Suv Prediction using Machine Learning in Python

A MINOR-I PROJECT REPORT

Submitted by

SUHAIL RIZVI 2020-310-221

in partial fulfilment for the award of the degree of

B. TECH COMPUTER SCIENCE & ENGINEERING

Under the supervision of

Ms.Pooja Gupta

Department of Computer Science & Engineering
School of Engineering Sciences & Technology

JAMIA HAMDARD

New Delhi-110062

(2023)

DECLARATION

I, **Mr. Suhail Rizvi** a student of **Bachelors of Technology Computer Science & Engineering (B. Tech CSE), Enrolment No: 2020-310-221)** hereby declare that the Project/Dissertation entitled "**Suv Prediction using Machine Learning in Python"** which is being submitted by me to the Department of Computer Science, Jamia Hamdard, New Delhi in partial fulfilment of the requirement for the award of the degree of **Bachelors of Technology Computer Science & Engineering (B. Tech CSE),** is my original work and has not been submitted anywhere else for the award of any Degree, Diploma, Associateship, Fellowship or other similar title or recognition.

SUHAIL RIZVI

2020-310-221

Date: April 2023

Place: New Delhi

ACKNOWLEDGEMENT

I express my sincere thanks to Ms. Pooja Gupta (Dept. of Computer Science and Engineering), my project in charge, who guided me through the project for her valuable suggestions and guidance for completing the project. This project has been a success only because of her guidance.

I deeply express my sincere thanks to our Head of Department **Dr. Farheen Siddiqui** for encouraging me and allowing me to present the project on the topic "**Suv Prediction using Machine Learning in**

Python" at the department premises for the partial fulfilment of the requirements leading to the award of B.Tech degree.

I am also thankful to the whole computer science and engineering department for providing the technical support to carry out the project work, letting us utilize all the necessary facilities of the institute and providing guidance at each& every step during the project work.

Suhail Rizvi
2020-310-221

INDEX

OBJECTIVE

The objective of this machine learning project is to develop a predictive model that can accurately predict the likelihood of suv prediction in employees age and salary using a employees dataset. The dataset contains 400 observations and five variables - Employee ID, Gender, Age, EstimatedSalary, and Purchased (dependent variable). The Purchased variable is binary, indicating whether an employee has purchased an SUV or not. The Age variable ranges from 18 to 60, and the Salary variable . We also visualize the results using graphs and charts to gain insights into the predictions.

The aim is to build a machine learning model that can accurately predict whether an employee will purchase an SUV based on their age and salary.
Machine learning analytics in manufacturing field is one of the major research fields in Information Technology with a lot of challenges. The goal of this research is to design a categorical solution to decide whether a employee is eligible and interested to purchase a sport-utility vehicle (SUV) based on the available data from the previous records collected from the banks.

The project may also involve feature selection, model evaluation, and optimization to achieve the highest possible accuracy and interpretablity in the predictions.

INTRODUCTION

In recent years, the market for sports utility vehicles (SUV) has grown significantly. One of the key factors that influence SUV sales is the income level of potential buyers. In this project, we aim to predict SUV purchases based on the employee status, age, and salary of potential buyers.

The dataset we will be using for this project contains information on the employee status, age, salary, and SUV purchase of a group of individuals. Our objective is to build a machine learning model that can predict whether or not a person will purchase an SUV based on their employee status, age, and salary.

To achieve this, we will use various machine learning algorithms such as logistic regression, decision trees, and random forests. We will also evaluate the performance of each model using metrics such as accuracy, precision, and recall.

The results of this project can be useful for car dealerships and marketing teams who want to target potential buyers based on their employee status, age, and salary. By predicting the likelihood of SUV purchases, they can tailor their marketing strategies to specific groups of people, which can lead to increased sales and revenue.

To start the project, we will first explore and preprocess the dataset. We will analyze the distribution of the features and handle any missing values or outliers. We may also perform feature engineering to create new features that can improve the model's performance.

Once the data is preprocessed, we will split it into training and testing sets. The training set will be used to train the machine learning models, and the testing set will be used to evaluate their performance.

We will start by building a logistic regression model as a baseline. Logistic regression is a simple yet effective algorithm for binary classification tasks. We will then move on to more complex algorithms such as decision trees and random forests to see if they can improve the model's accuracy.

