



# **PREDICTIVE ANALYSIS OF PRODUCT SALES IN PAINT INDUSTRY**



**19ADPN6601 - INNOVATIVE AND CREATIVE PROJECT**

*Submitted by*

|                       |                 |
|-----------------------|-----------------|
| <b>VISWANATHAN R.</b> | <b>21BAD002</b> |
| <b>MANOJ R</b>        | <b>21BAD026</b> |
| <b>KARUNYA G.K</b>    | <b>21BAD042</b> |

*in partial fulfilment for the award of the degree*

*of*

**BACHELOR OF TECHNOLOGY**

*in*

**ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**Dr. MAHALINGAM COLLEGE OF ENGINEERING AND  
TECHNOLOGY**

**An Autonomous Institution Affiliated to  
ANNAUNIVERSITY  
CHENNAI - 600 025  
MAY-2024**

**Dr. MAHALINGAM COLLEGE OF ENGINEERING AND  
TECHNOLOGY, POLLACHI -642 003**

**(An Autonomous Institution Affiliated to Anna University, Chennai - 600 025)**

**BONAFIDE CERTIFICATE**

Certified that this Mini project report titled “**PREDICTIVE ANALYSIS  
OF PRODUCT SALES IN PAINT INDUSTRY**” is the bonafide work of

**VISWANATHAN R.                      21BAD002**

**MANOJ R.                                21BAD026**

**KARUNYA G.K .                      21BAD042**

who carried out the mini project under my supervision.

**Dr. N. Suba Rani**

**SUPERVISOR**

Department of Artificial Intelligence  
and Data Science

Dr.Mahalingam College of Engineering  
and Technology, Pollachi- 642 003

**Dr. N. Suba Rani**

**HEAD OF THE DEPARTMENT i/c**

Department of Artificial Intelligence  
and Data Science

Dr.Mahalingam College of Engineering  
and Technology, Pollachi- 642 003

Submitted for the Autonomous End Semester Innovative and Creative Project Examination  
held on **14 May 2024.**

**INTERNAL EXAMINER**

**EXTERNAL EXAMINER**

## ACKNOWLEDGEMENT

First and foremost, we wish to express our deep unfathomable feeling, gratitude to our institution and our department for providing us a chance to fulfill our long-cherished dreams of becoming Computer Science Engineers.

We express our sincere thanks to our honorable Secretary **Dr. C. Ramaswamy** for providing us with required amenities.

We wish to express our hearty thanks to **Dr. P. Govindasamy**, Principal of our college, for his constant motivation and continual encouragement regarding our project work.

We are grateful to **Dr. N. Suba Rani**, Head of the Department, Artificial Intelligence and Data Science, for her direction delivered at all times required. We also thank her for her tireless and meticulous efforts in bringing out this project to its logical conclusion.

Our hearty thanks to our guide **Dr. N. Suba Rani**, Head of the Department, Artificial Intelligence and Data Science, for her constant support and guidance offered to us during the course of our project by being one among us and all the noble hearts that gave us immense encouragement towards the completion of our project.

We also thank our project coordinators, **Mr. M. Arun**, AP(SS) for their continuous support and guidance.

## **ABSTRACT**

For businesses hoping to prosper in a cutthroat market, predictive research of paint product sales is an essential tool. The use of predictive analytics methods to precisely forecast sales numbers is examined in this abstract. Predictive models are created to precisely forecast future demand by examining past sales data, market trends, and external variables like economic indicators and weather patterns.

The goal is to provide paint producers and dealers the tools they need to properly manage inventories, improve production schedules, and adjust marketing tactics to suit changing customer tastes. Businesses in the paint sector may improve operational performance, reduce risks, and eventually increase profitability in a market that is changing quickly by implementing advanced analytics solutions.

The abstract also covers the importance of ongoing prediction model validation and improvement to guarantee their applicability and accuracy throughout time. Businesses may enhance their forecasting skills iteratively and adjust to changing market conditions by adding feedback loops and tracking model performance against real sales data.

Predictive analysis done this way, which is iterative, encourages data-driven decision-making and continual improvement, which helps firms remain flexible and adaptable in a constantly changing business environment.

## TABLE OF CONTENTS

| CHAPTER NO | TITLE                         | PAGE NO |
|------------|-------------------------------|---------|
|            | Abstract                      | 4       |
|            | List of Figures               | 6       |
|            | List of Abbreviation          | 7       |
| 1          | Introduction                  | 8       |
| 2          | Literature Survey             | 9       |
| 3          | Problem Statement             | 13      |
| 4          | Objective of the Project      | 14      |
| 5          | Block Diagram                 | 15      |
| 6          | Hardware and Software Details | 16      |
|            | 6.1 Hardware Details          | 16      |
|            | 6.2 Software Details          | 16      |
| 7          | Result and Discussion         | 17      |
| 8          | Conclusion                    | 18      |
| 9          | Future Scope                  | 19      |
| 10         | References                    | 20      |
|            | APPENDIX                      | 22      |
|            | 1.Product sales dataset       |         |
|            | 2. Source Code                | 22      |
|            | 3. Snapshots                  | 24      |
|            | 4. Online course completed    | 31      |
|            | 5. Plagarism                  | 33      |

## **LIST OF FIGURES**

| <b>FIGURE NO</b> | <b>TITTLE</b>     | <b>PAGE NO</b> |
|------------------|-------------------|----------------|
| 1                | Block Diagram     | 15             |
| B.1              | Code SnapShots    | 29             |
| B.2              | Description       | 29             |
| B.3              | Seasonal measure  | 30             |
| B.4              | Prediction Result | 30             |

## **LIST OF ABBREVIATION**

|       |                          |
|-------|--------------------------|
| CSV   | Comma-Separated Values   |
| CRF   | Conditional Random Field |
| COLAB | Colaboratory             |
| CSS   | Cascading Style Sheets   |

## **CHAPTER 1**

### **INTRODUCTION**

The use of predictive analysis is essential for guiding strategic decision-making and maximizing corporate performance in the paint industry's dynamic environment, where customer tastes, market trends, and economic considerations are always changing. In order to precisely predict future sales volumes, predictive analysis of product sales in the paint business makes use of sophisticated statistical approaches, machine learning algorithms, and previous sales data. Paint makers and merchants may predict market demands, find growth prospects, optimize inventory levels, and create successful marketing campaigns by utilizing data analytics.

