Universidad de Los Andes

Maestría en Inteligencia Analítica de Datos

Proyecto Despliegue de soluciones analíticas

Autores:

Jose Daniel Garcia Correa

Cristian Mauricio Romero Buitrago

Cesar Daniel Ramírez Cely

Hugo Ruiz Bautista

Departamento de Ingeniería Industrial

2024

Resumen del problema

La deserción laboral, o el abandono de una organización por parte de sus empleados, es un reto significativo que afecta a empresas de todos los sectores. Este fenómeno no solo repercute en la productividad y en el clima organizacional, sino que también genera elevados costos asociados a la contratación y formación de nuevo personal. La competencia en el mercado actual exige a las empresas no solo captar talento sino también conservarlo, siendo esta retención de personal un factor esencial para mantener la continuidad operativa y el éxito a largo plazo de las organizaciones. Comprender los motivos que llevan a los empleados a renunciar es fundamental para que las empresas puedan actuar de manera estratégica, minimizando la rotación no deseada.

El dataset *IBM HR Analytics Employee Attrition & Performance* constituye una valiosa fuente de datos al incluir variables demográficas, características laborales y evaluaciones de desempeño de empleados de una empresa, ofreciendo una base rica para estudiar los factores que inciden en la deserción. Con esta información, se pueden identificar patrones y correlaciones en la salida de empleados, permitiendo que los departamentos de recursos humanos formulen políticas informadas y estrategias de retención. Estas políticas, a su vez, pueden optimizar la experiencia laboral, generando un ambiente de trabajo atractivo que impulse tanto el compromiso como la productividad de los empleados.

Este análisis, al revelar factores clave de deserción, busca ofrecer una respuesta a la pregunta central del proyecto: ¿Cuáles son los factores determinantes en la decisión de los empleados de abandonar la organización y cómo pueden estos hallazgos contribuir a diseñar estrategias efectivas de retención?

- **Pregunta de negocio**

¿Cuáles son los factores clave que influyen en la deserción laboral de los empleados en nuestra organización, y qué estrategias pueden implementarse para mejorar la retención del personal?

Este proyecto busca identificar los factores clave detrás de la deserción laboral y ofrecer estrategias para mejorar la retención de empleados. Utilizando el dataset de IBM, se realizará un análisis exhaustivo de atributos personales y laborales de los empleados, seguido por la creación de modelos predictivos para anticipar la probabilidad de deserción. Los hallazgos permitirán desarrollar recomendaciones estratégicas para recursos humanos, respaldadas por visualizaciones y un dashboard interactivo que facilitará el seguimiento de las métricas de retención.

- **Descripción del conjunto de datos**

El proyecto utiliza el dataset "IBM HR Analytics Employee Attrition & Performance", que incluye atributos personales y laborales de los empleados, tales como edad, género, nivel educativo, posición, salario, satisfacción laboral y antigüedad.

Modelos desarrollados

En el contexto del análisis de deserción laboral, los modelos predictivos de Random Forest y Naive Bayes son altamente aplicables debido a su capacidad para trabajar con datos de características demográficas y laborales variadas, como los presentes en el dataset de IBM HR Analytics Employee Attrition & Performance. Estos modelos permiten detectar patrones y correlaciones que pueden ayudar a predecir la probabilidad de que un empleado deje la organización, proporcionando así una base sólida para el diseño de estrategias de retención de personal efectivas.

Random Forest es una técnica avanzada de aprendizaje supervisado que utiliza una combinación de múltiples árboles de decisión para incrementar la precisión y confiabilidad de las predicciones. En lugar de depender de un solo árbol, Random Forest construye varios árboles sobre diferentes subconjuntos de datos y luego promedia sus resultados, reduciendo así el impacto de variaciones o sesgos individuales en los datos. Este enfoque es particularmente ventajoso en el análisis de deserción laboral, pues la deserción puede estar influenciada por una amplia variedad de factores interrelacionados, como la edad, el nivel educativo, el salario y la satisfacción laboral, que no siempre son fácilmente identificables mediante métodos simples de clasificación. Además, la naturaleza robusta de Random Forest frente al sobreajuste lo convierte en un modelo adecuado para datos de alta dimensionalidad y para realizar predicciones que se mantienen consistentes en nuevos conjuntos de datos. Este modelo también permite identificar las variables con mayor peso en las decisiones de deserción, lo cual facilita una interpretación valiosa de los factores críticos que contribuyen a la rotación, proporcionando a los líderes de recursos humanos una herramienta sólida para tomar decisiones informadas en la gestión del talento.

