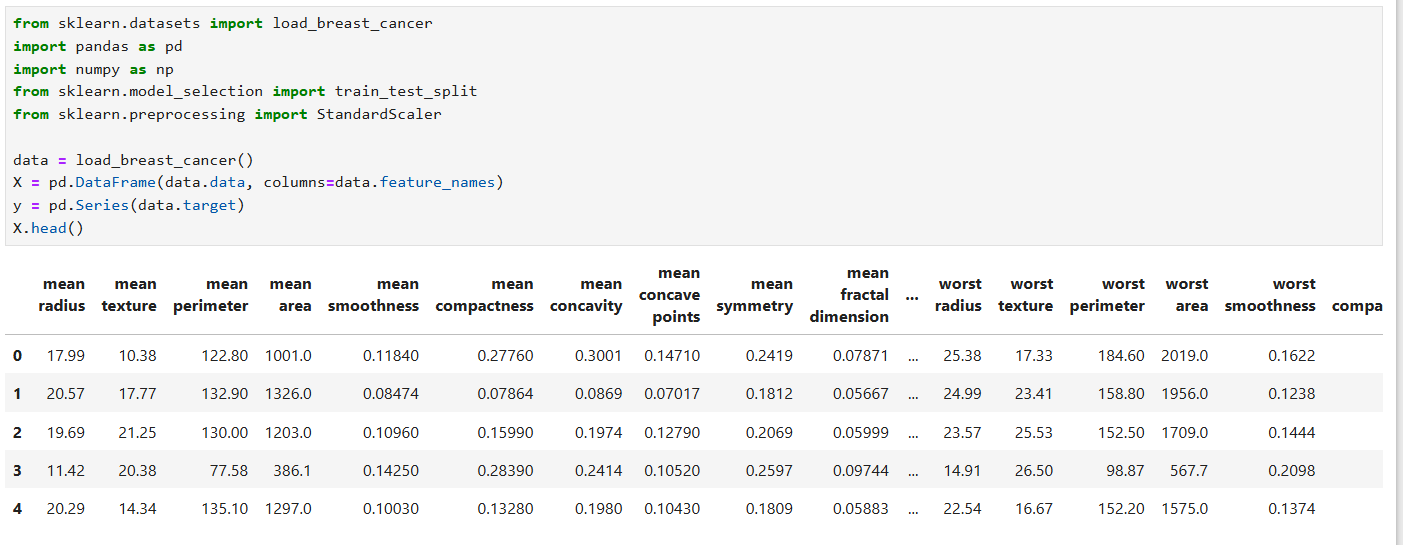
**Formative Assessment: Supervised Learning**

