**A PRACTICAL INDUSTRIAL TRAINING REPORT**

**On**

**IMPLEMENTATION OF DATA SCIENCE**

**Submitted to**

**KIIT Deemed to be University**

**In Partial Fulfillment of the Requirement for the Award of**

**BACHELOR’S DEGREE IN**

**COMPUTER SCIENCE AND ENGINEERING**

**BY**

**NITESH KUMAR MANDAL 21053300**

**UNDER THE GUIDANCE OF**

**Dr. Suchismita Rout**

****

**SCHOOL OF COMPUTER ENGINEERING**

**KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY**

**BHUBANESWAR, ODHISA- 751024**

**October 2024**

**KIIT Deemed to be University**

**SCHOOL OF COMPUTER ENGINEERING**

**BHUBANESWAR, ODHISA- 751024**

****

**CERTIFICATE**

**This is to certify that the project entitled**

**“Implementation of Data Science”**

**Submitted by**

**Nitesh Kumar Mandal 21053300**

**is a record of bonafide work carried out by him, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering ) at KIIT Deemed to be University, Bhubaneswar. This work is done during year 2023-2024, under my guidance.**

**Date: 04/10/2024**

**Dr. Supriyo Mandal**

**Project Guide**

**Acknowledgement**

I would like to express my profound gratitude to **Dr. Supriyo Mandal** for his expert guidance and continuous encouragement throughout my journey in completing the **Implementation of Data Science**. His invaluable support and insights were instrumental in helping me navigate the complexities of data science, from foundational concepts to practical applications. This experience has significantly enriched my understanding and skills in the field.

**Nitesh Kumar Mandal**

**Abstract**

This report explores the **Implementation of Data science** techniques on five different datasets, each posing a unique challenge and scope for predictive modeling and data analysis. The datasets used are Iris Flower Classification, Unemployment Rate in India, Car Price Prediction, Spam Mail Detection, and Sales Prediction based on advertising data. Through Python-based implementations, the report outlines the step-by-step procedures for data cleaning, model building, evaluation, and visualization, followed by conclusions drawn from the analysis of each dataset. This report summarizes the knowledge gained during the course, practical skills developed, and key takeaways from each module.

**CONTENT**

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | Introduction | | Page  NO. |
| 2 | Implementation of Data Science Techniques | |  |
|  | 2.1 | Iris Flower Classification |  |
|  | 2.2 | Unemployment Rate in India |  |
|  | 2.3 | Car Price Prediction |  |
|  | 2.4 | Spam Mail Detection |  |
|  | 2.5 | Sales Prediction using Advertising Data |  |
| 3 | Key Takeaways | |  |
| 4 | Conclusion | |  |
| 5 | References | |  |

1. **INTRODUCTION**

The rise of data science in recent years has revolutionized various industries by enabling data-driven decision-making. This report examines the application of data science techniques on five datasets across different domains such as classification, regression, and prediction. Each dataset represents a real-world problem, solved using Python's data science libraries like Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn.

The primary goals of the implementations are:

To explore the data through descriptive statistics and visualizations.

To build predictive models and assess their accuracy using relevant metrics.

To derive meaningful insights and discuss practical applications of the results.

**2. Implementation of Data Science Techniques**

This section outlines the analysis performed on each dataset, including data preprocessing, model building, and evaluation. The methods used will involve machine learning algorithms such as decision trees, random forests, logistic regression, and others where applicable.

