Escuela de Administración

Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Herramienta de data analytics para pruebas de auditoría interna del proceso misional del SENA

Presentado por:

Yeimy Paola Niño Castañeda

Bogotá, D.C.

2023

Escuela de Administración

Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Herramienta de data analytics para pruebas de auditoría interna del proceso misional del SENA

Presentado por:

Yeimy Paola Niño Castañeda

Bajo la dirección de:

Herbert Jair Bermudez Sosa

Maestría en Business Analytics

Escuela de Administración

Escuela de Ingeniería Ciencia y Tecnología

14 de julio de 2023

Bogotá, D.C., Colombia

2023

Contenido

Agradecimientos

A mi tutor del proyecto de grado, profesor Jair Bermudez Sosa, por su apoyo, guía y dedicación en la asesoría y revisión técnica de las diferentes entregas durante la ejecución del proyecto empresarial de la Maestría de Business Analytics y a la Oficina de Control Interno del Servicio Nacional de Aprendizaje Sena por su apoyo en la consecución de la información durante el desarrollo del proyecto.

Yeimy Paola Niño Castañeda

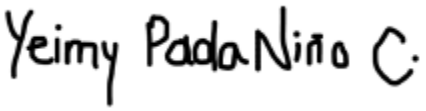
Dedicatoria

A mi familia y amigos por su apoyo y acompañamiento incondicional, por creer en mí y en este reto académico.

Declaración de originalidad y autonomía

Declaro bajo la gravedad del juramento, que he escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por mi propia cuenta y que, por lo tanto, su contenido es original.

Declaro que he indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

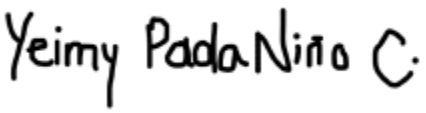


Yeimy Paola Niño Castañeda

Firmado en Bogotá, D.C. el 14 de julio de 2023

Declaración de exoneración de responsabilidad

Declaro que la responsabilidad intelectual del presente trabajo es exclusivamente de su autor. La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.



Yeimy Paola Niño Castañeda

Firmado en Bogotá, D.C. el 14 de julio de 2023

Lista de figuras

Lista de tablas

Abreviaturas

AUC: Área bajo la curva ROC

CRISP-DM: Cross-Industry Standard Process for Data Mining

ETL: Extracción, transformación y carga y carga de datos

IDE (Integrated Development Enviroment): Entorno de Desarrollo Integrado

SENA: Servicio Nacional de Aprendizaje SENA

Glosario

CRISP-DM Cross-Industry Standard Process for Data Mining: es un método probado para orientar sus trabajos de minería de datos. (IBM, 2021).

Decisiones Data Driven: (decisiones basadas en datos): se caracterizan por agrupar bases de datos robustas para tomar decisiones certeras y eficaces a nivel organizacional reemplazando así la toma de decisiones tradicional basada en intuiciones, observaciones y opiniones. (Ministerio de Tecnologías de la Información y las Comunicaciones, 2020, pág. 53)

ETL (Proceso de extracción, transformación y carga): es una canalización de datos que se usa para recopilar datos de varios orígenes. A continuación, transforma los datos según las reglas de negocio y los carga en un almacén de datos de destino. El trabajo de transformación en ETL tiene lugar en un motor especializado y, a menudo, implica el uso de tablas de almacenamiento provisional para conservar los datos temporalmente a medida que estos se transforman y, finalmente, se cargan en su destino. La transformación de datos que tiene lugar a menudo conlleva varias operaciones como filtrado, ordenación, agregación, combinación de datos, limpieza de datos, desduplicación y validación de datos. (Raunakjhawar, s. f.).

Herramientas de Analytics: aquellas que tienen como objetivo el tratamiento, inspección y transformación de los datos para obtener conclusiones que sirvan de base para la toma de decisiones en el ámbito de la auditoría. Su aplicación en el trabajo permite detectar errores, tendencias y fraudes gracias a las distintas técnicas estadísticas de análisis que llevan incorporadas. (Instituto de Censores Jurados de Cuentas de España, 2019).

Metodología ágil SCRUM: Scrum es un marco de trabajo liviano que ayuda a las personas, equipos y organizaciones a generar valor a través de soluciones adaptativas para problemas

complejos. En pocas palabras, Scrum requiere un Scrum Master para fomentar un entorno donde:

Un Product Owner ordena el trabajo de un problema complejo en un Product Backlog.

El Scrum Team convierte una selección del trabajo en un Increment de valor durante un Sprint.

El Scrum Team y sus interesados inspeccionan los resultados y se adaptan para el próximo Sprint. (Schwaber & Sutherland, 2020).

Resumen Ejecutivo

Debido al incremento del uso de analítica de datos para la toma de decisiones en las organizaciones, se hace necesario aplicar estas acciones en Entidades del Sector público para analizar los datos históricos y comunicar conclusiones acertados sobre el comportamiento de estos, por esta razón se diseñó para el área de auditoría interna de la Entidad estudiada una herramienta de Data Analytics, que permita profundizar en el análisis de la información y los datos que genera la Entidad a partir del desarrollo de pruebas de auditoría, previas a la etapa de desempeño de las auditorías en campo para obtener posibles alertas preventivas, realizar análisis descriptivo de las bases de datos, visualización de la información y aplicar modelos predictivos supervisados, lo cual generará valor y calidad en las auditorías efectuadas a los procesos misionales, exactamente al procedimiento de la formación profesional integral de los aprendices y así mismo, contribuirá al logro del objetivo estratégico de la Oficina de Control Interno de la Entidad y al rol que se ejecuta desde la tercera línea de defensa.

Palabras clave

Analítica de datos, modelos predictivos, auditoría interna, deserción de aprendices.

Abstract

Data analysis tool for internal audit testing of the missionary process

Due to the increased use of data analytics for decision making in organizations, it's necessary to apply these actions in the Public Sector Entities to analyze historical data and communicate correct conclusions about their behavior, for this reason it will be designed for Internal Audit Area of the Entity that I studied a data analytics tool that allows deepening the analysis of the information and data generated by the Entity from the development of audit tests, prior to the performance stage of the field audits to obtain possible preventive alerts, perform a descriptive analysis of the databases, visualization of the information and apply supervised predictive models, which will generate value and quality in the audits carried out on the mission processes, exactly to the procedure of the comprehensive professional training of the apprenticeships and thus contribute to the achievement of the strategic objective of the Office of Internal Control of the Entity and the role that is executed from the third line of defense.

Keywords

Data analytics, predictive models, internal audit, trainee desertion.

1. Introducción

La transformación digital de las Entidades de Gobierno y el uso de tecnologías emergentes en el sector público está incidiendo en que se desarrollen las decisiones a nivel estratégico usando decisiones basadas en datos (Decisiones Data Driven), las cuales están reemplazando las decisiones tradicionales basadas en la intuición, por lo cual se están migrando a tecnologías que permiten generar valor público.

Para promover el uso de estas tecnologías, las Entidades están implementando analítica de datos que les permiten transformar un conjunto de datos en información requerida para formular políticas públicas, es por esto que la perspectiva de las áreas de auditoría interna es que se efectúen análisis más profundos basados en la interacción de la integridad de la data, lo cual se traduce en añadir valor a la organización y lograr los objetivos estratégicos de la misma.

> Al fortalecer la analítica de datos en las áreas de auditoría interna se mejoraría la calidad de resultados comunicados, se involucraría la auditoría hacia los hechos que pueden ocurrir hacia adelante, se promovería el monitorio continuo y la mitigación de riesgos mediante el uso de tableros de control, así como también se construirían indicadores que permitan medir la efectividad de los controles. (Echeverría, 2019)

Actualmente, las áreas de auditoría interna cada vez más están implementando herramientas de Analytics “con el objetivo de hacer el tratamiento, inspección y transformación de los datos para obtener conclusiones que sirvan de base para la toma de decisiones en el ámbito de la auditoría” (Instituto de Censores Jurados de Cuentas de España, 2019), su aplicación reduce el tiempo de la auditoría en las diferentes etapas de estas, más aún en la etapa de aplicación de

pruebas y recolección de evidencias suficiente y competente, amplía el alcance de las auditorías al poder realizar análisis de la población completa de datos y ayuda a determinar un mayor juicio profesional del auditor para obtener conclusiones de mayor calidad.

Para desarrollar un modelo de análisis de datos, las áreas de auditoría interna tienen grandes retos, como, por ejemplo:

> El cambio cultural de los auditores que participan en los trabajos, así como el análisis de las capacidades analíticas del equipo de auditoria junto con el replanteamiento de las metodologías y enfoque de las auditorías para crear una visión y hoja de ruta de la estrategia técnica-analítica. (Deloitte, 2016)

Teniendo en cuenta lo anterior, en este proyecto se busca diseñar una herramienta de data analytics para aplicarla a las pruebas de auditoría de procesos misionales referentes a la formación profesional integral de la Entidad estudiada, en el cual se implementen las seis etapas del clico de vida de la ciencia de datos, así como la aplicación de analítica predictiva mediante la evaluación de tres modelos.

2. Objetivo

Diseñar una herramienta analítica que analice los programas donde se puede presentar deserción de estudiantes para emitir oportunamente alertas preventivas que permitan agregar valor en la toma de decisiones estratégicas de la etapa de planificación de las auditorías internas de la vigencia 2023.

2.1 Objetivos específicos

2.1.1 Objetivos estratégicos

1. Contribuir en la implementación de Data Analytics en las pruebas de auditoría efectuadas por las Oficinas de Control Interno del sector público correspondientes al orden nacional.

2. Analizar la población incluida en las datas de los procesos misionales para tomar decisiones estratégicas en la etapa de planificación de las auditorías, es decir, previo a la ejecución de auditorías en campo.

2.1.2 Objetivos de investigación

1. Evaluar el proceso misional de la Entidad mediante la visualización de datos en herramientas de Business Analytics y la aplicación de modelos predictivos supervisados.

2.1.3 Objetivos de Proyecto

1. Generar valor al procedimiento de trabajos de aseguramiento de la Oficina de Control Interno a partir del manejo de los datos mediante visualización en tableros de control.

2. Aplicar modelo predictivo supervisado para predecir los programas en los cuales se puede presentar deserción de estudiantes durante la vigencia 2023.

3. Alcance

El alcance del proyecto está delimitado para el proceso misional del Servicio Nacional de Aprendizaje SENA, delimitado al procedimiento de formación profesional de aprendices el cual se encuentra en cabeza de la Dirección de Formación Profesional, en este se encuentran el procedimiento de certificación académica de aprendices del cual se genera la deserción de aprendices de los procesos formativos. Así mismo, está delimitado al proceso de evaluación y control que es desarrollado por la Oficina de Control Interno de la Entidad, el cual impacta al procedimiento de trabajos de aseguramiento o de auditoría interna.

El alcance de análisis está delimitado a un análisis descriptivo, dado se aplicará estadística descriptiva para informar las tendencias de los datos, así mismo, está delimitado a un análisis predictivo dado que se diseñará una herramienta analítica mediante la aplicación de un modelo predictivo supervisado en la Oficina de Control Interno de la Entidad.

El alcance del análisis descriptivo y predictivo está delimitado a dos regionales de la Entidad, específicamente a las regionales Antioquia y Distrito Capital, por ser las regionales que presentan mayor deserción en los niveles de formación denominados curso especial, técnico y tecnólogo.

El alcance de los datos históricos para entrenar el modelo está delimitado para la vigencia 2022 y los datos potenciales para realizar las predicciones de los programas de formación en los cuales se puede presentar mayor deserción está delimitado para la vigencia 2023.

4. Metodología

4.1 Metodología para el proyecto de datos

Para el proyecto de datos se definió el uso de la metodología CRISP-DM, esta metodología según Galán Cortina (2015):

> Incluye un modelo y una guía, estructurados en seis fases, algunas de las cuales son bidireccionales, es decir que de una fase en concreto se puede volver a una fase anterior para poder revisarla, por lo que la sucesión de fases no tiene que ser ordenada desde la primera hasta la última. (p. 21)

A continuación, se realiza una breve descripción de las fases que se van a aplicar al diseño de la herramienta de data analytics para pruebas de auditoría interna del proceso misional:

4.1.1 Metodología CRISP-DM

Se usará la metodología CRISP-DM Cross-Industry Standard Process for Data Mining, para evaluar el ciclo de vida de los datos mediante la aplicación de las seis etapas en el proyecto, para las datas correspondientes a deserción de aprendices de las vigencias 2022 y 2023. Las seis etapas de pueden observar en la Figura 1., así:

Figura 1.

Metodología CRISP-DM

Nota. La figura muestra las fases de la metodología CRISP-DM. Fuente: Instituto de Ingeniería del Conocimiento (2021).

Las fases de la metodología son las que se describen a continuación y la aplicación de estas fases se va a desarrollar en el numeral 8.1 de este documento.

La primera fase es el entendimiento del negocio, en la que se requiere hacer un entendimiento del contexto de la Empresa o negocio donde se va a desarrollar el proyecto, la segunda fase es el entendimiento de los datos para verificar la calidad de los datos y el tipo de información que se va a usar para desarrollar el proyecto (data estructura o no estructurada). Como tercera fase tenemos la preparación de los datos, en el cual se ajustan los datos y se realizan las ETL extracción, transformación y carga para luego poder iniciar con el modelado de los datos en la cuarta fase, definir el modelo predictivo a utilizar en el proyecto y correr el

modelo con los datos seleccionados; posteriormente, se encuentra la quinta fase o fase de evaluación, en la cual se evalúa el modelo con mayor calidad o mayor ajuste y finalmente, se encuentra la fase de despliegue, en la cual se implementa el modelo analítico predictivo en la empresa o negocio definido en el proyecto.