**Due on 04th October 24**

**Submitted by Aiswarya Jayaprakash**

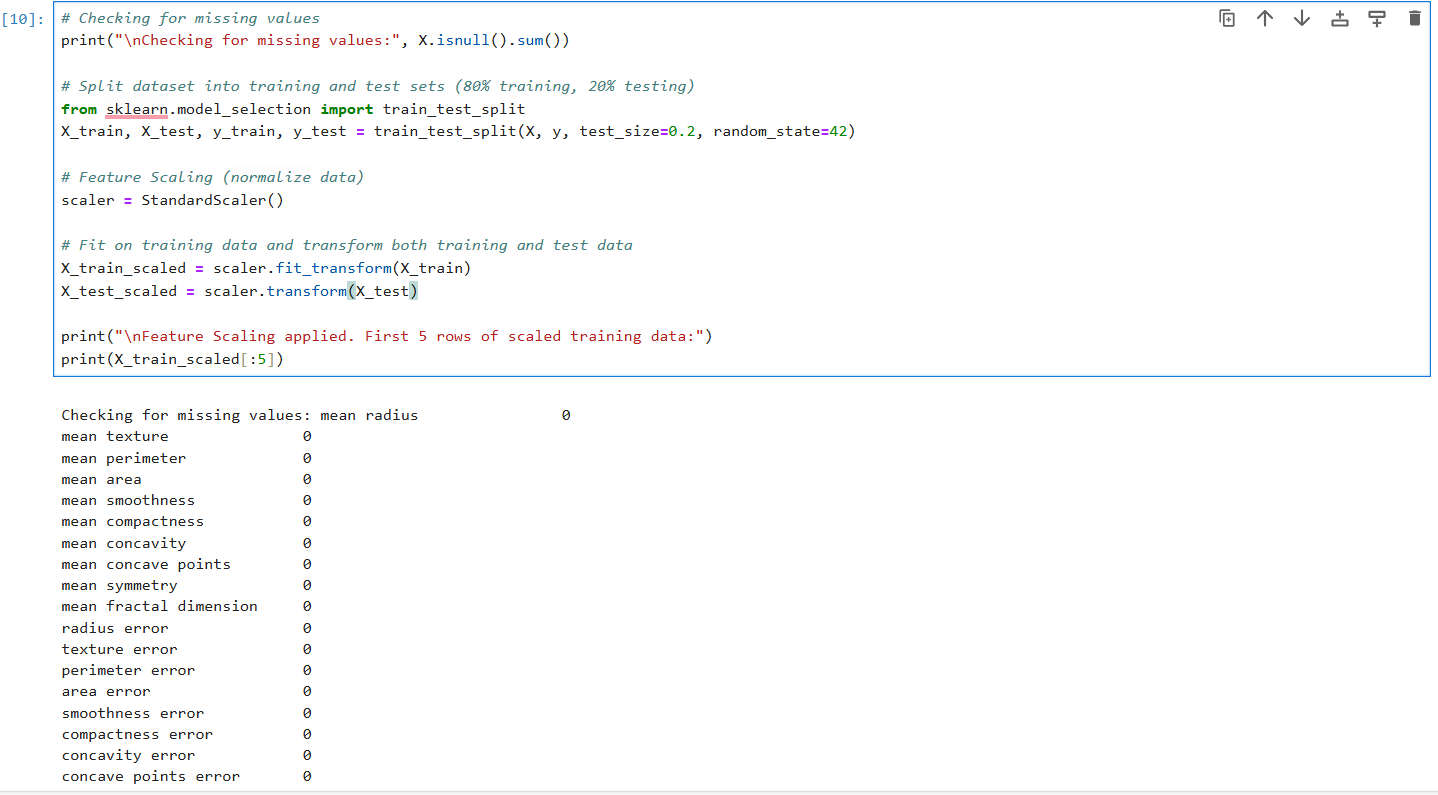
**Objective:**  
The objective of this assessment is to evaluate your understanding and ability to apply supervised learning techniques to a real-world dataset.  
  
**Dataset:**  
Use the breast cancer dataset available in the sklearn library.  
  
**Key components to be fulfilled :**  
  
**1. Loading and Preprocessing (2 marks)**

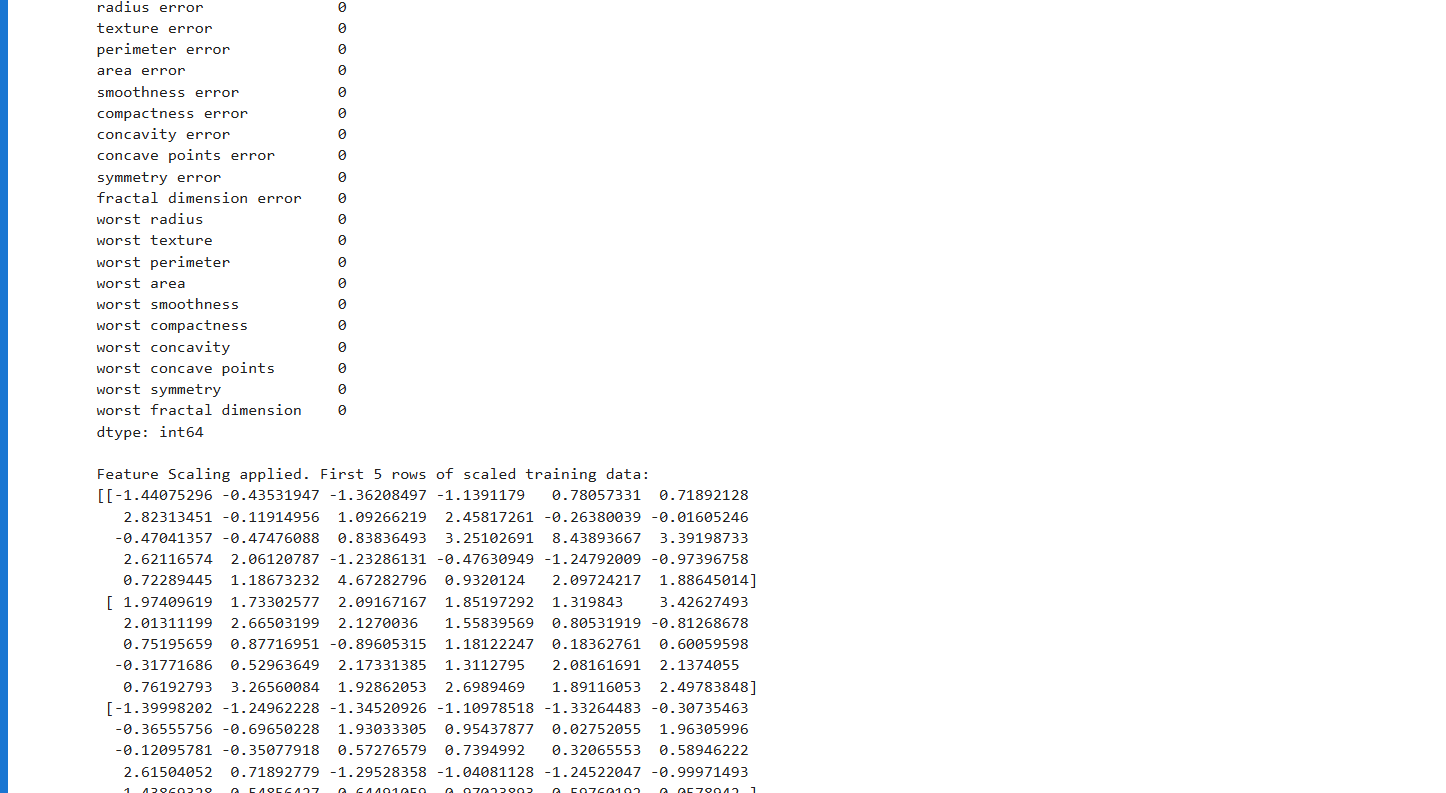
* **Load the breast cancer dataset from sklearn.**