**2.1 Iris Flower Classification**

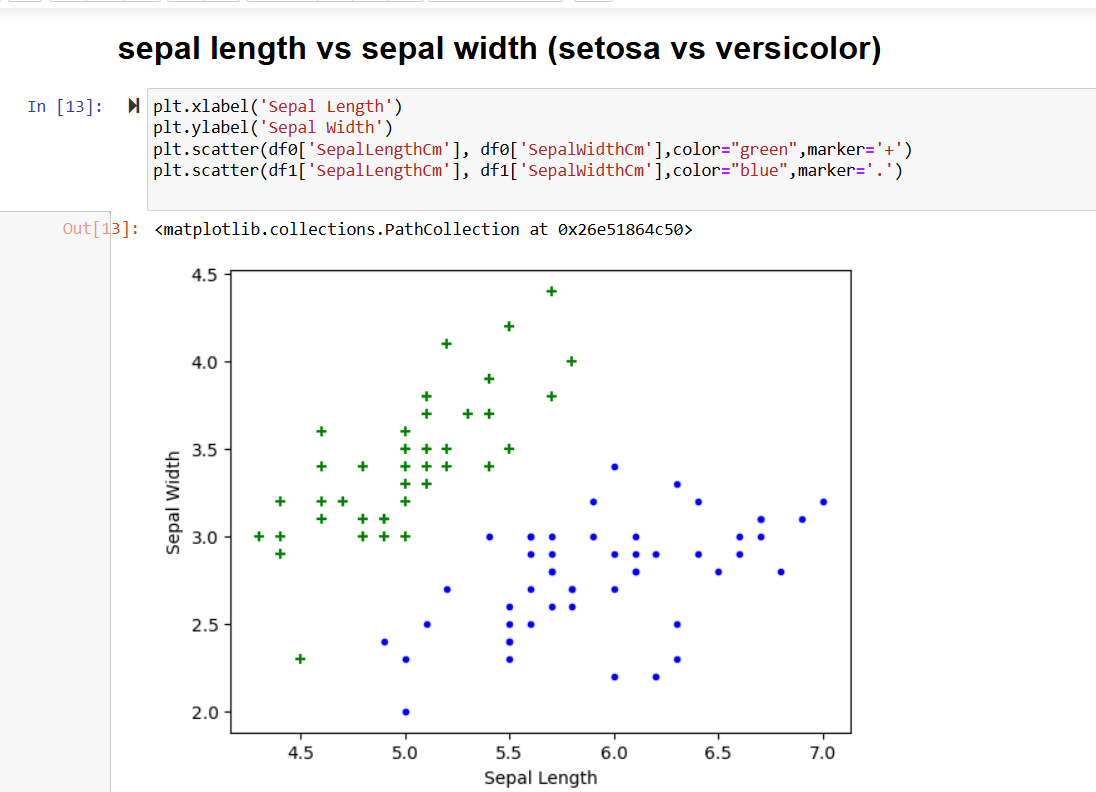
Objective:

The Iris dataset is one of the most famous datasets used in classification tasks. The goal is to classify iris flowers into three species: Setosa, Versicolor, and Virginica.

**Steps:**

Data Loading and Visualization

Feature Analysis (Sepal Length, Sepal Width, Petal Length, Petal Width)



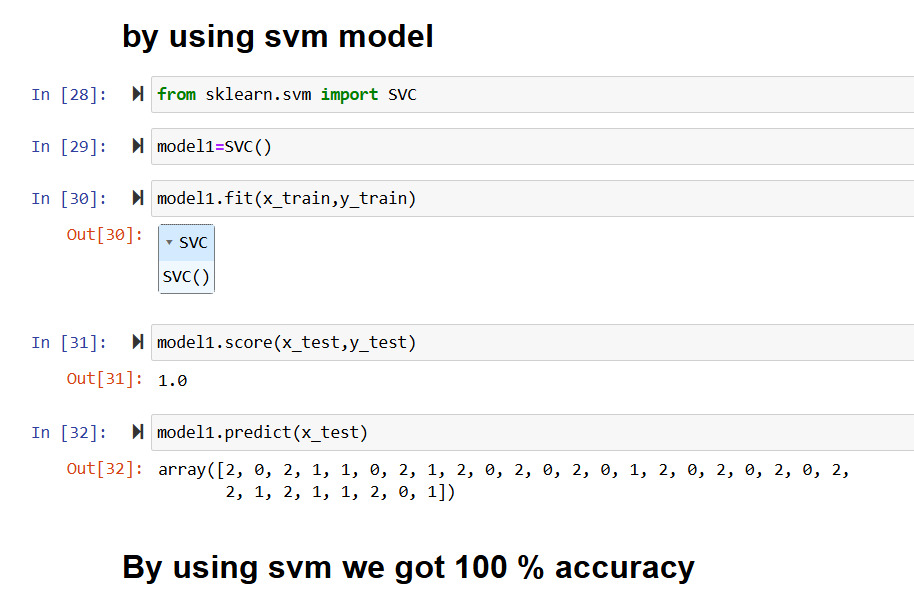
Model Building (Support Vector Machine-SVM)

Model Evaluation (Confusion Matrix, Accuracy, Heatmap)

**Results:**

Feature Importance Visualization

Accuracy Score for SVM (100%)



