Proyecto IA_Seguros_Prudential

1. **Descripción y objetivos del proyecto.**
2. **Preparación de los datos.**
3. **Análisis exploratorio de datos.**
4. **Preprocesamiento de datos y modelado predictivo.**
5. **Conclusión.**

1. Descripción y objetivos del proyecto.

Proyecto IA_Seguros_Prudential

Link del proyecto
https://www.kaggle.com/c/prudential-life-insurance-assessment/data

Este conjunto de datos contiene información sobre solicitantes de seguros para la compañía Prudential. Se busca crear un algoritmo que perfile solicitantes en una escala de 8 niveles.

El conjunto de datos proporcionado contiene variables que describen los atributos de los solicitantes de seguros de vida. La tarea consiste en predecir la variable "Response" para cada ID en el conjunto de prueba. "Response" es una medida ordinal de riesgo que tiene 8 niveles.

2. Preparación de los datos.

En primer lugar, se importan las bibliotecas necesarias para trabajar con los datos, como pandas, numpy, StandardScaler, MinMaxScaler y train_test_split. También se utiliza la biblioteca de Google Colab para montar la unidad de Google Drive.

Luego se descargan los datos directamente del portal kaggle.com con el objetivo de que cualquier persona que ejecute el notebook pueda tener acceso a ellos sin necesidad de contar con ellos de manera local o de adjuntarlos al cuaderno de Google Colab. Esto se logra en conjunto con la librería “files” y “drive” de google.colab para montar la unidad de Google Drive y las librerías kaggle y jovian para traer los datos desde el sitio de Kaggle. Inicialmente, el conjunto de datos es descargado en un archivo comprimido .zip en el almacenamiento asociado al entorno de ejecución del cuaderno de Google Colab y después se descomprime el

archivo en los diversos archivos contenidos en el mismo almacenamiento con ayuda de la ejecución de líneas de código que usan las librerías “os” y “zipfile”. Los datos ya están separados en dos conjuntos de entrenamiento y prueba. Los archivos se almacenan en el entorno de ejecución de Google Colab, siempre y cuando se haya autorizado el acceso de la cuenta de Google a este.

Finalmente, se descomprimen los archivos .zip y se guardan en el entorno de ejecución. Se utiliza el bucle for para asegurarse de que solo se extraigan los archivos .zip y se ignoren otros archivos presentes en el entorno de ejecución.

3. Análisis exploratorio de datos.

En el EDA (Análisis Exploratorio de Datos), se llevará a cabo una exploración de los datos a disposición para entender mejor el contenido del conjunto de datos, los patrones, relaciones y características de estos, y detectar posibles problemas y

oportunidades para mejorar el modelo de IA.

En primer lugar, se realizó un análisis exploratorio de datos para comprender mejor las características del conjunto de datos. Se observó que el conjunto de datos tiene un tamaño de:

Cantidad de filas en el archivo de entrenamiento:59381.

Cantidad de columnas en el archivo de entrenamiento:128.

En este conjunto de datos, se le proporcionan más de cien variables que describen los atributos de los solicitantes de seguros de vida. La tarea es predecir la variable "Respuesta" para cada Id en el conjunto de prueba. "Respuesta" es una medida ordinal de riesgo que tiene 8 niveles.

Variable	**Descripción**
Id	Un identificador único asociado a una solicitud.
Product_Info_1-7	Un conjunto de variables normalizadas relacionadas con el producto solicitado.
Ins_Age	Edad normalizada del solicitante.
Ht	Altura normalizada del solicitante.

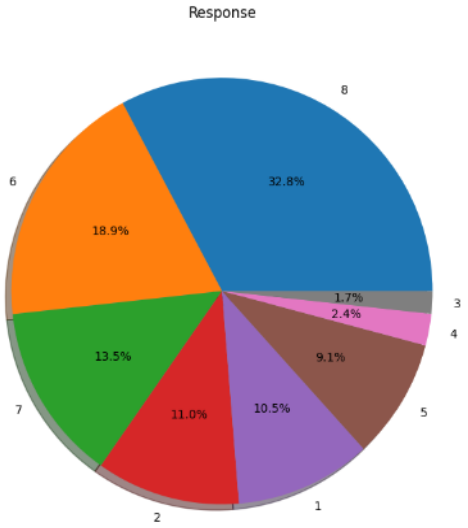
Wt	Peso normalizado del solicitante.
BMI	Índice de masa corporal normalizado del solicitante.
Employment_Info_1-6	Un conjunto de variables normalizadas relacionadas con la historia laboral del solicitante.
InsuredInfo_1-6	Un conjunto de variables normalizadas que proporcionan información sobre el solicitante.
Insurance_History_1-9	Un conjunto de variables normalizadas relacionadas con la historia de seguros del solicitante.
Family_Hist_1-5	Un conjunto de variables normalizadas relacionadas con la historia familiar del solicitante.
Medical_History_1-41	Un conjunto de variables normalizadas relacionadas con la historia médica del solicitante.
Medical_Keyword_1-48	Un conjunto de variables ficticias relacionadas con la presencia o ausencia de una palabra clave médica asociada con la solicitud.
Response	Esta es la variable objetivo, una variable ordinal relacionada con la decisión final asociada con una solicitud.

