Trabajo Práctico Inicial Plataforma de Gestión de Recursos Humanos con ERP y CRM con IA

Universidad Nacional de General Sarmiento

Integrantes del Equipo 6

Gaspar Abel Aquino
Gmail: abelaquino02@gmail.com
Cristian Jurajuria
Gmail: cljurajuria23@gmail.com
Lautaro Emanuel Moreno
Gmail: lemoreno2002@gmail.com
Rodrigo Montoro
Gmail: rodrigo.montoro@hotmail.com
Pablo Samuel Da Silva
Gmail: polilladasilva0@gmail.com

INVESTIGACIÓN TEÓRICA (Parte 1):

En un entorno empresarial cada vez más digitalizado, tecnologías como los sistemas ERP (Enterprise Resource Planning), CRM (Customer Relationship Management) e Inteligencia Artificial (IA) están transformando la gestión de Recursos Humanos (RRHH). Este trabajo investigativo explora cómo estas herramientas optimizan procesos clave, como el reclutamiento, la formación y la retención de talento, mejorando la eficiencia y la toma de decisiones. Además, se analizan las ventajas y desafíos de su implementación, ofreciendo una visión clara de su impacto en la competitividad y sostenibilidad.

ERP

Enterprise Resource Planning (ERP) es un software de gestión integral, diseñado como una estructura de datos centralizada para que todos los usuarios de la empresa accedan a los mismos datos a partir de procesos comunes, utilizado para la gestión de actividades empresariales cotidianas, como contabilidad, finanzas, gestión de inventarios e ingresos, gestión de proyectos, marketing, la fabricación y operaciones de la cadena de suministro. Incluye herramientas de análisis de datos y gestión financiera para planificar, presupuestar y notificar resultados financieros.

Pueden ser locales o en la nube, respaldando la gestión financiera, los recursos humanos y la producción. Son adaptables a diferentes sectores, mejorando la eficiencia y optimizando los recursos de la empresa. También se integran con aplicaciones de front-office para crear vistas holísticas de los clientes, incluidas soluciones de gestión de relaciones con clientes (CRM).

El ERP utilizado en RRHH permite una gestión integral del departamento, optimizando los procesos por medio de las herramientas interconectadas, facilitando la creación de informes conjuntos.
Además, posee un carácter escalable ya que añade módulos específicos como gestión de nóminas, evaluación, selección, portal del empleado o reporting (mejora la comunicación interna) y análisis de datos. Algunos ejemplos reales son: Workday HCM (Nissan, Amazon), SAP SuccessFactors con servicios (Volkswagen, ExxonMobil, Nestlé), Oracle Fusion Cloud HCM, Microsoft Dynamics 365 Human Resources, etc.

El ERP para RRHH aporta las siguientes ventajas:

- Mayor focalización en tareas departamentales que aportan valor añadido, al automatizar las tareas de trabajo puramente mecánico como nómina, reclutamiento, capacitaciones y gestión de desempeño.
- Control del gasto, por medio de la agilización de los procesos, el análisis y control del mismo.
- Análisis de datos al detalle y reportes para evaluar el desempeño de los empleados y su productividad.
- Comunicación interna entre el empleado y la empresa, permite trabajar de forma autónoma
- Escalabilidad, se adapta a cualquier tipo de empresa.

Entre los desafíos de implementar ERP para RRHH se encuentran los siguientes:

- Costo del software elevado. Además, tiene una curva de aprendizaje profunda y se necesita capacitación del personal.
- Implementación compleja. Puede llevar meses.
- Si es basado en la nube, depende de la conectividad a internet y la seguridad del proveedor.

CRM

CRM es la sigla utilizada para Customer Relationship Management y se refiere al conjunto de prácticas, estrategias comerciales y tecnologías enfocadas en la relación con el cliente.
El CRM almacena información de clientes actuales y potenciales (como nombre, dirección, número de teléfono, etc) e identifica sus actividades y puntos de contacto con la empresa. Esto incluye visitas de los clientes al sitio, llamadas telefónicas realizadas, intercambios por correo electrónico y varias otras interacciones.
Recopila e integra datos valiosos para preparar y actualizar a los equipos con información personal de los clientes, sus historiales de compra y sus preferencias.
También entre sus funcionalidades clave destacan la automatización de ventas, servicio al cliente y análisis y reporting (que proporciona informes y dashboards en tiempo real para analizar el comportamiento del cliente, el rendimiento de ventas y la efectividad de las campañas de marketing).

Cuando se trata de tipos de CRM, hay dos caminos que las empresas pueden tomar, según sus necesidades y presupuesto: CRM en la Nube o CRM Cloud (no se instala en una computadora ni requiere mantenimiento, también puede llamarse software como servicio) y CRM Local (se aloja en un servidor físico de la empresa y requiere manutención por parte de un equipo de TI propio).

Los sistemas CRM se aplican en la gestión de RRHH para centralizar y optimizar procesos clave almacenando toda la información referente a los empleados de una empresa (historial de cada empleado, evaluaciones de desempleo, datos de contacto, asistencias y ausencias, y detalles sobre salarios, beneficios y compensaciones). Además, integrados con plataformas de aprendizaje, apoyan el desarrollo profesional y la formación. También optimizan la comunicación interna, la evaluación del desempeño y la gestión de beneficios, ofreciendo una visión integral y personalizada de cada empleado.