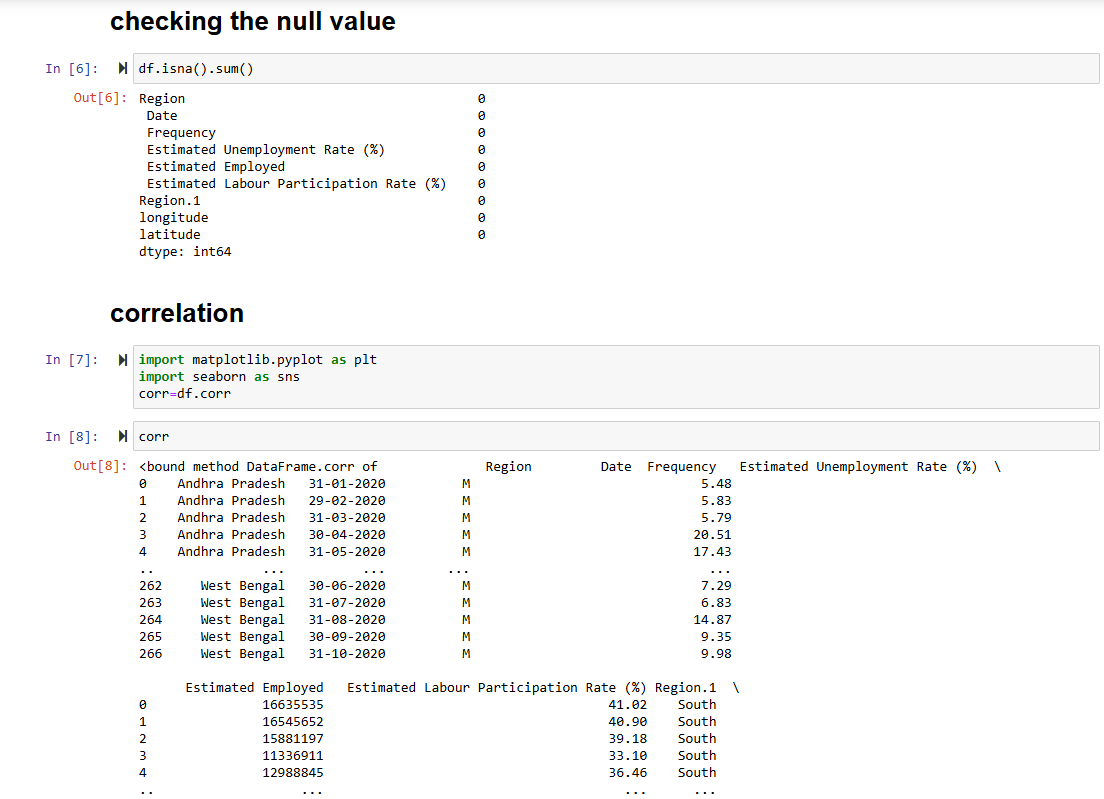
**2.2 Unemployment Rate in India**

Objective:

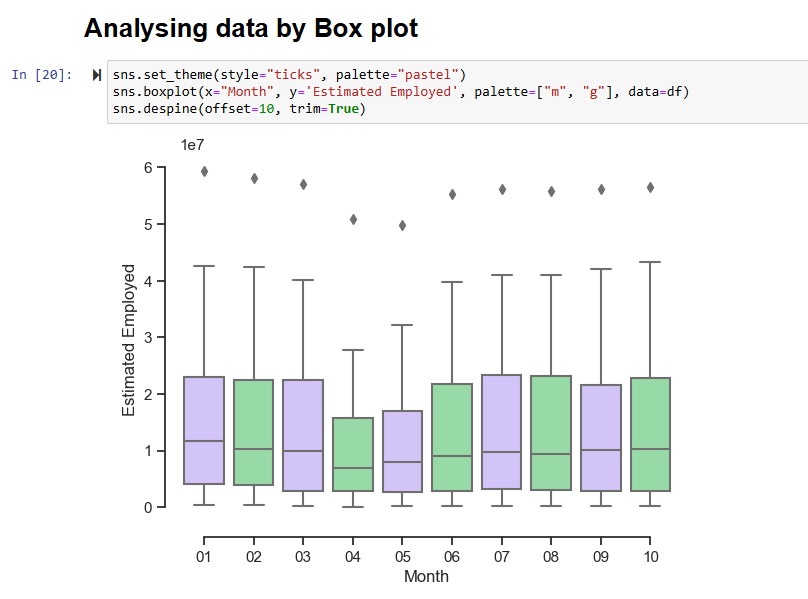
The aim of this analysis is to understand the trends in unemployment in India, identify patterns, and predict future unemployment rates.