Por otro lado, Naive Bayes es un modelo probabilístico que utiliza el teorema de Bayes para realizar predicciones basadas en la probabilidad condicional de los atributos, bajo la asunción de independencia entre ellos. Aunque esta suposición de independencia puede parecer restrictiva, Naive Bayes ha demostrado ser notablemente efectivo y eficiente en tareas de clasificación binaria, como en el análisis de deserción laboral, donde se busca clasificar empleados en función de su probabilidad de dejar la empresa. Su simplicidad y bajo costo computacional hacen que sea ideal para análisis rápidos y fácilmente interpretables, lo cual es valioso en escenarios de toma de decisiones ágiles. Además, Naive Bayes permite analizar tanto variables categóricas como continuas, lo cual resulta útil cuando se trabaja con datos diversos de empleados que incluyen características como la

antigüedad, el género, el puesto de trabajo y las evaluaciones de desempeño. Dado que este modelo también genera una visión clara de cómo factores específicos afectan la probabilidad de deserción, es útil para comprender los riesgos individuales asociados a cada factor, permitiendo a las organizaciones identificar rápidamente perfiles de empleados que pueden requerir intervenciones específicas para mejorar su retención.Ambos modelos tienen beneficios particulares en este proyecto. Random Forest ofrece una precisión elevada y es útil para entender la importancia de los factores predictivos, mientras que Naive Bayes permite una interpretación sencilla y es ideal para realizar clasificaciones rápidas con alta eficiencia. La combinación de ambos enfoques puede proporcionar una perspectiva completa y versátil para predecir la deserción, permitiendo a los líderes de recursos humanos implementar políticas preventivas y estrategias de retención basadas en datos sólidos y modelos interpretables.

El codigo realizado carga un conjunto de datos sobre recursos humanos, donde se analiza la variable Attrition (rotación de empleados). Primero, se realiza un preprocesamiento de los datos: se convierten las variables categóricas en números con LabelEncoder, y se normalizan las características numéricas utilizando MinMaxScaler. Posteriormente, se seleccionan un subconjunto de características relevantes y se divide el conjunto de datos en entrenamiento (75%) y prueba (25%).

El código luego entrena dos modelos de clasificación diferentes: un Random Forest y un Naive Bayes. Ambos modelos se entrenan con los datos de entrenamiento, y sus predicciones se evalúan utilizando métricas como la exactitud, el área bajo la curva ROC, el F1-score, la precisión y el recall. Estos modelos y sus métricas de rendimiento se registran en MLflow para un seguimiento y análisis posterior. Finalmente, se imprimen los resultados de la evaluación de cada modelo.

Tablero desarrollado

El tablero diseñado para desplegarse mediante DASH está estructurado en dos pestañas principales, ofreciendo una interfaz interactiva y visualmente atractiva que permite explorar y analizar los factores clave de la deserción laboral, además de realizar simulaciones para entender cómo ciertas características influyen en la probabilidad de abandono de los empleados.

Pestaña 1: Visualización de Datos Generales

Esta pestaña proporciona una vista general de la distribución demográfica y organizacional de los empleados, con gráficos interactivos que permiten a los usuarios explorar los datos desde diferentes perspectivas. Los componentes incluidos son:

Distribución de Género: Un gráfico circular que muestra el porcentaje de empleados masculinos y femeninos dentro de la organización, proporcionando un panorama sobre la equidad de género.