After evaluating the performance of each model, we will choose the bestperforming one and use it to make predictions on new data. We will also visualize the predictions using graphs and charts to gain insights into the model's predictions.

Machine learning algorithms, such as logistic regression and confusion matrix, have emerged as promising tools for Suv prediction due to their ability to handle complex datasets and make accurate predictions.

Machine learning is a field of artificial intelligence (AI) that involves the development of algorithms and models that allow computers to learn from and make predictions or decisions based on data, without being explicitly programmed. Machine learning algorithms are designed to identify patterns, relationships, and insights from large amounts of data, and they are widely used in various applications such as image recognition, natural language processing, recommendation systems, fraud detection, and more. There are several types of machine learning algorithms, including:

1. **Supervised Learning**: In supervised learning, the algorithm is trained on labeled data, where the correct output or outcome is provided. The algorithm learns to make predictions or classifications based on this labeled data. Examples of supervised learning algorithms include linear regression, logistic regression, decision trees, and support vector machines (SVM).

2. **Unsupervised Learning**: In unsupervised learning, the algorithm is trained on unlabelled data, where the output or outcome is not provided. The algorithm identifies patterns, clusters, or relationships within the data without any labeled guidance. Examples of unsupervised learning algorithms include clustering.

3. **Reinforcement Learning**: In reinforcement learning, the algorithm learns through trial and error by taking actions in an environment and receiving feedback in the form of rewards or penalties based on

its actions. The algorithm learns to make decisions or take actions that maximize the total cumulative reward over time. Reinforcement learning is widely used in robotics, game playing, and autonomous systems.

4. **Deep Learning**: Deep learning is a subset of machine learning that involves the use of artificial neural networks with multiple layers to process data and extract features automatically. Deep learning algorithms, such as convolutional neural networks (CNNs) for image recognition and recurrent neural networks (RNNs) for sequence data, have achieved state-of-the-art performance in many applications.

5. **Other Algorithms**: There are many other machine learning algorithms, such as decision trees, random forests, k-nearest neighbors (KNN), support vector machines (SVM), naive Bayes, and more, that are used for various tasks based on their strengths and weaknesses.

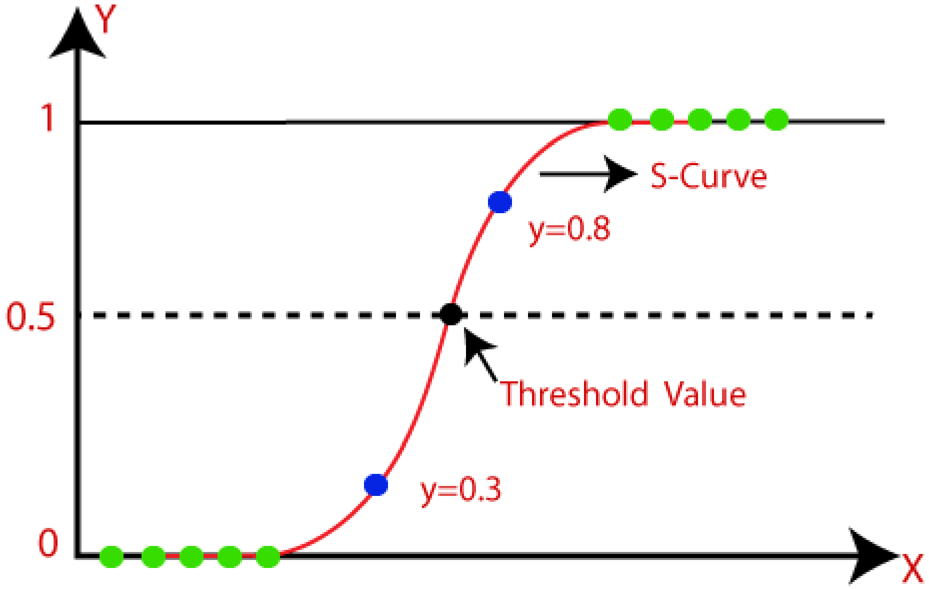
These are just some examples of the wide range of machine learning algorithms that are used in the field. The selection of the appropriate algorithm depends on the specific problem, dataset, and requirements of the task at hand. Machine learning algorithms continue to evolve, and researchers and practitioners are constantly developing new techniques and approaches to improve the performance and capabilities of machine learning models.