**Steps:**

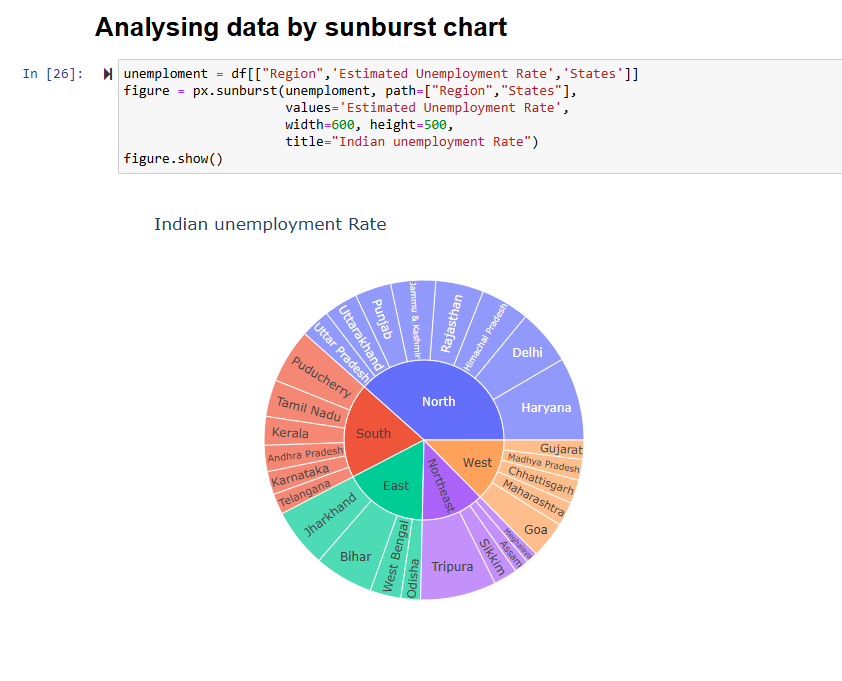
Data Cleaning (Handling missing values)



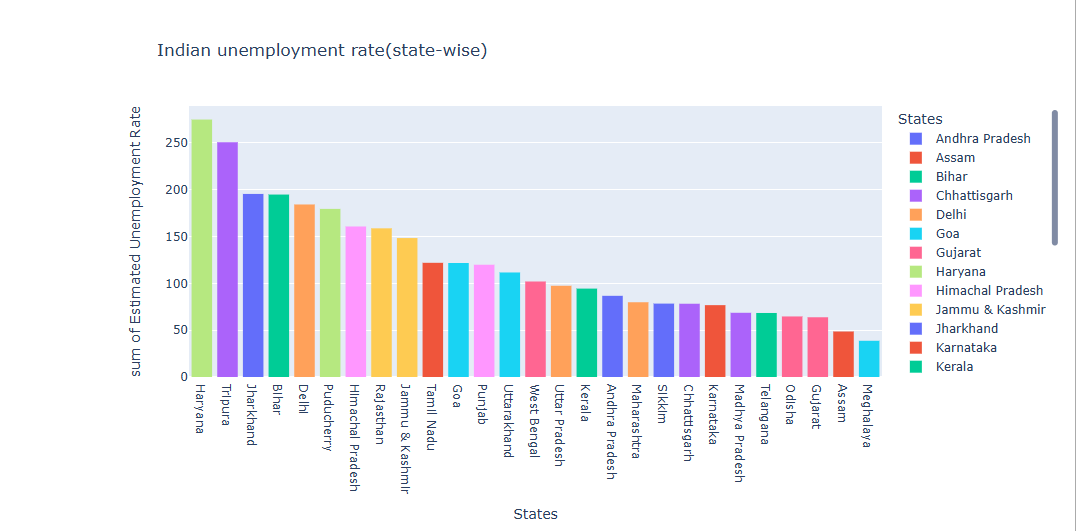
Exploratory Data Analysis by Using Various Graphs, Plots (Box plot, scatter plot, sunburst chart)

