**Practical 01**

**Aim:**

To build a logistic regression model to predict whether a person has diabetes using the Pima Indians Diabetes dataset and to compare the model performance before and after selecting the top 5 most important features using feature selection techniques.

**Theory:**

**Feature Selection and Filtering**

Feature selection is the process of identifying and selecting the most relevant features (variables) from a dataset that contribute the most to predicting the target variable. This reduces model complexity, speeds up training, and can improve model performance. Common methods include:

* **Filter Methods** (e.g., SelectKBest)
* **Wrapper Methods** (e.g., Recursive Feature Elimination)
* **Embedded Methods** (e.g., feature importance from tree models)

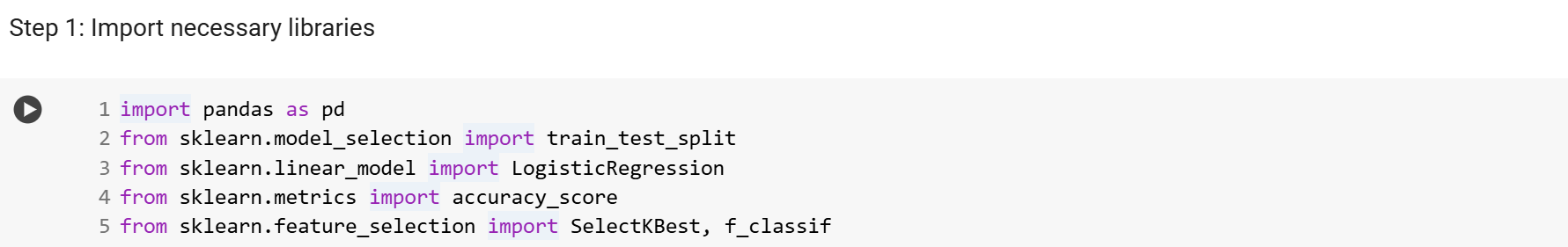
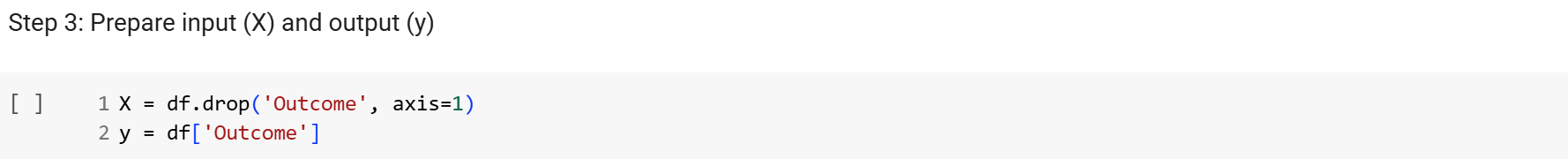
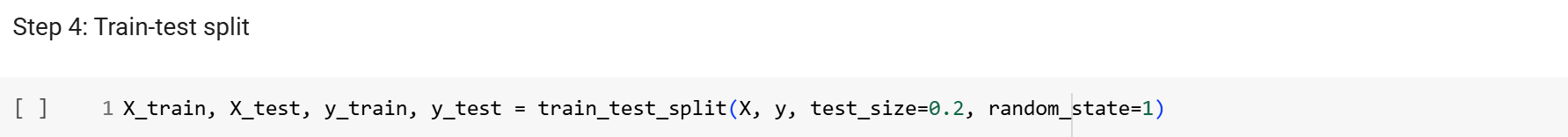
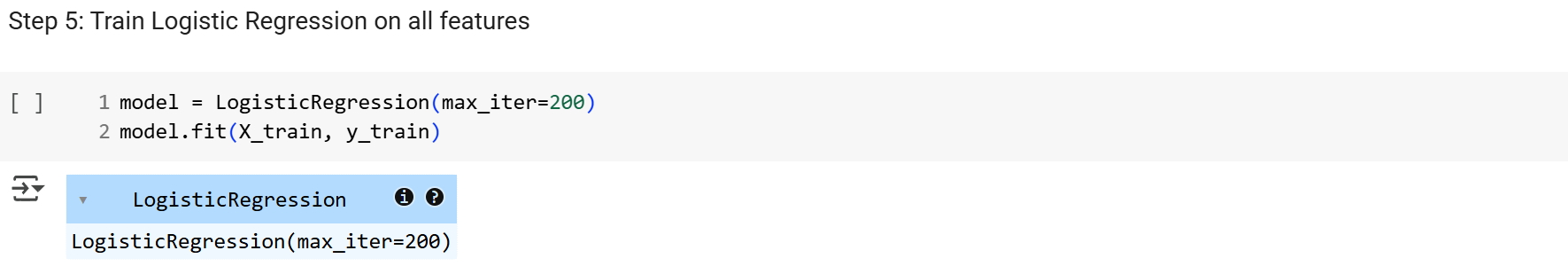
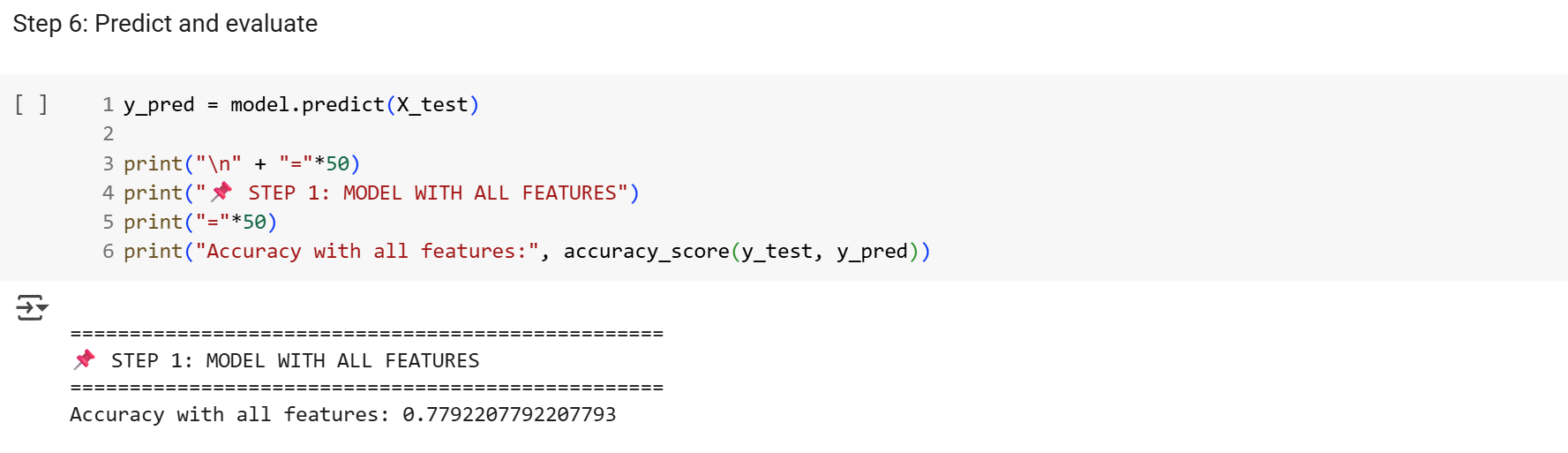
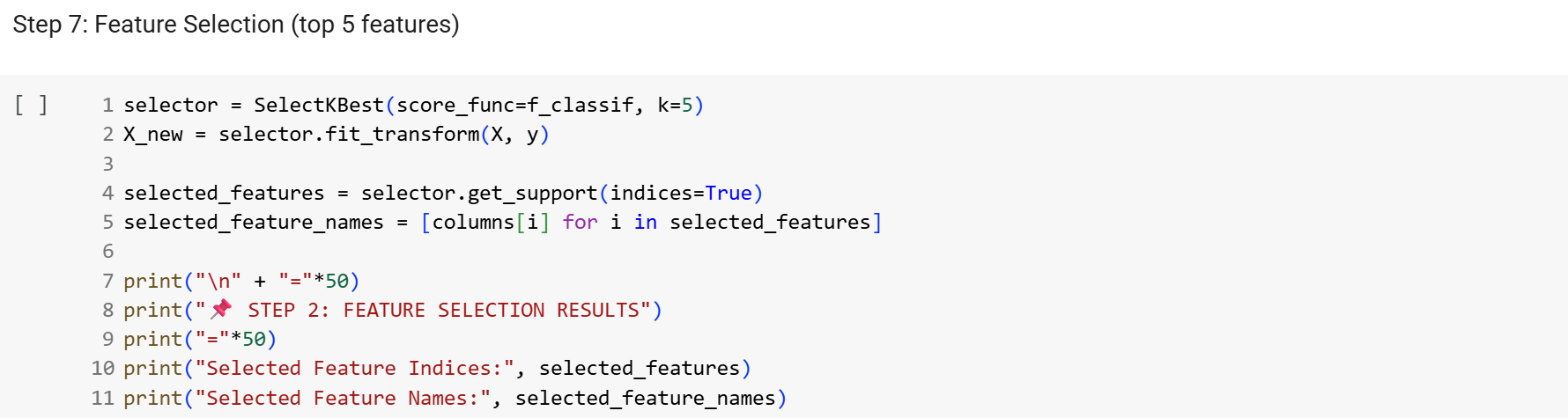
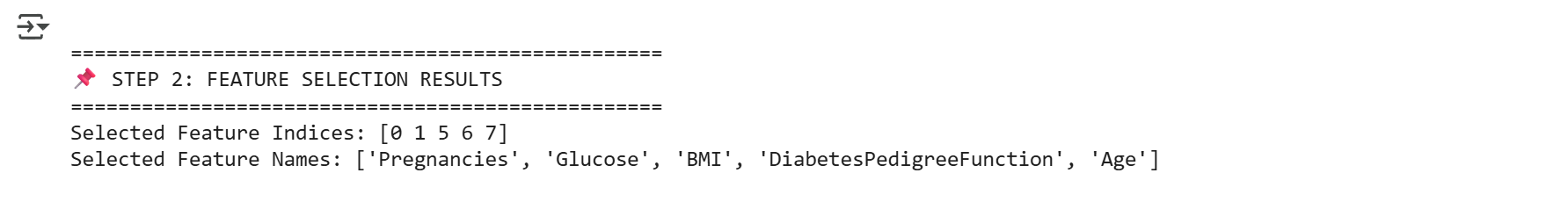
In this practical, we use **SelectKBest** with the **f\_classif** scoring function, which uses ANOVA F-statistics to select the features most associated with the target outcome.

