser que el sexo sea un atributo relevante para determinar la probabilidad de que un estudiante desierte.

Gráfico

Descripción generada automáticamente

Al parecer existe una alta probabilidad de que las personas casadas deserten, sin embargo, hay que tener en cuenta que la población casada es muy poca y esto puede llevar a comportamientos muy extremos.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Se visualizan diferencias entre los porcentajes de personas que desertan según la etnia a la que pertenece, pudiendo existir una mayor deserción cuando se pertenece a pueblos RROM (Gitanos), sin embargo, la cantidad de registros es muy baja y podrían darse resultados que no son estadísticamente correctos.

Al parecer existen programas con un mayor grado de deserción, siendo los tres que tienen una mayor cantidad de registros historicos:

- administración de empresas turísticas y hoteleras con el 21%

- finanzas y relaciones internacionales con 19.5%

- historia con 18.3%

Además de una baja cantidad de deserción en programas como:

- ciencia de datos con 3.4%

- Filosofia y arqueología con 7.4%

- conservación y restauración del patrimonio cultural mueble con el 8%

Por lo tanto, el atributo del programa puede estar fuertemente relacionado con la deserción.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Dado que los programas influyen en los datos al revisarlos por facultad, vemos comportamientos muy similares y que se pueden relacionar a la deserción, siendo la facultad de "administración de empresas turísticas y hoteleras" la que más estudiantes desertados presenta y la de menor cantidad es la de "dirección de matemáticas"

Gráfico, Gráfico de barras

Descripción generada automáticamente

Parece ser que los semestres en los que más desertan los estudiantes son los cuatro primeros (obviar el de semestre cero que es un registro) y el de mayor porcentaje de deserción tiene es el semestre 11.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Según el comportamiento observado en los box plot, los desertores ven menos materias que los que no desertan, esto tiene lógica dado que si te gradúas vas a ver una cantidad elevada de materias. Además, se obeserva como los que desertan en un 2% de los casos ven más de 50 materias, así que a deserción no solo ocurren en los primeros semestres.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Si bien el promedio de desertores es menor al de los que no desertan, no se ve  un comportamiento tan drástico, por lo que se podría pensar que la mayoría de los desertores no lo hacen por que tengan malas notas, ya que menos del 25% presentan un promedio menor al 3.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Como era de esperarse los que desertan pierden más materias que los que no lo hacen, pero esto tambien puede deberse a las normas de la universidad.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Las personas que desertan inscriben menos créditos que los que no, en el 75% de los casos que desertaron lo hicieron antes de inscribir 80 créditos, de manera que un estudiante que supera esta barrera tiene menos probabilidad de que desierte.

Das las revisiones mostradas y el entendimiento de los datos, se decide ejecutar el modelado con algunos de los atributos iniciales y otros generados mediante ingeniería de atributos, siendo estos:

* 'Sexo',
* 'Estado Civil',
* 'Discapacidad',
* 'Departamento Nacimiento',
* Programa',
* 'Facultad',
* 'Tipo Plan',
* 'PROMEDIO\_1er\_SEMESTRE',
* 'PROMEDIO\_2do\_SEMESTRE',
* 'Desertor',
* 'Cantidad de materias vistas',
* 'Promedio de las materias',
* 'Cantidad de materias perdidas',
* 'Cantidad de creditos inscritos',
* 'Semestre Ubicación'

# Preparación de los datos

La preparación y transformación de los datos para ejecutar el entrenamiento del modelado se da en el archivo PreparacionDatos.ipynb, y estos son algunas de las daciones tomadas y las actividades allí ejecutadas.

Las transformaciones seleccionadas para los datos son las siguientes:

1. Datos numéricos a los que se les aplica MinMaxScaler:

    - PROMEDIO\_1er\_SEMESTRE

    - PROMEDIO\_2do\_SEMESTRE

    - Cantidad de materias vistas

    - Promedio de las materias

    - Cantidad de materias perdidas

    - Cantidad de creditos inscritos

2. Datos categóricos a los que se les aplica la creación de variables dummy:

    - Sexo

    - Estado Civil

    - Discapacidad

    - Departamento Nacimiento

    - Programa

    - Facultad

    - Tipo Plan

    - Semestre Ubicación

Finalmente, los datos preparados son almacenados en el archivo con nombre “datos\_transformados.csv”

# Modelado

El modelado de los datos se da en el archivo “modelado.ipynb”, en este se ejecutan tareas como la creación de los sets de datos de entrenamiento y prueba, así como el entrenamiento de los modelos seleccionados: Regresión logística, Árbol de decisión y SVC.

Se ejecuta un entrenamiento inicial partiendo de los hiperparametros por defecto y con el acercamiento de cross validation, de manera que se obtengan modelos entrenados con la totalidad de los datos y reflejen un mejor comportamiento durante su uso.

Los resultados del recall, que es la métrica seleccionada se muestran a continuación.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Como se puede observar los mejores resultados para el entrenamiento se dan para el modelo de árbol de decisión (DT) con un recall que super el 60%

El comportamiento de este modelo se mejora al ejecutarlo en los datos de prueba, mostrando un incremento en la métrica de cerca de un 28%