****

****

****

**Results:**

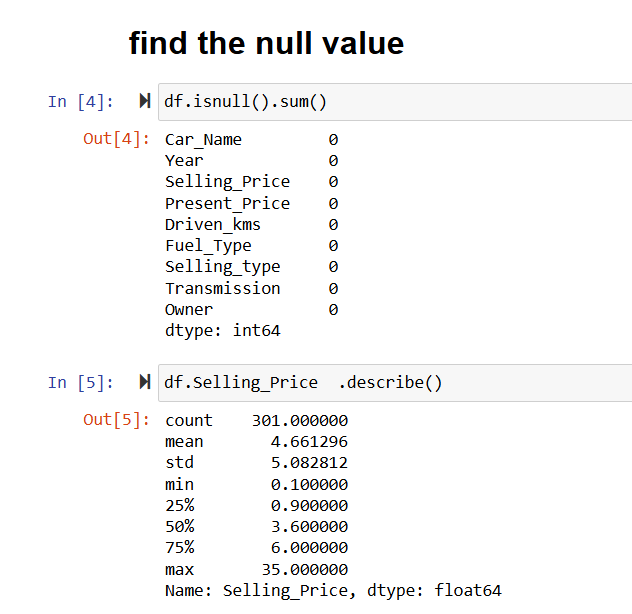
Indian unemployment rate (state-wise)

**3.3 Car Price Prediction**

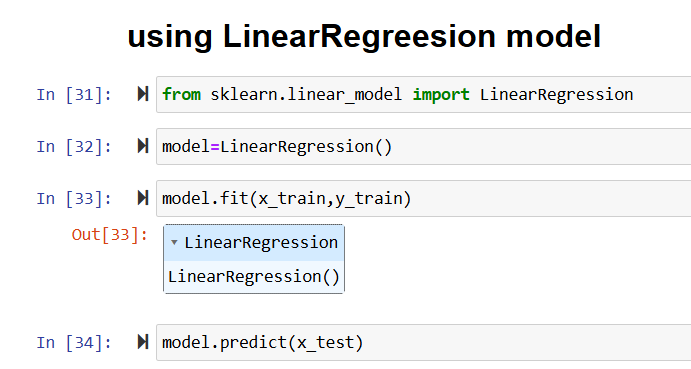
**Objective:** Predict the prices of cars based on their attributes such as engine size, horsepower, and other features.

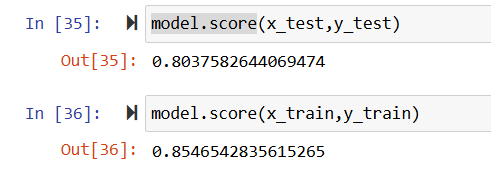
**Steps:**

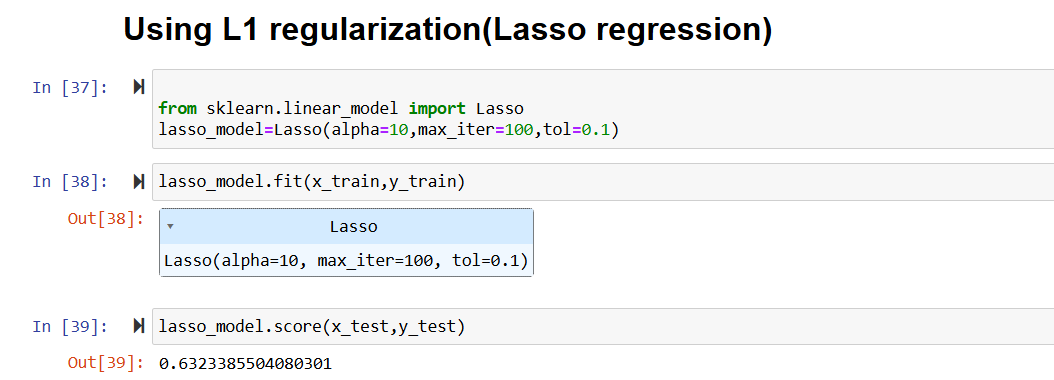
* Data Preprocessing (Missing Value Imputation)