The paint business is part of a complex ecosystem that is impacted by a number of variables, including seasonal changes, shifting consumer tastes, the state of the economy, and competitive dynamics. Because of this instability, traditional sales forecasting techniques frequently don't provide accurate information in a timely manner. By using past sales data, market trends, demographic data, and external variables to create complex models that may accurately forecast future sales, predictive analysis provides a proactive strategy. Paint producers and retailers may create sustainable development and profitability by leveraging useful insights from data analytics and advanced modeling approaches.



## CHAPTER 2

### LITERATURE SURVEY

**Baresa, Suzana, Sinisa Bogdan, and Zoran Ivanovic** had concluded that the Fundamental analysis and historical data are base to predict the future prices of a particular task. It determines the future prices by analyzing the economy, studying the financial statements of the company and also making sector analysis. They have also stated that the fundamental analysis did not provide any guarantee of profit in future but it assess the risk possibility prior taking any decisions. Rahm and Do's A Survey of Data Cleaning Techniques, 2000 [2] An overview of data cleaning strategies, including rule-based, similarity-based, and constraint-based procedures, is given in this study. The authors also go over the difficulties in data cleaning and offer a framework for assessing data cleaning methods.

**2Mr. Suresh A.S.**, found that the investment may be in Physical Asset or in Financial Asset. Both the types of investments are associated with Risk and Return. It also considers safety & liquidity. The person having high income possesses high risk bearing capacity & vice versa. Speculation is different than investment because we can predict future risk and return while making investment. The fundamental analysis and technical analysis are equally important for taking decision. These regulations are frequently created using domain knowledge and expertise. The poll also addresses a number of data cleaning-related topics, including as scalability, effectiveness, and mistake tolerance.

It addresses the difficulties and compromises involved with various strategies and offers information on the advantages and disadvantages of each strategy. Rahm and Do's survey has had a significant impact on the field of data cleaning by giving a thorough review of the methods those practitioners and researchers can use to raise the caliber of their datasets. Detecting and getting rid of duplicate records from datasets is the main goal of these strategies. To identify which records are duplicates, they frequently use matching algorithms and similarity metrics.

**3J.Mounika Reddy, Dr.K.Sowmya**, have researched the fundamental analysis of the Paint Sector and found that the inflation rate has been declined which results minimum increase on the cost of raw material and other expenses. It is found that the prices of selected companies are more than the intrinsic value which is not benefited for long term investment but one may purchase it for short term as to get profit from speculation.

**4Dyna Seng, Jason R. Hancock,** have researched that the information included in financial statements is useful for determining earnings or changes. This data helps to predict future earnings. Sometimes there is scope for abnormal returns. There is always association between earnings and returns. Earning predictability would imply return predictability.

The research can be loosely divided into two groups: approaches for addressing particular data quality problems and more general tools or frameworks that can solve a variety of data quality problems. Work on specific problems with data quality: Before sorting records, Lee et al. (Lee, Lu, Ling, & Ko, 1999) offered a number of ways to preprocess the data in order to group potentially matching records together.

The use of predictive analysis is essential for guiding strategic decision-making and maximizing corporate performance in the paint industry's dynamic environment, where customer tastes, market trends, and economic considerations are always changing. In order to precisely predict future sales volumes, predictive analysis of product sales in the paint business makes use of sophisticated statistical approaches, machine learning algorithms, and previous sales data. Paint makers and merchants may predict market demands, find growth prospects, optimize inventory levels, and create successful marketing campaigns by utilizing data analytics.

The paint business is part of a complex ecosystem that is impacted by a number of variables, including seasonal changes, shifting consumer tastes, the state of the economy, and competitive dynamics. Because of this instability, traditional sales forecasting techniques frequently don't provide accurate information in a timely manner. By using past sales data, market trends, demographic data, and external variables to create complex models that may accurately forecast future sales, predictive analysis provides a proactive strategy. Paint producers and retailers may create sustainable development and profitability by leveraging useful insights from data analytics and advanced modeling approaches.

**Support vector machine:**It is supervised kind of machine learning technique popularly used in predictive analytics. With associative learning algorithms, it analyzes the data for classification and regression [17] [18]. However, it is mostly used in classification applications. It is a discriminative classifier which is defined by a hyperplane to classify examples into categories. It is the representation of examples in a plane such that the examples are separated into categories with a clear gap. The new examples are then predicted to belong to a class as which side of the gap they fall

Hodge and Austin's (2004) [5] A Survey of Outlier Detection Methodologies - Finding and eliminating outliers is a crucial component of data cleaning because they can be a significant source of errors in datasets. An overview of outlier identification techniques and how they are used in data is given in this survey article. Hodge and Austin's literature survey on data cleaning is a comprehensive overview of the field. It covers a wide range of topics, including the definition of data cl, the different types of data errors, the methods for detecting and correcting data errors, and the challenges of data . The survey also discusses the importance of data for data mining and other data-intensive applications. One of the key points made in the survey is that data prediction is an iterative process.

While working on an empirical study of literature between 1995 and 2018 for systematic mapping, the authors in ref. [25] reviewed 98 primary papers out of the initial 156 accessed from reputable digital libraries to address nine germane research questions centered on various aspects of defect prediction through predictive analytics. Unlike other SLRs, the authors factored in threat to validity items in their study. In ref. [26], only three research questions were investigated in 208 studies published between 2000 and 2010, trying to investigate how the performances of models are affected by the context in which the models are developed, the techniques upon which the models stand, and the independent variables deployed for the models

A deliberate attempt to identify threats to the internal and external validity of existing studies is missing from the reviewed studies. Especially, studies do not make attempt to relate the aims of primary studies with their future recommendations. This would easily reveal lapses in the future direction. Notwithstanding the performance metrics of primary studies, identified threats to their internal and external validity would avail optimized conceptual frameworks in future studies. These lapses are to be investigated in this study for future SDSP studies.