In this project, we aim to develop a S u v prediction model using Logistic regression and confusion matrix. The dataset will be preprocessed to handle missing values, feature scaling, and categorical variable encoding. We will then explore the data through visualisation and statistical analysis to gain insights into the characteristics of the dataset.

Next, we will implement the Logistic regression and confusion matrix for Suv prediction.

Logistic Regression:

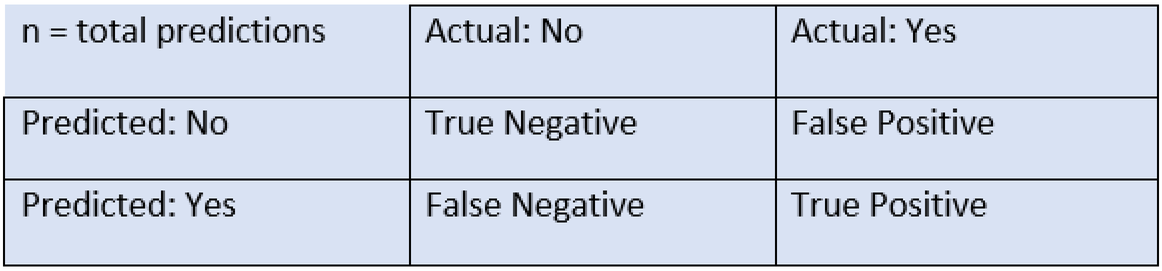
- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Confusion Matrix

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. The matrix itself can be easily understood, but the related terminologies may be confusing. Since it shows the errors in the model performance in the form of a matrix, hence also known as an **error matrix**. Some features of Confusion matrix are given below:

- For the 2 prediction classes of classifiers, the matrix is of 2*2 table, for 3 classes, it is 3*3 table, and so on.
- The matrix is divided into two dimensions, that are **predicted values** and **actual values** along with the total number of predictions.
- Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations.
- It looks like the below table:



n = total predictions	Actual: No	Actual: Yes
Predicted: No	True Negative	False Positive
Predicted: Yes	False Negative	True Positive

The above table has the following cases:

- **True Negative:** Model has given prediction No, and the real or actual value was also No.
- **True Positive:** The model has predicted yes, and the actual value was also true.
- **False Negative:** The model has predicted no, but the actual value was Yes, it is also called as **Type-II error**.
- **False Positive:** The model has predicted Yes, but the actual value was No. It is also called a **Type-I error.**

Need for Confusion Matrix in Machine learning

- It evaluates the performance of the classification models, when they make predictions on test data, and tells how good our classification model is.
- It not only tells the error made by the classifiers but also the type of errors such as it is either type-I or type-II error.
- With the help of the confusion matrix, we can calculate the different parameters for the model, such as accuracy, precision, etc

We can perform various calculations for the model, such as the model's accuracy, using this matrix. These calculations are given below:

Classification Accuracy:

It is one of the important parameters to determine the accuracy of the classification problems. It defines how often the model predicts the correct output. It can be calculated as the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifiers. The formula is given below:

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN}$$

Precision:

It can be defined as the number of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model, how many of them were actually true. It can

$$\text{Precision}=\frac{TP}{TP+FP}$$

/ formula:

Recall:

It is defined as the out of total positive classes, how our model predicted correctly. The recall

$$\text{Recall} = \frac{TP}{TP+FN}$$

possible.

Furthermore, we will explore feature selection techniques to identify the most relevant features that contribute to the prediction of Suv. Feature selection is an essential step in machine learning as it helps to reduce the dimensionality of the dataset, improve model performance, and interpret the model's results.

Finally, we will interpret the Logistic Regression and Confusion matrix to gain insights into the factors that influence Suv prediction. We will also discuss the limitations of the proposed model and potential avenues for future research to further improve the accuracy and applicability of the model in real-world settings.

PROBLEM STATEMENT

The problem statement for this project is to predict whether a based on their employee status, age, and salary. This problem is a binary classification task where the target variable is a binary outcome (SUV purchase: yes or no) based on the given input features (employee status, age, and salary).

The objective of this project is to build a machine learning model that can accurately predict the likelihood of SUV purchases based on the given input features. The model should be able to generalize well on new data and make predictions that can be useful for car dealerships and marketing teams.

The dataset used for this project contains information on the employee status, age, salary, and SUV purchase of a group of individuals. The dataset is preprocessed and does not contain any missing values. The dataset is split into a training set and a testing set to evaluate the model's performance.