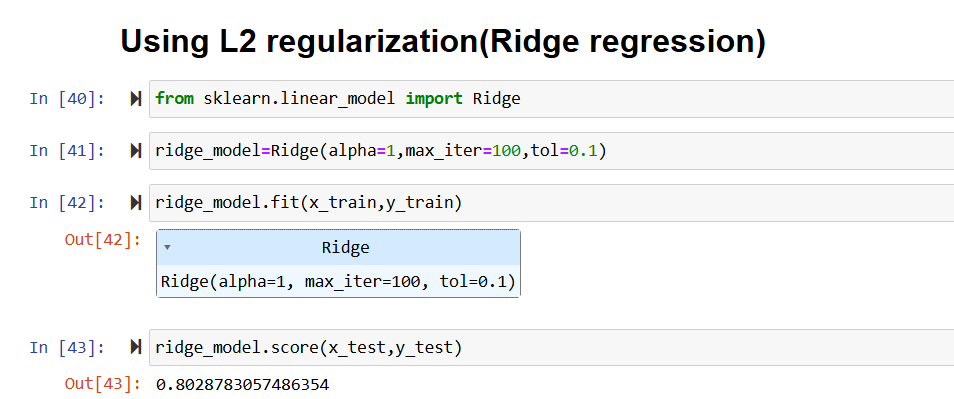


* Feature Selection (Correlation Heatmap)
* Model Building (Linear Regression, L1 regularization (Lasso), L2 regularization(Ridge regression) )





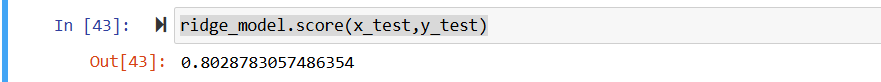




* Model Evaluation

**Results:**

Here we are getting more accuracy in L1 regularization (Lasso) than other model so we can use this model as for car price prediction.



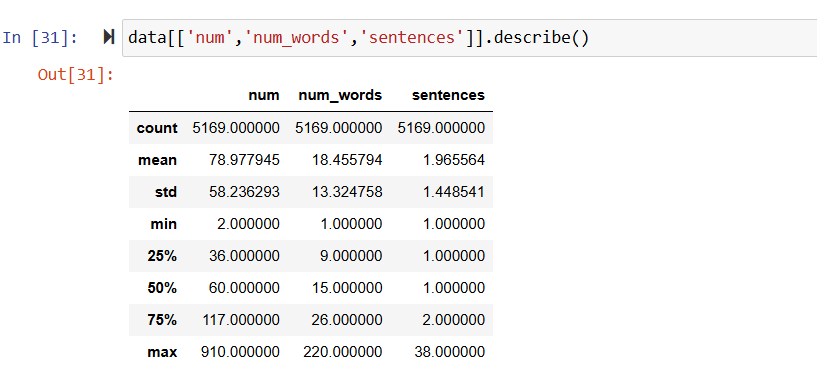
#### ****2.4 Spam Mail Detection****

**Objective:**

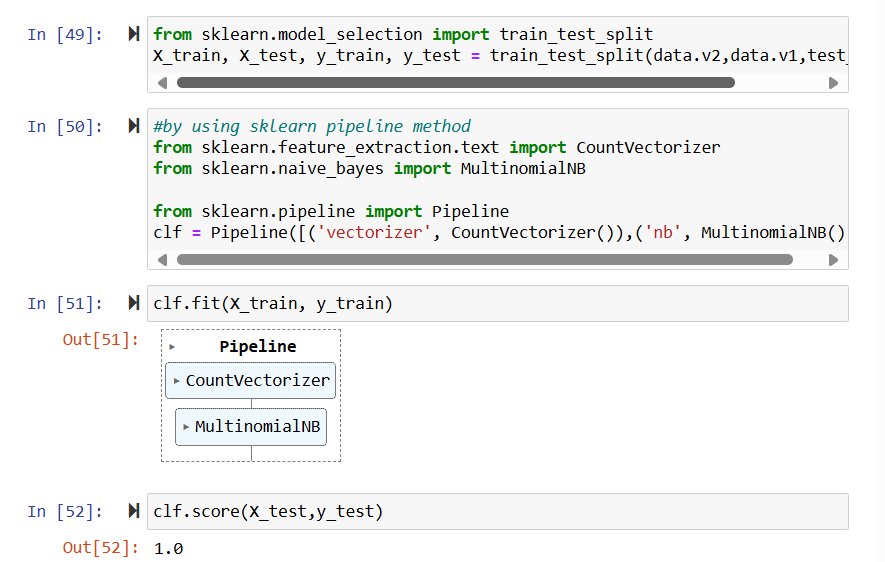
Classify emails as spam or not spam using text classification techniques.