El CRM para RRHH aporta las siguientes ventajas:

- Mejor comunicación interna. Información de los empleados y un historial de interacciones. Evaluación de desempeño constante.
- Automatización y gestión de candidatos y reclutamiento mediante pipelines y comunicación automatizada.
- Onboarding de nuevos empleados con recordatorios y tareas personalizadas
- Retención de clientes por medio del análisis de datos.

Entre los desafíos de implementar CRM para RRHH se encuentran los siguientes:

- El personal que lo utilizará deberá contar con un periodo de aprendizaje y adaptación para acostumbrarse adecuadamente a el sistema.
- Cumplir con la Ley de Protección de Datos. El CRM almacena muchos datos personales de clientes, de ahí que deban cuidarse al máximo y ser precavidos a la hora de utilizarlos (el usuario debe haber dado antes su consentimiento). Además, los usuarios deben contar con la posibilidad de darse de baja de servicios, como recibir notificaciones o avisos.
- Gastos del software y del equipo técnico. Precios elevados que se convierten en una barrera infranqueable

IA en RRHH

La inteligencia artificial (IA) es un campo de la ciencia enfocado en desarrollar sistemas capaces de razonar, aprender y actuar con un nivel de inteligencia comparable al humano o en escenarios donde la escala de datos supera la capacidad de análisis manual. En el ámbito empresarial, la IA se ha convertido en una herramienta clave, con aplicaciones en análisis de datos, generación de predicciones, procesamiento de lenguaje natural, recomendaciones y recuperación inteligente de información.

Uno de los sectores donde la IA está transformando los procesos es Recursos Humanos (RRHH). Tradicionalmente, los líderes de RRHH han evolucionado de una función meramente administrativa a un rol estratégico dentro de las organizaciones. Sin embargo, aún destinan gran parte de su tiempo a tareas repetitivas, como la revisión de currículums, la programación de entrevistas y el procesamiento de documentos. La IA permite automatizar estas actividades, optimizando la gestión del talento y permitiendo a los profesionales de RRHH enfocarse en iniciativas estratégicas que agreguen valor a la organización.

Desde el reclutamiento hasta la retención de empleados, la IA mejora la eficiencia en distintos procesos, entre ellos:

- Análisis de currículums: Herramientas como LinkedIn Recruiter y HireVue utilizan IA para analizar grandes volúmenes de CVs y extraer información clave, como experiencia laboral y habilidades, facilitando la selección de candidatos.
- Evaluación de candidatos: Plataformas como Pymetrics emplean IA y juegos cognitivos para evaluar habilidades y características de personalidad, alineándose con los requisitos del puesto.
- Selección y entrevistas: Sistemas como X0PA AI predicen la idoneidad de un candidato basándose en datos históricos, mientras que chatbots de IA agilizan la interacción con postulantes, respondiendo preguntas y programando entrevistas de manera automática.
- Gestión del desempeño: Herramientas como Workday y SAP SuccessFactors analizan métricas de productividad y satisfacción, permitiendo una evaluación más objetiva del talento dentro de la empresa.

Ventajas y Desafíos de la IA en RRHH

El uso de la IA en la gestión del capital humano conlleva tanto ventajas como desafíos. Entre los beneficios destacan la automatización de tareas, la reducción de tiempos de contratación, el mejor control de la plantilla, la optimización de los flujos de trabajo, el registro preciso de datos y la predicción de tendencias para futuras contrataciones. Sin embargo, también presenta desafíos como la excesiva dependencia de la tecnología, posibles sesgos en los algoritmos, preocupaciones sobre la seguridad y confidencialidad de los datos y el impacto en el empleo y en el desarrollo de habilidades humanas.

INVESTIGACIÓN TEÓRICA (Parte 2):

El Machine Learning es una rama de la inteligencia artificial que permite a los sistemas aprender a partir de datos y tomar decisiones sin ser programados. Para desarrollar modelos de aprendizaje automático en Python, una de las bibliotecas más utilizadas es Scikit-learn, que cuenta con herramientas eficientes para la clasificación, regresión, agrupamiento y detección de anomalías.

En esta investigación, exploramos el uso de Scikit-learn y analizaremos las siguientes técnicas de Machine Learning:

- Regresión Logística y Árboles de Decisión: Utilizados para problemas de clasificación.
- Regresión Lineal: Aplicada a predicciones numéricas.
- K-means: Algoritmo de agrupamiento para segmentación de datos.
- Isolation Forest: Técnica de detección de anomalías.

Cada una de estas técnicas será detallada en términos de su funcionamiento y aplicaciones.

Scikit learn:

Scikit-learn es una librería de código abierto que facilita el desarrollo de modelos de aprendizaje automático con herramientas simples pero efectivas. Implementada en Python, esta librería se basa en otras como NumPy, SciPy y Matplotlib para su funcionalidad. Entre sus herramientas más destacadas están los algoritmos para clasificación, regresión, detección y agrupamiento.