Así como, se va a utilizar la metodología CRISP-DM para el proceso de análisis de los datos, se van a usar modelos predictivos supervisados para desarrollar el análisis predictivo, por lo cual a continuación se van a explicar los modelos y la razón teórica por la cual fueron seleccionados para aplicarlos en la predicción de deserción de programas de formación.

4.1.2 Modelos supervisados aplicados

Para el análisis se definió utilizar tres modelos predictivos supervisados, es decir, modelos que tienen una variable objeto definida, de los cuales se elegirá el modelo con mayor ajuste o AUC.

Se definió usar tres modelos predictivos, así: el árbol de clasificación, un gradient boosting y un random forest.

4.1.2.1 Árboles de decisión o de clasificación.

> Los árboles de decisión o de clasificación de acuerdo con Rivera Vergaray (2021)
>
> Proveen de una herramienta de clasificación muy potente. Su uso en el manejo de datos la hace ganar en popularidad dadas las posibilidades que brinda y la facilidad con que son comprendidos sus resultados por cualquier usuario. El árbol en sí mismo, al ser obtenidos, determinan una regla de decisión. (p. 6)

Los árboles de decisión fueron utilizados en un artículo en el cual se estudió la detección de estudiantes con riesgo de deserción académica, para el cual Rivera Vergaray, (2021) expresa:

> Luego de aplicar y evaluar los modelos aplicados al conjunto de datos extraídos de la base de datos del sistema académico, los mejores resultados se obtuvieron con el KNN y el árbol de decisión. En el modelo KNN se obtuvo un Accuracy de 88,844% y un error de 11,156%, y en el modelo de árbol de decisión se obtuvo un Accuracy de 88,4% y un Error de 11,6%. (p. 6)

Por lo anterior, en este artículo que también estudió deserción de estudiantes el modelo que obtuvo mayor exactitud fue el árbol de decisión, así mismo, en el estudio hacia la construcción de un modelo predictivo de deserción académica basado en técnicas de minería de datos se aplicó “el algoritmo J48 se obtuvo el modelo predictivo que describe las reglas y o condiciones que son causas de deserción en el caso estudiado” (Castro Sotomonte et al., 2016, p. 49).

En el estudio modelo predictivo de deserción estudiantil basado en árboles de decisión, se definió como modelos predictivos a utilizar el “Classification and Regression Tree (CART) de la herramienta R, se construyó un árbol con cuatro niveles de profundidad y mismo número reglas, que evalúan a los posibles desertores. Llevando a concluir que las variables nivel y notas tienen mayor influencia en la deserción”(Cuji et al., 2017, p.17). Lo anterior, usando el algoritmo J48, con la herramienta WEKA. De otra parte, los árboles de clasificación según Cardona Taborda et al. (2016):

> El algoritmo J48 implementado en Weka es una versión del clásico algoritmo de árboles de decisión C45 propuesto por Quilan. El proceso de construcción del árbol comienza por el nodo raíz, el que tiene asociados todos los ejemplos o casos de entrenamiento. Lo primero es seleccionar la variable o atributo a partir de la cual se va a dividir la muestra de entrenamiento original (nodo raíz), buscando que en los subconjuntos generados haya una mínima variabilidad respecto a la clase. Este proceso es recursivo, es decir, una vez que se haya determinado la variable con la que se obtiene la mayor homogeneidad respecto a la clase en los nodos hijos, se vuelve a realizar el análisis para los nodos hijos. Aunque en el límite este proceso se detendría cuando todos los nodos hojas contuvieran casos de una misma clase, no siempre se desea llegar a este extremo, para lo cual se implementan métodos de pre-poda u post-poda de los árboles. (p. 145)

Adicional a lo anterior, en el estudio de determinación de variables predictivas de deserción inicial para generar un sistema de alerta temprana. Análisis sobre una muestra de estudiantes beneficiarios de la beca de nivelación académica en una universidad pública en Chile se seleccionó según Contreras, (2021).

> Método de detección automática de interacción de chicuadrado (CHAID), entre otras razones por la naturaleza categórica de la variable dependiente. El método consiste en un rápido algoritmo de árbol estadístico y multidireccional que explora datos de forma rápida y eficaz, creando segmentos y perfiles respecto del resultado deseado y permitiendo la detección automática de interacciones mediante chi-cuadrado. (p.12)

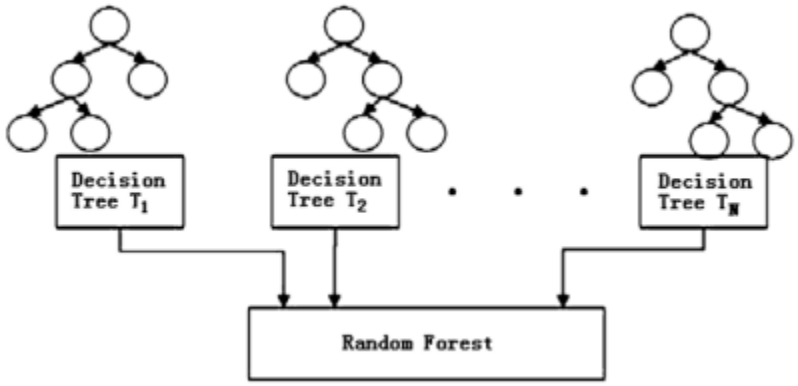
De la información incluida en los artículos anteriores, inferimos que usando un modelo predictivo de árbol de clasificación o de decisión se pueden segmentar, clasificar grupos de datos y predecir un evento, así como reducir la dimensión de los datos y ser usados en un modelo predictivo de deserción.

4.1.2.2 Random Forest.

El modelo predictivo Random Forest o bosques aleatorios "consiste en generar N árboles de decisión con M atributos del total de ellas y dividir el total entre los N árboles, de donde se tienen N subconjuntos para trabajar cada uno con un árbol con atributos distintos" (Aguilar Vilca y Camargo Ramos, 2021, p. 44). A continuación, se presenta gráficamente la definición de Random Forest en la Figura 2.

Figura 2.

Modelo Random Forest



Nota. La figura muestra el modelo predictivo supervisado random forest. Fuente: Tesis Sistema inteligente basado en redes neuronales, máquina de soporte vectorial y random forest para la predicción de deserción de clientes en microcréditos de bancos (2021).

El modelo predictivo Random Forest fue utilizado en el artículo denominado un análisis multinomial y predictivo de los factores asociados a la deserción universitaria, en el cual después de aplicar el modelo y los indicadores de capacidad de ajuste, según Fernández Martín et al. (2019):

> El random forest muestra la mejor combinación. Predice correctamente un 83% de lo que clasifica como estudiantado desertor (verdaderos positivos) y detecta un 34% (sensibilidad) de la totalidad. Es importante resaltar que el random forest se obtiene a partir de una combinación de árboles, que de forma independiente clasifican a cada sujeto en desertor y no desertor. La clasificación final se genera a partir de una combinación de todas las clasificaciones obtenidas de cada árbol. Para este algoritmo la selección de los nodos de los árboles se obtiene de una selección aleatoria de dos variables. (p. 17)

En el artículo consultado denominado modelo de deserción estudiantil según Manríquez Pacheco, (2022):

> Implementación la realizaremos utilizando la librería de scikit-learn. La aplicación del modelo Random Forest, cada árbol del conjunto se construye a partir de una muestra aleatoria extraída con reemplazo del conjunto de entrenamiento. Luego se divide cada nodo durante la construcción de un árbol, la mejor división se encuentra entre todas las entradas o un subconjunto aleatorio de tamaño max_features. (p.18)

Por otra parte, según Manríquez Pacheco (2022) expone:

> El objetivo de esta aleatoriedad es reducir la varianza del estimador del modelo. A partir de lo anterior es que los modelos Random Forest consiguen bajos indicadores de varianza al combinar los árboles y la implementación de scikit-learn combina clasificadores promediando su predicción probabilística, en lugar de permitir que cada clasificador vote por una sola clase. (p. 18)

En la aplicación del estudio anterior, los resultados de precisión del Random Forest fueron favorables, dado que "en el modelo de Random Forest los resultados del F1 score fueron de un 81% de precisión" (Manríquez Pacheco, 2022, p. 36).

Según lo anterior, es factible utilizar un modelo predictivo Random Forest para predecir deserción de estudiantes, dado que estos modelos según la evaluación de sus resultados tienen precisiones entre 81% y 83%.

4.1.2.3 Gradient boosting.

El Gradient boosting es un modelo que también utiliza árboles de decisión o de clasificación y "ensambla varios árboles de manera secuencial y realiza varias repeticiones mientras se va corrigiendo el error" (Quintero, 2022, p. 58).

Según el trabajo denominado Diseño de un modelo predictivo de la deserción estudiantil de postgrado en una institución de educación superior el uso de Gradient boosting Jurado Mantilla, (2019) presentó:

> Una validación cruzada de 3 capas muestra una exactitud del 94.83% con una sensibilidad del 96.23% y una especificidad del 80%, por lo tanto, el modelo sí muestra una concordancia entre los valores de la predicción con los reales. Por otro lado, se observa que la curva ROC del modelo muestra un área bajo la curva de 0.94 y de acuerdo con la literatura previamente mencionada, se considera al modelo con una alta exactitud para realizar la predicción. (p. 61)

Por lo anterior, según la literatura el Gradient Boosting o GMB es acto para ser utilizado en modelos de deserción de estudiantes.

De acuerdo con la revisión de literatura correspondiente a los tres modelos (árboles de decisión, random forest y gradient boosting), se establece que estos modelos son aptos para ser usados con datos correspondientes a deserción de estudiantes, por lo cual en el numeral 8.1.7 Modelado, se va a definir el modelo que presentó mayor ajuste y con el cual se va a realizar la predicción de deserción de los programas de formación.

5. Cronograma

El cronograma se desarrollará bajo la metodología ágil SCRUM, atendiendo las fases del ciclo de vida de la ciencia de los datos, así:

Tabla 1.

Cronograma de ejecución del proyecto

PLANEACIÓN ESCENARIO ÁGIL							
Historia de usuario	**Prioridad (Alta, media, baja)**	**Criterios de aceptación**	**Tareas**	**Sprint**	**Fecha Proyectada**		
					Fecha Inicial	**Fecha Final**	**Duración**
FASE I: CONOCIMIENTO DEL NEGOCIO							
1. Entendimiento Proceso de Formación Profesional Integral y de las pruebas de auditoría.	Alta	1. Efectuar revisión del Procedimiento de Formación Profesional Integral y de la normatividad asociada, así como a las auditorías efectuadas.	1. Documento marco de descripción del Proceso de Formación Profesional Integral.	Sprint 1	28/05/2022	28/06/2022	30 días
2. Definir el equipo SCRUM que va a participar en el proyecto	Alta	3. Definir el product backlog y el sprint backlog para el Proyecto Empresarial.	1. Realizar el cronograma del Proyecto Empresarial en Metodología SCRUM	Sprint 2	25/05/2022	25/05/2022	1 día
3. Identificar el origen de los datos	Alta	4. Analizar el tipo de bases de datos (estructurada, no estructurada y semi estructurada)	1. Realizar el Diccionario de Datos del dataset “Deserción de aprendices”.		07/06/2022	10/07/2022	4 días
			2. Realizar Matriz de Riesgos de Seguridad del Proyecto Empresarial		18/06/2022	24/06/2022	5 días
FASE II: ADQUISICIÓN Y ENTENDIMIENTO DE LA DATA							
1. Solicitud y entendimiento tipo de bases de datos	Media	1. Solicitud al proveedor de la información de las bases de datos con las características definidas previamente.	1. Obtener de parte del proveedor de la información las Bases de datos correspondientes al proceso Misional.	Sprint 3	19/05/2022	07/06/2022	13 días
	Media	2. Realizar limpieza de dato o ETL	2. Explorar y analizar las bases de datos correspondientes a Certificación		08/08/2022	05/09/2022	30 días

			Académica y deserción de estudiantes.				
	Media	3. Analizar el tipo de datos de acuerdo con el procedimiento de Formación Profesional Integral	3. Cruzar y limpiar las bases de datos, analizarlas y consolidarlas		06/09/2022	13/09/2022	5 días
FASE III: DESARROLLO Y ANÁLISIS							
1. Análisis estadístico y visualización de la data	Baja	1. Realizar análisis de datos descriptivo	1. Realizar visualización de datos en Tableros de Control de Power BI	Sprint 4	17/02/2023	07/03/2023	1 mes
		2. Analizar los tipos de modelos que se van a utilizar	2. Investigar los tipos de modelos predictivos que se pueden utilizar para el proyecto				
		2. Ejecutar los modelos predictivos	Elegir el mejor modelo según el mayor AUC		08/03/2023	05/04/2023	1 mes
		3. Extraer conclusiones iniciales de la data e identificar alertas con alcance preventivo.	2. Obtener los resultados de la predicción		10/04/2023	14/04/2023	8 días
FASE IV: TOMA DE DESICIONES							
1. Usar la herramienta de Data Analytics	Baja	1. Implementación de herramienta analítica y explicación de los insight identificados.	1. Establecer las alertas preventivas identificadas en los Centros de Formación.	Sprint 5	17/04/2022	17/05/2022	1 mes

Nota. Fases de elaboración del proyecto. Fuente: Elaboración propia

6. Descripción de la Situación organizacional donde se realizará el proyecto (Contexto)

6.1 Contexto empresarial

El proyecto se aplicará en el Servicio Nacional de Aprendizaje SENA, Entidad de Orden Nacional adscrita al Ministerio de Trabajo, cuya misión es ofrecer y ejecutar formación profesional integral, para la incorporación de las personas en actividades productivas.