The Breast Cancer dataset is a built-in dataset in the sklearn library which consist of 569 samples with 30 features and a binary classification target (malignant or benign).

* **Preprocess the data to handle any missing values and perform necessary feature scaling.**

****

****

* Explain the preprocessing steps you performed and justify why they are necessary for this dataset.

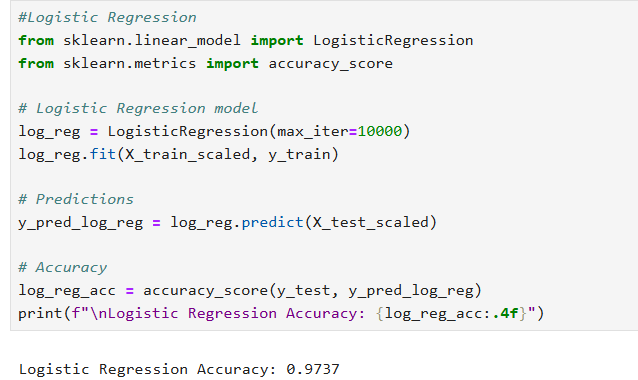
The missing values in the dataset is checked using X.isnull().sum() and it is confirmed that the dataset does not contain missing values. However, if there were any, they could be handled using strategies like **mean imputation** or **removal** of missing rows.

**Feature Scaling:** Some algorithms (like SVM and k-NN) are sensitive to the magnitude of features, which can impact their performance. Therefore, scaling is essential. Hence, we apply StandardScaler to standardize features with zero mean and unit variance. Further, we split the data into 80% training and 20% test sets using train\_test\_split.

**2. Classification Algorithm Implementation (5 marks)**

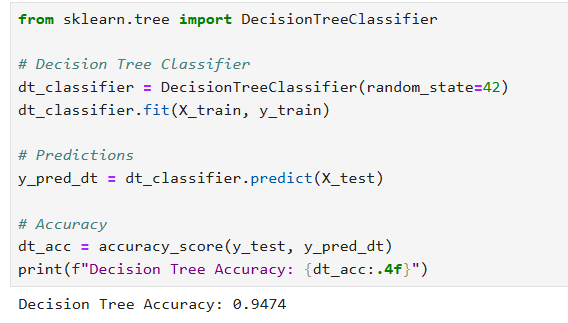
* Implement the following five classification algorithms:  
  **1. Logistic Regression**

Logistic regression is a linear binary classification model suitable for datasets where the relationship between features and output is approximately linear. It calculates probabilities of an observation belonging to one of two classes using the logistic function and is suitable for binary classification problems like this. It uses the sigmoid function to map predictions to a binary outcome. Logistic regression is efficient for binary classification and can give insights into feature importance. The accuracy score is computed to evaluate the model’s performance.

****

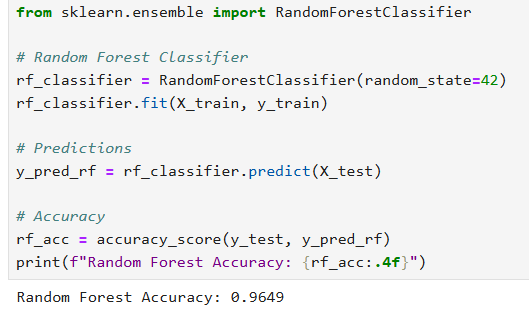
**2. Decision Tree Classifier**

A non-linear model that splits the data into subsets based on feature thresholds. They work by recursively partitioning the data**.** It is easily interpretable but can overfit small datasets. Decision trees don't require feature scaling, as they are not distance-based algorithms.