Distribución por Años de Trabajo: Un gráfico de barras que representa la cantidad de empleados según sus años de antigüedad en la empresa, lo que puede ayudar a identificar patrones de retención y riesgo de deserción en relación con la experiencia laboral.

Distribución por Departamento: Un gráfico de barras que indica el número de empleados en cada departamento, facilitando la identificación de áreas con mayores desafíos de retención o patrones de alta deserción.

La primera pestaña permite explorar de manera visual los atributos clave relacionados con los empleados y sus características, lo que contribuye a un entendimiento inicial de los datos y su contexto.

Pestaña 2: Ingreso de Información y Predicción de Deserción

La segunda pestaña está diseñada como un formulario interactivo donde los usuarios pueden ingresar datos de un trabajador individual, tales como:

Variables Demográficas: Edad, género, estado civil.

Variables Laborales: Departamento, nivel educativo, años de trabajo, salario, satisfacción laboral, distancia desde casa al trabajo.

Una vez ingresada la información, el sistema procesa los datos mediante modelos predictivos previamente entrenados para calcular la probabilidad de deserción del empleado en cuestión. Los resultados se presentan de manera clara, acompañados de recomendaciones específicas que el departamento de recursos humanos puede considerar para mejorar la retención de dicho empleado.

Tablero desplegado

El diseño del tablero está directamente alineado con los objetivos del proyecto, que busca identificar factores clave detrás de la deserción laboral y formular estrategias efectivas de retención.

La primera pestaña responde a la necesidad de explorar y comprender patrones generales en la deserción mediante visualizaciones interactivas de los datos demográficos y laborales, permitiendo identificar tendencias y áreas críticas dentro de la organización.

La segunda pestaña complementa este análisis general con un enfoque personalizado, permitiendo simular cómo las características individuales de un empleado influyen en su probabilidad de abandonar la organización. Esto facilita la toma de decisiones basada en datos para recursos humanos.

Al integrar estas funcionalidades, el tablero no solo apoya la visualización de los hallazgos del análisis, sino que también proporciona herramientas prácticas para formular estrategias de retención, optimizando la experiencia laboral de los empleados y fortaleciendo el compromiso organizacional.