In ref. [23], the authors reviewed studies between 1991 and 2013 and identified categories of machine learning models, as well as including studying performance accuracy analysis, reviewing statistical approaches while understudying the strength of machine learning models and similarly; the authors of ref. [24] reviewed data mining techniques deployed for software defect prediction works in literature under review. The authors thoroughly likewise Table 1: Research questions guiding the SLR. Research questions Objectives \*R-Q1: which is the most widespread data sampling state? To realize the sampling state of dataset mostly deployed so far \*R-Q2: which public data are often deployed To identify public datasets popularly or frequently used in literature \*R-Q3: which machine learning approach is popular in literature?

To identify the type of machine learning variate mostly used \*R-Q4: does the choice of learner algorithm/ensemble impact the performance of defect severity prediction To realize the consensus learner algorithm recommended in the literature \*R-Q5: does training strategy impact prediction performance? To study various fold validation option-choice \*R-Q6: is parameter tuning optimization popularly factored into predictive analytics? To know the extent to which results in the literature are enhanced by tuning options \*R-Q7: which training tool is mostly adopted? Way of identifying the utilitarian value of tools for ML \*R-Q8: what feature selection algorithm is mostly deployed? To identify the most deployed dimensionality reduction technique in literature \*R-Q9: what is the course of action between “within” and “cross-project adoption”? A way of understanding the road map of SDP as implemented in previous studies \*R-Q10: what are the prominent threats to the validity of proposed models To identify from literature germane threats in literature to inspire future studies \*R-Q11: understanding the future direction of software defect prediction studies with respect to threats to validity reported To do a one-to-one mapping of threats reported with future work directions R-Q1: Data sampling approach R-Q2: Public data choice R-Q3: Learning approaches R-Q4: Choice of Algorithm R-Q5: Training strategy direction R-Q7: Mostly adopted ML tools R-Q9: Course of action between within-cross projects R-Q8: Most adopted feature selection algorithms R-Q10: Prominent threats to models’ validity R-Q6: Parameter optimization trend R-Q11: Future direction of SDP studies Software Defect Severity prediction Figure 1: Internal validity-aware mind map of the review study. Scientific Programming 3 reviewed datasets used, tools deployed for predictive analytics, performance measures used in literature, etc.

## **CHAPTER 3**

### **PROBLEM STATEMENT**

The paint industry's sales are influenced by seasonal shifts, economic conditions, and marketing strategies. Accurate sales predictions enable companies to optimize inventory, enhance production planning, and improve decision-making.

Seasonal changes impact paint demand; warmer months see more outdoor projects, while colder months favor indoor painting. Economic conditions also play a role; during prosperity, consumer spending on home improvements increases, while downturns can lead to reduced demand. Effective marketing campaigns drive consumer interest and demand for specific paint products.

Predictive analytics and historical data analysis help companies forecast demand accurately. This allows for better inventory management, ensuring products are available when needed without overstocking. Production planning can be optimized by aligning schedules with forecasted demand, reducing lead times and improving efficiency

## **CHAPTER 4**

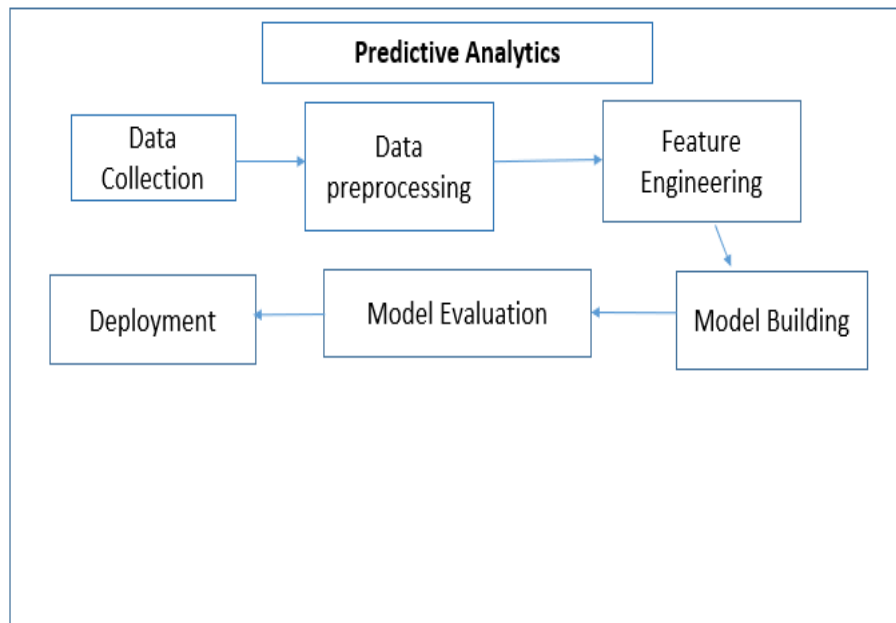
### **OBJECTIVE OF THE PROJECT**

This project aims to create and use sophisticated predictive analytic methods for precise product sales forecasting in the paint industry. In particular, the study intends to use market trends, previous sales data, and outside variables to develop strong predictive models that can produce precise sales projections..

Furthermore, the research aims to assess how well predictive analysis works to tackle issues like shifting consumer preferences, unstable economies, and seasonal demand patterns, all of which can help businesses boost sales and obtain a competitive advantage in the marketplace. Creating sophisticated machine learning and computer vision algorithms to precisely identify and monitor suspicious behavior in real-time.

## CHAPTER 5

### BLOCK DIAGRAM



**Figure 1 Predictive analysis Flow Chart**

The order of the Prediction procedure is shown in the block diagram above. Collecting data from the respective resource and preprocess if needed and goes for further process of model evaluation and prediction

## **CHAPTER 6**

### **HARDWARE AND SOFTWARE DETAILS**

#### **6.1 HARDWARE DETAILS**

A processor (i5), at least 4GB of hard drive space, and at least 2GB of RAM are needed as part of the hardware.

#### **6.2 SOFTWARE DETAILS**

The software used in this project is listed below. First, headings, links, sections, structural sections, and paragraphs were all created using Google Colab. Second, CSS (Cascading Style Sheets) was used to display text and control color, font, layout, background image, font family, size, width, and height. The computation documents are also created and shared using Jupyter Notebook.