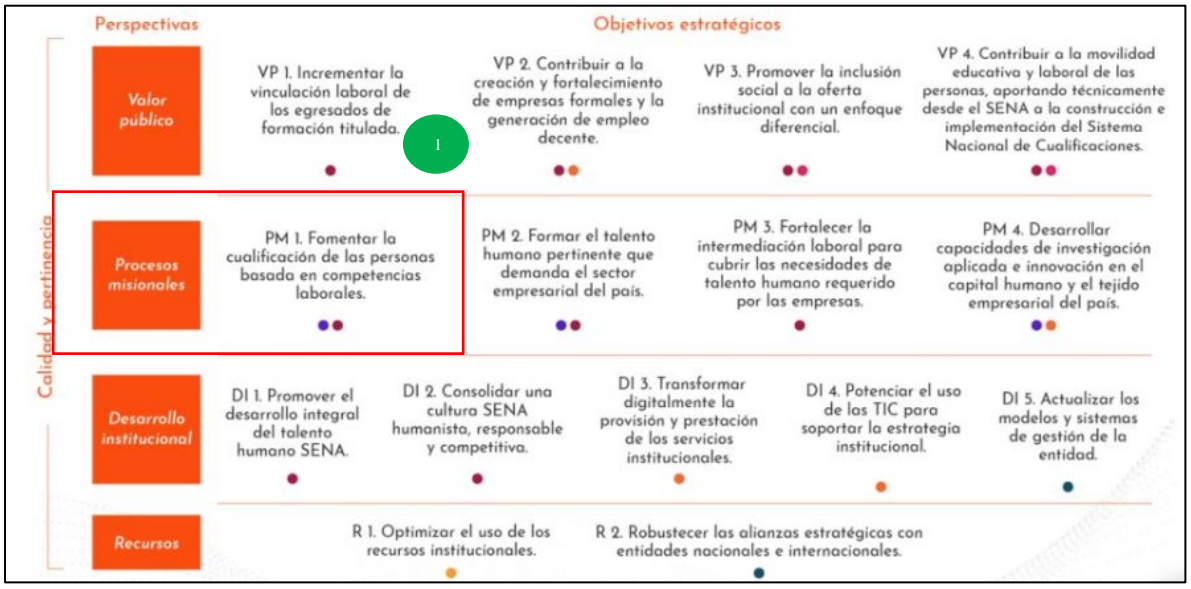
A continuación, vamos a revisar cómo está organizada estratégicamente la Entidad y en que parte del proceso impacta el proyecto que se ha diseñado:

6.1.1 Mapa estratégico de la Entidad

La Entidad cuenta con el Plan Estratégico Institucional 2019-2022 que está compuesto por cuatro perspectivas (valor público, procesos misionales, desarrollo institucional y recursos) que permiten desarrollar los objetivos estratégicos que se encuentran diseñados para cumplir con el Plan Nacional de Desarrollo. Dentro de la perspectiva de procesos misionales, se encuentra el objetivo “Fomentar la cualificación de las personas basada en competencias laborales” (Servicio Nacional de Aprendizaje SENA, 2023) de este objetivo se desagregan las acciones registradas en el plan de acción del año 2022 (ver Figura 3), en el cual se visualizan las metas de certificación y de retención.

Figura 3.

Mapa estratégico SENA



Nota. perspectiva y objetivo estratégico que van a impactar el proyecto. Fuente: Servicio Nacional de Aprendizaje SENA (2023).

6.1.2 Estructura orgánica

La estructura orgánica del servicio nacional de aprendizaje SENA está enmarcada en la ley 119 de 1994 y el Decreto 249 de 2004 modificado por el Decreto 2520 de 2013. El proyecto de analítica de datos se desarrollará con datos que impactan la Dirección de Formación Profesional y la Oficina de Control Interno de la Entidad.

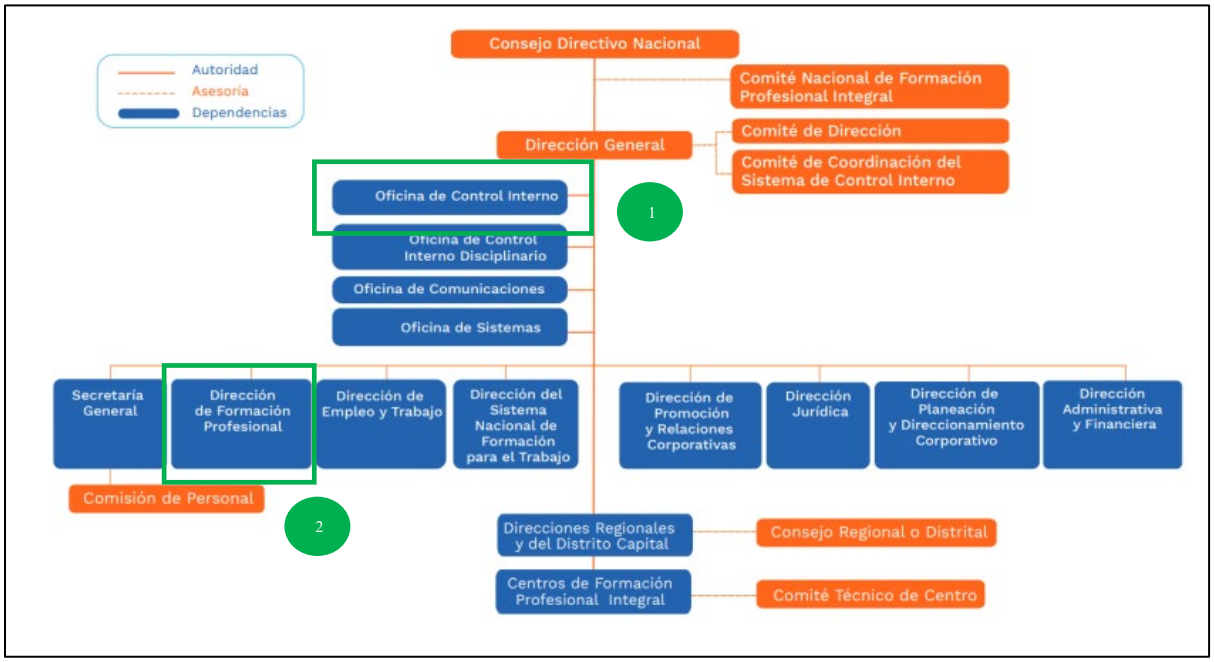
La Dirección de Formación Profesional tiene como función principal “diseñar, administrar y orientar la formación profesional integral a través de estrategias y programas de formación por competencias, asegurando el acceso, pertinencia y calidad”(SENA, 2023) y a la

Oficina de Control Interno le corresponde “diseñar, dirigir, organizar y evaluar los resultados del Sistema de Control Interno del SENA, de conformidad con las normas vigentes”(SENA, 2023).

A continuación, podemos visualizar en el organigrama las dependencias mencionadas anteriormente en las cuales se aplicará el proyecto de analítica de datos:

Figura 4.

Estructura Orgánica del SENA



Nota. Organigrama en el cual se presentan las dependencias Oficina de Control Interno y Dirección de Formación Profesional en las cuales se va a desarrollar el proyecto de analítica de datos. Fuente: Servicio Nacional de Aprendizaje SENA (2023).

6.1.3 Composición de la Entidad

La Entidad está compuesta por 33 regionales que se refiere a los 33 departamentos a nivel nacional incluyendo Distrito Capital y 117 Centros de Formación a nivel nacional de diferentes especialidades y sectores de la economía.

6.1.4 Mapa de procesos de la Entidad

Dentro del proyecto se tendrán en cuenta los procesos misionales y de evaluación, en razón a que el proyecto se va a desarrollar en los siguientes campos:

6.1.4.1 Proceso de evaluación.

Se tendrá en cuenta el proceso de gestión de evaluación y control y el procedimiento de trabajos de aseguramiento que corresponde a la Oficina de Control Interno de la Entidad.

6.1.4.2 Proceso misional.

Se tendrá en cuenta el proceso de gestión de formación profesional integral que incluye el procedimiento de certificación académica y de deserción de aprendices asociados a la Dirección de Formación Profesional Integral.

La Certificación de Formación Integral, es un procedimiento en el cual la Entidad expide documentos académicos, estos pueden ser títulos, certificaciones, actas de grado etc, a los aprendices que culminaron de forma satisfactoria su formación, por lo cual teniendo en cuenta la certificación de aprendices se puede establecer la información referente a la deserción de aprendices de los programas de formación titulada y complementaria, dicha información es consolida a nivel nacional.

Ahora trataremos específicamente los dos temas de estudio, la certificación de formación profesional y la deserción de aprendices, como se detalla a continuación:

6.1.4.2.1 Certificación de Formación Profesional.

"Acto administrativo por el cual el SENA otorga títulos o certificados a los Aprendices que culminan satisfactoriamente el proceso de formación profesional integral y a las personas que demuestran su Competencia laboral en el Proceso de Evaluación y Certificación para el Trabajo"(SENA, s. f., 2023).

Los Certificados del SENA pueden ser de tres tipos, el primero asociado a títulos de modalidad de formación titulada, es decir, trabajador especializado, técnico, especialización técnica, tecnólogo, ingeniero técnico, y especialización tecnológica. El segundo, certificados en modalidad de formación complementaria, que son otorgados cuando se imparten actualizaciones de competencias específicas de un sector productivo y el tercero que son certificados de participación, que están ligados a eventos de divulgación tecnológica.

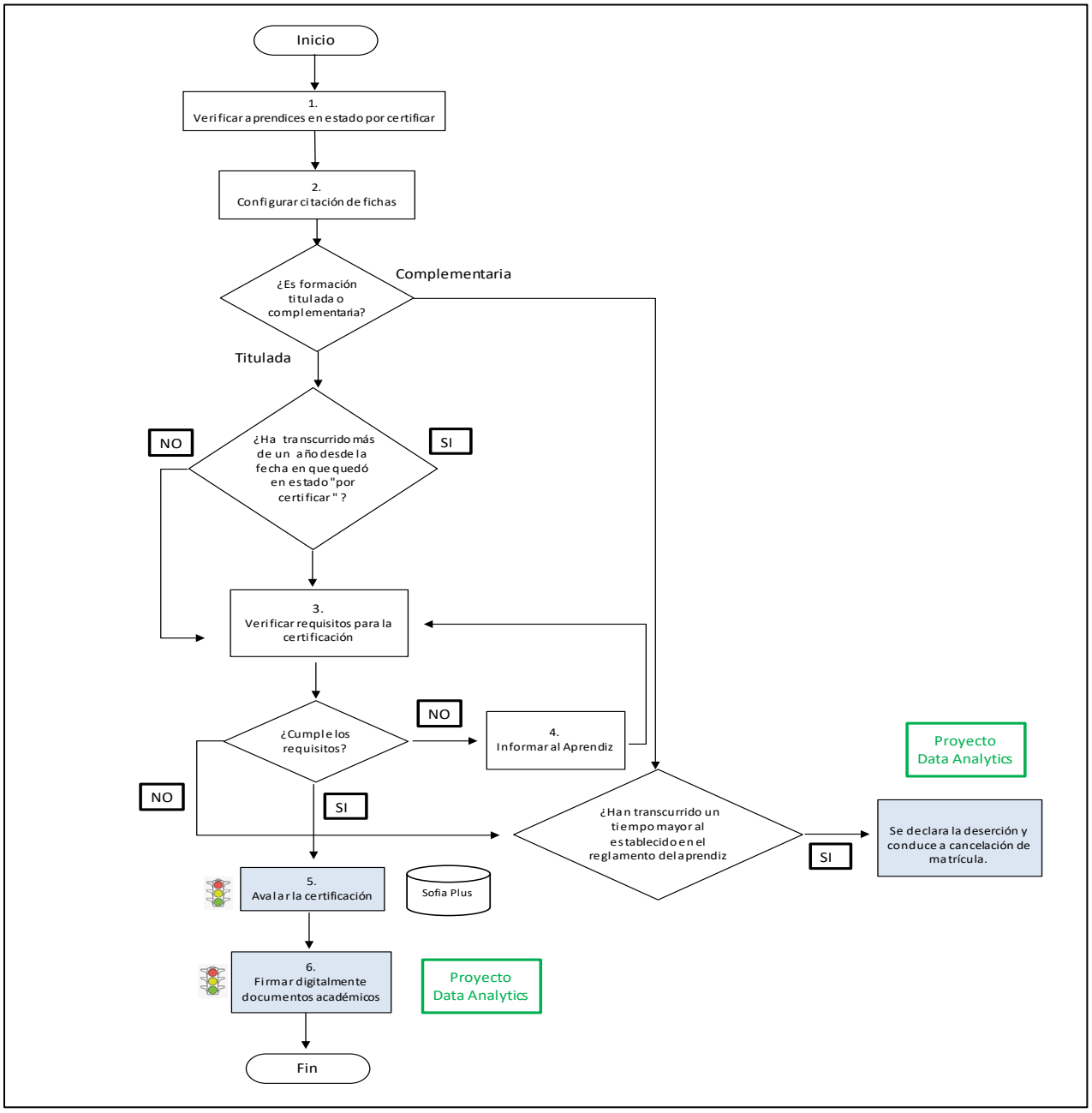
6.1.4.2.2 Deserción del proceso de formación.

