FACULTAD DE CIENCIAS PURAS Y NATURALES

CARRERA DE INFORMÁTICA

PROYECTO INTELIGENCIA ARTIFICIAL

TITULO	**Capstone Project-IBM Employee Attrition Prediction**
AUTOR	**Nombre y Apellido**
	Churqui Mamani Angel
FECHA	20/05/2023

Carrera	Informática Asignatura
Asignatura	INF - 354
Docente	Lic. Moises Silva

PROYECTO DE INTELIGENCIA ARTIFICIAL

Tópico: Cualquier temática del área médica afrontada por IA

Dataset: No es restrictiva. Pero lo mínimo tratar es una tabla de al menos por 1500 filas

Proceso:

- Descripción clara del objetivo de investigación a partir del dataset elegido
- Descripción detallada de los campos del dataset
- Proceso básico de análisis de datos:
  - Preprocesamiento (al menos una valida, otros dos por ver los resultados, si no se aplica justifique porque), Balanceo de datos
  - Selección del clasificador (acorde a los datos supervisado, no supervisado). El clasificador puede, pero no necesariamente depender del preprocesamiento.
    Justificar el clasificador (máximo 2 planas con fuente ISBN, DOI)
  - Primera ejecución: Confiabilidad, matriz de confusión
  - Splits: al menos 100 asignaciones, la mediana de la confiabilidad
    Académico (primera ejecucion) 80(train)/20(test) – Investigación 50/50 (segunda ejecución)
  - Primer Código: Github, kaagle, codelab
  - Aplicar Componentes principales (PCA), determinar la cantidad óptima para mejorar o llegar al resultado anterior. Al menos unas 5 ejecuciones. (12 columnas, 10, 11, 9, 5, 3)
    Explicar cómo funciona PCA (algebra lineal)
  - Escribir un artículo de al menos 4 paginas
  - Segundo Código: Github, kaagle, codelab
  - Artículo que resuma lo realizado (min 3 hojas)

Otras: Relacionadas a clusters, algoritmos genéticos, agentes inteligentes, otras.

Fecha presentación: 19 de junio – 7:00 am. **Revisiones**

2T o recuperatorio: 20 de junio

1. **Topico**

El tema del dataset para el Proyecto Capstone "IBM Employee Attrition Prediction" se centra en predecir la atrición de empleados en una empresa utilizando datos proporcionados por IBM.

2. **Objetivo del dataset elegido**

El objetivo general del dataset es desarrollar un modelo predictivo de atrición de empleados utilizando el dataset "IBM Employee Attrition Prediction" proporcionado por IBM, con el fin de identificar a los empleados con mayor riesgo de abandonar la empresa y tomar medidas preventivas para retenerlos.

El siguiente dataset se puede usar para los siguientes objetivos específicos:

- Realizar un análisis exploratorio de datos para comprender las características del dataset y identificar patrones, correlaciones y tendencias relevantes relacionadas con la atrición de empleados.
- Preprocesar los datos, realizando la limpieza de datos, manejando los valores faltantes y transformando las variables categóricas en numéricas si es necesario.
- Seleccionar las características más relevantes para la predicción de la atrición de empleados, descartando aquellas que no aporten valor al modelo.
- Entrenar y evaluar diferentes modelos de aprendizaje automático, como regresión logística, árboles de decisión, bosques aleatorios o redes neuronales, para predecir la atrición de empleados.
- Optimizar el modelo seleccionado ajustando los hiperparámetros y aplicando técnicas de validación cruzada para mejorar su rendimiento.

3. **Dataset**

3.1 Campos del Dataset

Dentro de nuestro Dataset “IBM Employee Attrition Prediction” tenemos los siguientes campos o variables:

NOMBRE DE LA COLUMNA	**DESCRIPCION**
Age	Edad del empleado.
Attrition	Atrición del empleado. Indica si el empleado ha dejado la empresa o no. Puede tener valores como "Yes" (Sí) o "No" (No).
BusinessTravel	Frecuencia de viajes de negocios del empleado.
DailyRate	Tasa diaria de salario del empleado.