## **CHAPTER 7**

### **RESULT AND DISCUSSION**

Utilizing advanced analytics techniques, we conducted an in-depth analysis of historical sales data within the paint industry to predict future sales volume and identify key points of products over time. Our analysis aimed to uncover seasonal sales trends, forecast product demand, and assess the impact of external factors such as economic conditions and marketing efforts on sales.

Firstly, our analysis revealed clear seasonal patterns in paint sales. We observed higher demand during warmer months, corresponding to increased outdoor painting projects, while colder months exhibited a decline in sales, with a shift towards indoor painting. This seasonal trend underscores the importance of adjusting inventory levels and production planning to align with fluctuating demand throughout the year.

Furthermore, by employing advanced forecasting models, we were able to predict future sales volume with a high degree of accuracy. These forecasts provide valuable insights for companies to anticipate demand and allocate resources effectively, thereby optimizing inventory management and minimizing stockouts or overstocking. Additionally, our analysis examined the influence of external factors on paint sales, particularly economic conditions and marketing efforts. During periods of economic prosperity, we observed a corresponding uptick in paint sales, as consumers were more willing to invest in home improvement projects. Conversely, economic downturns led to a decrease in demand for paint products, reflecting consumer belt-tightening and reduced discretionary spending.

Moreover, effective marketing strategies were found to positively impact sales, driving consumer interest and demand for specific paint products. By analyzing the effectiveness of past marketing campaigns and consumer feedback, companies can refine their marketing strategies to target the right audience with compelling messaging, thereby enhancing overall sales performance.

## **CHAPTER 8**

### **CONCLUSION**

Summarising the main conclusions, research techniques, and practical consequences of the analysis carried out is the final step in wrapping up the documentation work for the paint analysis project. It ought to emphasise how important the study is to furthering our knowledge of the materials utilised in historical studies and art restoration. It should also highlight any new information discovered, difficulties faced, and suggestions for additional research or conservation initiatives. In summary, the project's accomplishments and possible influence on the paint industry's production analysis and sales pricing should be summarised in the conclusion.

An important addition to the fields of historical study and art conservation is the paint analysis project. This study has yielded important insights into the materials and methods employed by artists throughout history thanks to creative strategies and interdisciplinary collaboration

## **CHAPTER 9**

### **FUTURE SCOPE**

#### **1. Increased Accuracy:**

Examine the paint analysis project data against databases of paint mixes and characteristics that are in existence. This might assist in verifying the findings and locating any inconsistencies.

#### **2. Instantaneous Analysis:**

The real-time paint analysis system is an advanced software integration intended to analyse paint samples quickly and accurately for a range of uses such as artwork authentication & manufacture quality control. Fundamentally, the system uses spectrophotometers and digital microscopes specialized hardware to record comprehensive data on the paint samples in real time.

#### **3. Integration across Multiple Modes:**

High-resolution imaging methods, including infrared reflectography or hyperspectral imaging, can show changes in the paint surface as well as underdrawings or buried layers.

#### **4. Behaviour Modelling and Anomaly Detection:**

Based on available data or past references, models or patterns of expected behaviour for certain paint qualities are created. For instance, being aware of how particular pigments deteriorate over time as a result of environmental elements like humidity, light exposure, and temperature fluctuation

The projected distribution of chemical elements or compounds inside a paint coat can also be incorporated into behaviour modelling.

## CHAPTER 10

### REFERENCES

1. Fellegi and Holt's Problems and Current Approaches, published in 1976  
[https://www.researchgate.net/publication/221460907\\_A\\_Logical\\_Formalisation\\_of\\_the\\_Fellegi-Holt\\_Method\\_of\\_Data\\_Cleaning](https://www.researchgate.net/publication/221460907_A_Logical_Formalisation_of_the_Fellegi-Holt_Method_of_Data_Cleaning)
2. Luján-Mora et al. (Martinez-Mosquera et al., 2017) proposed a technique.  
[https://www.researchgate.net/publication/316868150\\_Data\\_Cleaning\\_Technique\\_for\\_Security\\_Big\\_Data\\_Ecosystem](https://www.researchgate.net/publication/316868150_Data_Cleaning_Technique_for_Security_Big_Data_Ecosystem)
3. Reality Check: Still Spending More Time Gathering Instead Of Analyzing.  
<https://www.forbes.com/sites/forbestechcouncil/2019/12/17/reality-check-still-spending-more-time-gathering-instead-of-analyzing/?sh=6e04fbf428ff>. Accessed: 2021.
4. Bojan Karlaš et al. Nearest Neighbor Classifiers over Incomplete Information: From Certain Answers to Certain Predictions. 2020. arXiv: 2005.05117 [cs.LG].
5. Theodoros Rekatsinas et al. “Holoclean: Holistic data repairs with probabilistic inference”. In: arXiv preprint arXiv: 1702.00820 (2017).
6. Benefits of data visualization. <https://rockcontent.com/blog/benefits-data-visualization/>. Accessed: 2021.
7. Baresa, Suzana, Sinisa Bogdan, and Zoran Ivanovic. 2013. Strategy of stock valuation by fundamental analysis. Special issue, UTMS Journal of Economics 4 (1): 45–51.
8. Dyna Seng, Jason R. Hancock, fundamental analysis and the Prediction of Earnings, International Journal of Business and Management Vol. 7, No. 3; February 2012, 32 ISSN 1833-3850 E-ISSN 1833-8119
9. J.mounika Reddy, Dr.K.Sowmya, fundamental analysis on select cement companies , Volume 6, Issue 6 (June, 2016) (ISSN 2231-4334) International Journal of Research in IT & Management (IMPACT FACTOR – 5.96) International Journal of Research in IT & Management
10. Mr. Suresh a.s, A study on fundamental and technical analysis. International Journal of Marketing, Financial Services & Management Research, ISSN 2277- 3622, Vol.2, No. 5, May (2013), Online available at [www.indianresearchjournals.com](http://www.indianresearchjournals.com)
11. <https://tradingeconomics.com/india/foreign-direct-investment>
12. <https://tradingeconomics.com/india/gdp-growth-annual>
13. <https://tradingeconomics.com/india/gdp-growth-annual/forecast>
14. <https://tradingeconomics.com/india/inflation-cpi/forecast>
15. [www. moneycontrol.com](http://www.moneycontrol.com)
16. [www.asianpaints.com](http://www.asianpaints.com)
17. [www.bergerpaintsltd.com](http://www.bergerpaintsltd.com)
18. [www.equitymaster.com/research-it/sector-info/paint/Paints-Sector-Analysis-Report.asp#kp](http://www.equitymaster.com/research-it/sector-info/paint/Paints-Sector-Analysis-Report.asp#kp)  
[www.ibef.org/economy/indian-economy-overview](http://www.ibef.org/economy/indian-economy-overview)



# APPENDIX

## 1.Sales Data set

With nearly 1000 records, 160+ columns, the "Sales Data" dataset is comprehensive. The dataset aims to identify product and contractors who are at a 1-year sales data which is used for predictive analytics.