"se considera deserción en el proceso de formación: a) Cuando el Aprendiz injustificadamente no se presente por tres (3) días consecutivos al Centro de Formación o empresa en su proceso formativo. b) Cuando al terminar el periodo de aplazamiento aprobado por el Sena, el Aprendiz no reingresa al programa de formación. c) Cuando transcurridos dos (2) años, contados a partir de la fecha de terminación de la etapa lectiva del programa, el Aprendiz no ha presentado la evidencia de la realización de la etapa productiva" (SENA, 2020).

En la Figura 5. se observa el procedimiento de certificación académica y la parte del proceso en el cual impacta la deserción.

Figura 5.

Procedimiento de certificación de Aprendices



Nota. Identificación de la parte del procedimiento donde se va a aplicar el proyecto de data analytics. Fuente: Servicio Nacional de Aprendizaje SENA (2022).

6.1.5 KPI´s de Certificación y de deserción del proceso de formación

De acuerdo con el proyecto que se va a realizar de la herramienta de data analytics se identificaron los KPIs que impactan el proyecto, por lo cual se profundizó en el origen de los mismos, por tratarse de sector público, esta medición está definida desde el documento CONPES 3654 de política de rendición de cuentas de la rama ejecutiva a los ciudadanos del Departamento Nacional de Planeación en el que tratan sobre la rendición de cuentas interna dado que según (DNP Departamento Nacional de Planeación & Departamento Administrativo de la Función Pública, 2010):

> "los sistemas de control interno constituyen un soporte para este tipo de rendición de cuentas. Entre entidades públicas del poder ejecutivo se produce también la rendición de cuentas interna, dado que se genera información, explicaciones y posibilidades de premios o sanciones. Esta rendición de cuentas se puede dar desde los Ministerios hacia la Presidencia, de las entidades vinculadas y adscritas hacia los Ministerios cabeza de sector y de los Ministerios y entidades hacia entidades de planificación y presupuesto"(DNP Departamento Nacional de Planeación & Departamento Administrativo de la Función Pública, 2010). (p.21)

Por lo anterior, mediante el artículo 343 de la Constitución Política de Colombia de 1991 se estableció que "La entidad nacional de planeación que señale la ley, tendrá a su cargo el diseño y la organización de los sistemas de evaluación de gestión y resultados de la administración pública, tanto en lo relacionado con políticas como con proyectos de inversión, en las condiciones que ella determine."(Asamblea Nacional Constituyente, 1991).

Para el Servicio Nacional de Aprendizaje SENA los indicadores están suscritos anualmente en el plan de acción, es por esto que, para la certificación total de formación titulada y complementaria, los indicadores se miden así:

Certificación = Meta anual / Número de aprendices certificados

Por otra parte, para la deserción de aprendices la tasa de deserción se mide, así:

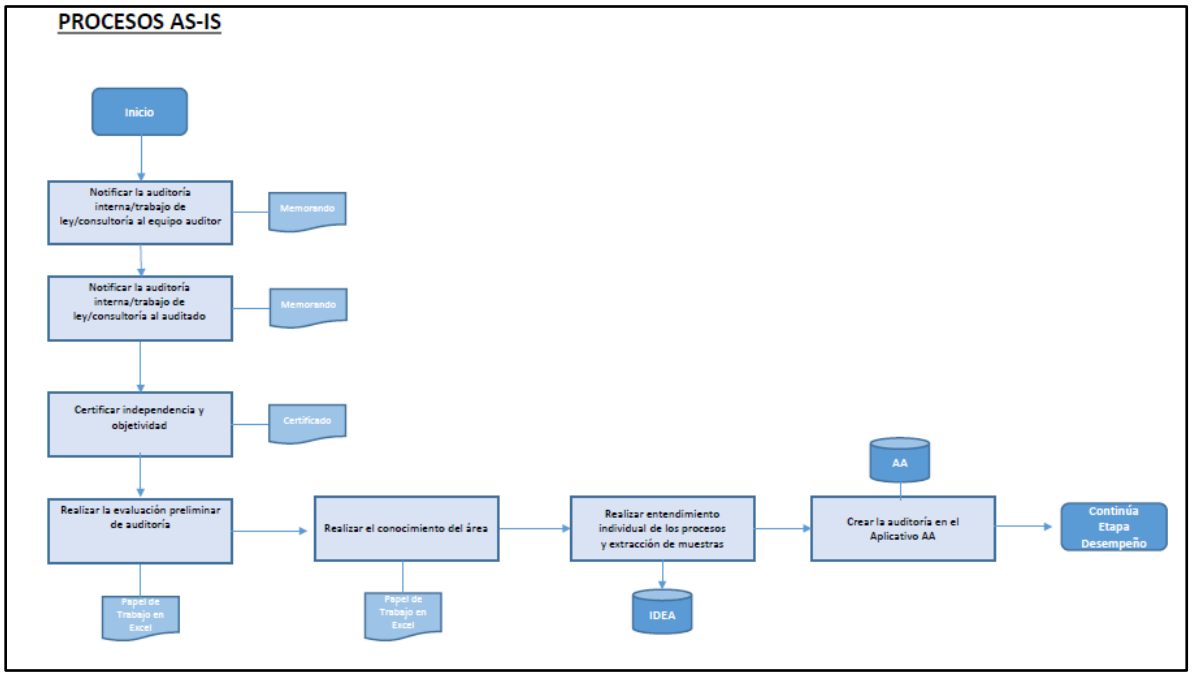
Cálculo deserción: (Desertores / Cupos en formación) * 100 = tasa de deserción.

6.1.6 Contexto Auditoría Interna Entidad

De acuerdo con el análisis de la situación actual del proceso de evaluación y el procedimiento de auditoría interna para trabajos de aseguramiento se realizó el análisis AS-IS correspondiente a la arquitectura del negocio para la etapa de planificación de la auditoría, cuyo fin es mapear el estado del proceso y determinar elementos como la meta de negocio de la Oficina de Control interno que es el porcentaje de cumplimiento sobre las auditorías aprobadas por el Comité Institucional de Control Interno de la Entidad, así mismo, se observó que las aplicaciones que actualmente se usan son AutoAudit en la cual se gestionan las auditorías, IDEA que es una herramienta estadística y Excel en el que se manejan las bases de datos por proceso y por regional. A continuación de visualiza gráficamente en la Figura 6 este análisis:

Figura 6.

Análisis AS-IS proceso de control y evaluación



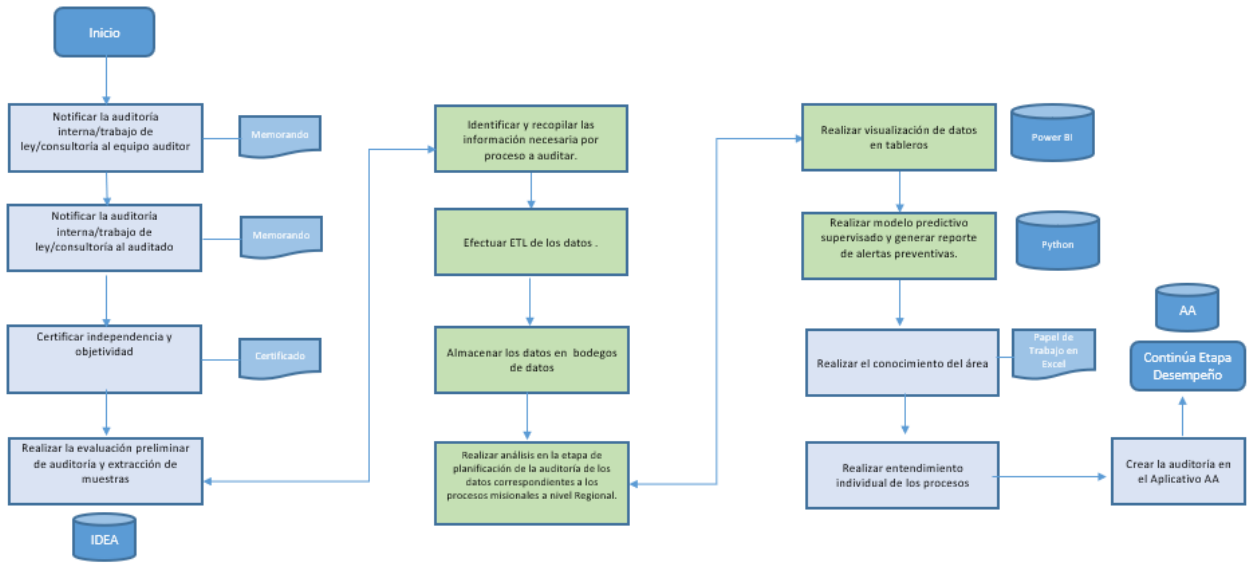
Nota. Análisis de la arquitectura del negocio en el cual mediante el mapeo de procesos se demuestra la situación actual del proceso de control y evaluación de la Oficina de Control Interno. Fuente: elaboración propia (2023).

Como se observa en el análisis AS-IS la Oficina de Control Interno de la Entidad no ha implementado dentro de su proceso analítica de datos para determinar durante la etapa de planificación puntos claves y relevantes que pueden ser objeto de revisión.

Adicionalmente, se efectuó el análisis de arquitectura TO-BE, es decir, después de la aplicación del proyecto, que mejoras se proponen al proceso de control evaluación, en el marco del procedimiento de trabajos de aseguramiento de la Oficina de Control Interno de la Entidad, por cuanto la mejora que se propone es desde la identificación y recopilación de información según el proceso a auditar hasta la ejecución de modelos predictivos supervisados y la generación de alertas preventivas a la Dirección de Formación Profesional, como se visualiza en la Figura 7.

Figura 7.

Análisis TO-BE proceso de control y evaluación



Nota. Análisis de la arquitectura TO-BE del negocio en el cual mediante el mapeo de procesos se demuestra la situación que se quiere lograr con la aplicación del proyecto en el proceso de control y evaluación de la Oficina de Control Interno. Fuente: elaboración propia (2023).

6.1.7 Analítica de Datos en la Auditoría Interna

Según el avance de la auditoría interna, las oficinas de control interno deben realizar auditorías avanzadas, es decir, auditorías que agreguen valor a las Entidades de Gobierno y que integren el análisis de riesgos y de controles con las tecnologías emergentes para lograr "Cambiar muchos hallazgos no relevantes a menos hallazgos pero de mayor impacto y cambiar un enfoque de revisión de actividades pasadas o investigar hechos cumplidos e involucrar una mirada hacia adelante"(Echeverría, 2019).

Ahora bien, el desafío de la auditoría interna es agregar pruebas que efectúen el análisis de datos por medio de herramientas de inteligencia de negocios y de modelos predictivos que permitan identificar oportunamente en los procesos objeto de revisión las falencias que se presentan, así como, "responder de una manera importante y notable a las expectativas de los stakeholders de las organizaciones, pasando del aseguramiento de riesgos hacia la prevención de su materialización, bajo un monitoreo continuo a través de tableros de control"(Echeverría, 2019).

El beneficio que trae la implementación de data analytics en las áreas de auditoría interna puede significar la planificación de auditorías de forma más ágil, "para mejorar su acceso a los datos y desarrollar conocimientos clave antes del trabajo de campo" (Deloitte, 2016); sin embargo, los desafíos gerenciales que enfrentan las áreas de auditoría interna para aplicar en su proceso la analítica de datos son significativos, por cuanto, "se requieren conocimientos tanto técnicos como comerciales para utilizar la analítica, pero también se requiere conocimiento específico del dominio para usar el análisis de manera efectiva" (Islam & Stafford, 2022).

7. Descripción de la situación estudio de caso y/o problemática empresarial y método y/o estrategia a aplicar para su solución

La identificación del problema empresarial se realizó de acuerdo con la metodología del libro The analytics lifecycle toolkit : a practical guide for an effective analytics capability (Nelson, 2018), para lo cual se establecieron las siguientes actividades:

7.1 Proceso para el marco del problema (Problema empresarial)

7.1.1 Identificar el problema

Actualmente dentro de la Oficina de Control de la Entidad no se aplica analítica de datos en las auditorías del proceso misional ni en los análisis que se desarrollan en la etapa de planeación de la auditoría, es decir, previo a la ejecución de las auditorías en campo.