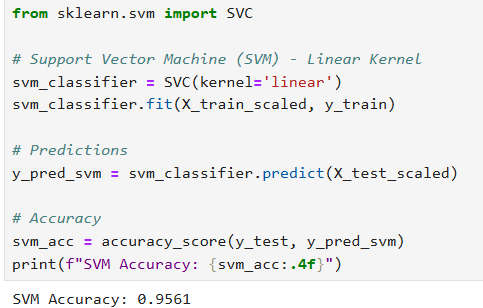
**3. Random Forest Classifier**

An ensemble of decision trees that reduces overfitting by averaging predictions across multiple trees. Generally, yields high accuracy. The algorithm combines many weak decision tree learners into a strong ensemble and this process in known as bagging.



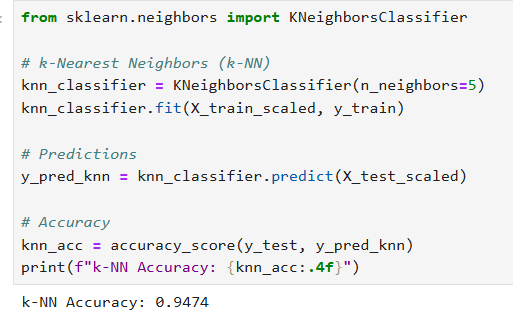
**4. Support Vector Machine (SVM)**

**SVM f**inds the optimal hyperplane that best separates the classes. It works well with high-dimensional data. SVM requires standardized data since it is sensitive to the scale of features.



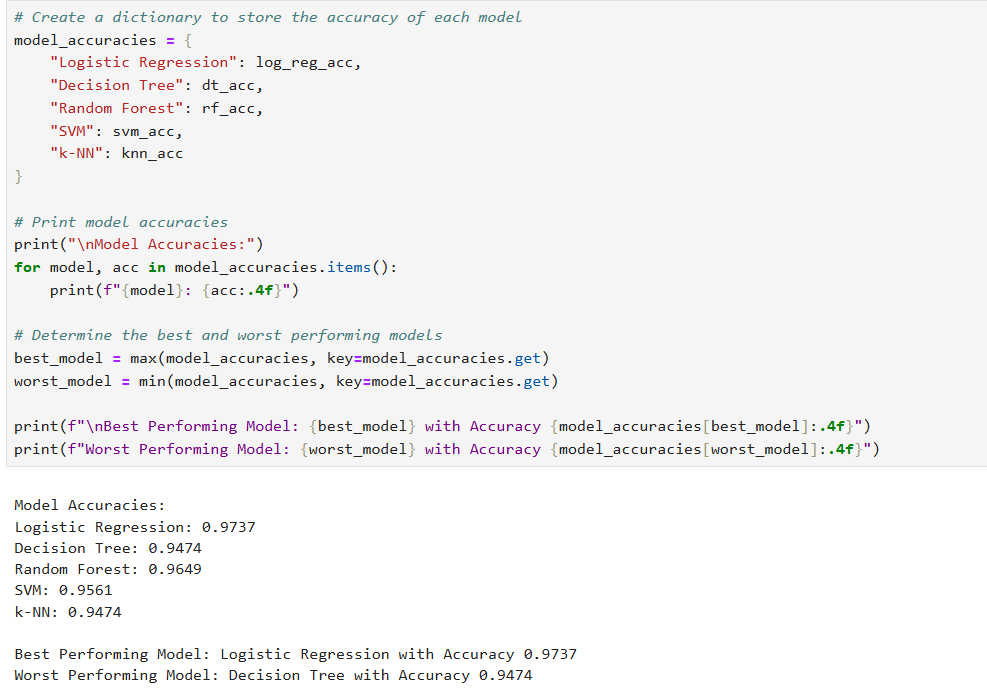
**5. k-Nearest Neighbors (k-NN**)

A non-parametric algorithm that classifies a data point based on the majority class among its nearest neighbors. It is simple to implement and effective for smaller datasets but sensitive to noisy data. Feature scaling is required for k-NN because it is distance-based.



**3. Model Comparison (2 marks)**

* Compare the performance of the five classification algorithms.



* Which algorithm performed the best and which one performed the worst?

After running all these models, the accuracies among them were compared in which the **best performing model is Logistic Regression with an accuracy of 0.9737**  and **worst** performing models is decision tree with an accuracy of 0.9474