PRUEBA TÉCNICA PARA CIENTIFICO DE DATOS

El propósito de esta prueba es medir sus capacidades de programación en Python, modelado, preparación y análisis de datos, dominio de herramientas de visualización y estadística descriptiva.

Se adjunta una base (archivos CSV con codificación UTF-8) para desarrollar la prueba técnica. Este contiene las alertas evaluadas en los últimos 7 años con su decisión.

El objetivo de la prueba es implementar un modelo de clasificación calendarizable, es decir desarrollado en el Orquestador 2.0 cumplimiento los lineamientos de Bancolombia, que determine la decisión de cierre las alertas. Adicional, realizar un análisis descriptivo de esta información Power BI, donde se diseñe un modelo de datos tipo estrella o copo de nieve, y otro tablero en Power BI donde se puedan observar los resultados del modelo de clasificación y métricas e indicadores que me permitan determinar la calidad del mismo. Al final del ejercicio, nos debe entregar en un repositorio Git:

A. El código Python del modelo

B. El proyecto calendarizable

C. Los dos proyectos PBI

D. Un archivo de texto (.pdf) que contenga una descripción del proceso de desarrollo de la prueba y las conclusiones del ejercicio.

Debe preparar una exposición, de máximo 20 minutos, donde explique el proceso y los resultados.

Le agradecemos su participación en este ejercicio y le deseamos muchos éxitos en el desarrollo de este.

1.1. Desarrollo

1.1.1. Creación del código en Python

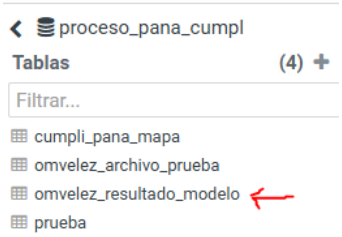
El desarrollo se crea a través de la aplicación Anaconda Navigator sobre Jupyterlab, donde se escribe en código de Python, en ella se generan dos archivos .py.

- Prueba_Tecnica

  El archivo comprende el proceso de ejecución del modelo de clasificación "RandomForest" y de un análisis exploratorio de variables la cual comprende unos gráficos de las variables categóricas, booleanas y predictiva.

- Prueba_Tecnica_Carga

  El archivo contiene el código con el cual se sube la información base del desarrollo como lo es el archivo "alertas 1.csv" y el archivo resultado del modelo creado denominado "Resultado_Prueba_Tecnica" a la Landing Zone, en la zona "proceso_pana_cumpli"



.

1.1.2. Creación de los Power Bi

Se crean dos Power Bi en el desarrollo denominados así:

- ANALISIS_DESCRIPTIVO

  En Panel de control concentra lo correspondiente a un análisis descriptivo de la tabla insumo del desarrollo "alertas 1.xlsx" que fue cargada a la LZ y leído a través de una conexión ODBC Impala, en el Panel de control está compuesto por cuatro pestañas (Menu-Cantidad-Monto-Analisis Variables) el panel de control en su menú se ve así:

- RESULTADO_MODELO

  En el Panel de control encontramos el resultado del modelo en los datos de prueba correspondiente al 30% de la data total, teniendo indicadores como lo son la efectividad del modelo y la matriz de confusión, así mismo un conteo de fallos y de aciertos, también se identifican filtros que permiten relacionar el resultado del modelo.

1.2. Resultado

1.2.1. El resultado del modelo es cargado en un repositorio GITHUP, adicional se crea un documento de desarrollo en formato PDF para dar entendimiento del proceso que llevo.

1.3. Conclusiones

1.3.1. El tiempo de creación del modelo es limitado por consiguiente no se alcanza a realizar mayor volumen de pruebas con diferentes modelos permitiendo ajustar un poco mas la eficiencia del modelo que en sus datos de prueba non arroja un valor de 93.17%.

El resultado del modelo se identifica eficiente en el pronóstico de la categoría Alerta Normalizada, sin embargo, para pronosticar las categorías de Operación Inusual Normalizada y Operación Sospechosa, por consiguiente, se hace necesario agregar un volumen superior de registros con estos valores para equilibrar el resultado del modelo.

Es necesario verificar si el modelo puede ser ajustado agregando otras variables descriptivas del cliente como lo son los estados financieros, variables que pueden ayudar en el diagnostico.

La herramienta Power Bi es de gran ayuda al momento de identificar el resultado del modelo, en la cual se haya que si los clientes tienen Prensa Negativa el pronostico del mismo se eleva en las categorías de Operación Inusual Normalizada y Operación Sospechosa.