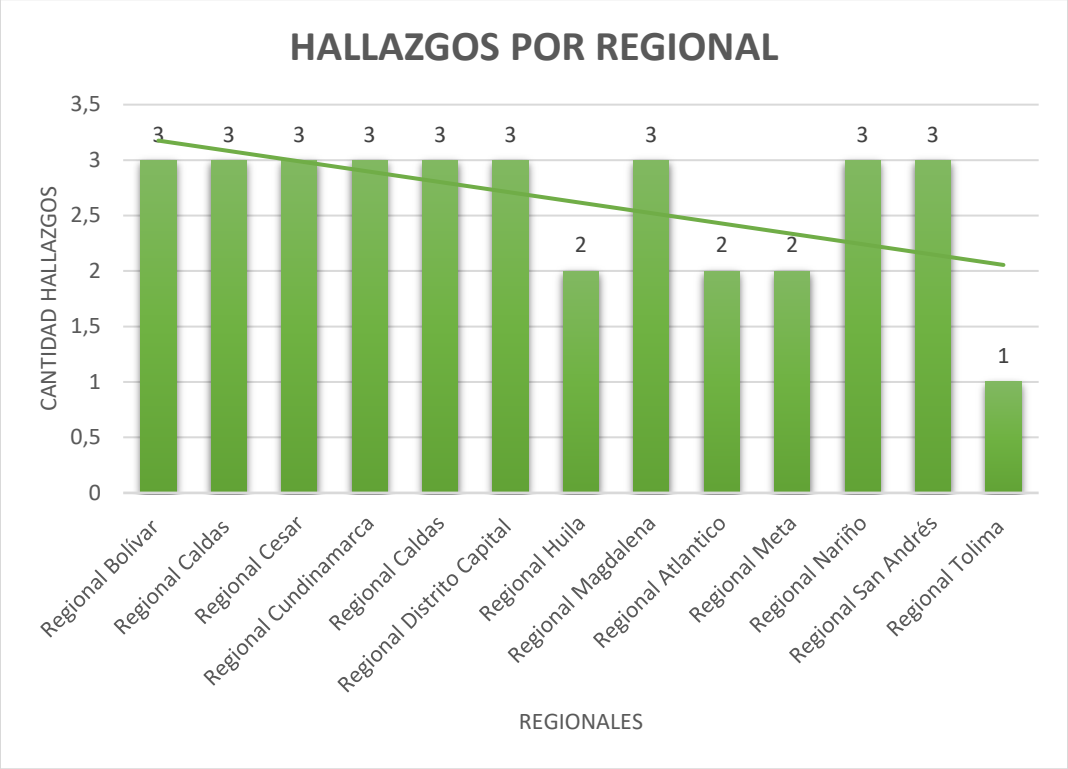
Así mismo, no se ha implementado data analytics para realizar un análisis a la completitud de la población del proceso misional, específicamente a la certificación académica de estudiantes, que permita eliminar los sesgos que se producen con el uso de muestras estadísticas en la ejecución de auditorías, así como identificar los programas de formación donde se presenta mayor tasa de deserción lo cual puede influir en el cumplimiento de metas de certificación académica de formación titulada y complementaria en la Entidad.

De otra parte, se han generado 34 hallazgos desde el año 2020 hasta el 2022, en los Centros de Formación a nivel nacional correspondientes al procedimiento de certificación académica de estudiantes los cuales se realizan posterior a que ocurran los hechos, por cuánto no se generan oportunamente alertas preventivas que agreguen valor y determinen oportunamente si se va a presentar deserción de aprendices en los programas de formación a nivel nacional dado

que esta situación afecta directamente los indicadores anuales de la certificación académica. A continuación, se visualizan estos casos en la Figura 8:

Figura 8.

Hallazgos generados sobre certificación académica



Nota. Hallazgos generados inoportunamente sobre hechos cumplidos. Fuente: elaboración propia (2023).

Así mismo, realizando el análisis descriptivo de los datos correspondientes a la deserción de aprendices durante la vigencia 2022, se evidenció que a nivel nacional se matricularon 35 millones de aprendices, de los cuales se presentó deserción de 13 millones de aprendices, es decir, que el 37% de los aprendices desertaron de la formación, como se visualiza a continuación:

Figura 9.

Aprendices desertores vs aprendices matriculados

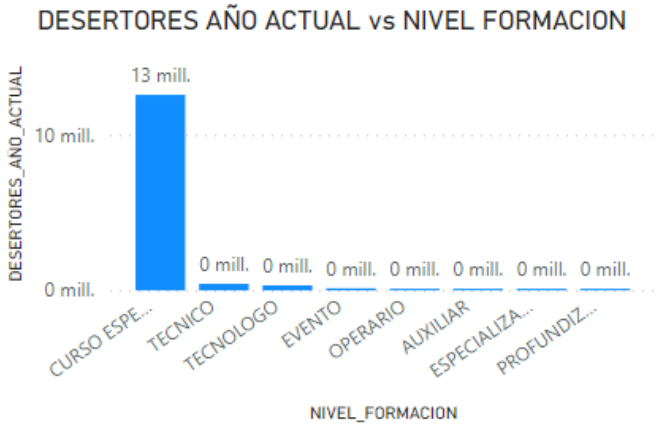
Nota. Visualización de datos data deserción de aprendices. Fuente: elaboración propia (2023).

Adicionalmente, aplicando el KPI de la tasa de deserción de aprendices en la vigencia 2022, se evidenció que para la modalidad de formación a distancia se presentó el 0,02 % de deserción, en la modalidad presencia se presentó deserción del 6,69% y en la modalidad virtual un 30,75%, siendo la modalidad virtual en la que más se presenta deserción de aprendices.

Los niveles de formación en los que más se presentó deserción a nivel nacional en la vigencia 2022 fueron el curso especial con 12.541.602, el técnico con 411.949 y el tecnólogo con 314.476, aprendices respectivamente como se muestra en la Figura 10.

Figura 10.

Desertores según nivel de formación

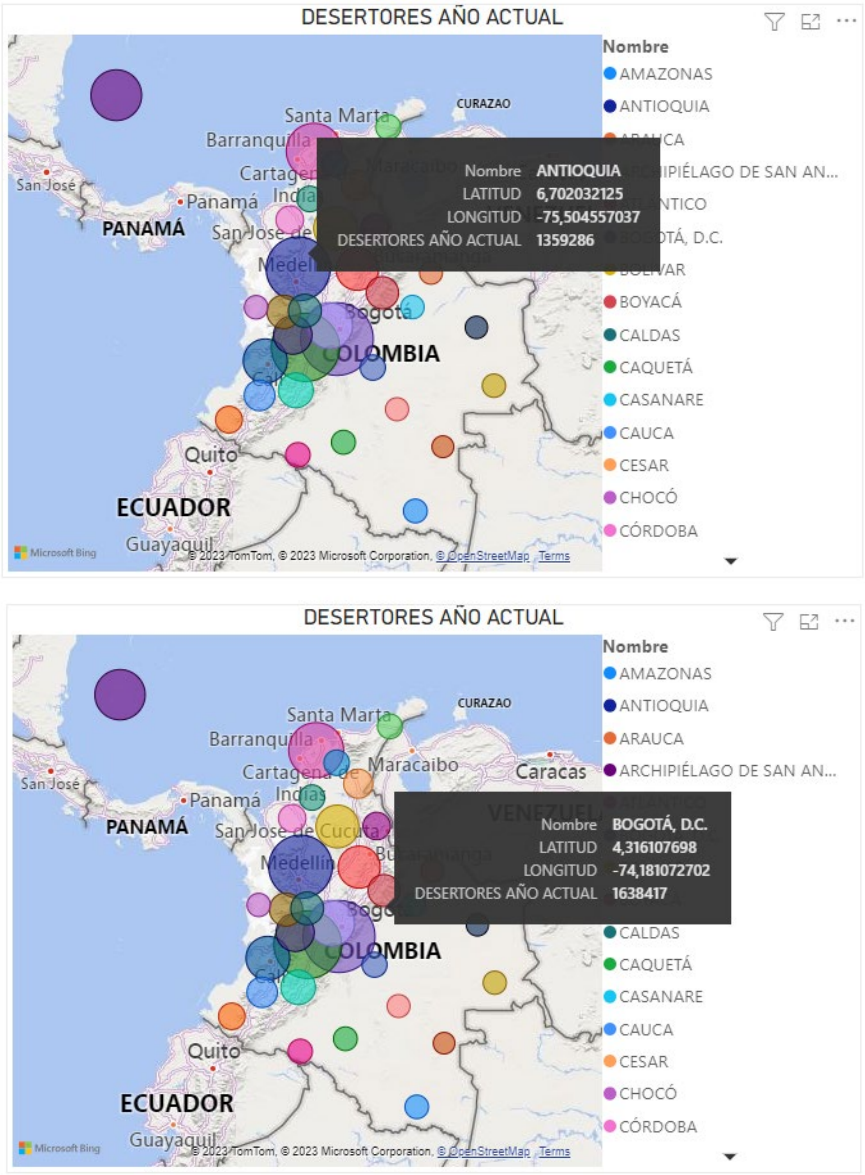


Nota. Visualización de datos data deserción de aprendices comparado con el nivel de formación. Fuente: elaboración propia (2023).

A nivel nacional la mayor deserción se presentó en las regionales de Antioquia y Distrito Capital, con 1.359.286 y 1.638.417, respectivamente como se observa en la Figura 11.

Figura 11.

Desertores Regionales de Antioquia y Distrito Capital



Nota. Visualización de datos data deserción de aprendices regional Antioquia y Distrito Capital. Fuente: elaboración propia (2023).

7.1.2 Formular la pregunta

¿Por qué la Oficina de Control Interno no aplica analítica de datos en las pruebas de la etapa de planificación de las auditorías que ejecuta para evaluar el proceso misional de la Entidad?

7.1.3 Definir causas raíz

- Falta de planeación estratégica para incluir data analytics en la etapa de planeación, en el proceso de evaluación y control, es decir, el proceso asociado a la Oficina de Control Interno.
- Falta de implementación de una auditoría integrada o avanzada, basada en data analytics y monitoreo del riesgo de los procesos misionales en tiempo real.

7.2 Proceso para el marco del problema (Problema analítica)

7.2.1 Identificar el problema

La Oficina de Control Interno no realiza los análisis de la etapa de planificación de las auditorías internas basada en la aplicación de data analytics.

7.2.3 Formular la pregunta

7.2.3.1 Pregunta analítica descriptiva.

¿Existe una herramienta analítica en la Oficina de Control Interno para efectuar las pruebas de auditoría del proceso misional de la Entidad, pertenecientes al proceso de formación profesional integral asociados al procedimiento de deserción de aprendices?

7.2.3.2 Pregunta analítica predictiva.

¿Existe un modelo analítico en la Oficina de Control Interno para predecir cuáles son los programas donde se puede presentar deserción de estudiantes para emitir oportunamente alertas preventivas durante la etapa de planificación de las auditorías?

7.2.4 Definir causas raíz

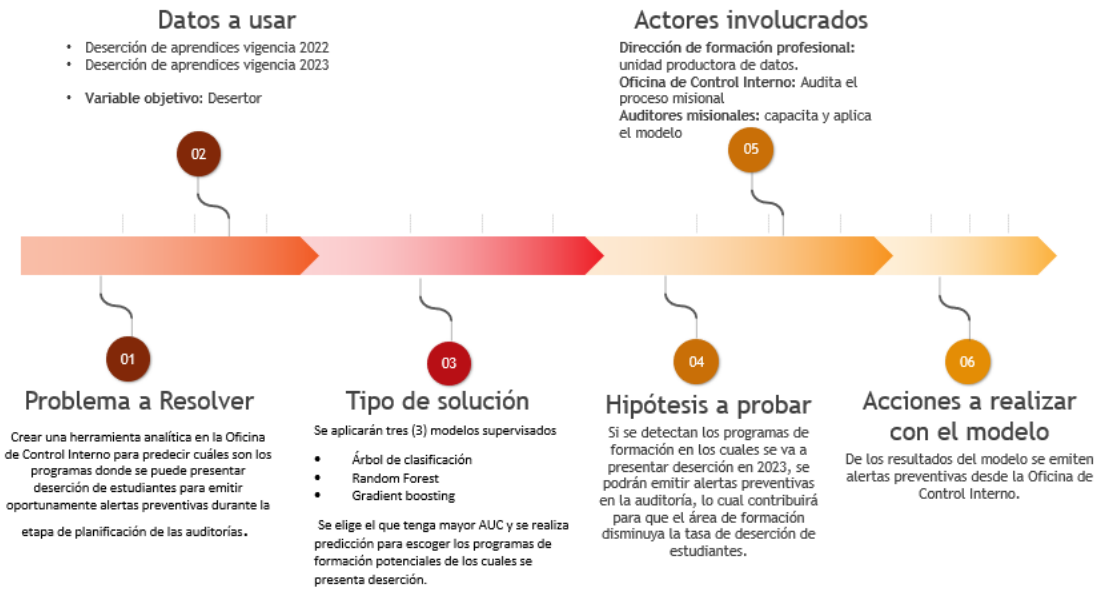
- Falta de conocimiento del ciclo de vida de los datos para aplicarlos en las pruebas de las auditorías misionales.
- Aplicación de pruebas de auditoría sobre muestras estadísticas las cuales incluyen sesgos o errores de muestreo.
- Falta de conocimiento de data analytics por parte de los auditores del proceso misional.

8. Descripción de las alternativas, estrategias y/o acciones que se toman en el análisis de la solución a la problemática

La solución a la problemática identificada se presenta mediante la metodología CANVAS, así:

Figura 12.

Metodología Canvas



Nota. Metodología Canvas en la cual se resume en 6 etapas el modelo analítico predictivo. Fuente: elaboración propia (2023).

Como se ha descrito en el documento, ya se ha identificado el problema y la metodología a usar para plantear la solución del problema analítico, por lo cual se presenta un análisis mediante la metodología Canvas la cual se describe en 6 etapas, que son el problema a resolver, los datos a usar, el tipo de solución, la hipótesis a probar, los actores involucrados y las acciones a realizar con el modelo.

Por otra parte, para realizar la ejecución del proyecto se implementará la metodología CRISP-DM, por cuanto se va a ejecutar usando las etapas del ciclo de vida de la analítica, así:

8.1 Aplicación de metodología CRISP-DM

8.1.2 Comprensión del Negocio

La comprensión del negocio se evaluó en el estudio del contexto de la Entidad en la cual se va a ejecutar el proyecto, descrita en el numeral 6 de este proyecto.

8.1.3 Entendimiento de los datos

8.1.3.1 Deserción de estudiantes.

La base de datos se tomó de la página web www.datosabiertos.gov.co que es la plataforma nacional de datos abiertos o públicos de Colombia.

La data analizada correspondió a la deserción de la formación profesional integral a nivel nacional, contiene la información relacionada a 117 centros de formación y 33 regionales a nivel nacional:

8.1.3.2 Diccionario de Datos.

15 columnas y 936.886 filas. Ver Anexo Técnico Anexo 1.