El propósito de Scikit-learn es permitir el desarrollo de modelos de aprendizaje automático, los cuales pueden abordar problemas supervisados y no supervisados. Algunas de sus funcionalidades clave son:

- Procesamiento previo de datos: Ajustar y transformar los datos que se usarán para entrenar el modelo antes de aplicar los algoritmos.
- Selección de características: Identificar únicamente las características relevantes para el problema.
- Implementación de algoritmos: Para resolver problemas de clasificación, regresión y agrupamiento
- Evaluación de modelos: Medir la capacidad predictiva de los modelos
- Ajuste de modelos: Optimizar el rendimiento de modelos en datos calculados.

En el aprendizaje no supervisado, el modelo trabaja sin conocer de antemano las respuestas correctas ni los valores ideales. Por este motivo, se utilizan métodos como el agrupamiento o la detección de patrones para descubrir estructuras ocultas en los datos. Por otro lado, en el aprendizaje supervisado, el modelo se entrena con datos que ya incluyen la respuesta correcta. Esto significa que se le enseña al modelo la relación entre una entrada y su resultado esperado, permitiendo aplicar algoritmos para resolver problemas específicos.

Para desarrollar un modelo de machine learning se debe pasar por una serie de etapas, de las cuales se identifican:

1. Preparar los datos con los que aprenderá el modelo. Se debe cargar el dataset, limpiarlo y analizarlo, adaptándolo según las necesidades que se tenga, desde normalizar hasta verificar datos faltantes, es decir, se hace un análisis exploratorio y se hace un preprocesamiento de

datos.

Ejemplo: Se utiliza pd.read_csv("datos.csv") para cargar un archivo CSV con información de ventas, se revisa el contenido y se ajustan las variables numéricas para asegurar la calidad del análisis.

2. Seleccionar un modelo que se adecue a la necesidad, esto va a depender si se tiene un problema supervisado o no supervisado. Por ejemplo, para un problema supervisado puede elegirse un modelo de clasificación.

   Ejemplo: Para predecir si un cliente realiza una compra, se puede elegir un modelo de regresión logística o un árbol de decisión, ya que se trata de una clasificación.
3. Entrenar el modelo, en esta etapa se dividen los datos, un conjunto se usa para entrenarlo y el otro para evaluar el modelo. El modelo aprende a relacionar entradas con las salidas o detectar patrones en los datos.

   Ejemplo: Utilizando la función train_test_split de Scikit-learn, se separa el 70% de los datos para entrenar y el 30% restante para validar la capacidad del modelo de clasificar correctamente a los clientes.
4. Evaluar del modelo, con el conjunto de datos que se apartó para la validación, se mide su capacidad para generalizar y predecir nuevos datos, evaluando su generalidad y precisión.

   Ejemplo: Tras entrenar el modelo, se evalúa su capacidad predictiva midiendo la precisión con la que clasifica a los clientes en el conjunto de prueba.
5. Probar el modelo, con el modelo ya entrenado y evaluado, se le cargan nuevos datos que no estaban en el dataset inicial para probar el comportamiento con datos reales. En esta etapa se asegura que el modelo sea útil y fiable.

   Ejemplo: Una vez validado, el modelo se utiliza en una campaña de marketing en tiempo real para predecir y segmentar a los clientes según su comportamiento de compra.

Regresión logística:

La Regresión Logística es un modelo estadístico para estudiar las relaciones entre un conjunto de variables cualitativas Xi y una variable cualitativa Y. Utilizado para la clasificación y el análisis predictivo, permite predecir la probabilidad de que ocurra un evento utilizando 1 en caso afirmativo y 0 en caso contrario, mediante la optimización de los coeficientes de regresión. El resultado varía entre 0 y 1. Si el valor supera un umbral (definido acorde al criterio de clasificación deseado, generalmente es 0,5) es probable que ocurra el evento, mientras que si está por debajo, no es así.

La función que mejor cumple estas condiciones es la función Sigmoide, que transforma cualquier número en un valor acotado entre 0 y 1. Se escribe como:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

O también definida como

$$P(Y=1) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_nX_n)}}$$

donde *P(Y = 1)* es la variable dependiente que define la probabilidad de que un evento ocurra, *Xi* son las variables predictoras (independientes, son los datos de entrada), *b0* es el valor base cuando *X = 0* y *bi* son los coeficientes que determinan cuánto afecta cada *Xi*. Los *bi* son estimados a través de la

estimación de máxima verosimilitud (MLE). Este método se calcula utilizando el producto de todas las probabilidades individuales. Para simplificarlo, se utiliza el logaritmo de verosimilitud. Es decir:

$$L(b_0, b_1, ..., b_n) = \prod_{i=1}^{n} P(Y_i)$$

$$\log L = \sum_{i=1}^{n} [Y_i \log(P(Y_i)) + (1 - Y_i) \log(1 - P(Y_i))]$$

Por último, para encontrar valores de los coeficientes que sean óptimos, se utilizan algoritmos numéricos como el Gradiente Descendente que ajusta los coeficientes minimizando la función de pérdida (métrica que mide que tan bien o mal funciona el modelo).

$$J(b) = -\frac{1}{n} \sum_{i=1}^{n} [Y_i \log(P(Y_i)) + (1 - Y_i) \log(1 - P(Y_i))]$$

Los coeficientes son ajustados en cada iteración hasta que el modelo se ajusta lo mejor posible a los datos.

