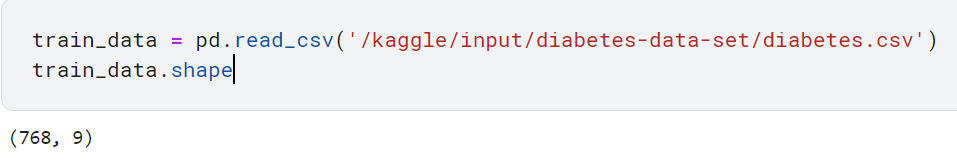
AI Based Diabetes Prediction System

* Introduction

Phase 3 outlines the process and analysis for a technology project that involves loading and preprocessing a diabetes dataset. The primary goals of this project are to perform data preprocessing, feature selection, and classification using machine learning techniques.

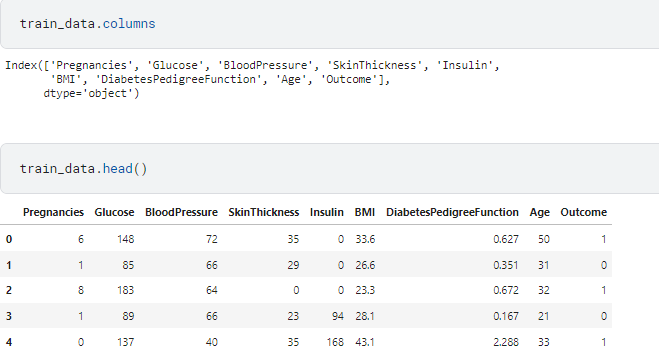
* Dataset Loading and Preprocessing

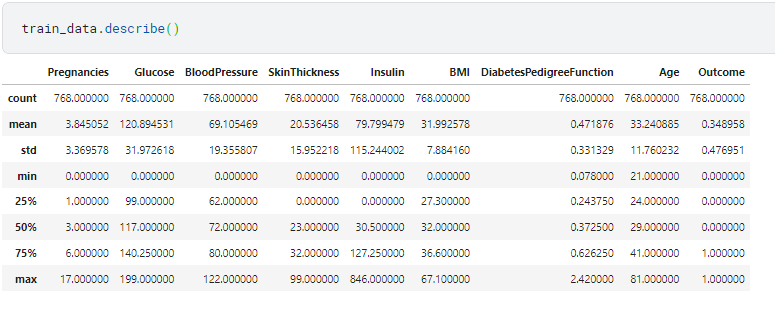
We start by loading the diabetes dataset from a CSV file using the Pandas library. The dataset is stored in train\_data.



* Data exploration:

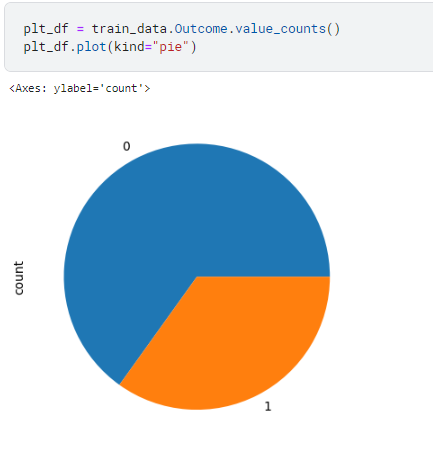
Explore the dataset to understand its structure and features. This step helps to gain insights into the data and decide which features are relevant for our prediction model





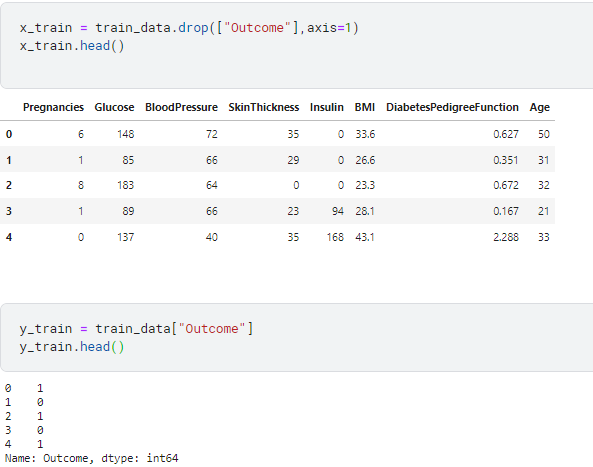
* Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a crucial step in understanding the characteristics of the dataset and identifying patterns or trends. In this project, we conducted the following EDA tasks:



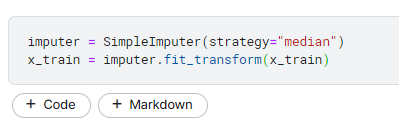
* Spliting data into X and Y:

We then split the data into features (x\_train) and the target variable (y\_train).