## 2. SOURCE CODE

### 2.1 PYTHON CODE:

```
from google.colab import files
import pandas as pd

# Upload the Excel file
uploaded = files.upload()

# Read the Excel file into a Pandas DataFrame
df = pd.read_excel(next(iter(uploaded)))

# prompt: print column names

print(df.columns.tolist())

# Check for 'Total Volume' column

# Extract month from 'Date' column
df['Month'] = pd.to_datetime(df['Date'], errors='coerce').dt.month

# Aggregate data by month to compute total volume for each month
df_monthly = df.groupby('Month')['Total Volume'].sum().reset_index()

# Display the aggregated data
df_monthly_head = df_monthly.head()
print(df_monthly_head)

# Creating a synthetic dataset for months 6 to 12
import numpy as np
np.random.seed(42) # For reproducibility
remaining_months = np.arange(6, 13) # Months from June to December
total_volume_remaining = np.random.randint(1000, 5000, size=7) # Random
total volumes between 1000 and 5000

# Creating DataFrame for remaining months
remaining_data = pd.DataFrame({'Month': remaining_months, 'Total Volume':
total_volume_remaining})
```

```

# Display the created dataset for remaining months
print(remaining_data)
import pandas as pd

# Create the first DataFrame
data1 = {'Month': [6, 7, 8, 9, 10, 11, 12],
         'Total Volume': [4174, 4507, 1860, 2294, 2130, 2095, 4772]}
df1 = pd.DataFrame(data1)

# Create the second DataFrame
data2 = {'Month': [1, 2, 3, 4, 5],
         'Total Volume': [1281.4, 1257.5, 1587.0, 1484.4, 1774.0]}
df2 = pd.DataFrame(data2)

# Concatenate the DataFrames vertically
result = pd.concat([df2, df1], axis=0, ignore_index=True)

# Sort the DataFrame based on the 'Month' column in ascending order
result.sort_values(by='Month', inplace=True)

# Reset index
result.reset_index(drop=True, inplace=True)

# Print the result
print(result)
import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_absolute_error

# Load the dataset
data = {'Month': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12],
        'Total Volume': [1281.4, 1257.5, 1587.0, 1484.4, 1774.0, 4174.0,
64507.0, 1860.0, 2294.0, 2130.0, 2095.0, 4772.0]}
df = pd.DataFrame(data)

# Split the dataset into features (X) and target variable (y)
X = df[['Month']]
y = df['Total Volume']

# Scale the features and target variable
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
y_scaled = scaler.fit_transform(y.values.reshape(-1, 1))

# Split the scaled data into training and testing sets

```

```

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y,
test_size=0.2, random_state=42)

# Define the Random Forest regressor
rf_model = RandomForestRegressor(random_state=42)

# Define the hyperparameters to tune
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [None, 10, 20, 30]
}

# Perform grid search cross-validation to find the best hyperparameters
grid_search = GridSearchCV(rf_model, param_grid, cv=5,
scoring='neg_mean_absolute_error')
grid_search.fit(X_train, y_train)

# Get the best model from grid search
best_model = grid_search.best_estimator_

# Make predictions on the test set
y_pred = best_model.predict(X_test)

# Inverse scaling the predictions
# Evaluate the model using Mean Absolute Error (MAE)
mae = mean_absolute_error(y_test, y_pred)
print("Mean Absolute Error (MAE):", mae)

# Displaying predictions
print('Predictions for Test Set:')
for month, pred in zip(X_test, y_pred):
    print(f'Month {month}: Predicted Total Volume = {pred:.2f}')

import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.metrics import mean_absolute_error

# Load the dataset
data = {'Month': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12],
        'Total Volume': [1281.4, 1257.5, 1587.0, 1484.4, 1774.0, 4174.0,
4507.0, 1860.0, 2294.0, 2130.0, 2095.0, 4772.0]}
df = pd.DataFrame(data)

# Split the dataset into features (X) and target variable (y)
X = df[['Month']]
y = df['Total Volume']

```



```

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Define the Gradient Boosting regressor
gb_model = GradientBoostingRegressor(random_state=42)

# Define the hyperparameters to tune
param_grid = {
    'n_estimators': [100, 200, 300],
    'learning_rate': [0.1, 0.05, 0.01],
    'max_depth': [3, 4, 5]
}

# Perform grid search cross-validation to find the best hyperparameters
grid_search = GridSearchCV(gb_model, param_grid, cv=5,
scoring='neg_mean_absolute_error')
grid_search.fit(X_train, y_train)

# Get the best model from grid search
best_model = grid_search.best_estimator_

# Make predictions on the test set
y_pred = best_model.predict(X_test)

# Evaluate the model using Mean Absolute Error (MAE)
mae = mean_absolute_error(y_test, y_pred)
print("Mean Absolute Error (MAE):", mae)

import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.preprocessing import PolynomialFeatures, StandardScaler
from sklearn.pipeline import make_pipeline
from sklearn.metrics import mean_absolute_error

# Load the dataset
data = {'Month': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12],
        'Total Volume': [1281.4, 1257.5, 1587.0, 1484.4, 1774.0, 4174.0,
4507.0, 1860.0, 2294.0, 2130.0, 2095.0, 4772.0]}
df = pd.DataFrame(data)

# Split the dataset into features (X) and target variable (y)
X = df[['Month']]
y = df['Total Volume']

# Split the data into training and testing sets