Tipos de Regresión Logística:

Binaria:

En este enfoque, la respuesta solo tiene dos posibles resultados (0, 1), es decir, redondea la respuesta a los valores más cercanos. Por lo general, la respuesta menor a 0,5 se redondea a 0 y las respuestas mayores a 0,5 se redondean a 1. Es el enfoque más utilizado dentro del modelo de regresión logística. Se utiliza la función sigmoide, vista anteriormente, y se clasifica según el valor de la probabilidad. Es decir:

Si $P(Y=1) > 0.5 \rightarrow$ Clasificamos como 1.

Si $P(Y=1) < 0.5 \rightarrow$ Clasificamos como 0.

Si $P(Y=1) = 0.5 \rightarrow$ El modelo está indeciso y necesita de un ajuste de umbral o más variables predictoras.

Es utilizado comúnmente en la predicción de si un correo es spam o no, si un tumor es maligno o no, la detección de fraude en tarjetas de crédito (Y puede ser fraude (1) o no fraude (0), Xi pueden ser el monto de la transacción, ubicación geográfica, hora de la compra, historial), diagnóstico de cáncer de mama (Y puede ser tumor maligno (1) o benigno (0), Xi pueden ser tamaño del tumor, textura de la celda y la edad del paciente), etc.

Multinomial:

En este tipo de modelo de regresión, la variable dependiente tiene tres o más respuestas finitas posibles sin un orden especificado. Funciona mapeando los valores de resultado con diferentes

valores entre 0 y 1. En este caso, la función utilizada es softmax para calcular probabilidades de cada clase. Esta función tiene una cantidad k de clases y para cada clase k sucede lo siguiente:

$$P(Y=k)=\frac{e^{(b_{0k}+b_{1k}X_1+b_{2k}X_2+\ldots+b_{nk}X_n)}}{\sum_{j=1}^{K} e^{(b_{0j}+b_{1j}X_1+b_{2j}X_2+\ldots+b_{nj}X_n)}}$$

Asegurando que todas las probabilidades suman 1. Se elige la clase con mayor probabilidad.

Se utiliza, por ejemplo, para la clasificación de tipos de vehículos (Y puede ser auto (0), moto (1) o camión (2), Xi pueden ser número de ruedas, potencia del motor y peso del vehículo), reconocimiento de emociones en texto para analizar mensajes (Y puede ser feliz (0), enojado (1) o triste (2), Xi pueden ser palabras clave, uso de signos de puntuación y longitud del mensaje), entre otros casos.

Ordinal:

La regresión logística ordinal es un tipo especial de regresión multinomial para problemas en los que los números representan rangos en lugar de valores reales. Es decir, cada respuesta tiene tres o más posibles resultados pero en este caso, tienen un orden definido. Se utiliza la función llamada modelo de umbrales, definida como:

$$P(Y\leq k)=\frac{1}{1+e^{-(b_k-(b_1X_1+b_2X_2+\ldots+b_nX_n))}}$$

Permite clasificar definiendo un conjunto de umbrales que dividen las clases de manera ordenada.

Por ejemplo, puede usarse para predecir la respuesta a una pregunta de una encuesta en la que se pide a los clientes que clasifiquen su servicio como malo (0), regular (1), bueno (2) o excelente (3) en función de un valor numérico como el número de artículos que le han comprado a lo largo del año. Otro ejemplo puede ser la predicción del riesgo crediticio, con posibles respuestas como bajo riesgo (0), medio riesgo (1) y alto riesgo (2) en función de las variables predictoras X como los ingresos mensuales, el historial de pagos y las deudas actuales. También se puede usar para escalas de calificación del 1 al 5, A a la F, etc.

Árboles de decisión:

Los árboles de decisión son un algoritmo de aprendizaje supervisado no paramétrico, se utiliza tanto para tareas de clasificación como de regresión. Con una estructura jerárquica de árbol, consta de un nodo raíz (el comienzo del árbol), ramas, nodos internos (conocidos como nodos de decisión) y nodos hoja. Es una representación fácil de digerir de la toma de decisiones, es decir que cualquier grupo de una organización puede comprender mejor la razón de una decisión.

El aprendizaje de árboles de decisión realiza una búsqueda para identificar los puntos de división óptimos dentro de un árbol, repitiendo la división de forma descendente y recursiva hasta que todos los nodos o la mayoría de los registros estén clasificados con etiquetas de clase específicas.

Cuanto más grande es un árbol de decisión, más complejo se vuelve. Es por eso que en ocasiones es mejor realizar una poda de las ramas que poseen características menos importantes. Es

decir, que entre más chico sea el árbol, más facilidad para alcanzar nodos de hojas puras, es decir, puntos de datos en una sola clase hay.

Existen dos métodos para elegir el mejor atributo en cada nodo, la ganancia de información y la impureza de Gini. Estos actúan como criterio de división para modelos de árbol de decisión y ayudan a evaluar la calidad de cada condición de prueba y lo bien que podrá clasificar las muestras en una clase.