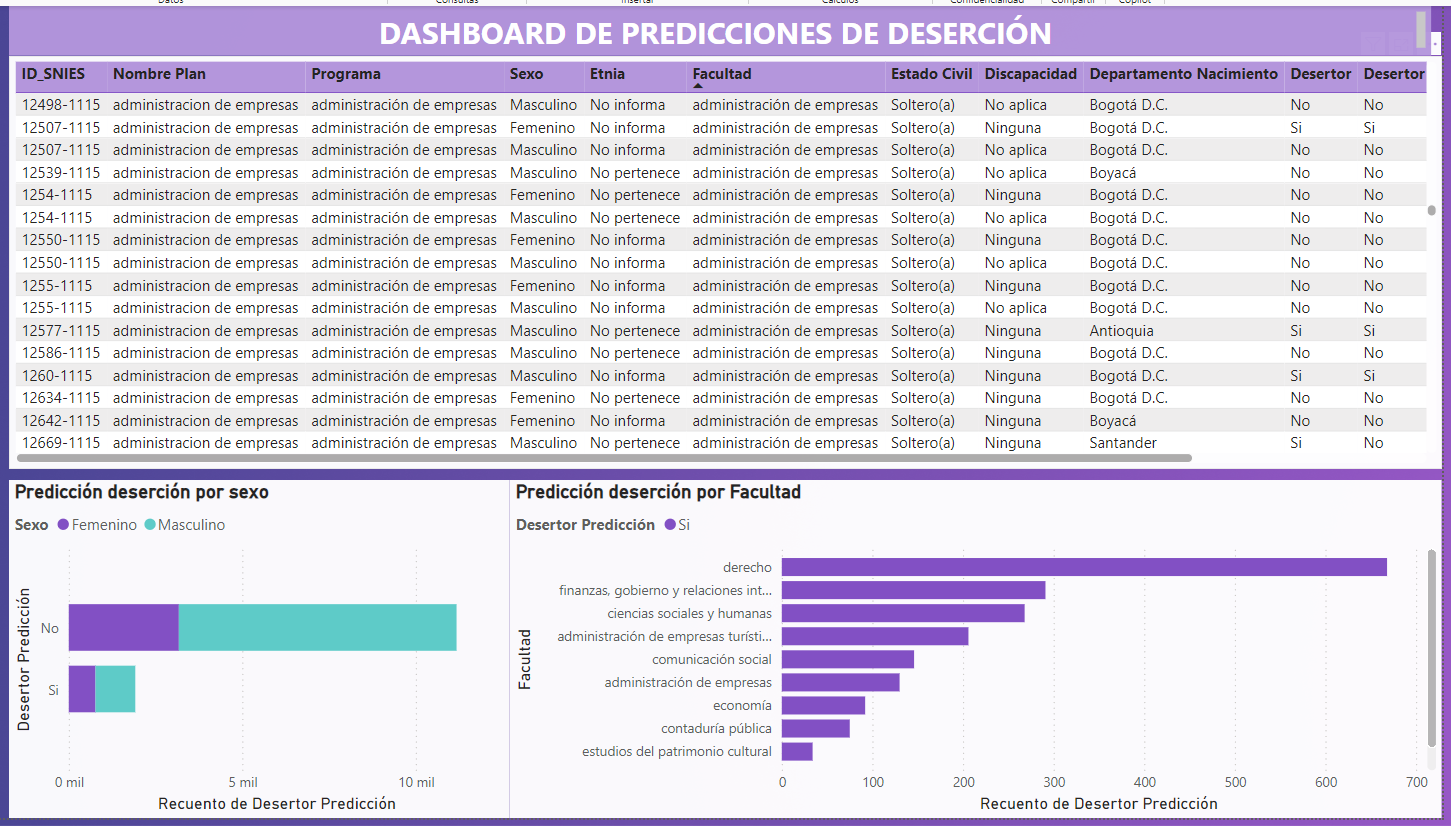
Se decide ejecutar el entrenamiento de los mismos algoritmos pero incluyendo Grid Search, de manera que se busca optimizar los hiperparametros y así alcanzar un mayor recall. Sin embargo, los resultados tanto en el set de entrenamiento como en el de pruebas no superan al modelo de árbol de decisión estándar.

El mejor modelo entrenado fue un árbol de decisión con los hiperparámetros por defecto. Alcanzado un recall de 89.1% en el set de datos de prueba. Este es utilizado para la generación de las predicciones y probabilidades para todo el set de datos y los resultados son almacenados en el archivo “datos\_solucion.csv”

# Implementación

A manera de implementación se crea un dashboard en Power BI con el nombre “Dashboard predicciones de deserción.pbix”, en este se muestra un tabal con los datos más relevantes de los estudiantes, incluyendo la predicción de deserción realizada y su probabilidad de ocurrencia.

Este dashboard permite navegar en los distintos atributos de los estudiantes e irlos comparando contra la predicción de deserción realizada con el modelo de árbol de decisión entrenado.



# Feedback

Como feedback de los procesos realizados, se tendrían los siguientes:

* El proceso de limpieza de los datos se pudo haber mejorado si se revisaran los datos nulos inicialmente, dado que se habrían descartados algunos atributos de manera temprana.
* Se crean atributos adicionales a partir de los datos de las materias y las notas, igualmente se pudieron haber generado algunos relacionados a los créditos.
* En la selección de los modelos se debe incluir el RandomForestClassifier, dado que puede presentar mejores resultados.
* Se requiere revisar la importancia de los atributos y probar a entrenar modelos más simples o complejos que obtengan mejores resultados en los sets de entrenamiento y pruebas.

# Deuda técnica

Como deuda técnica se tiene:

* La necesidad de operacionalizar todo el flujo, desde la recolección de los datos desde la fuente que se establezca, pasando por la preparación de los mismos, el entrenamiento del modelo y validación respecto a la alteración de la métrica, con el fin de realizar actividades de despliegue automático en caso de alteraciones en los valores o limites que se establezcan.
* Se debe establecer la forma en que el modelo se va a desplegar y consumir para la alimentación del dashboard de Power BI, se puede pensar en un servicio API de consumo de datos, procesamiento y carga a un repositorio de consulta.
* Se debe incluir el despliegue del dashboard, de manera que los usuarios puedan acceder a este y actualizar los datos a media que estos se sumen.