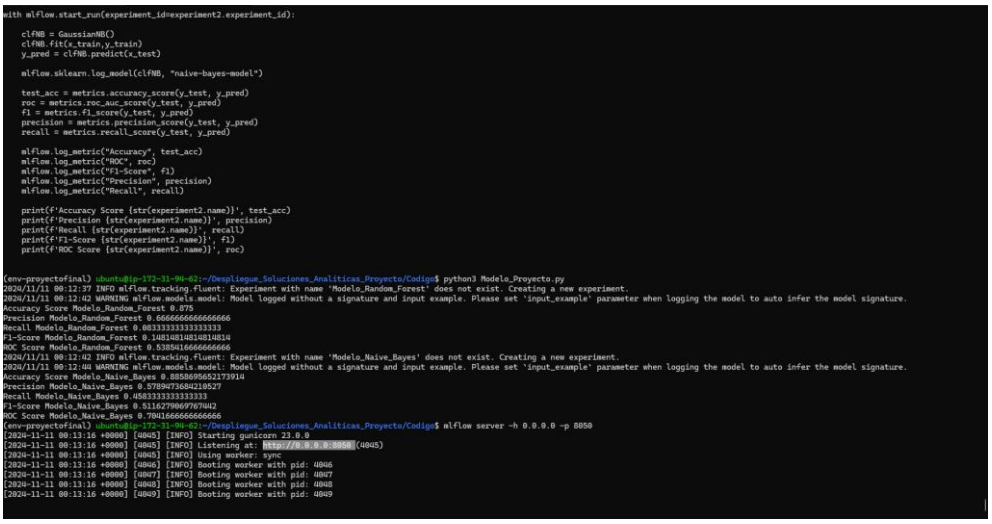
Repositorios GitHub

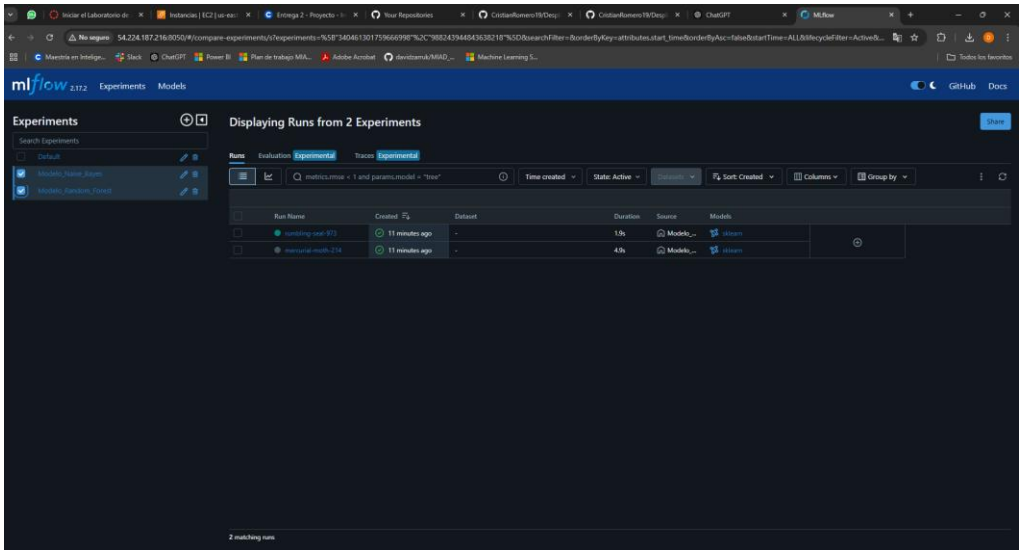
https://github.com/CristianRomero19/Despliegue_Soluciones_Analiticas_Proyecto?tab=readme-ov-file%23despliegue_soluciones_anal-ticas_proyecto