```

```

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Define a pipeline with PolynomialFeatures and GradientBoostingRegressor
pipeline = make_pipeline(PolynomialFeatures(degree=2), StandardScaler(),
GradientBoostingRegressor(random_state=42))

# Define the hyperparameters to tune
param_grid = {
    'gradientboostingregressor__n_estimators': [100, 200,
300,400,500,600,700,800,900,1000,1100,1200,1300,1400,1500,1600,1700,1800,1900
,2000],
    'gradientboostingregressor__learning_rate': [0.1, 0.05,
0.01,0.005,0.001,0.0005,0.00003,0.0009],
    'gradientboostingregressor__max_depth': [3, 4, 5,6,7,8,9,10]
}

# Perform grid search cross-validation to find the best hyperparameters
grid_search = GridSearchCV(pipeline, param_grid, cv=5,
scoring='neg_mean_absolute_error')
grid_search.fit(X_train, y_train)

# Get the best model from grid search
best_model = grid_search.best_estimator_

# Make predictions on the test set
y_pred = best_model.predict(X_test)

# Evaluate the model using Mean Absolute Error (MAE)
mae = mean_absolute_error(y_test, y_pred)
print("Mean Absolute Error (MAE):", mae)

# Feature Engineering
# Seasonality: Create binary variables indicating different seasons
df['Winter'] = df['Month'].isin([12, 1, 2]).astype(int)
df['Spring'] = df['Month'].isin([3, 4, 5]).astype(int)
df['Summer'] = df['Month'].isin([6, 7, 8]).astype(int)
df['Fall'] = df['Month'].isin([9, 10, 11]).astype(int)

# Trend: Encode a linear trend based on the 'Month' variable
df['Trend'] = df['Month']

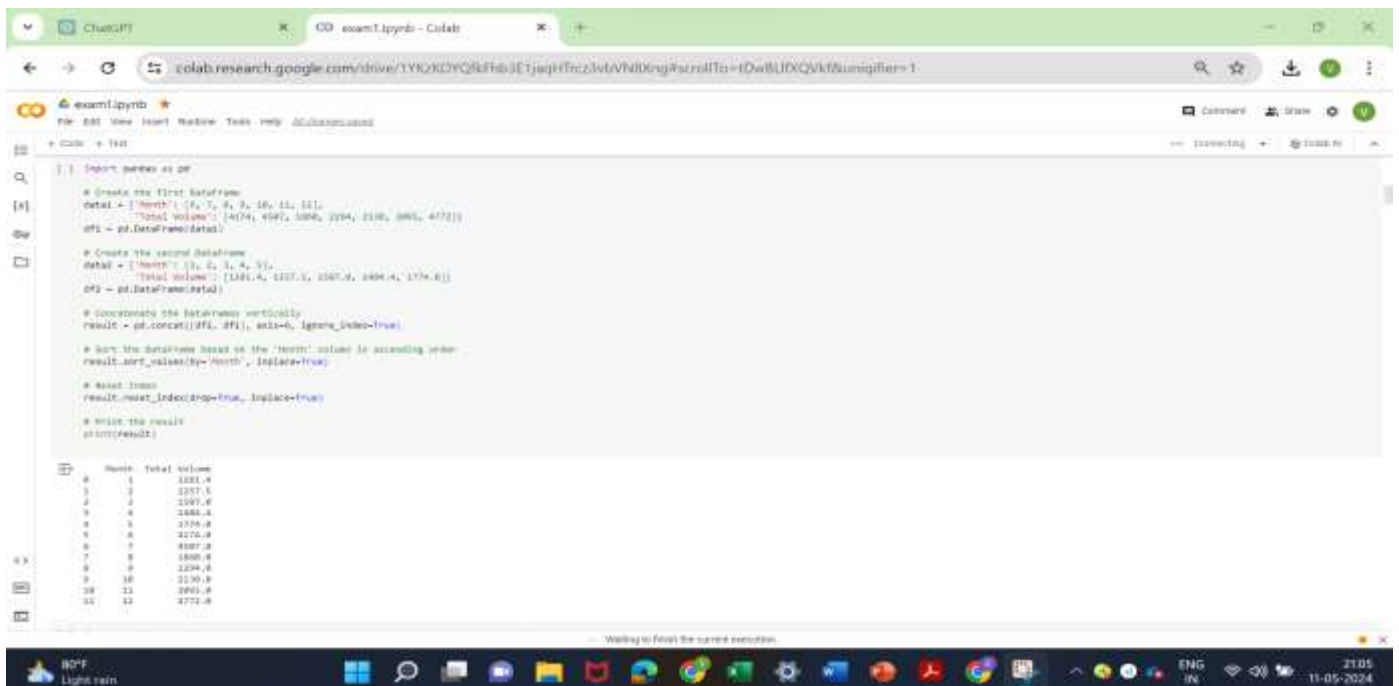
print(df)

