**Data Analyst Internship Report**

**Introduction**

|  |  |
| --- | --- |
| Name | Vedant Maladkar |
| Project Title | * E-commerce Furniture Dataset 2024 * Iris classification |
| Domain | Data Analyst |
| Tools | Python, Machine Learning |

**Project 1-** E-commerce Furniture Dataset 2024

**Objective-**

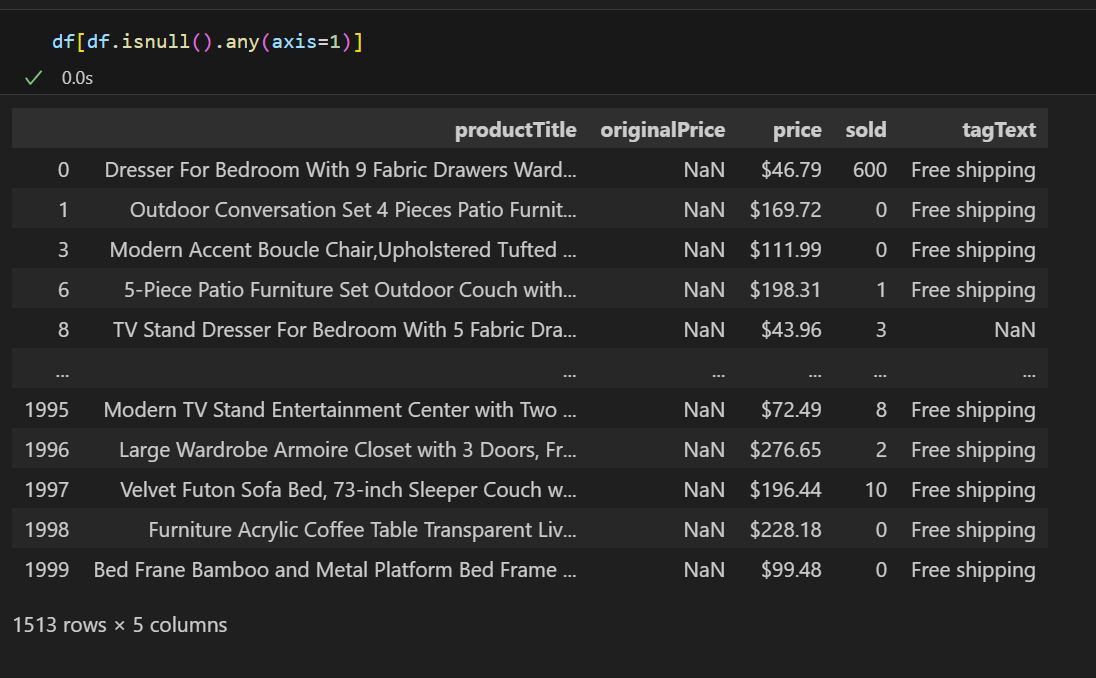
The aim is to build a logistic regression model that helps us predict whether a furniture product will be a high-selling item (sold > 50 units) on the bases of its price and shipping option.

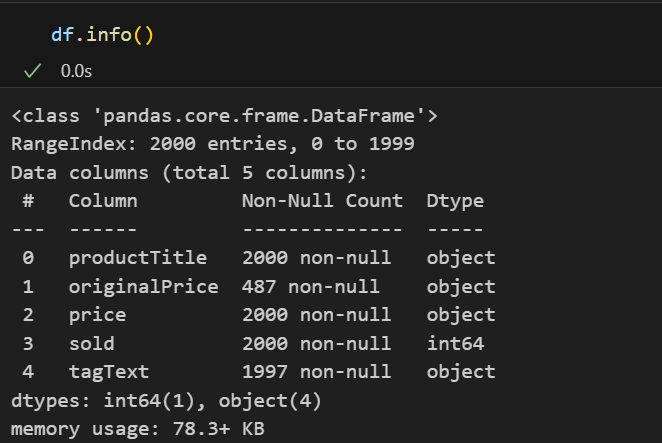
This model enables e-commerce businesses to identify high-selling products for targeted marketing, better inventory decisions, and competitive pricing strategies.