Para la ganancia de la información, la entropía es un concepto que mide la impureza de los valores de la muestra, definida con la fórmula:

$$\text{Entropy}(S) = -\sum_{c \in C} p(c)\log_2 p(c)$$

donde S representa un conjunto de datos, c las clases en conjunto S y p(c) la proporción de puntos de datos que pertenecen a la clase c respecto del número total de puntos de datos del conjunto S. Sus valores pueden estar entre 0 y 1, si todas las muestras del conjunto S pertenecen a una clase, la entropía será igual a cero. Si las muestras se dividen en 2 clases, la entropía será 1. Se debe utilizar el atributo con la menor entropía posible para seleccionar la mejor característica.

La ganancia de información representa la diferencia en la entropía antes y después de una división en un atributo. Cuanto mayor sea la ganancia, mejor división habrá.

En el caso de la impureza de Gini, es la probabilidad de clasificar incorrectamente un punto de datos aleatorio del conjunto de datos si se etiquetara en función de la distribución de clases del conjunto de datos. Se define con la siguiente fórmula:

$$\text{Gini Impurity} = 1 - \sum_i (p_i)^2$$

Tipos de Árboles de decisión:

ID3 (Iterative Dichotomiser 3): Aprovecha la entropía y la ganancia de información como métricas para evaluar las divisiones de los candidatos.

C4.5: Considerado una iteración posterior de ID3, utiliza la ganancia de información o los ratios de ganancia para evaluar los puntos de división dentro de los árboles de decisión.

CART (Árboles de clasificación y regresión): Utiliza la impureza de Gini (mide la frecuencia con la que se clasifica erróneamente un atributo elegido al azar) para identificar el atributo ideal para dividir.

Algunos ejemplos de sus usos son:

Diagnóstico médico con ID3, C4.5: Se quiere predecir si un paciente tiene diabetes en función de sus síntomas y pruebas médicas (edad, IMC, glucosa, presión). Se selecciona el mejor atributo en cada paso utilizando la ganancia de información y se generan reglas como:

Nivel de glucosa > 140 → hay que seguir evaluando.

IMC > 30 → es probable que tenga diabetes.

Clasificación de imágenes con CART: Se quiere identificar si una imagen contiene un rostro humano o no. Se usan datos como la simetría facial, distancia entre ojos, color de piel y la relación ancho/alto del rostro. CART crea un árbol binario que clasifica imágenes en "rostro" o "no rostro" y se utilizan divisiones como:

Imagen simétrica → Seguir evaluando

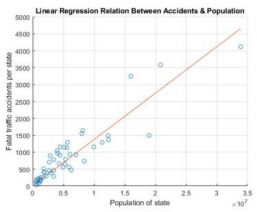
Existen características de ojos, naríz y boca → clasificar como rostro

Regresión Lineal (Predicción Numérica)

La regresión lineal es una técnica estadística ampliamente utilizada que permite modelar y analizar la relación entre una variable dependiente y una o más variables independientes. En su forma más simple, asume una relación lineal entre estas variables y busca encontrar la mejor línea recta que se ajuste a los datos observados. Sin embargo, su aplicabilidad va más allá de la simple relación lineal; es una herramienta poderosa en el análisis de datos y se aplica en diversos campos, desde la economía y las ciencias sociales hasta la medicina y la ingeniería. En la economía, se utiliza para estudiar la relación entre variables económicas, como el crecimiento del PIB y el consumo de energía. En la medicina, se puede utilizar para predecir el riesgo de enfermedades en función de factores de riesgo conocidos. En el marketing, la regresión lineal puede ayudar a comprender cómo las variables de precio, promoción y distribución afectan las ventas de un producto.

La regresión lineal permite identificar y cuantificar la relación entre variables, lo que resulta fundamental para la toma de decisiones informadas en una amplia variedad de situaciones. Al comprender cómo se relacionan las variables, podemos predecir y estimar valores futuros, evaluar el impacto de los cambios en las variables independientes y analizar la significancia de las relaciones observadas.

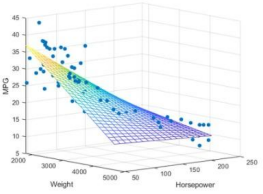
$$Y = \beta_0 + \beta_1 X + \epsilon$$



Regresión lineal simple:

Ejemplo de regresión lineal simple que muestra cómo predecir el número de accidentes de tráfico mortales en un estado (variable de respuesta ***Y***) en comparación con la población del estado (variable predictora ***X***).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$



Regresión lineal múltiple:

Ejemplo de regresión lineal múltiple, que predice las millas por galón (MPG) de diferentes coches (variable de respuesta ***Y***) en función del peso y la potencia (variables predictivas ***X***).

En cuanto a la parte práctica de este trabajo, se podría utilizar el modelo de regresión lineal múltiple para resolver la opción que nos compete (Opción 6), ya que tenemos una variable dependiente (“canal adecuado para notificaciones”) y dos variables independientes (“edad” y “antigüedad”).

Regresión lineal vs. regresión logística

Al igual que la regresión lineal, la regresión logística también se utiliza para estimar la relación entre una variable dependiente y una o más variables independientes, pero se utiliza para hacer una predicción sobre una variable categórica frente a una continua. La unidad de medida también difiere de la regresión lineal, ya que produce una probabilidad, pero la función logit transforma la curva S en línea recta.
Si bien ambos modelos se utilizan en el análisis de regresión para hacer predicciones sobre resultados futuros, la regresión lineal suele ser más fácil de entender. La regresión lineal tampoco requiere un tamaño de muestra tan grande como la regresión logística, que necesita una muestra adecuada para representar los valores en todas las categorías de respuesta.