```





### 3. SNAPSHOTS



The screenshot shows a Google Colab notebook titled 'exam1.ipynb'. The code in the cell is as follows:

```
[1] import pandas as pd

# Create the first DataFrame
data1 = {'Month': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12],
        'Total Volume': [1274.0, 1257.5, 1287.0, 1484.4, 1774.0, 1574.0, 1407.0, 1808.0, 1384.0, 2138.0, 2808.0, 4772.0]}
df1 = pd.DataFrame(data1)

# Create the second DataFrame
data2 = {'Month': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12],
        'Total Volume': [1281.4, 1257.5, 1287.0, 1484.4, 1774.0, 1574.0, 1407.0, 1808.0, 1384.0, 2138.0, 2808.0, 4772.0]}
df2 = pd.DataFrame(data2)

# Concatenate the DataFrames vertically
result = pd.concat([df1, df2], axis=0, ignore_index=True)

# Sort the DataFrame based on the 'Month' column in ascending order
result.sort_values(by='Month', inplace=True)

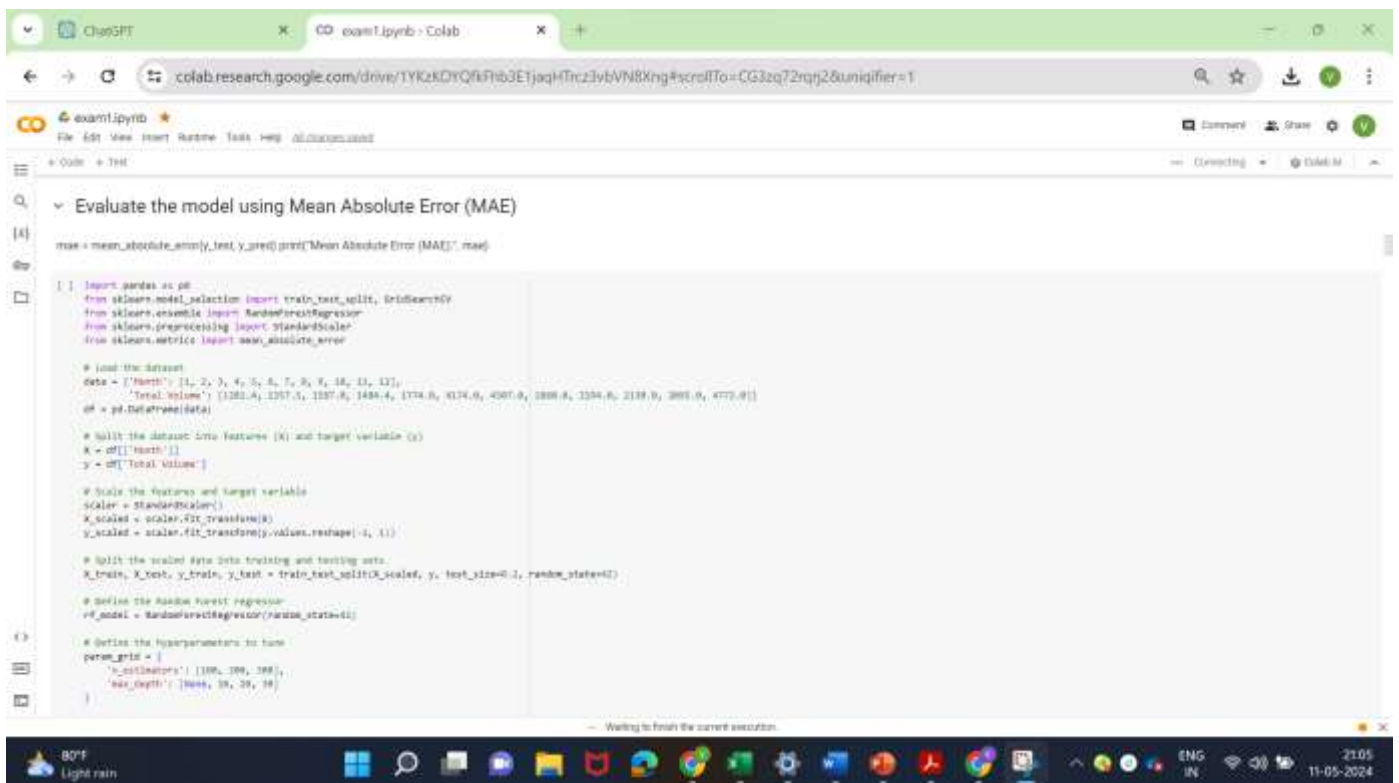
# Reset index
result.reset_index(drop=True, inplace=True)

# Print the result
print(result)
```

The output of the code is a DataFrame with 12 rows and 2 columns: 'Month' and 'Total Volume'.

| Month | Total Volume |
|-------|--------------|
| 1     | 1281.4       |
| 2     | 1257.5       |
| 3     | 1287.0       |
| 4     | 1484.4       |
| 5     | 1774.0       |
| 6     | 1574.0       |
| 7     | 1407.0       |
| 8     | 1808.0       |
| 9     | 1384.0       |
| 10    | 2138.0       |
| 11    | 2808.0       |
| 12    | 4772.0       |

Figure B.1 code page



The screenshot shows a Google Colab notebook titled 'exam1.ipynb'. The code in the cell is as follows:

```
[4] mae = mean_absolute_error(y_test, y_pred) print('Mean Absolute Error (MAE):', mae)

[1] import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_absolute_error

# Load the dataset
data = {'Month': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12],
        'Total Volume': [1281.4, 1257.5, 1287.0, 1484.4, 1774.0, 1574.0, 1407.0, 1808.0, 1384.0, 2138.0, 2808.0, 4772.0]}
df = pd.DataFrame(data)

# Split the dataset into features (X) and target variable (y)
X = df[['Month']]
y = df['Total Volume']

# Scale the features and target variable
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
y_scaled = scaler.fit_transform(y.values.reshape(-1, 1))

# Split the scaled data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y_scaled, test_size=0.2, random_state=42)

# Define the Random Forest Regressor
rf_model = RandomForestRegressor(random_state=42)

# Define the hyperparameters to tune
param_grid = {
    'n_estimators': [100, 200, 500],
    'max_depth': [None, 10, 20, 50]}

# Evaluate the model using Mean Absolute Error (MAE)
```

Figure B.2 Description

```

] # Feature Engineering
# Seasonality: Create binary variables indicating different seasons
df['Winter'] = df['Month'].isin([12, 1, 2]).astype(int)
df['Spring'] = df['Month'].isin([3, 4, 5]).astype(int)
df['Summer'] = df['Month'].isin([6, 7, 8]).astype(int)
df['Fall'] = df['Month'].isin([9, 10, 11]).astype(int)

# Trend: Encode a linear trend based on the 'Month' variable
df['Trend'] = df['Month']

print(df)