**Dataset Description-**

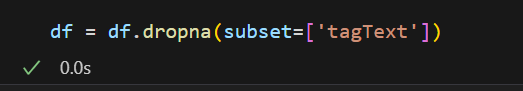
This dataset is named 'E-commerce Furniture Dataset 2024’ and it contains historical record of furniture products sold. This dataset contains 2000 entries and multiple columns such as ‘productTitle’, ‘originalPrice’, ‘price’, ‘sold’, ‘tagText’.

**Exploratory Data Analysis (EDA)-**

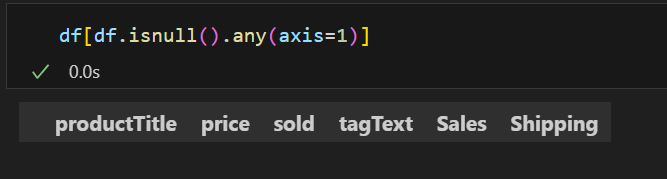




Using this I remove the ‘originalPrice’ column as more than 1000 values are null and removed the missing values in ‘tagText’ column.

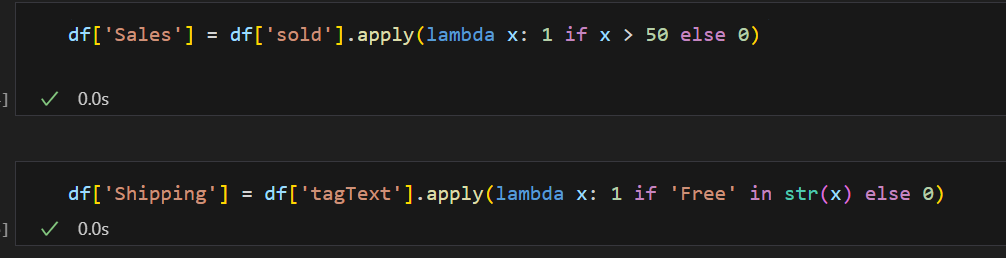


After doing this I got

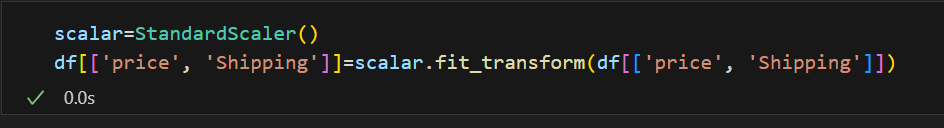


This shows that there are no null values in this dataset.