Por otra parte, se realizó un entendimiento de la base de datos a usar, según el alcance del modelo predictivo, en la cual se visualiza el conteo de datos, la media estándar, rangos y los valores mínimos y máximos como se observa en la Figura 10.

Figura 13.

Entendimiento de la data de deserción de aprendices formación profesional integral

	TOTAL	DESERTOR
count	19708	0
mean	13	NaN
std	16	NaN
min	0	NaN
25%	0	NaN
50%	0	NaN
75%	24	NaN
max	200	NaN

	TOTAL	DESERTOR
count	311042	311042
mean	20	0
std	22	0
min	-757	0
25%	3	0
50%	15	0
75%	27	0
max	987	1

his

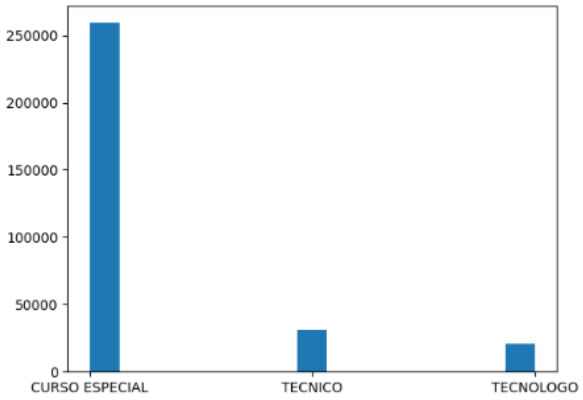
Nota. Información para conocer estadísticas de la distribución de las variables numéricas de la base potencial e histórica. Fuente: elaboración propia (2023).

Se realiza un entendimiento de la data histórica, es decir la data correspondiente a la deserción de aprendices de la formación profesional de la vigencia 2022 y de la data potencial, es decir, la data de los aprendices que se encuentran en formación profesional integral para la vigencia 2023 y que corte a 31 de marzo de 2023 aún no han desertado de los programas de formación. Ver Figura 14 y 15.

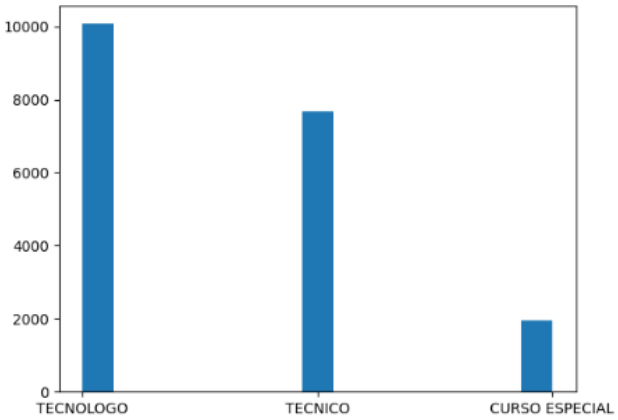
Figura 14.

Data histórica y potencial de deserción por nivel de formación

```
plt.hist(df_historico[["NIVEL_FORMACION"]], bins = 15)
plt.show()
```



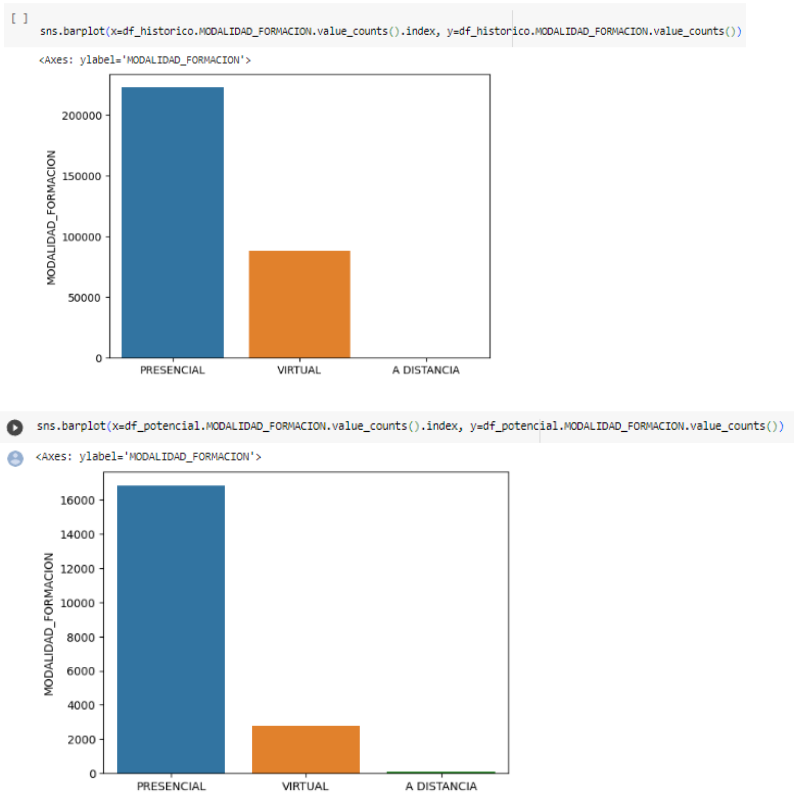

```
plt.hist(df_potencial[["NIVEL_FORMACION"]], bins = 15)
plt.show()
```



Nota. análisis con diagramas de barras para variables categóricas y comparación de las poblaciones potenciales e históricas para ver si la base actual de desertores potenciales tiene diferencias especiales con la base de deserciones anteriores de aprendices. Fuente: elaboración propia (2023).

Figura 15.

Data histórica y potencial de deserción por modalidad de formación



Nota. análisis con diagramas de barras para variables categóricas y comparación de las poblaciones potenciales e históricas para ver si la base actual de desertores potenciales tiene diferencias especiales con la base de deserciones anteriores de aprendices. Fuente: elaboración propia (2023).

8.1.4 Preparación de los datos

8.1.4.1 Duplicados.

Las siguientes columnas presentan datos duplicados, sin embargo, corresponden a datos duplicados de información que se recopila de los programas de formación anualmente, por cuanto puede ser objeto de duplicación dependiendo de la regional en la cual se generó.

Tabla 2.

Columnas con data duplicada

NOMBNRE COLUMNA EN DATA
CODIGO_REGIONAL
NOMBRE_REGIONAL
CODIGO_CENTRO
NOMBRE_CENTRO
IDENTIFICADOR_UNICO_FICHA
CODIGO_PROGRAMA
VERSION_PROGRAMA
NOMBRE_PROGRAMA_FORMACION
NIVEL_FORMACION
MODALIDAD_FORMACION
TOTAL_APRENDICES_MATRICULADOS
DESERTORES_AÑO_ACTUAL

Nota. Nombre de las columnas en la data deserción de la formación profesional integral que contiene información duplicada. Fuente: Elaboración propia.

8.1.4.2 Vacíos.

Se evidenciaron 91 registros vacías sin nombre de programa de formación que pertenecen CÓDIGO DEL PROGRAMA 924100, por lo cual estos campos se van a rellenar con la palabra “NO APLICA” no se van a eliminar porque tienen información del total de aprendices y del total aprendices desertados.

Se eliminaron 61 registros correspondientes al programa de formación OMI 1.38 SENSIBILIZACION CON RESPECTO AL MEDIO MARINO. SECCION A-II/1, dado que las columnas denominadas TOTAL_APRENDICES_MATRICULADOS y DESERTORES_AÑO_ACTUAL no contenían ningún tipo de información.

8.1.4.3 Limpieza de datos.

Se efectúo la siguiente limpieza de datos, así:

Tabla 3.

ETL a la data deserción de la formación profesional integral

NOMBNRE COLUMNA EN DATA	ETL
NOMBRE_CENTRO	Se eliminaron comillas (“) a 936.886 registros. Se eliminaron (.)en 9.712 registros.
FECHA_INICIO_FICHA	Se eliminaron comillas (“) a 1.873.772 registros.
FECHA_TERMINACION_FICHA	Se eliminaron comillas (“) a 1.873.772 registros.
CODIGO_PROGRAMA	Se eliminaron comillas (“) a 1.873.772 registros.
NOMBRE_PROGRAMA_FORMACION	Se eliminaron comillas (“) a 1.873.772 registros.
NIVEL_FORMACION:	Se eliminaron comillas (“) a 1.873.772 registros.
MODALIDAD_FORMACION	Se eliminaron comillas (“) a 1.725.862 registros.
NOMBRE_PROGRAMA_FORMACION	Se eliminaron dos puntos (:) a 26.462 registros.
NOMBRE_PROGRAMA_FORMACION	Se eliminaron puntos (.) a 189.382 registros.

Nota. ETL realizada a la data deserción de la formación profesional integral extraída de la plataforma de datos abiertos. Fuente: Elaboración propia

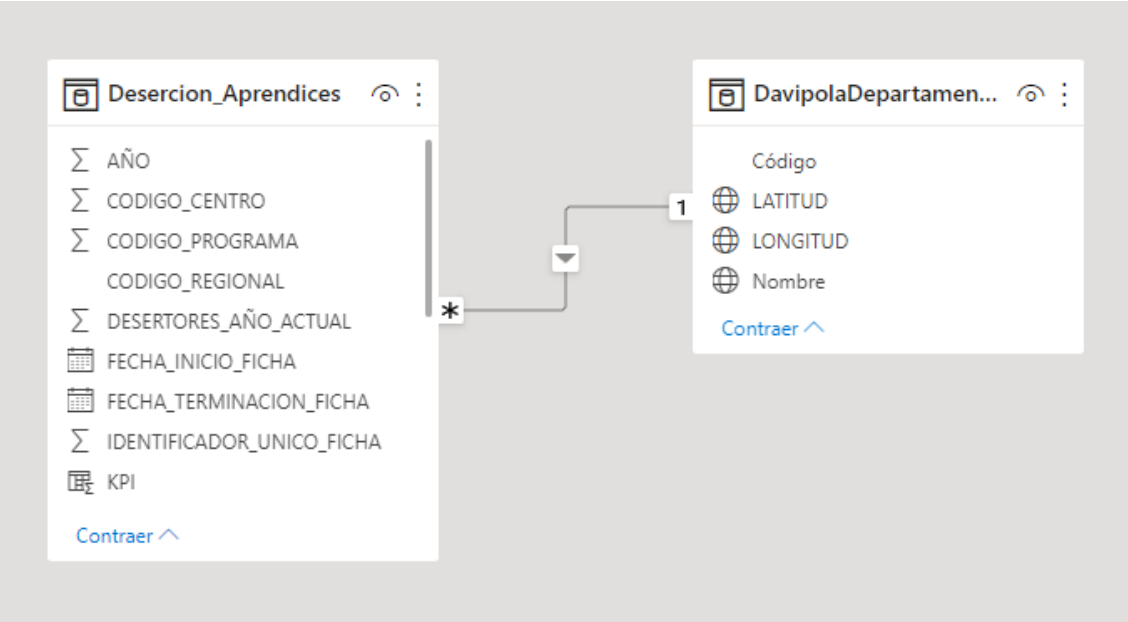
8.1.5 Visualización de datos

La visualización de datos se presenta mediante un tablero de control en la herramienta Power BI en la cual se pueden observar la cantidad de aprendices matriculados, desertados, la tasa de deserción, la modalidad de formación, la regional y centro de formación para verificar a nivel nacional la data.

Para realizar el tablero se usó la base de datos de deserción de aprendices de la formación profesional y la base de datos davipola por departamentos exportada del DANE para realizar la georreferenciación, por lo cual la relación de entidades de las bases de datos es la siguiente:

Figura 16.

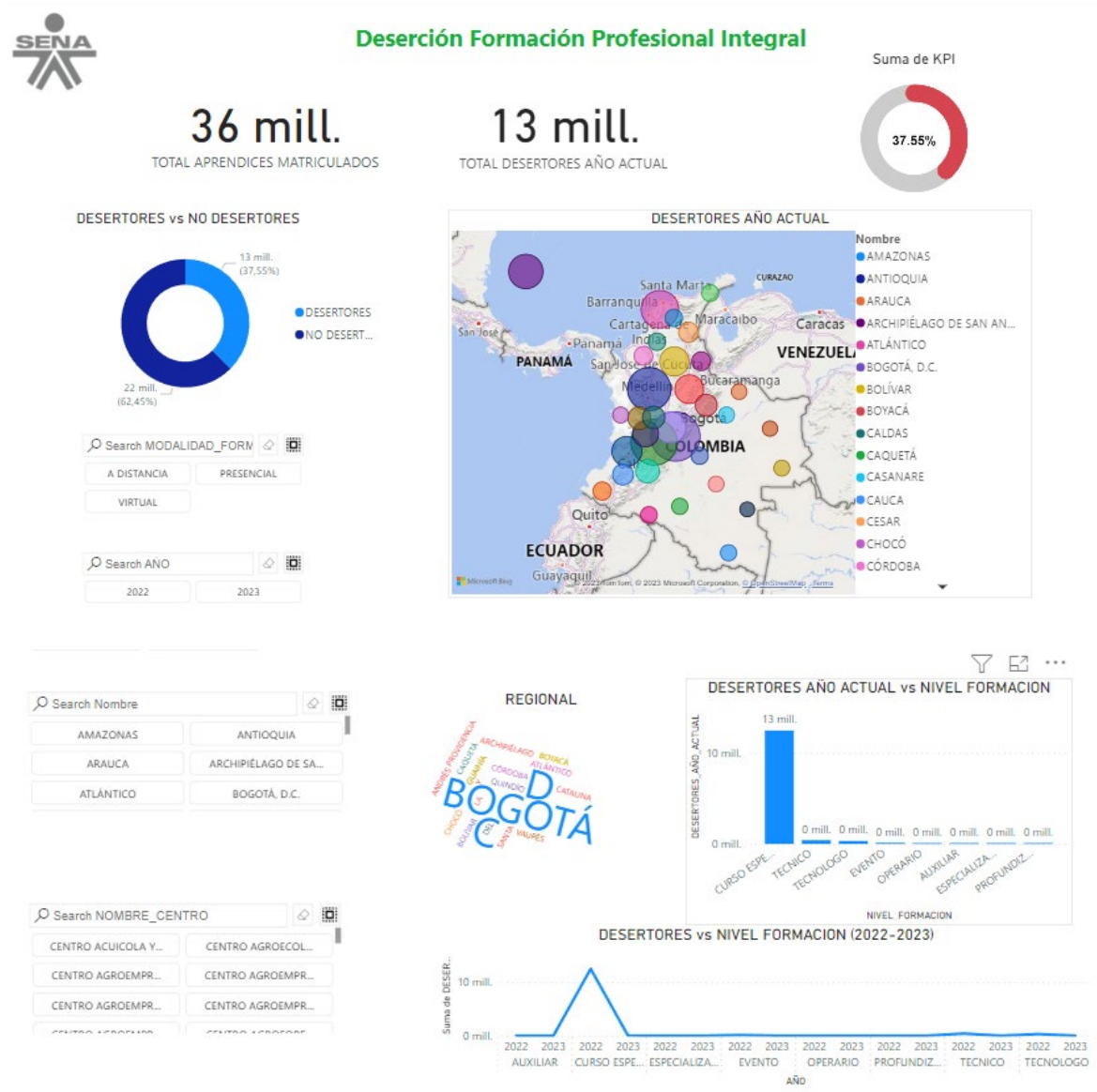
Relación de entidades de las bases de datos



Nota. Entidades que se usaron para efectuar la visualización de datos en la herramienta Power BI. Fuente: elaboración propia (2023).