**Steps:**

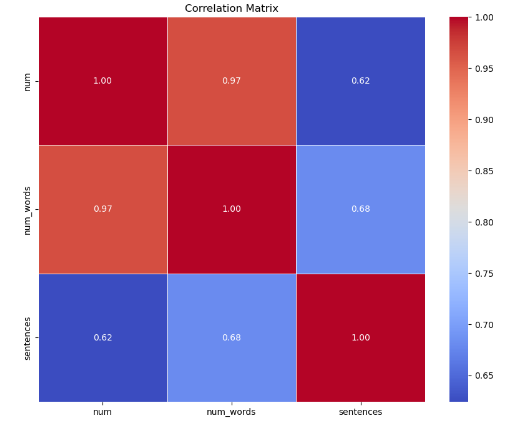
* Text Data Preprocessing



* Model Building (Sklearn pipeline method)

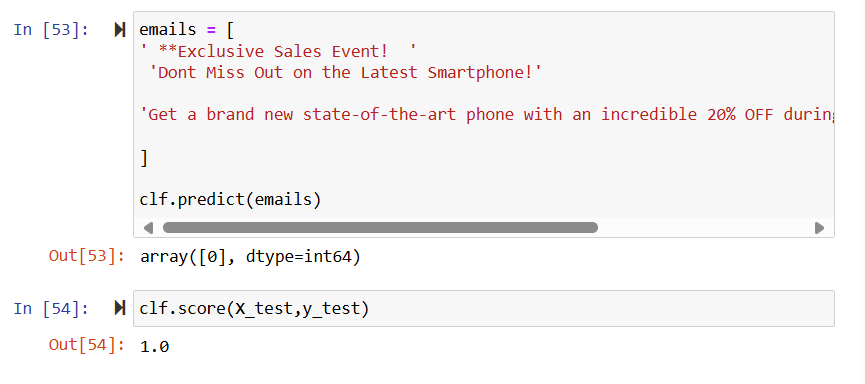


* Model Evaluation
* Model was evaluated by using correlation Matrix



**Results:**

* We achieved the accuracy of 100%

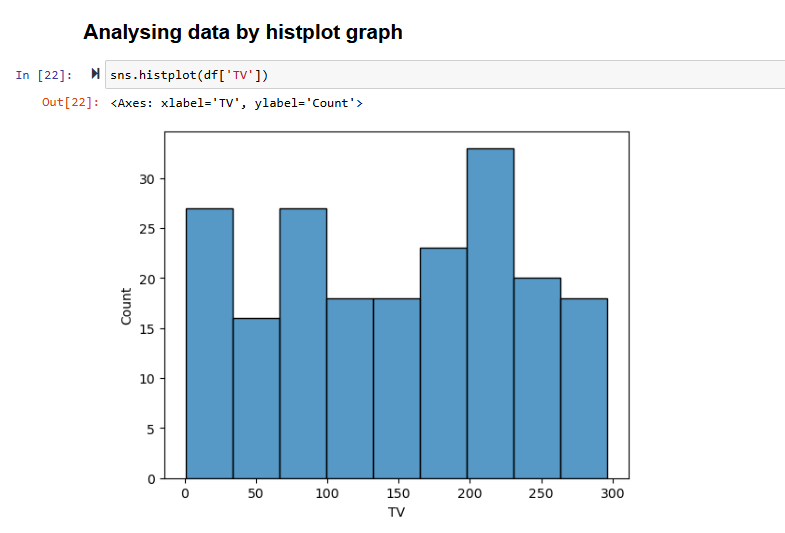


**2.5 Sales Prediction using Advertising Data**

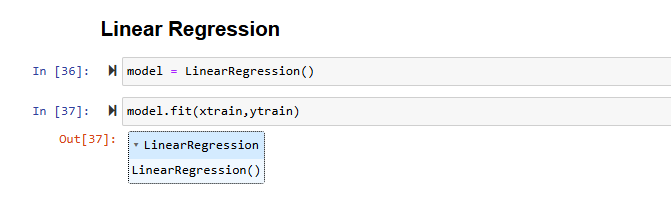
**Objective:** Predict sales based on advertising spends on TV, Radio, and Newspaper.