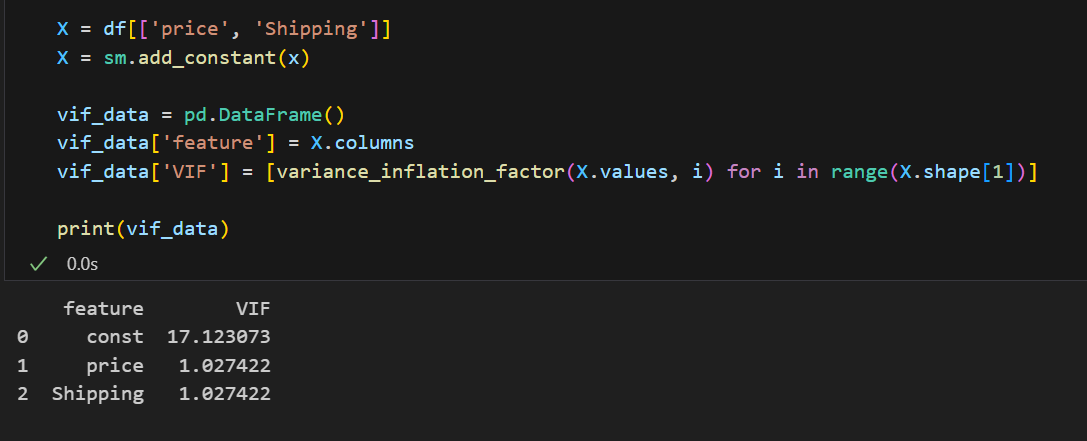
**Feature Engineering-**



* This creates a new column ‘Sales’ where the value in binary form such that (1 if Quantity > 50) or else 0.
* The same goes for ‘Shipping’ where ‘Free Shipping’ is 1 and the other is 0.



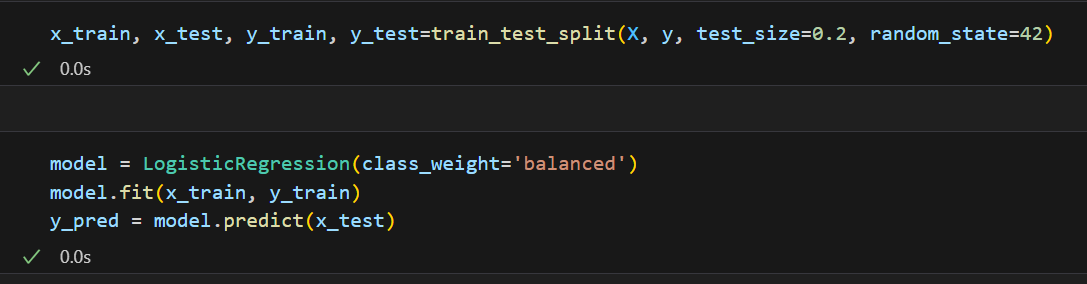
* This normalizes the ‘price’ and ‘Shipping’ using Standard Scalar



* This uses VIF analysis to remove the multi-colinear features.

**Model Building-**

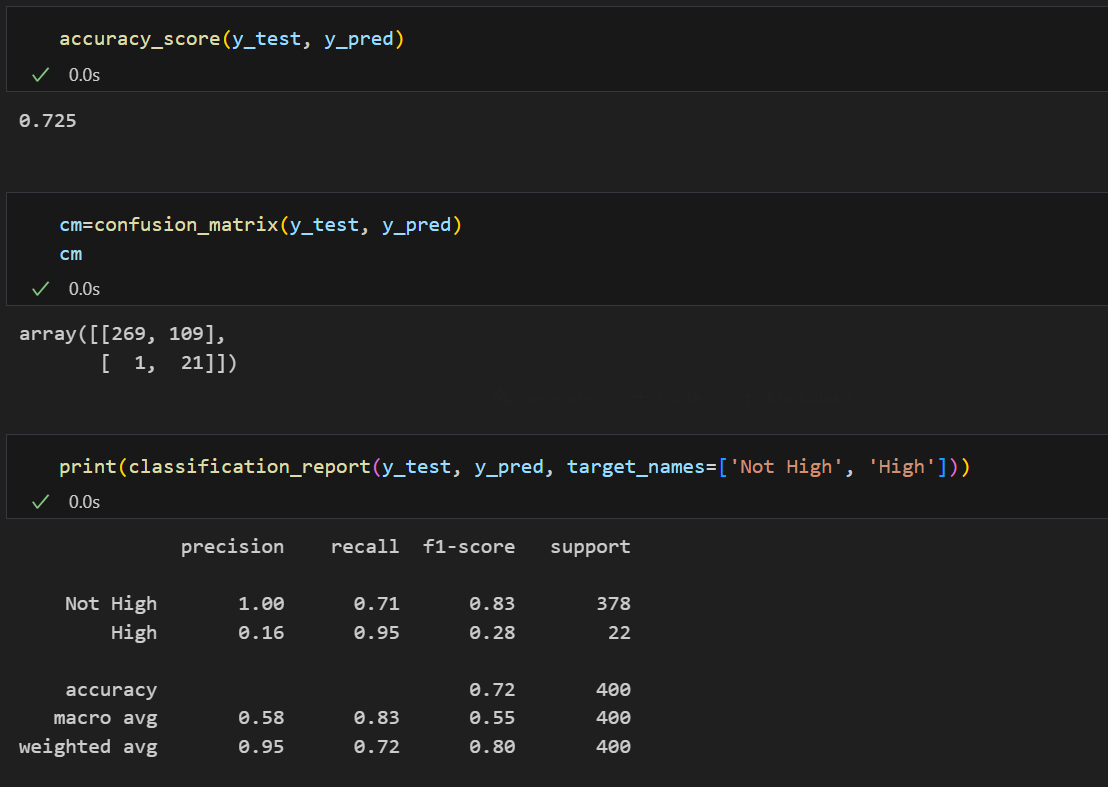
We used Logistic Regression to classify products are high-selling or not. The dataset is split in 80-20 training and testing sets. Here classweight is balanced because it automatically adjusts for class imbalance by assigning higher weights to minor class.