Figura 17.

Tablero de Control Deserción Formación Profesional Integral



Nota. Visualización de datos de deserción de aprendices a nivel nacional año 2022 y 2023, en la cual se observa la tasa de deserción, los desertores a nivel nacional, la modalidad y los niveles de deserción. Fuente: elaboración propia (2023)

De acuerdo con la visualización de los datos en la herramienta Power BI, se analizó el alcance del modelo de predicción, el cual se va a enfocar en las Regionales de Antioquia y Distrito Capital, dado que son dos de las regionales que más presentaron deserción de cursos especiales, técnicos y tecnólogos con un 21% para una cantidad de 2.793.623 sobre el total de aprendices desertados de 13.268.022.

A continuación, se presenta la información detallada por nivel de formación:

Tabla 4.

Deserción de la formación profesional integral Regional Antioquia y Distrito Capital

NIVEL DE FORMACIÓN	ANTIOQUIA	DISTRITO CAPITAL
CURSO ESPECIAL	1.226.922	1.470.570
TECNICO	35.722	31.099
TECNOLOGO	10.217	19.093
Total	**1.272.861**	**1.520.762**

Nota. data deserción de la formación profesional integral regional Antioquia y Distrito Capital por nivel de formación. Fuente: Elaboración propia

8.1.6 Modelado

8.1.6.1 Bases de datos

a. Base de datos de deserción de aprendices de la formación profesional integral de la vigencia 2022 por nivel de formación de cursos especiales, técnicos y tecnólogos correspondiente a la regional Distrito y a la Regional Antioquia.

b. Base de datos para predecir las potenciales deserciones de aprendices de la formación profesional integral de la vigencia 2023.

8.1.6.2 Variable objeto (y) del modelo.

La variable objeto del modelo es DESERTOR variable binaria (0 y 1) que define si el aprendiz desertó o no desertó del programa de formación en la vigencia 2022.

8.1.6.3 Variable (x) del modelo.

Tabla 5.

Variables x del modelo

VARIABLE X DEL MODELO
CODIGO_REGIONAL'
CODIGO_CENTRO'
TOTAL_APRENDICES_MATRICULADOS'
NIVEL_FORMACION_CURSO ESPECIAL'
NIVEL_FORMACION_TECNICO'
NIVEL_FORMACION_TECNOLOGO'
MODALIDAD_FORMACION_A DISTANCIA'
MODALIDAD_FORMACION_PRESENCIAL'
MODALIDAD_FORMACION_VIRTUAL'

Nota. Variables x del modelo predictivo de deserción de aprendices según programa de formación. Fuente: Elaboración propia.

Para las variables 'NIVEL_FORMACION', 'MODALIDAD_FORMACION' y 'MODALIDAD_FORMACION' se realizó un trabajo para convertirlas a variables dummy con el objetivo de poder entrenar el modelo.

8.1.6.4 Herramientas de analítica.

- Python (Google Colaboratory)

8.1.7 Modelación

A continuación, se presenta la aplicación de los 3 modelos predictivos para identificar los programas de formación en los cuales se puede presentar deserción de aprendices, para lo cual se dividió la población en entrenamiento y testeo, con un 75% para el entrenamiento de datos y 25% para el testeo.

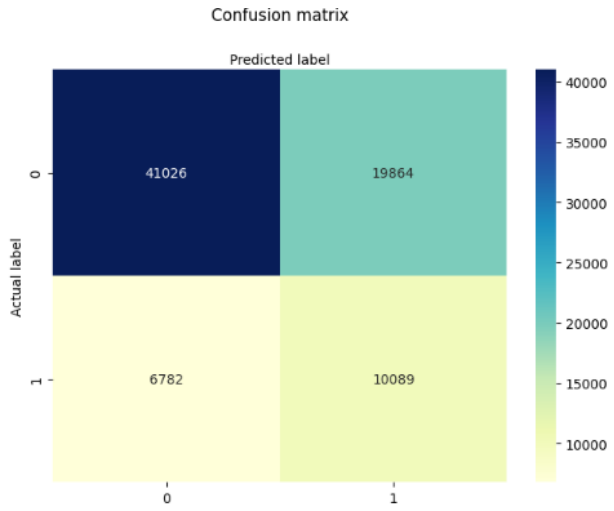
8.1.7.1 Árbol de clasificación.

Se calculó el modelo de árbol de clasificación para predecir la población de prueba y se calculó la matriz de confusión que valora el modelo predictivo, en la cual se observa que las estimaciones correctas del modelo son 41.026 y 10.089 tanto para los verdaderos positivos como para los verdaderos negativos y por el contrario la otra diagonal indica 6.782 falsos negativos y 19.864 falsos positivos.

Así mismo, se definieron los hiperparámetros del árbol de clasificación y se entrenó el modelo con el mejor hiperparámetro max_depth= 7.

Figura 18.

Matriz de confusión y variables importantes modelo árbol de clasificación



Fuente: elaboración propia (2023)

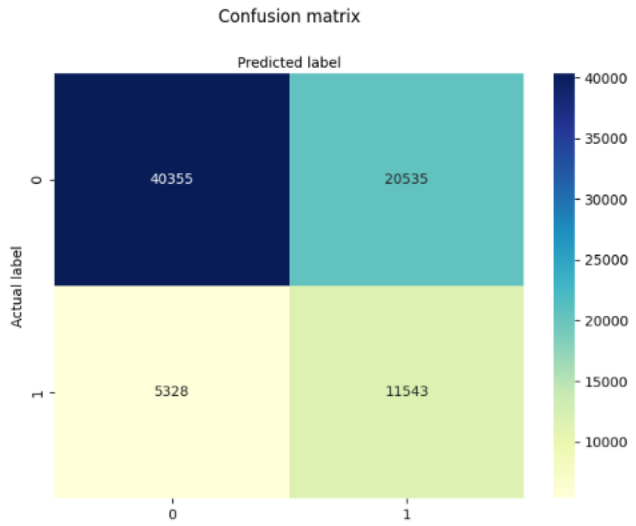
8.1.7.2 Gradient Boosting.

Se calculó el modelo de gradient boosting para predecir la población de prueba y se calculó la matriz de confusión que valora el modelo predictivo, en la cual se observa que las estimaciones correctas del modelo son 40.355 y 11.543 tanto para los verdaderos positivos como para los verdaderos negativos y por el contrario la otra diagonal indica 5.328 falsos negativos y 20.535 falsos positivos.

Así mismo, se definieron los hiperparámetros del gradient boosting y se entrenó el modelo con los mejores hiperparámetro así: n_estimators=300, learning_rate=0.5, max_depth=7 y max_features='auto'.

Figura 19.

Matriz de confusión y variables importantes modelo Gradient Boosting



Fuente: elaboración propia (2023)

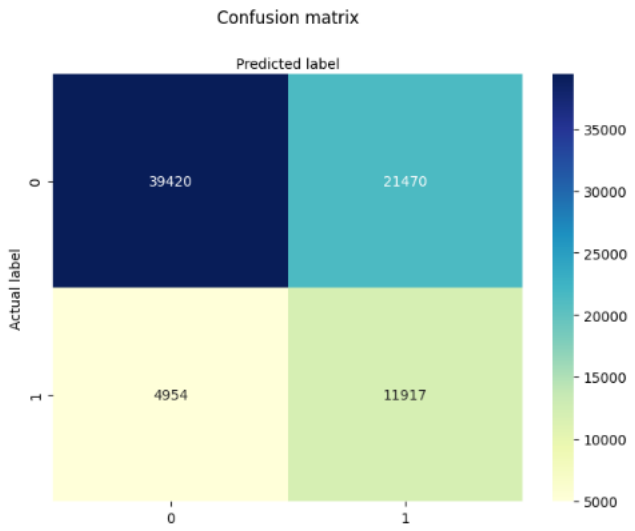
8.1.7.3 Random Forest.

Se calculó el modelo de random forest para predecir la población de prueba y se calculó la matriz de confusión que valora el modelo predictivo, en la cual se observa que las estimaciones correctas del modelo son 39.420 y 11.917 tanto para los verdaderos positivos como para los verdaderos negativos y por el contrario la otra diagonal indica 4.954 falsos negativos y 21.470 falsos positivos.

Así mismo, se definieron los hiperparámetros del random forest y se entrenó el modelo con los mejores hiperparámetro así: n_estimators=200, max_depth=None, max_features='log2', min_samples_split=10, min_samples_leaf=4.

Figura 20.

Matriz de confusión y variables importantes modelo Random Forest



Fuente: elaboración propia (2023)

8.1.8 Evaluación

Se evaluaron los modelos con varias métricas como, la precisión, la sensibilidad, el valor F1, la exactitud y la curva ROC, obteniendo los siguientes resultados:

Tabla 6.

Métricas de evaluación de los modelos

MÉTRICAS	ÁRBOL DE CLASIFICACIÓN	GRADIENT BOOSTING	RANDOM FOREST
Precisión (Precision)	0.337	0.360	0.357
Sensibilidad (Recall)	0.598	0.684	0.706
Valor F1 (F1 Score)	0.431	0.472	0.474
Exactitud (accuracy)	0.842 +/- 0.002	0.844 +/- 0.002	0.851 +/- 0.002
Curva ROC (Receiver Operating Characteristic) ROC (AUC-ROC	0.883	0.912	0.917

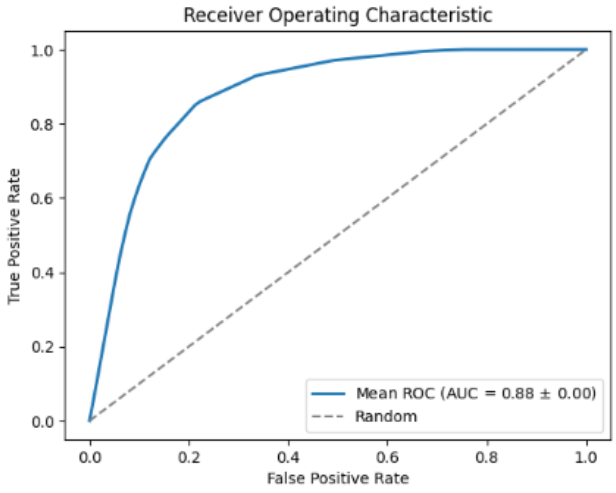
Nota. Métricas para evaluar los modelos predictivos

8.1.9 Selección de modelo para realizar la predicción

De acuerdo con la comparación de los tres modelos, el que presentó mejores resultados según las métricas de evaluación, fue el modelo random forest, obteniendo mayor resultado sobre los demás con una sensibilidad (Recall) de 0.706, un valor en F1 (F1 Score) de 0.474, una exactitud de 0,851 y un AUC o área bajo la curva de 0.917. A continuación de presentan las gráficas de la curva ROC o área bajo la curva:

Figura 21.

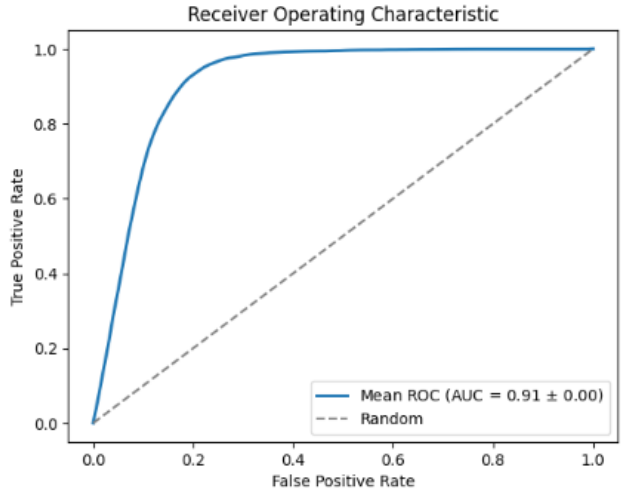
Cálculo de la curva ROC y el AUC para modelo de árbol de clasificación.