**Steps:**

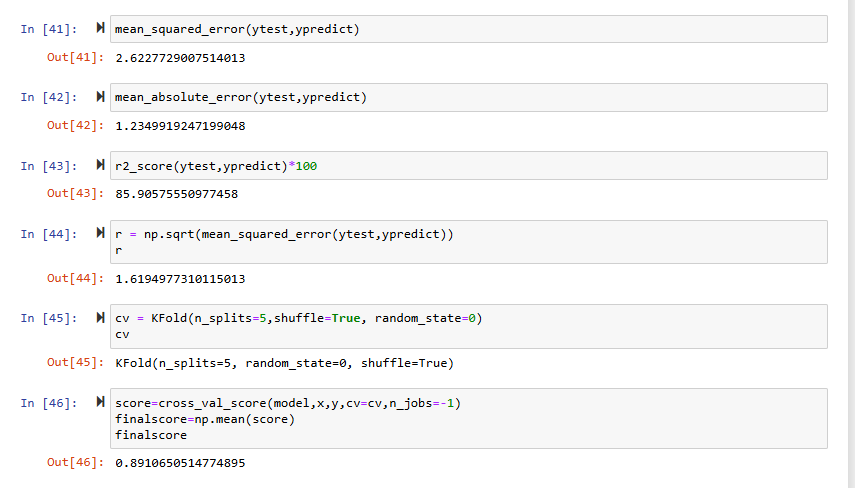
* Exploratory Data Analysis (Using various graphs)



* Model Building (Linear Regression)

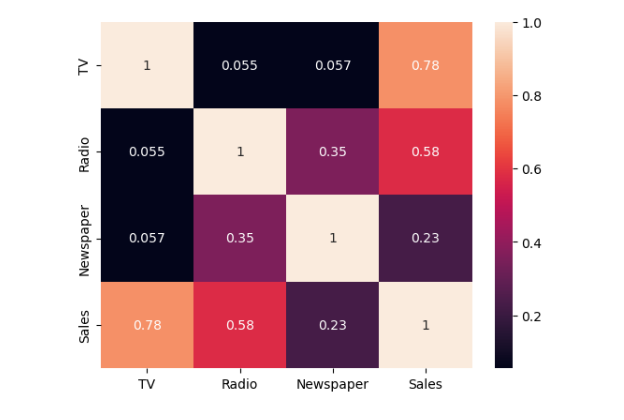


* Model Evaluation (Mean Squared Error, Mean Absolute Error, R-squared Score)



**Results:**

* Final Score was 89.10% for sales Prediction.
* Heatmap showing the overall sales using different source or mode of Advertisements.



**3. Key Takeaways**

* Iris Flower Classification

Through this exercise, I understood how to visualize data distributions and how different machine learning models can achieve varying levels of accuracy depending on the dataset. The importance of choosing the right algorithm for the right task was evident here.

* Unemployment Rate in India

This section enhanced my skills in time-series analysis, which is crucial for economic data.

* Car Price Prediction

This section taught me the importance of feature selection, where irrelevant features could degrade model performance. It also provided hands-on experience with regression models and evaluating performance using metrics like Mean Absolute Error and R-squared.

* Spam Mail Detection

This task highlighted the significance of text preprocessing and the use of text vectorization. I also learned the difference in performance between different classifiers when dealing with text-based classification problems.

* Sales Prediction using Advertising Data

This dataset allowed me to understand how multiple variables affect the target variable (sales). I learned how to deal with multicollinearity and use Linear regression to improve the model's robustness.

**4. CONCLUSION**

The process of applying data science techniques across these five datasets has really highlighted how powerful predictive analytics can be in different fields. Whether it was classifying flowers with the Iris dataset or forecasting unemployment trends in India, each project came with its own set of challenges and learning experiences. Using Python's data science libraries made it easier to handle the data, build models, and evaluate their performance effectively.

Throughout this journey, I’ve gained valuable hands-on experience with key techniques like classification, regression, and time-series forecasting. I also learned how crucial data visualization, preprocessing, and feature selection are in shaping accurate models. Each dataset revealed important insights, showing just how impactful data science can be when applied to solve real-world problems.

**References**

* [**https://www.coursera.org/learn/python-for-applied-data-science-ai**](https://www.coursera.org/learn/python-for-applied-data-science-ai)
* [**https://www.coursera.org/learn/what-is-datascience**](https://www.coursera.org/learn/what-is-datascience)
* [**https://www.w3schools.com/python**](https://www.w3schools.com/python)
* [**https://github.com/NiteshKrMandal/Oasis\_infobyte\_intership**](https://github.com/NiteshKrMandal/Oasis_infobyte_intership)
* [**https://www.canva.com/design/DAFSZMvBiCI/d8sj\_Z-sTrmOYK\_pTlvb9w/view?utm\_content=DAFSZMvBiCI&utm\_campaign=designshare&utm\_medium=link&utm\_source=publishsharelink#1**](https://www.canva.com/design/DAFSZMvBiCI/d8sj_Z-sTrmOYK_pTlvb9w/view?utm_content=DAFSZMvBiCI&utm_campaign=designshare&utm_medium=link&utm_source=publishsharelink#1)