Isolation Forest(deteccion de anomalias)

El Isolation Forest es un algoritmo no supervisado diseñado para detectar anomalías en datos, especialmente útil en conjuntos de alta dimensión. Su enfoque se basa en la idea de que las anomalías son valores atípicos y, por lo tanto, más fáciles de aislar que los datos normales.

¿Qué es una anomalía?

Las anomalías (o valores atípicos) son datos que se desvían significativamente del comportamiento esperado. Detectarlas es clave en áreas como seguridad, análisis de fraudes y mantenimiento predictivo.

¿Cómo funciona Isolation Forest?

A diferencia de otros métodos que modelan datos normales, este algoritmo aísla anomalías construyendo múltiples árboles de decisión de forma aleatoria, estos son los pasos de su funcionamiento:

1. **Construcción de isolation forests**

El algoritmo Isolation Forest crea múltiples Isolation Trees, que son árboles diseñados específicamente para aislar puntos de datos en lugar de clasificarlos.

A diferencia de los árboles de decisión tradicionales, que buscan categorizar datos, los Isolation Trees dividen los datos repetidamente de forma aleatoria, seleccionando características y valores de división al azar. Este proceso continúa hasta que cada punto de datos queda aislado. Como las anomalías suelen diferir del resto, tienden a ser aisladas en menos pasos, lo que facilita su detección.

2. **División de características aleatorias**

Los Isolation Trees aplican un enfoque aleatorio en cada nodo: primero seleccionan una característica al azar del conjunto de datos y luego eligen un valor de división dentro de su rango de valores. Esta aleatoriedad evita que las anomalías queden ocultas en ciertas ramas, ya que, al ser diferentes de la mayoría de los datos, tienden a aislarse más rápido en el proceso.

3. Aislamiento de puntos de datos

Los puntos de datos recorren las ramas del Isolation Tree según sus características. En cada división, si su valor es menor que el umbral elegido, siguen la rama izquierda; si es mayor, van a la derecha. Este proceso se repite hasta que el punto queda aislado en un nodo hoja, lo que indica que ha sido completamente separado del resto del conjunto de datos.

4. Puntuación de anomalía

El principio clave de los Isolation Trees es la longitud del camino que sigue un punto de datos dentro del árbol.

- Las anomalías se aíslan más rápido, ya que suelen estar fuera del rango común de valores y requieren menos divisiones para llegar a un nodo hoja.
- Los puntos normales, en cambio, comparten características con otros datos y necesitan más divisiones antes de ser completamente aislados.

5. Cálculo del puntaje de anomalía

- Cada punto de datos es evaluado en todos los Isolation Trees del bosque
- Se registra la cantidad de divisiones necesarias para aislarlo en cada árbol.
- Se calcula un puntaje de anomalía promediando estas longitudes en todos los árboles.

6. Identificacion de anomalias

Los puntos de datos con trayectorias promedio más cortas se consideran anomalías, ya que fueron aislados con menos divisiones, lo que indica que se desvían del resto.

Isolation forest en Python

Dos de las principales implementaciones de **Isolation Forest** están disponibles en **Scikit Learn** y en **H2O**. Si bien las dos están muy optimizadas, existen pequeñas diferencias a la hora de utilizarlas.

- En la implementación de **Scikit Learn**, al entrenar el modelo, se tiene que especificar el porcentaje de anomalías que se espera en los datos de entrenamiento (contamination). Con este valor, el modelo aprende el valor a partir del cual una observación se considera anomalía. Al aplicar el método **predict()** se obtiene -1 si es anomalía (outlier) o 1 si es un dato normal (inliers). Para recuperar la métrica de anomalía en lugar de la clasificación, hay que emplear el método **score_samples()**. Este último devuelve el valor negativo de la distancia de aislamiento.
- En la implementación de **H2O**, el modelo sí devuelve la distancia de aislamiento como resultado del método **predict()**. Para determinar si una observación es una anomalía o no, es necesario identificar el valor límite a partir de los cuantiles de las distancias predichas para las observaciones de entrenamiento.

Casos de uso y aplicaciones de Isolation forest

Ciberseguridad

- Detección de intrusiones
- Detección de fraudes

Finanzas

- Fraudes con tarjetas de crédito
- Gestión de riesgos

Cuidado de la salud

- Detección de enfermedades raras
- Monitoreo de pacientes

Telecomunicaciones

- Supervisión del rendimiento de la red
- Predicción de abandono

Ventajas del Isolation Forest

Eficiencia

- Complejidad temporal lineal
- Requisitos bajos de memoria

Simplicidad e interpretabilidad

- Fácil de entender
- Puntuaciones de anomalías claras

Paralelización y velocidad

- Procesamiento paralelo
- Ejecución rápida

Versatilidad

- Amplia gama de aplicaciones
- Adaptabilidad

Limitaciones de los Isolation Forest

La alta dimensionalidad, el desequilibrio de datos, la sensibilidad de los parámetros y un preprocesamiento y una selección de características minuciosos son consideraciones esenciales. Además, si bien el algoritmo es eficiente, los grandes conjuntos de datos y los requisitos de procesamiento en tiempo real aún pueden plantear desafíos de rendimiento. Asimismo, la

interpretabilidad de los resultados y el manejo de tipos de datos mixtos o categóricos requieren estrategias bien pensadas para garantizar la detección precisa de anomalías.