Model Evaluation includes:

* Accuracy Score
* Confusion Matrix
* Classification Matrix

**Results-**



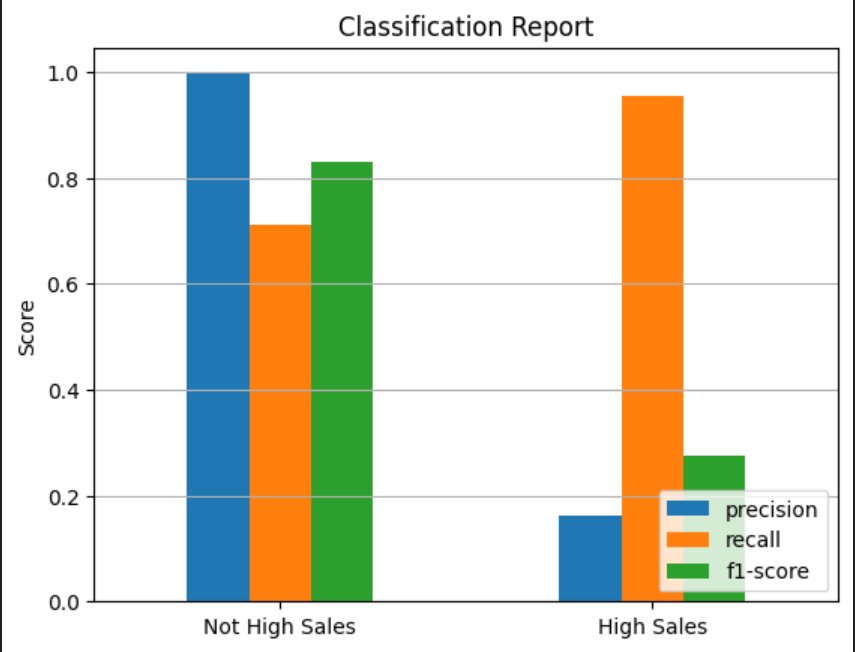
The Logistic Regression shows an accuracy score of approximately 73%

Confusion Matrix and Classification Reports showed:

* Very high recall rate for ‘High Sales’ which is ‘0.95’. This model identified 21 out of 22 high-selling products.
* Low precision for ‘High Sales’ of ‘0.16’. This tells us that the model flagged many products as high-selling which was wrong, indicating there is room for improving precision by reducing false positives.

**Plot-**

This plot show the classification report for both not high-selling and high-selling.



**Summary-**

The model effectively supports decision-making by ensuring high-selling products are correctly flagged, even if it occasional has some false positives. Future improvements can be made by including additional features such as product category, customer review rate, etc.