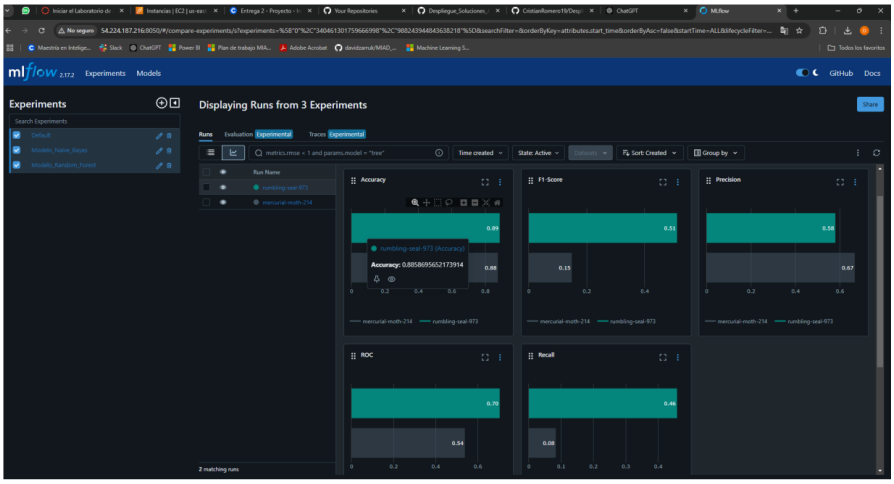
Video

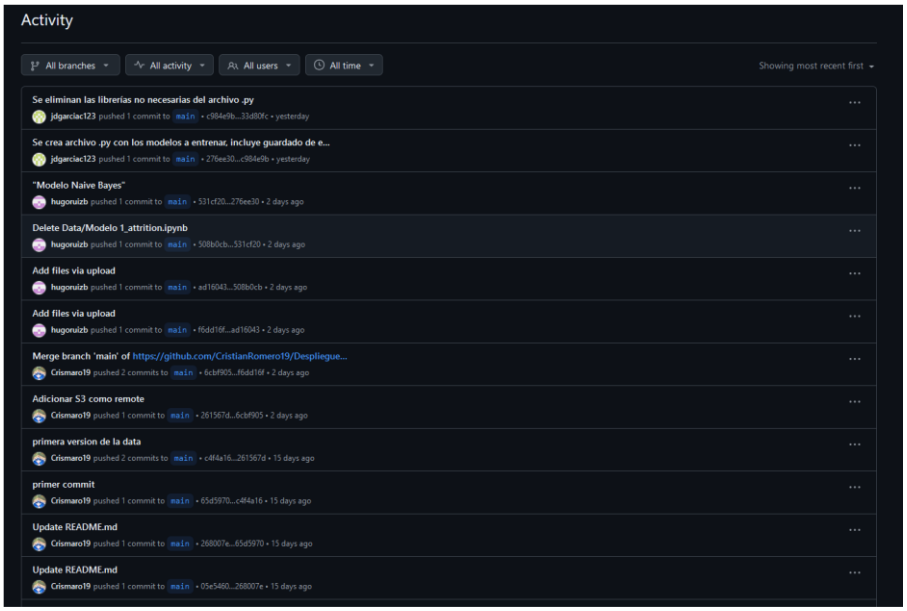
https://www.youtube.com/watch?v=QpYBBCrJmGg

Pantallazos MLFlow




mlflow 2.17.2
Experiments
Models
GitHub
Docs
Experiments
Displaying Runs from 2 Experiments
Share
Runs
Evaluation
Traces
Run Name
Created
Dataset
Duration
Source
Models
rumbling-seal-973
11 minutes ago
1.9s
mercurial-moth-214
11 minutes ago
4.9s
2 matching runs


mlflow 2.17.2
Experiments
Models
GitHub
Docs
Experiments
Displaying Runs from 3 Experiments
Share
Accuracy
F1-Score
Precision
ROC
Recall
2 matching runs


Activity
All branches
All activity
All users
All time
Showing most recent first
Se eliminan las librerías no necesarias del archivo .py
jdgarciac123 pushed 1 commit to main • c984e9b...33d80fc • yesterday
Se crea archivo .py con los modelos a entrenar, incluye guardado de e...
jdgarciac123 pushed 1 commit to main • 276ee30...c984e9b • yesterday
"Modelo Naive Bayes"
hugoruizb pushed 1 commit to main • 531cf20...276ee30 • 2 days ago
Delete Data/Modelo 1_attrition.ipynb
hugoruizb pushed 1 commit to main • 508b0cb...531cf20 • 2 days ago
Add files via upload
hugoruizb pushed 1 commit to main • ad16043...508b0cb • 2 days ago
Add files via upload
hugoruizb pushed 1 commit to main • f6dd16f...ad16043 • 2 days ago
Merge branch 'main' of https://github.com/CristianRomero19/Despliegue...
Crismaro19 pushed 2 commits to main • 6cbf905...f6dd16f • 2 days ago
Adicionar S3 como remote
Crismaro19 pushed 1 commit to main • 261567d...6cbf905 • 2 days ago
primera version de la data
Crismaro19 pushed 2 commits to main • c4f4a16...261567d • 15 days ago
primer commit
Crismaro19 pushed 1 commit to main • 65d5970...c4f4a16 • 15 days ago
Update README.md
Crismaro19 pushed 1 commit to main • 268007e...65d5970 • 15 days ago
Update README.md
Crismaro19 pushed 1 commit to main • 05e5460...268007e • 15 days ago

Reporte de trabajo en equipo

Cristian Romero se encargó de la realización del enlace entre la máquina virtual y el repositorio en GitHub, asegurando una integración eficiente de los archivos y facilitando la colaboración en el proyecto a su vez también realizó el despliegue del dash. Hugo Ruiz realizó los códigos para los diferentes modelos de predicción a su vez del montaje de la API, desarrollando y ajustando los algoritmos necesarios para el análisis. José García y Cesar Ramírez se encargaron del montaje del dash, para la visualización optima de la información y finalmente, entre todos realizaron aportes a los commits en GitHub, colaborando en las actualizaciones y mejoras continuas del repositorio.