K-Means (agrupamiento)

K-Means es un algoritmo de clustering (agrupamiento) no supervisado utilizado para dividir un conjunto de datos en grupos (clusters) basados en similitudes. El objetivo es minimizar la varianza interna agrupando puntos de datos en k clusters, donde cada punto pertenece al clúster con la media (centroide) más cercana.

Funcionamiento de K-Means

1. **Inicialización de centroides**
    - Se eligen aleatoriamente **K puntos** del conjunto de datos como **centroides iniciales** (uno por cada cluster).
    - Alternativamente, se pueden usar métodos más sofisticados como **K-Means++** para una inicialización más eficiente.
2. **Asignación de puntos a clusters**
    - Cada punto del dataset se asigna al **centroide más cercano**, usando una medida de distancia (generalmente **distancia euclidiana**).
    - Matemáticamente, se asigna cada punto.
3. **Actualización de centroides**
    - Se recalcula la posición de cada centroide como el **promedio (media)** de todos los puntos asignados a su clúster.

4. **Iteración hasta convergencia**
    - Se repiten los pasos 2 y 3 hasta que:
        - Los centroides ya no cambian significativamente.
        - Se alcanza un número máximo de iteraciones.
        - La función de costo (inercia) ya no mejora.

Aplicación de K-Means

Segmentación de Clientes (Marketing Analytics): Agrupa clientes automáticamente según similitudes (ej. gasto, frecuencia), sin necesidad de etiquetas previas. Su ventaja es que es rápido y escalable para bases de datos grandes, permitiendo identificar patrones ocultos (ej. "clientes leales" vs. "en riesgo").

Compresión de imágenes: Reduce una paleta de millones de colores a pocos colores representativos (centroides), manteniendo la esencia visual. Como ventaja minimizando la pérdida de calidad al preservar los colores más frecuentes (ej. útil para formatos como GIF o imágenes indexadas).

Análisis biológicos (genes/células): Agrupa genes o células con patrones de expresión similares, revelando subtipos (ej. células cancerosas vs. sanas). Funciona bien con datos de alta dimensión (ej. miles de genes) tras reducir dimensionalidad (con PCA).

Ventajas

- Simple y rápido para datasets no muy grandes.
- Eficiente computacionalmente (complejidad lineal en el número de puntos).
- Funciona bien con clusters esféricos y de tamaño similar.

Desventajas

- Sensible a la inicialización aleatoria (puede caer en mínimos locales).
- Requiere especificar K (número de clusters) de antemano.
- No funciona bien con clusters no esféricos o de densidad variable.
- No maneja outliers de manera robusta.

Conclusiones:

El uso de Scikit-learn en Python facilita la implementación de técnicas de Machine Learning, resolviendo problemas de clasificación, regresión, agrupamiento y detección de anomalías de manera eficiente. A lo largo de esta investigación, exploramos conceptos como la Regresión Logística, Árboles de Decisión, Regresión Lineal, K-means e Isolation Forest, analizando su funcionamiento y aplicaciones.
Estas herramientas son fundamentales en el análisis de datos y la toma de decisiones automatizada, demostrando su importancia desde la predicción de valores hasta la detección de fraudes.