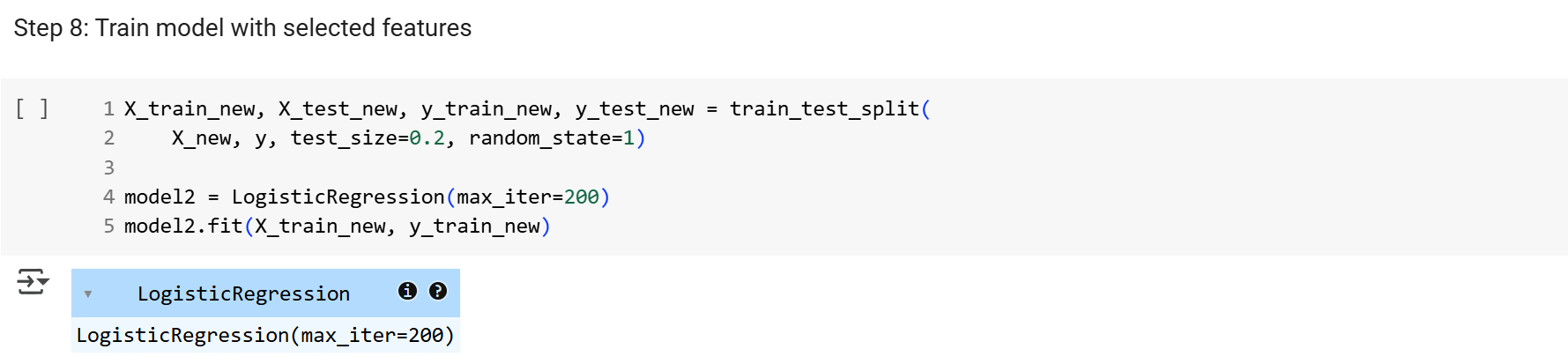
**Logistic Regression Model**

Logistic Regression is a supervised machine learning algorithm used for classification problems. It predicts the probability of a categorical dependent variable (binary outcome).  
Key points:

* It uses the logistic (sigmoid) function to map predicted values between 0 and 1.
* Suitable for binary classification (e.g., diabetes: yes/no).
* The decision boundary is based on probability thresholds (typically 0.5).

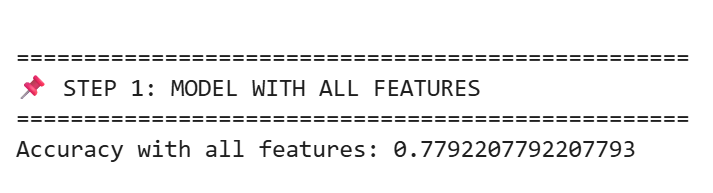
**Code:**

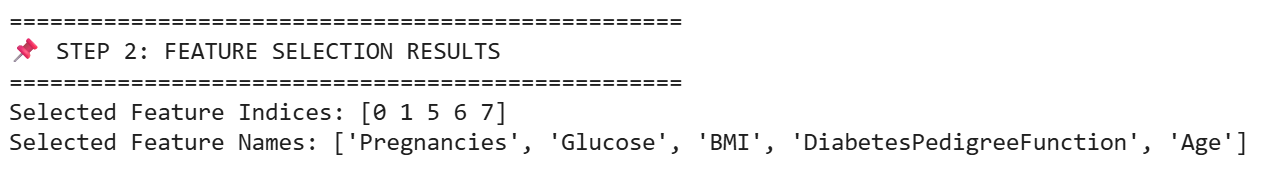
 

**Output Screenshots:**

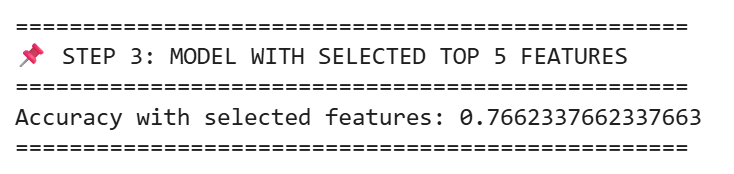
* **First accuracy (before feature selection)**



* **List of selected features**



* **Second accuracy (after feature selection)**



**Conclusion:**

From the above practical, we successfully built a logistic regression model to predict diabetes using all features and then after applying feature selection using SelectKBest.