**Reference-**

* Libraries - pandas, matplotlib, scikit-learn, statsmodels.
* Tools – Visual Studio Code

**Project 2-** Iris classification

**Objective-**

The aim of this project involves creating a Logistic Regression model to classify iris flowers into three species (Setosa, Versicolour, and Virginica) based on the length and width of their petals and sepals.

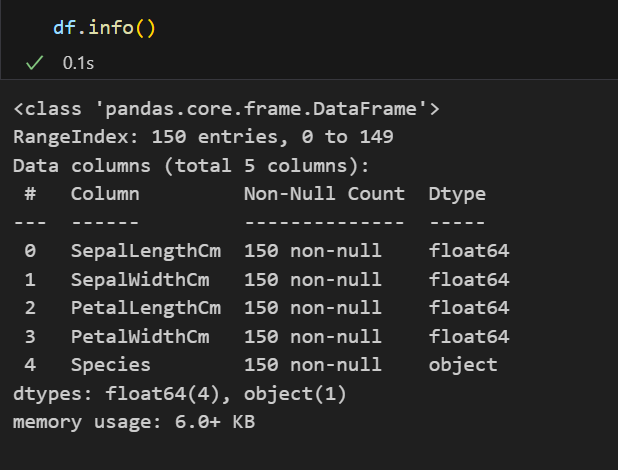
**Problem Statement-**

* The model should achieve a high level of accuracy in classifying iris species.
* The model's predictions should be consistent and reliable, as measured by cross-validation.

**Dataset Description-**

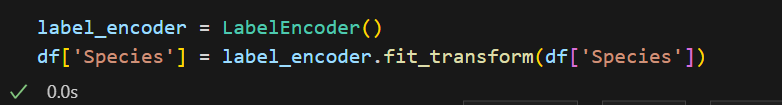
This dataset is named ‘Iris’. It contains a total of 150 entries. The columns of this dataset are 'SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm', 'Species'

**Exploratory Data Analysis (EDA)-**

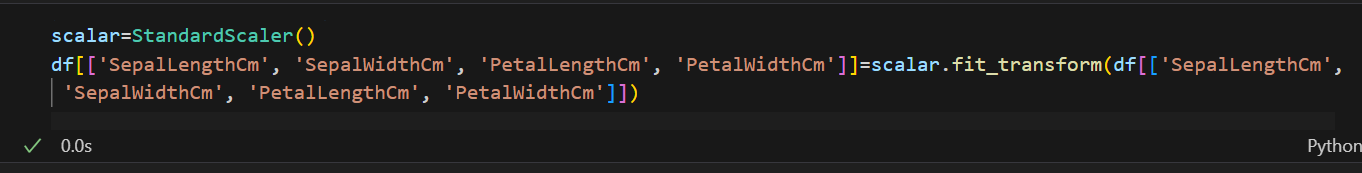


Seeing this we can say that the there are no null values and the dataset is already cleaned.