Department	Departamento en el que trabaja el empleado.
DistanceFromHome	Distancia desde el hogar al lugar de trabajo del empleado.
Education	Nivel educativo del empleado.
EducationField	Campo educativo o área de estudio del empleado.
EmployeeCount	Contador de empleados. Puede indicar el número de empleados en un grupo o unidad específica.
EmployeeNumber	Número de empleado. Un identificador único para cada empleado en la empresa.
ApplicationID	Identificador de solicitud o aplicación. Puede estar relacionado con el proceso de contratación o registro del empleado.
EnvironmentSatisfaction	Nivel de satisfacción del empleado con el entorno laboral.
Gender	Género del empleado.
HourlyRate	Tasa de salario por hora del empleado.
JobInvolvement	Nivel de involucramiento del empleado en su trabajo.
JobLevel	Nivel de puesto del empleado.
JobRole	Rol o puesto de trabajo del empleado en la empresa.
JobSatisfaction	Nivel de satisfacción laboral del empleado.
MaritalStatus	Estado civil del empleado.
MonthlyIncome	Ingreso mensual del empleado en términos de salario.
MonthlyRate	Tasa salarial mensual del empleado.
NumCompaniesWorked	Número de compañías en las que el empleado ha trabajado anteriormente.
Over18	Indicador de si el empleado es mayor de 18 años.
OverTime	Indicador de si el empleado trabaja horas extras.
PercentSalaryHike	Porcentaje de aumento salarial para el empleado.
PerformanceRating	Calificación de desempeño del empleado.
RelationshipSatisfaction	Nivel de satisfacción del empleado con las relaciones en el entorno laboral.
StandardHours	Horas estándar de trabajo.
StockOptionLevel	Nivel de opciones de compra de acciones del empleado.
TotalWorkingYears	Total de años de experiencia laboral del empleado.
TrainingTimesLastYear	Número de veces que el empleado recibió capacitación el año pasado.
WorkLifeBalance	Equilibrio entre el trabajo y la vida personal del empleado.
YearsAtCompany	Número de años que el empleado ha estado trabajando en la empresa.
YearsInCurrentRole	Número de años que el empleado ha estado en su puesto actual.
YearsSinceLastPromotion	Número de años desde la última promoción del empleado.

YearsWithCurrManager	Número de años que el empleado ha estado trabajando con su actual supervisor o gerente.
Employee Source	Fuente de reclutamiento o contratación del empleado.

Estos campos proporcionan información relevante sobre cada videojuego, como su nombre, plataforma, género, ventas en diferentes regiones, puntuaciones de críticos y usuarios, así como la clasificación por edad. Estos datos pueden ser utilizados para realizar análisis y exploraciones relacionadas con las ventas y la recepción de los videojuegos a lo largo del tiempo.

4. Retroalimentación

Problema empresarial

"La deserción en recursos humanos se refiere a la pérdida gradual de empleados a lo largo del tiempo. En general, la deserción relativamente alta es problemática para las empresas. Los profesionales de recursos humanos a menudo asumen un papel de liderazgo en el diseño de los programas de compensación de la empresa, la cultura laboral y los sistemas de motivación que ayudan a la organización a retener a los mejores empleados".

Nuestro papel es descubrir los factores que conducen a la deserción de los empleados a través del análisis exploratorio de datos y explorarlos mediante el uso de varios modelos de clasificación para predecir si es probable que un empleado renuncie. Esto podría aumentar en gran medida la capacidad de recursos humanos para intervenir a tiempo y remediar la situación para evitar el desgaste.

Si bien este modelo se puede ejecutar de forma rutinaria para identificar a los empleados que tienen más probabilidades de renunciar, el factor clave del éxito sería el elemento humano de comunicarse con el empleado, comprender la situación actual del empleado y tomar medidas para remediar los factores controlables que pueden evitar la deserción del empleado.

Análisis de RRHH

El análisis de recursos humanos (análisis de recursos humanos) es un área en el campo del análisis que se refiere a la aplicación de procesos analíticos al departamento de recursos humanos de una organización con la esperanza de mejorar el desempeño de los empleados y, por lo tanto, obtener un mejor retorno de la inversión. El análisis de recursos humanos no solo se ocupa de recopilar datos sobre la eficiencia de los empleados. En cambio, su objetivo es proporcionar información sobre cada proceso mediante la recopilación de datos y luego usarlos para tomar decisiones relevantes sobre cómo mejorar estos procesos.

Conjunto de datos

Este es un conjunto de datos hipotético creado por científicos de datos de IBM. El conjunto de datos tiene (23436R X 37C) que contiene tipos de datos numéricos y categóricos que

describen los antecedentes y las características de cada empleado; y etiquetados (aprendizaje supervisado) con si todavía están en la empresa o si se han ido a trabajar a otro lugar. Los modelos de Machine Learning pueden ayudar a comprender y determinar cómo estos factores se relacionan con el desgaste de la fuerza laboral.

Variables tontas

Una variable "tonta" o "dummy" es un término utilizado en programación para referirse a una variable que se declara, pero no se utiliza realmente en el código. Estas variables suelen ser creadas para cumplir con ciertos requisitos sintácticos o estructurales del lenguaje de programación utilizado. El propósito principal de una variable tonta es actuar como marcador o marcador de posición para mantener la integridad del código sin afectar su funcionalidad.

Regresiones logísticas

La regresión logística es un tipo de modelo de regresión utilizado para predecir variables categóricas binarias o multinomiales. A diferencia de la regresión lineal, que se utiliza para predecir variables continuas, la regresión logística se utiliza para modelar la relación entre un conjunto de variables independientes y una variable dependiente categórica.

La regresión logística utiliza una función logística, también conocida como función sigmoide, para transformar la salida en un rango entre 0 y 1, lo que se interpreta como la probabilidad de que una observación pertenezca a una determinada clase.