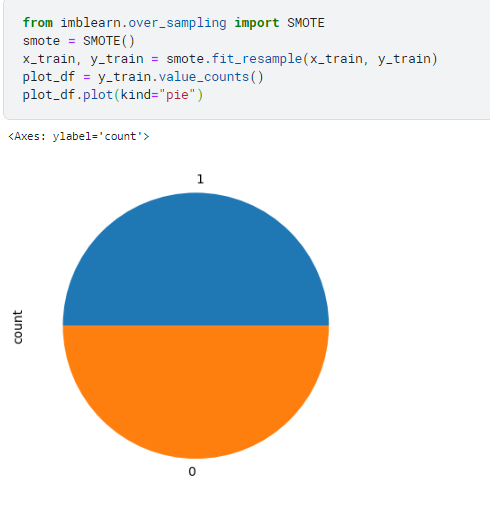
* Missing Value Imputation

For columns with missing values, we employ the SimpleImputer class to fill them with the median value. The imputed features are stored back in x\_train.



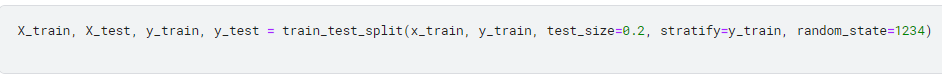
* Handling Class Imbalance

To address class imbalance in the target variable, we apply the Synthetic Minority Over-sampling Technique (SMOTE) from the imbalanced-learn (imblearn) library. SMOTE generates synthetic samples for the minority class to balance the dataset.



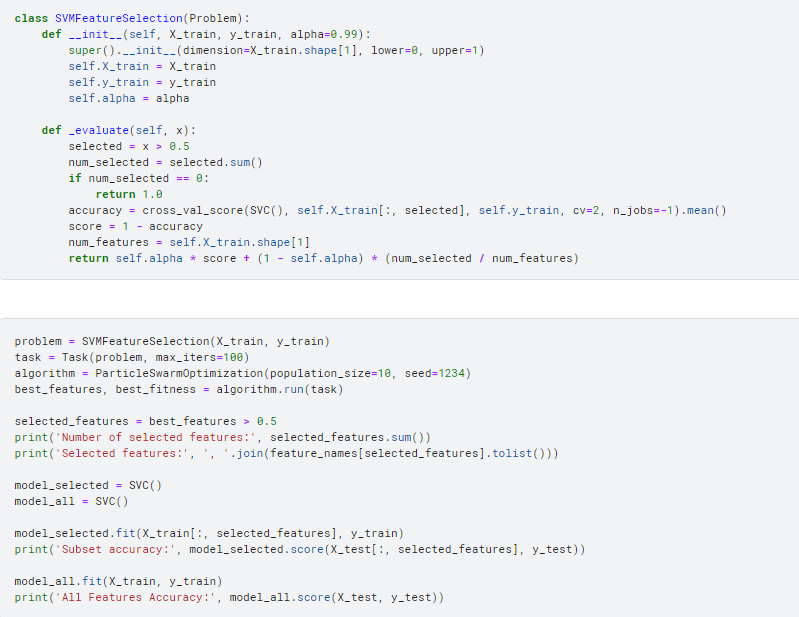
* Train-Test Split

We split the data into training and testing sets using the train\_test\_split function. The stratify parameter ensures that the class distribution is maintained in both sets.



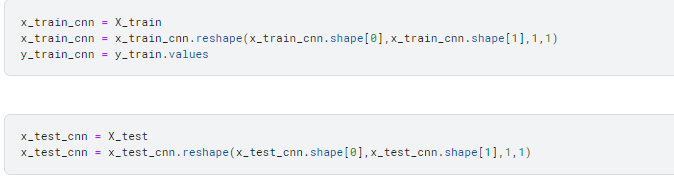
* Feature Selection

For feature selection, we use a Particle Swarm Optimization (PSO) algorithm. The algorithm is employed to find the subset of features that maximize a fitness function.



* Convolutional Neural Network (CNN)

The dataset is reshaped to be compatible with a Convolutional Neural Network (CNN) model. The reshaped training and testing sets are x\_train\_cnn and x\_test\_cnn, respectively.



* Conclusion

Phase 3 summarizes the key steps involved in the AI project, from data preprocessing and handling class imbalance to feature selection and model training. Further analysis, model evaluation, and fine-tuning may be required to achieve the project's objectives