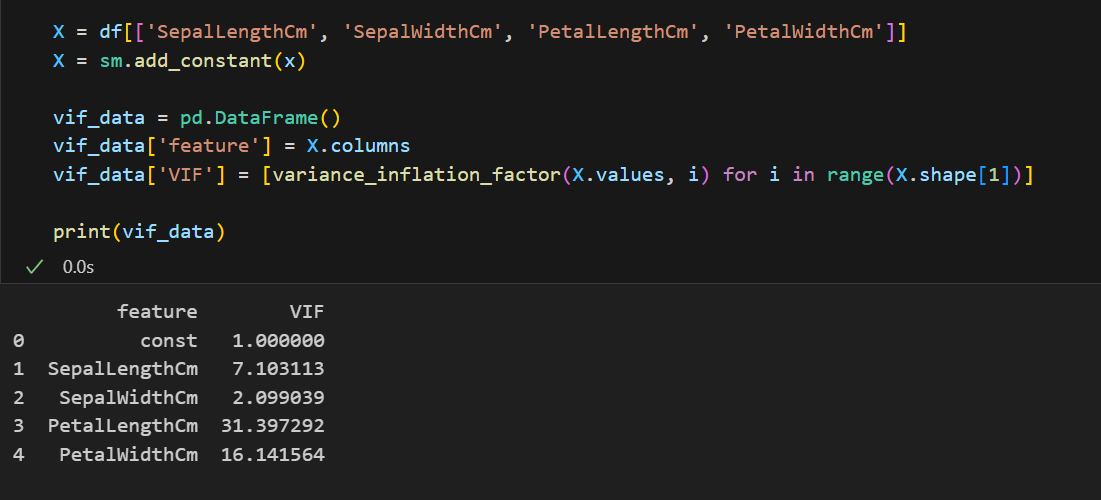
**Feature Engineering-**



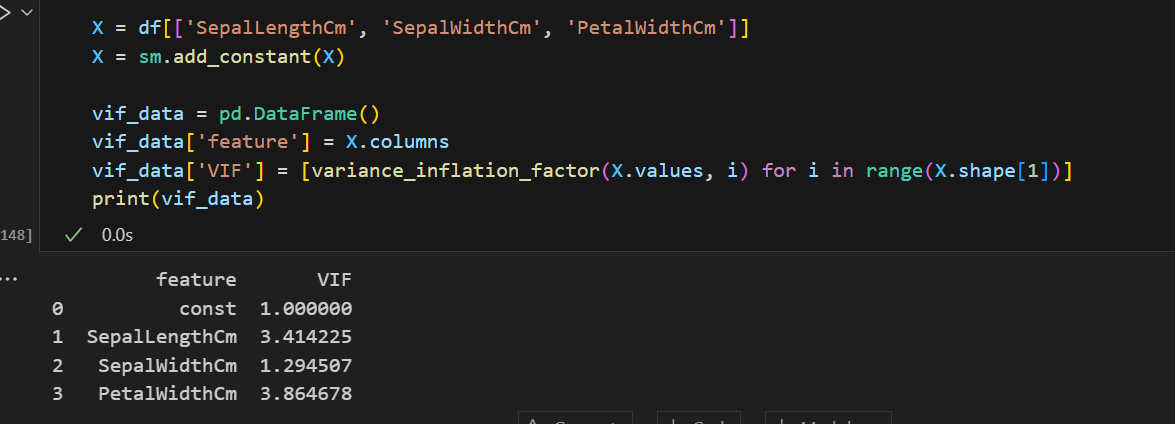
Using this the code automatically converts the ‘Species’ string values to the numerical values (iris-setosa, iris-virsicolour, iris-viginica) as 0, 1, 2



This normalized 'SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm' using Standard Scalar.

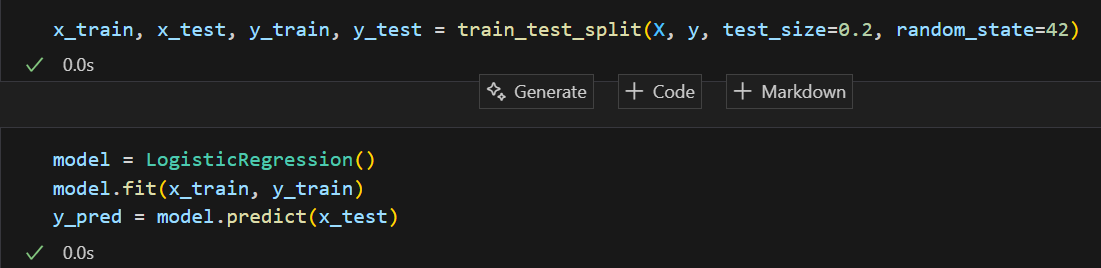


Here we are removing ‘PetalLengthCm’ since the value is multicollinear. After removing the column we get



