Laboratorio practico demostrando la funcionalidad correspondiente del Árbol de Decisión implementado Pruning

En base a la implementación se comprende la importación de múltiples librerías, mediante las cuales estas forman parte del lenguaje Python. El desglose de las ya mencionadas librerías corresponde denotarse por grupos mediante las aplican diferentes funcionalidades. Los grupos se desglosan en el grupo A conformado por Pandas, StringIO donde estos permiten realizar el control de Data Frames. El grupo B conformado por Skelarn, permite implementar el algoritmo correspondiente para el modelo. El grupo C se conforma por Seaborn, Matplotlib las cuales permiten generar el control de imágenes. Por otro lado dicho segmento de código se dispone acompañado mediante la implementación de DocStrings permitiendo explicar el funcionamiento de los mismos.

Avanzando con el procedimiento del laboratorio se considera el llamado y almacenamiento de los datos CSV, para lo cual este se debe encontrar especificado en el directorio similar al cuaderno de jupyter. Para almacenar el archivo CSV se implementa la asignación de una variable seguida del comando de read\_csv, el cual corresponde formar parte de la librería Pandas. Prosiguiendo se considera que el archivo csv no se encuentra limpio, por lo cual al disponer del conocimiento de los datos manejados se considera convertir dichos valores en blanco mediante la función strip. Por ultimo mediante la implementación de la variable se realiza la presentación de la misma, en busca del análisis de los datos.

Una vez se ha almacenado el archivo CSV dentro de la variable se corresponde realizar el mapeo de los datos. Por lo cual analizando los datos estos requieren ser transformados a valores numéricos, considerando que el modelo únicamente desempeña su papel con números. El referirse en transformar valores string a numéricos, se implementa mediante en base a la funcionalidad del método map(). Mediante el cual funciona obteniendo por parámetro los valores que se proceden a transformar. De esta forma se declaran los principales como son el genero y la clase de diabetes que pertenece.

Considerando que el DataFrame dispone únicamente de valores numéricos, este se procede a realizar el análisis del mismo permitiendo dividir en variables objetivas y subjetiva. Para lo cual se implementa un análisis en base a la predicción, de esta forma en base a las columnas. De tal manera para la variable mediante la cual se almacenan las características se toma únicamente los valores que puedan impactar al modelo. Para la variable en la cual se almacena la etiqueta se toma la columna mediante la cual permite clasificar las predicciones.

Con el DataFrame cargado en el cuaderno de Jupyter se considera realizar lo que corresponde la limpieza de los datos. Por lo cual se analiza la estructura de este DataSet, ya que se observa que existen valores en texto, por lo cual el modelo requiere se le envié únicamente valores numéricos. El procedimiento que se realiza corresponde en del dataFrame obtener cada columna y realizar un mapeo con la variable declaradas previamente con los valores numéricos.

Seguida la importación de las librerías se corresponde en leer el DataSet considerando que se ha agregado al cuaderno de jupyter la dirección de los datos. Este DataSet se considera realizar la lectura mediante la funcionalidad de la librería pandas la cual permite leer archivos csv, seguido del nombre del archivo. Como siguiente paso se corresponde en realizar la implementación para imprimir lo que corresponde la estructura del DataSet.

Se inicia el desarrollo del laboratorio correspondiente de árboles de decisión realizando la implementación de las librerías del lenguaje Python. Las cuales sirven tanto para el manejo de datos, cálculos matemáticos, lectura y escritura de DataFrames y el empleo de una gráfica para mostrar los resultados. Como librerías principales se disponen de pandas la cual permite el manejo de grandes archivos en formatos CSV, por lo cual se lo ha instanciado para esta manipulación. La segunda librería con importancia corresponde Sklearn el cual se implementa para obtener los diferentes métodos del árbol de decisión. La tercera librería corresponde Matplotlib la cual permite realizar graficas en 2D en base a los resultados obtenidos.

El requerir distribuir lo que corresponde los datos en variables para entrenamiento y test, se implementan mediante el método train\_test\_split(). El cual dicho método permite separar de toda la información según se le establezca, por lo cual en base a la practica realizada se distribuye en proporción al 70% para entrenamiento y el sobrante 30% para test. El realizar la distribución correspondiente de los datos, facilita la implementación del modelo considerando que su aprendizaje de etiquetado será mediante las variables de entrenamiento.

En este caso se realizará la implementación del modelo clasificador, para ello se implementa el llamado del mismo desde la librería Sklearn. Donde dicho método será implementado y almacenado en una variable, donde posterior permita realizar en base a dicha variable el llamado de los métodos para entrenar y predecir. Como se ha mencionado se implementara la variable del modelo para realizar el entrenamiento, de esta manera se llama el método fit(), donde este recibe por parámetros la asignación de las variables de entrenamiento. Para implementar lo que corresponde la predicción se corresponde llamar dicho método mediante la declaración predict(), la cual recibe por parámetro la variable X para test.