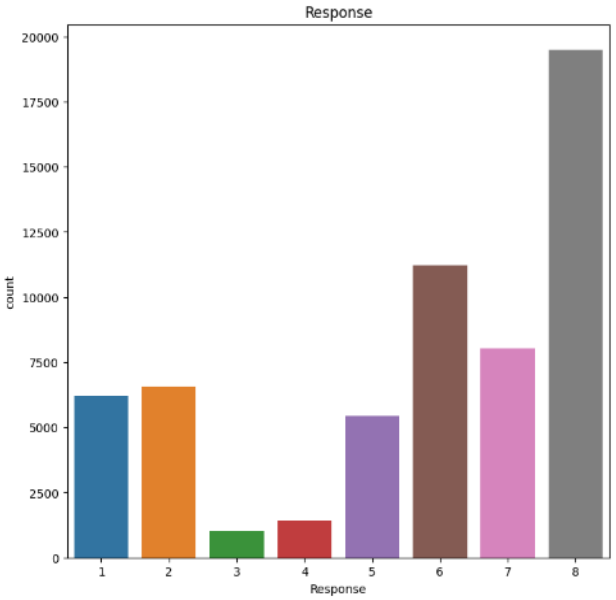
Distribución del conjunto de datos

El siguiente histograma muestra la distribución del conjunto de datos de los solicitantes por sus clasificaciones de riesgo determinadas (es decir, Response)

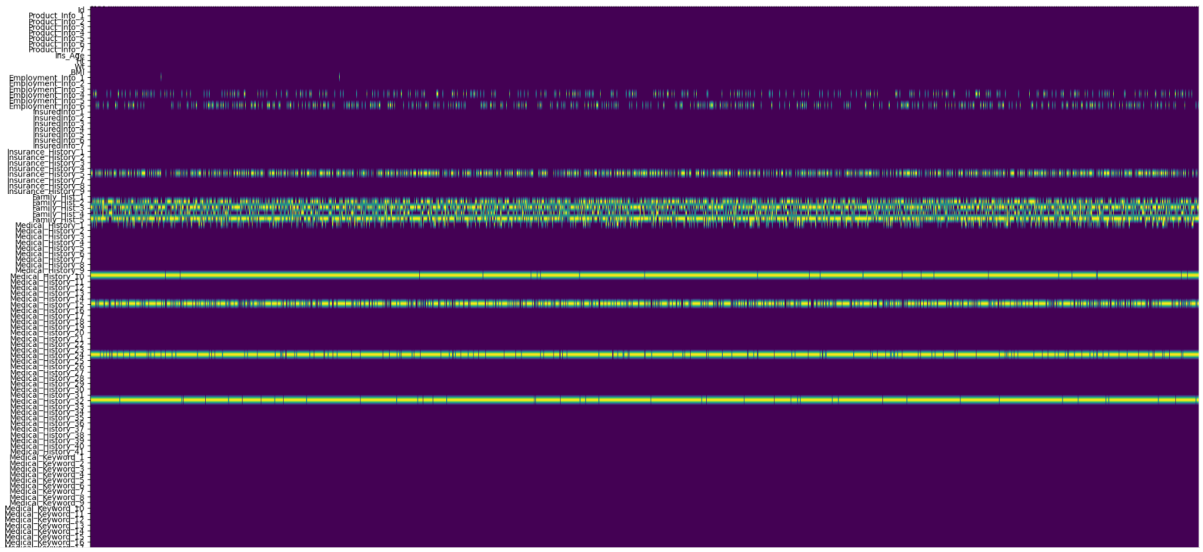
En la exploración del conjunto de datos se recurrió a un graficado de los datos faltantes usando Matplotlib y se obtuvo la gráfica a continuación. En ella se esquematizan en color púrpura lo datos existentes y en color verde-amarillo los datos faltantes. De la gráfica se puede concluir que no es necesario proceder a eliminar artificialmente datos puesto que el conjunto de datos ya no cuenta con varios de ellos y habrá que realizar el ejercicio de llenarlos dependiendo del tipo de dato, si se trata de variables categóricas se puede proceder llenando las zonas

vacías con la moda de los datos y si se trata de variables numéricas se puede proceder a llenar con la mediana o la media de los datos.





Así mismo se procedió a generar una matriz de correlación entre las columnas, recordando que la matriz de correlación de los datos en la siguiente tabla que muestra la relación entre cada par de variables en el conjunto de datos. De ella se observa que hay poca correlación entre ellas, pero sí se observa que hay cierto nivel de correlación entre las variables asociadas a las historias de aseguramientos previos, lo cual tiene sentido, puesto que las historias de usos de pólizas de seguro en su conjunto reflejan el estado de “asegurabilidad” de una persona, sea una persona sana o sea una persona con diversos cuadros de enfermedades.



4. Preprocesamiento de los datos y modelado preliminar

Se aplicaron técnicas de preprocesamiento de datos para preparar el conjunto de datos para su uso en el modelado. Se llenaron los valores faltantes utilizando la moda para variables categóricas y la mediana para variables numéricas.

Además, se utilizó OneHotEncoder para codificar las variables categóricas y transformarlas en variables binarias. Esto es necesario para que el modelo de regresión logística pueda interpretar correctamente las variables categóricas.

Luego, se tomaron tanto el conjunto de datos de entrenamiento como el conjunto de validación. El conjunto de entrenamiento se utilizó para ajustar el modelo, mientras que el conjunto de validación se utilizó para medir la precisión del modelo.

Para el modelado predictivo, se ajustó un modelo de Regresión Logística utilizando el conjunto de datos de entrenamiento y se midió su precisión utilizando el conjunto de validación. Se encontró que el modelo tenía una precisión de alrededor del 33%, lo cual es bajo. Se sugiere aumentar el número máximo de iteraciones del modelo para mejorar su precisión.

5. Conclusión.

Se encontró que el modelo utilizado tiene una precisión relativamente baja, lo que sugiere que se necesitan más investigaciones y ajustes para mejorar el rendimiento de este e incluso cambiar de modelo. Entre las posibles mejoras se encuentran la exploración de diferentes técnicas de selección de características y la optimización de los hiperparámetros del modelo.