**Model Building-**

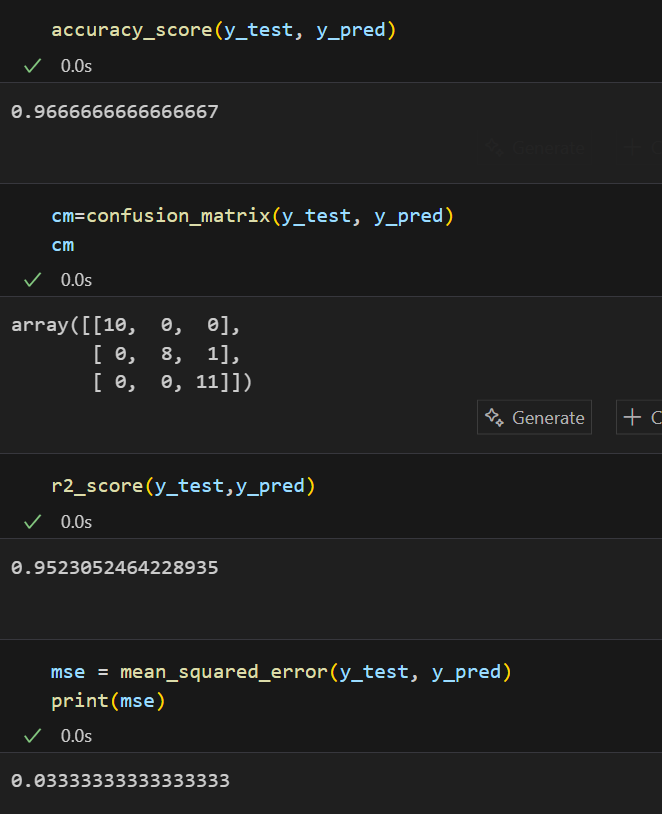
Here we use Logistic Regression and divide dataset to 80-20 for training and testing sets.