Fuente: elaboración propia (2023)

Figura 22.

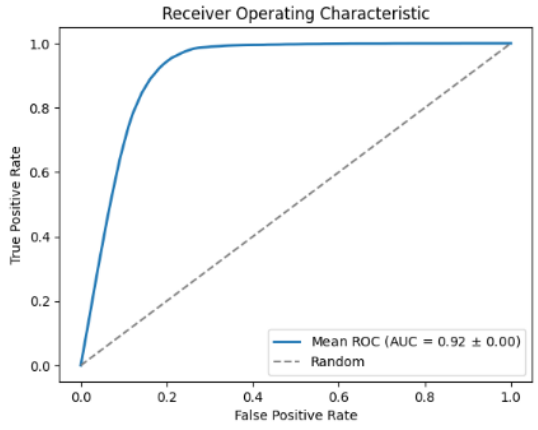
Cálculo de la curva ROC y el AUC para modelo de Gradient Boosting.



Fuente: elaboración propia (2023)

Figura 23.

Cálculo de la curva ROC y el AUC para modelo de Random Forest



Fuente: elaboración propia (2023)

Así mismo, como resultado del modelo se identificó que los programas de formación a los cuales es necesario que la Oficina de Control Interno realice alertas preventivas a la

Dirección de Formación profesional para que esta última, a su vez tome acciones y disminuya la tasa de deserción en estos programas de formación, corresponden a 202 programas de formación en los cuáles el modelo efectuó la predicción de los cuáles el 42% pertenece a cursos de nivel de formación técnica, el 36% a formación especial y 22 a formación en tecnólogo.

Tabla 7.

Predicción modelo deserción programas de formación

NIVEL FORMACIÓN	PREDICCIÓN PROGRAMAS DE FORMACIÓN
CURSO ESPECIAL	73
TECNICO	84
TECNOLOGO	45

Nota. Predicción según Programas de formación y niveles de formación que van a desertar en 2023. Fuente: Elaboración propia

8.1.10 Despliegue

Para el despliegue del modelo se requiere capacitar a los auditores misionales en el ciclo de vida de los datos y en analítica predictiva, así mismo, es fundamental poner en producción el modelo analítico en el segundo semestre de 2023 y realizar una auditoría al proceso misional usando el modelo.

9. Plan y recomendaciones de implementación y aplicación

Para aplicar la herramienta de data analytics se recomienda que la Oficina de Control Interno capacite a los auditores que participan en la auditoría del proceso misional de la Entidad, en el ciclo de vida de los datos, data analytics y analítica descriptiva y analítica predictiva, así como en el uso de herramientas de business analytics

9.1 Recurso humano

Para ejecutar el modelo se deben tener en cuenta en recurso humano las siguientes horas:

1. Horas hombre acompañamiento del Product Owner de la Oficina de Control Interno.

2. Horas hombre acompañamiento del Scrum Master Tutor del Proyecto.

3. Horas hombre acompañamiento del Development del Proyecto Empresarial.

9.2 Limitaciones en recursos de máquina

Uno de los temas que se deben revisar en la Oficina de Control Interno son las Máquinas/PC, dado que se requiere que carguen y ejecuten grandes volúmenes de datos para poder implementar analítica predictiva.

9. Conclusiones

De la comparación de los tres modelos (árbol de clasificación, Gradient Boosting y random forest), que son usados según la revisión teórica para predecir deserción de estudiantes el que obtuvo mayor precisión fue el Random Forest con un 0,92%, por lo cual fue el modelo utilizado para efectuar la predicción sobre los programas de formación en los cuales se puede presentar deserción en las Regionales de Antioquia y Bogotá según el alcance definido.

De acuerdo con el modelo predictivo supervisado, son 202 programas de formación en los cuales se va a presentar deserción de estudiantes en la vigencia 2023 y de los cuales es necesario que la Oficina de Control Interno realice alertas preventivas en las Regionales de Antioquia y Distrito Capital, de los cuales el 42% pertenece a nivel técnico, el 36% a cursos especiales y el 22% a tecnólogos.

Se propone aplicar el modelo de deserción de aprendices en la Entidad a nivel nacional, para identificar potenciales alertas preventivas a la Dirección de Formación Profesional y contribuir a la disminución de la tasa de deserción de la vigencia 2023.

Como se evidenció, el proyecto aporta un avance en la aplicación de analítica predictiva en las Oficinas de Control Interno de sector Nación, dado que el proyecto fue ejecutado en el Servicio Nacional de Aprendizaje SENA entidad adscrita al Ministerio de Trabajo de Colombia.

Para implementar el proyecto se recomienda que se capacite a los auditores que revisan el proceso misional en la Entidad, en herramienta de business analytics, en analítica descriptiva y predictiva y en todo el ciclo de vida de los datos.

Se recomienda tener en cuenta las horas hombre que se van a utilizar para implementar el proyecto, así como solicitar máquinas o PC para ejecutar grandes volúmenes de datos, dado que la próxima fase del proyecto se debe analizar datos de la Entidad a nivel nacional.

Referencias bibliográficas

Aguilar Vilca, D., & Camargo Ramos, J. C. (2021). *Sistema inteligente basado en redes neuronales, máquina de soporte vectorial y random forest para la predicción de deserción de clientes en microcréditos de bancos* [Universidad Nacional Mayor de San Marcos]. http://cybertesis.unmsm.edu.pe/bitstream/handle/20.500.12672/16390/Aguilar_vd.pdf?sequence=1&isAllowed=y

Asamblea Nacional Constituyente. (1991). *Constitución Política de Colombia 1991*. http://www.secretariasenado.gov.co/senado/basedoc/constitucion_politica_1991.html

Cardona Taborda, C. H., Gelves García, N., & Palacios Rozo, J. J. (2016). Análisis de datos mediante el algoritmo de clasificación 48, sobre un cluster en la nube de AWS. *Universidad Distrital Francisco José de Caldas*, *especial*, 145. http://revistas.udistrital.edu.co/ojs/index.php/REDES/index

Castro Sotomonte, J. E., Rodríguez Rodríguez, C. C., Montenegro Marín, C. E., Gaona García, P. A., & Castellanos, J. G. (2016). Hacia la construcción de un modelo predictivo de deserción académica basado en técnicas de minería de datos. *Revista Científica Universidad Distrital Francisco José de Caldas*, *26*, 49. https://doi.org/10.14483/23448350.11089

Contreras, C. (2021). Determinación de variables predictivas de deserción inicial para generar un sistema de alerta temprana. Análisis sobre una muestra de estudiantes beneficiarios de la beca de nivelación académica en una universidad pública en Chile. *Calidad en la educación*, *54*, 12-45. https://doi.org/10.31619/caledu.n54.828

Cuji, B., Gavilanes, W., & Sanchez, R. (2017). Modelo predictivo de deserción estudiantil basado en arboles de decisión. *Revista Espacios*, *38*(55), 17. http://ww.revistaespacios.com/a17v38n55/a17v38n55p17.pdf

Deloitte. (2016). *Internal audit analytics: The journey to 2020 Insights-driven auditing*. Deloitte Development. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/risk/us-risk-internal-audit-analytics-pov.pdf

DNP Departamento Nacional de Planeación & Departamento Administrativo de la Función Pública. (2010). *CONPES 3654 de 2010 Politica de Rendición de Cuentas de la Rama Ejecutiva a los ciudadanos*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=83124

Echeverría, F. (2019, septiembre 20). *Auditoría Interna Re Imaginada—KPMG Colombia*. KPMG. https://kpmg.com/co/es/home/insights/2019/09/auditoria-interna-re-imaginada.html

Fernández Martín, T., Solís Salazar, M., Hernández Jiménez, M. T., & Moreira Mora, T. E. (2019). Un análisis multinomial y predictivo de los factores asociados a la deserción universitaria. *Revista Electrónica Educare*, *23*(1), 73-97. https://doi.org/10.15359/ree.23-1.5

Galán Cortina, V. (2015). *Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario*. 21.

IBM. (2021, agosto 17). *IBM Documentation*. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Instituto de Censores Jurados de Cuentas de España. (2019). *La transformación digital en el sector de auditoría*. https://www.icjce.es/adjuntos/transf-digital.pdf

Islam, S., & Stafford, T. (2022). Factors associated with the adoption of data analytics by internal audit function. *Managerial Auditing Journal*, *37*(2), 193-193-223. Emerald Insight. https://doi.org/10.1108/MAJ-04-2021-3090

Jurado Mantilla, M. J. (2019). *Diseño de un modelo predictivo de la deserción estudiantil de postgrado en una institución de educación superior*. http://www.dspace.espol.edu.ec/xmlui/handle/123456789/48758

Manríquez Pacheco, R. (2022). *Modelo de Deserción Estudiantil*. https://repositorio.udd.cl/handle/11447/7490

Nelson, G. S. (2018). *The analytics lifecycle toolkit: A practical guide for an effective analytics capability* (1st ed.). Hoboken, New Jersey : Wiley.

Quintero, Y. A. (2022). *Diseño de un modelo predictivo para generar alertas tempranas de deserción universitaria en los programas de pregrado presenciales de la Facultad de Ingeniería de la Universidad de Antioquia.* [Universidad de Antioquia]. https://bibliotecadigital.udea.edu.co/bitstream/10495/29368/1/QuinteroYudy_2022_ModeloPredictivoDesercio%cc%81n.pdf

Raunakjhawar. (s. f.). *Extracción, transformación y carga de datos (ETL)—Azure Architecture Center*. Recuperado 23 de mayo de 2023, de https://learn.microsoft.com/es-es/azure/architecture/data-guide/relational-data/etl

Rivera Vergaray, K. (2021). Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica. *Revista Innovación y Software*, *2*(2), 6-13. https://www.redalyc.org/journal/6738/673870839001/html/

Schwaber, K., & Sutherland, J. (2020). *La Guía de Scrum La Guía Definitiva de Scrum: Las Reglas del Juego*. ScrumGuides.org. https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-Spanish-Latin-South-American.pdf

SENA. (s. f.). *Glosario*. SENA. Recuperado 4 de mayo de 2023, de https://www.sena.edu.co:443/es-co/ciudadano/Paginas/glosario.aspx

SENA. (2020). *Identificación causas de deserción 2020*. Servicio Nacional de Aprendizaje SENA. https://www.sena.edu.co/es-co/ciudadano/Documents/identificacion_causas_desercion_2020.pdf

SENA. (2023). *Estructura Orgánica del SENA*. Servicio Nacional de Aprendizaje SENA. https://www.sena.edu.co/es-co/sena/Documents/Estructura_Organica_SENA_%20PP_dao.pdf

Servicio Nacional de Aprendizaje SENA. (2023, abril 8). *Planeación estratégica*. SENA. https://www.sena.edu.co/es-co/sena/Paginas/planeacionEstrategica.aspx

Anexos Técnicos

Anexo A.

Diccionario de datos:

Campo	Descripción del Campo	Tipo
CODIGO_REGIONAL	Código de la regional	Unicode string [DT_WSTR] Texto
NOMBRE_REGIONAL	Nombre de la Regional (33 regionales a nivel nacional	Unicode string [DT_WSTR] Texto
CODIGO_CENTRO	Código del Centro de Formación	Unicode string [DT_WSTR] Texto
NOMBRE_CENTRO	Nombre del Centro de Formación	Unicode string [DT_WSTR] Texto
IDENTIFICADOR_UNICO_FICHA	Código de la Ficha (curso) con respecto al programa de formación.	Unicode string [DT_WSTR] Texto
FECHA_INICIO_FICHA	Fecha inicial de la ficha	Unicode string [DT_WSTR] Texto
FECHA_TERMINACION_FICHA	Fecha final de la ficha	Unicode string [DT_WSTR] Texto
CODIGO_PROGRAMA	Código del Programa (curso) de formación	Unicode string [DT_WSTR] Texto
VERSION_PROGRAMA	Versión del Programa (curso) de formación	Unicode string [DT_WSTR] Número
NOMBRE_PROGRAMA_FORMACION	Nombre del programa de formación	Unicode string [DT_WSTR] Texto
NIVEL_FORMACION	Títulos otorgados según el nivel de formación profesional integral. -Auxiliar -Especialización técnica -Especialización tecnológica	Unicode string [DT_WSTR] Texto

	-Operario -Profundización técnica -Técnico -Técnico Profesional -Tecnólogo	
MODALIDAD_FORMACION	Modalidad de formación Presencial Virtual A distancia	Unicode string [DT_WSTR] Texto
TOTAL_APRENDICES_MATRICULADOS	Total aprendices matriculados por programa de formación	Unicode string [DT_WSTR] Número
DESERTORES_AÑO_ACTUAL	Total aprendices que desertaron por programa de formación	Unicode string [DT_WSTR] Número
PERIODO	Vigencia al cual corresponden los datos	Unicode string [DT_WSTR] Número

Nota. Diccionario de datos correspondiente a la data de deserción de la formación profesional integral. Fuente: Elaboración propia

Anexo B.

Archivo con Tablero de Power BI con la visualización de la deserción profesional integral.

Anexo C.

Archivo con código Python del modelo predictivo de deserción de aprendices en programas de la formación profesional.

Anexo D.

Base de datos denominada “Deserción_Aprendices_Modelo_1”

Base de datos denominada “Deserción_Aprendices_Modelo_1_Potencial”