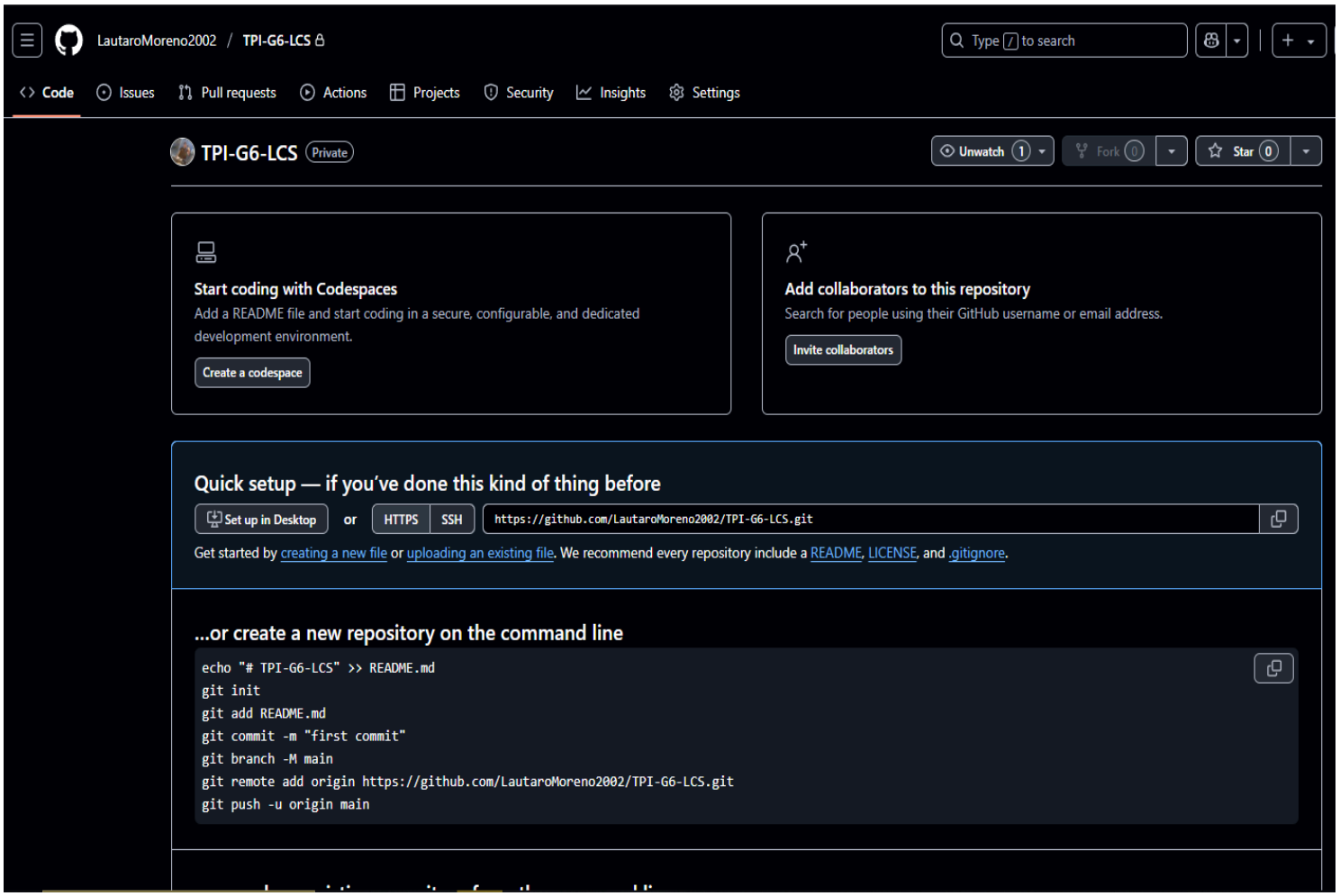
SOBRE NUESTRO REPOSITORIO:

Dejamos evidencia de la creación de nuestro repositorio en Github donde alojaremos el código de nuestro proyecto, propio de la parte práctica del Trabajo Práctico Inicial.

El link del repositorio es el siguiente:

https://github.com/LautaroMoreno2002/TPI-G6-LCS

Adicionalmente, dejamos adjunta una captura de pantalla donde se muestra el repositorio creado.



FUENTES:

INVESTIGACIÓN TEÓRICA (Parte 1):

https://www.oracle.com/ar/erp/what-is-erp/#:~:text=frecuentes%20sobre%20ERP-,Definici%C3%B3n%20de%20planificaci%C3%B3n%20de%20recursos%20empresariales%20(ERP),de%20la%20cadena%20de%20suministro

https://www.wolterskluwer.com/es-es/expert-insights/sistemas-erp-recursos-humanos-claves#:~:text=de%20Recursos%20Humanos?-,%C2%BFQu%C3%A9%20es%20ERP%20en%20Recursos%20Humanos?,del%20ERP%20de%20Recursos%20Humanos

https://www.salesforce.com/mx/crm/#que-hace-un-crm-scroll-tab

https://www.sesamehr.co/blog/gestion-de-equipos/crm-recursos-humanos/

https://cloud.google.com/learn/what-is-artificial-intelligence?hl=es-419

https://factorial.es/blog/inteligencia-artificial-rrhh/

https://betterfly.com/blohttps://glocalthinking.com/ia-rrhh/

https://glocalthinking.com/ia-rrhh/

https://www.seidor.com/es-ar/blog/ventajas-crm

https://www.qad.com/es-MX/blog.mx/-/blogs/ventajas-y-desventajas-de-un-erp

INVESTIGACIÓN TEÓRICA (Parte 2):

https://www.tokioschool.com/noticias/que-es-scikit-learn/

https://datascientest.com/es/scikit-learn-decubre-la-biblioteca-python

https://openwebinars.net/blog/como-entrenar-un-modelo-de-machine-learning-con-scikit-learn/

https://aws.amazon.com/es/what-is/logistic-regression/#:~:text=La%20regresi%C3%B3n%20log%C3%ADstica%20es%20un,que%20se%20muestra%20a%20continuaci%C3%B3n

https://www.ibm.com/docs/es/spss-statistics/saas?topic=regression-logistic

https://www.ibm.com/es-es/topics/logistic-regression

https://datascientest.com/es/que-es-la-regresion-logistica

https://www.ibm.com/es-es/think/topics/decision-trees

https://programas.uniandes.edu.co/blog/regresion-lineal

https://la.mathworks.com/discovery/linear-regression.html

https://cienciadedatos.net/documentos/py22-deteccion-anomalias-isolation-forest-python

https://www.geeksforgeeks.org/anomaly-detection-using-isolation-forest/

https://spotintelligence.com/2024/05/21/isolation-forest/#Use_Cases_and_Applications_of_Isolation_Forest

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

https://datarundown.com/k-means-clustering-pros-cons/