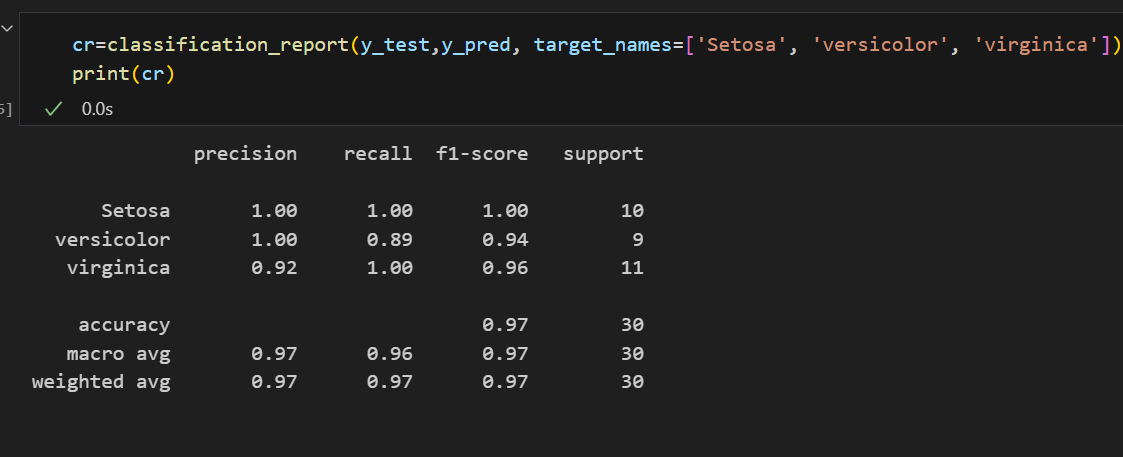


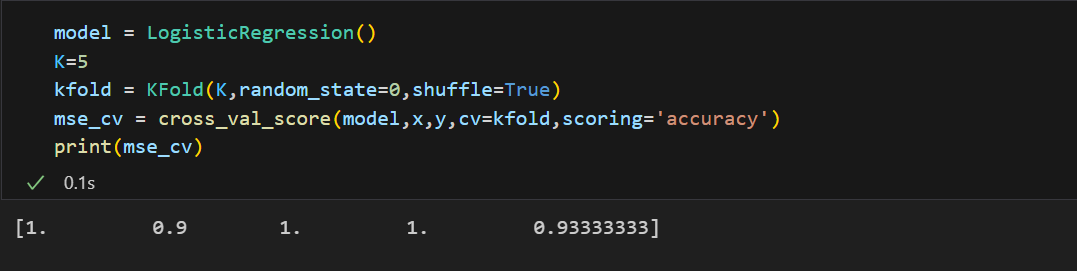
Model evaluation includes:

* Accuracy Score
* Confusion matrix
* R2-Score
* Mean Squared Error
* Classification report
* K-Fold Cross Validation

**Results-**







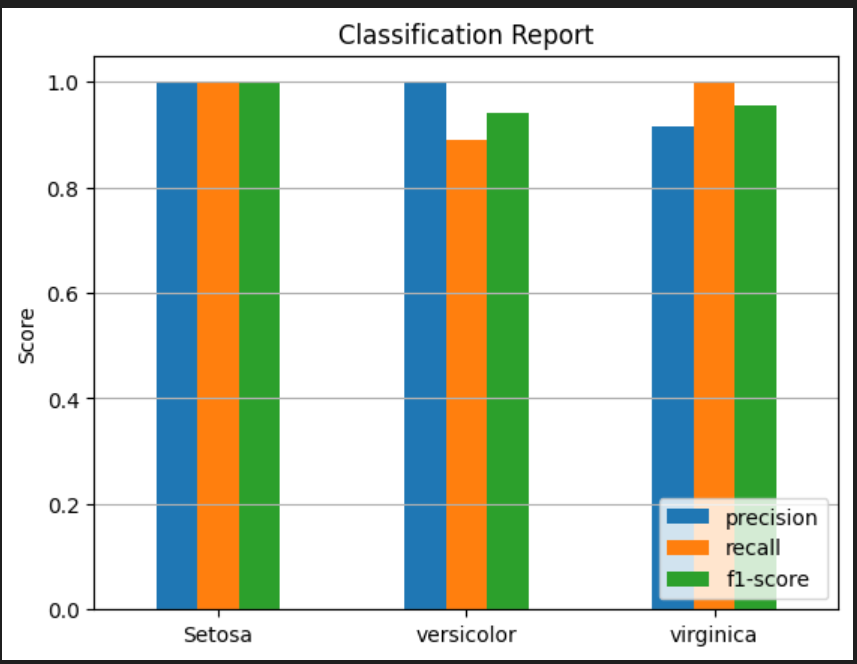
Logistic model accuracy is approximately 97%

Model evaluation shows:

* R2 score is approximately 0.95 showing strong model fit
* Mean Squared Error is 0.033 indicating low prediction error
* Setosa and Virginica showed no false negatives or false positives which means it predicted accurately with perfect recall
* Versicolor showed one error
* The 5 fold cross-validation conforms that I have an average accuracy of 96.67% conforming that the Logistic Regression model performs consistently across different subsets of iris.

**Plot-**

This plot show the classification report setosa, versicolor and virginica.



**Summary-**

* The logistic regression model classifies the iris species with high accuracy and perception of 96.67% showing strong performance.
* This report showed that setosa and virginica have perfect classification and versicolor have a very high classification.
* This is further verified using the 5-fold cross-validation where average accuracy is 96.67%.
* This project makes us understand how reliable is logistic regression for multi-class classification and has a strong foundation for more advanced machine learning applications.

**Reference-**

* Libraries - pandas, matplotlib, scikit-learn, statsmodels.
* Tools – Visual Studio Code