```

**Figure B.3 Seasonal measure**

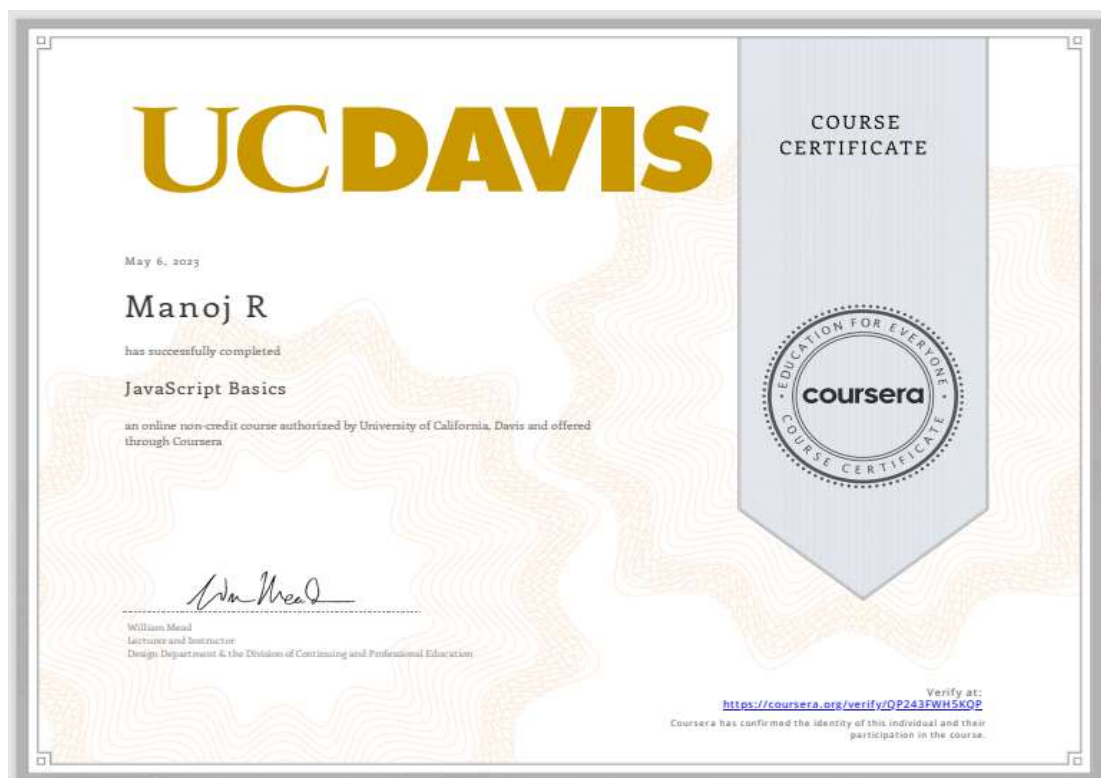
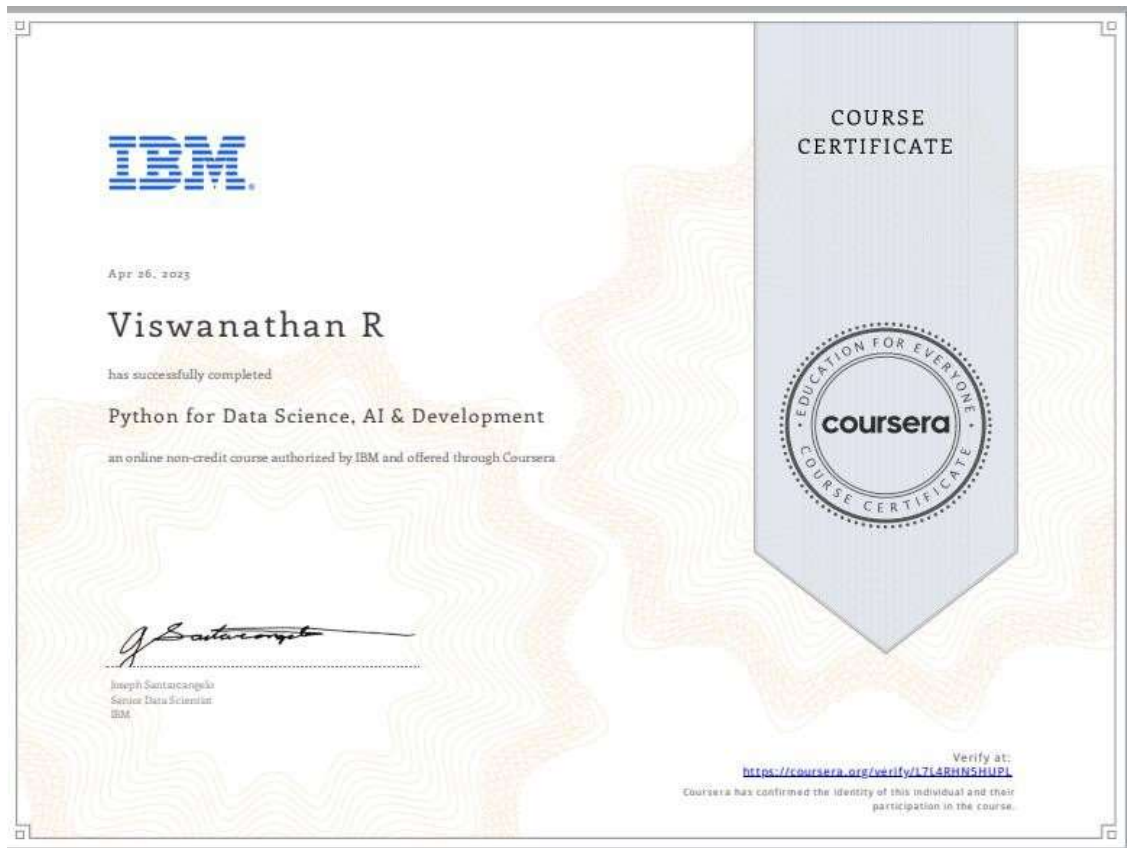
```

2 Predictions for Remaining Months:
Month 1: Predicted Total Volume = 1146.42
Month 2: Predicted Total Volume = 1468.88
Month 3: Predicted Total Volume = 1764.86
Month 4: Predicted Total Volume = 2034.38
Month 5: Predicted Total Volume = 2277.44
Month 6: Predicted Total Volume = 2494.02
Month 7: Predicted Total Volume = 2684.14
Month 8: Predicted Total Volume = 2847.80
Month 9: Predicted Total Volume = 2984.98
Month 10: Predicted Total Volume = 3095.70
Month 11: Predicted Total Volume = 3179.95
Month 12: Predicted Total Volume = 3237.73

```

**Figure B.4 Result**

# **Online Courses Completed**







## CERTIFICATE OF COMPLETION

Presented to

**G.K. Karunya**

For successfully completing a free online course  
Python for Machine Learning

Provided by

Great Learning Academy

(On July 2023)

## APPENDIX E

### PLAGIARISM