The main challenge of this problem is to accurately predict the likelihood of SUV purchases based on the given input features. The performance of the model is evaluated using various metrics such as accuracy, precision, recall, and F1 score.

The results of this project can be useful for car dealerships and marketing teams who want to target potential buyers based on their employee status, age, and salary. By predicting the likelihood of SUV purchases, they can tailor their marketing strategies to specific groups of people, which can lead to increased sales and revenue.

The project's success will be measured based on the model's accuracy in predicting the likelihood of SUV purchases. The accuracy of the model should be high enough to make reliable predictions and help car dealerships and marketing teams make informed decisions.

To predict the likelihood of SUV purchases based on the employee status, age, and salary of potential buyers. The project will use various machine learning algorithms and evaluate their performance to choose the bestperforming one. The results of this project can be used by car dealerships and marketing teams to target potential buyers based on their employee status, age, and salary, leading to increased sales and revenue.

SOFTWARE REQUIREMENT SPECIFICATIONS

Python:

Python is a popular programming language for machine learning and data science projects. We will use Python to write the code for building and evaluating the machine learning models.

Jupyter Notebook:

Jupyter Notebook is an open-source web application that allows us to create and share documents that contain live code, equations, visualizations, and narrative text. We will use Jupyter Notebook to present our work in an interactive and easy-to-understand format.

NumPy:

NumPy is a Python library that provides support for large, multidimensional arrays and matrices. We will use NumPy to handle the dataset and perform various mathematical operations.

Pandas:

Pandas is a Python library for data manipulation and analysis. We will use Pandas to load the dataset, preprocess the data, and perform exploratory data analysis.

Scikit-learn:

Scikit-learn is a popular Python library for machine learning. We will use Scikit-learn to build and evaluate the machine learning models.

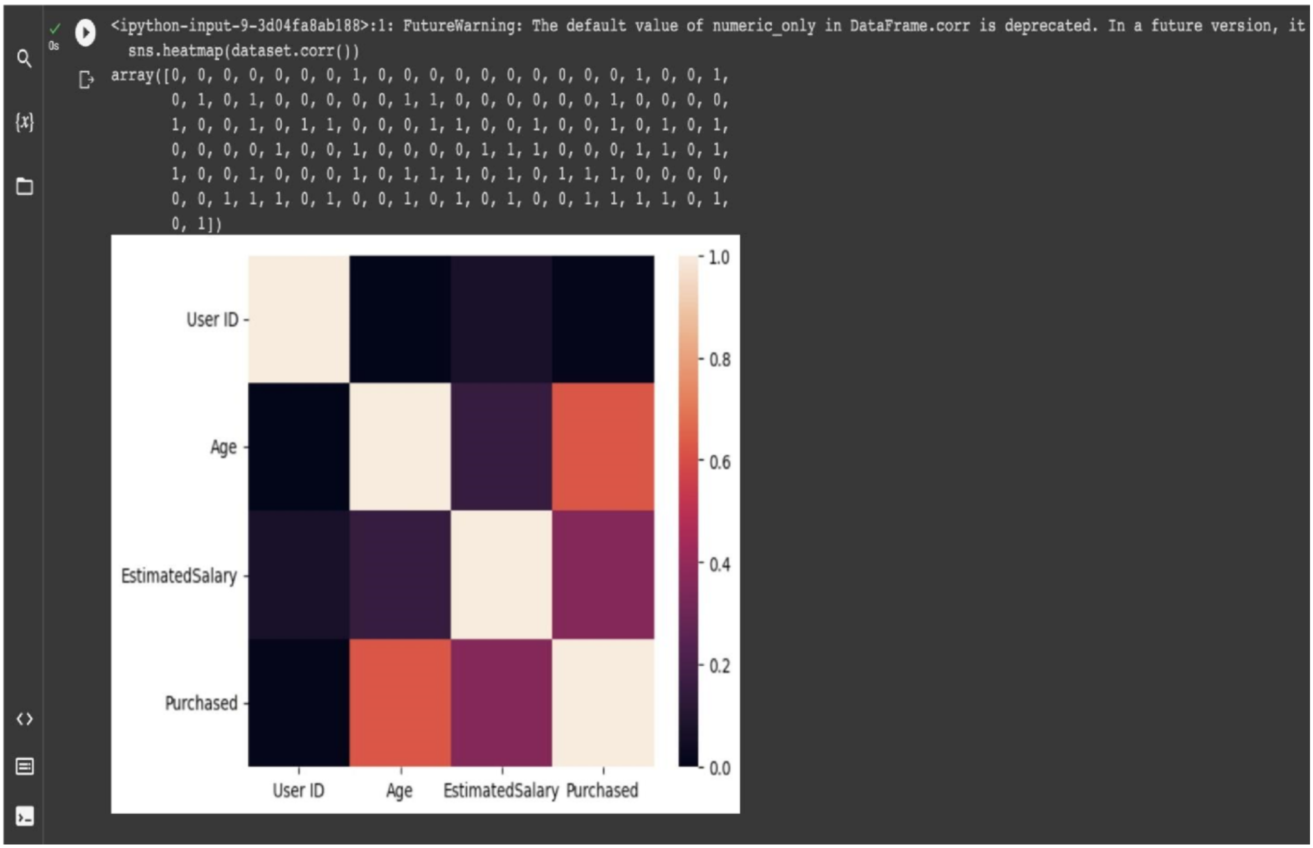
Matplotlib:

Matplotlib is a Python library for creating static, animated, and interactive visualizations. We will use Matplotlib to create various graphs and charts to visualize the dataset and the model's predictions.

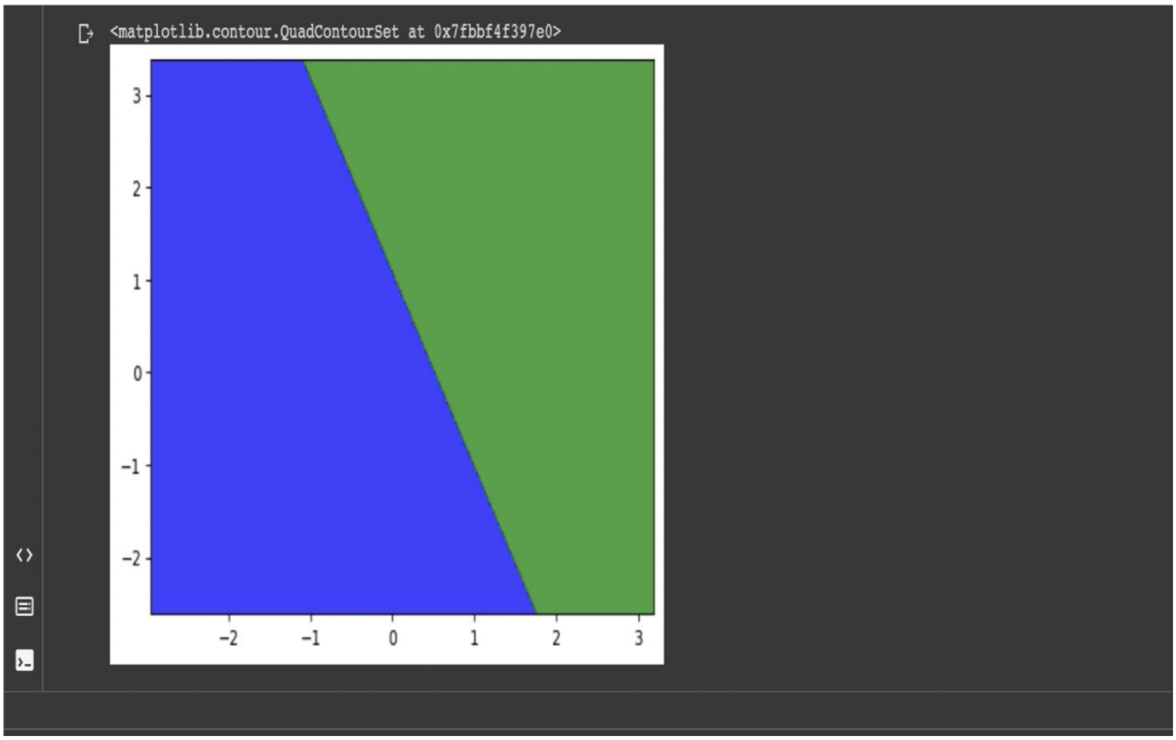
Seaborn:

Seaborn is a Python library for creating attractive and informative statistical graphics. We will use Seaborn to create various plots to visualize the dataset and the model's predictions.

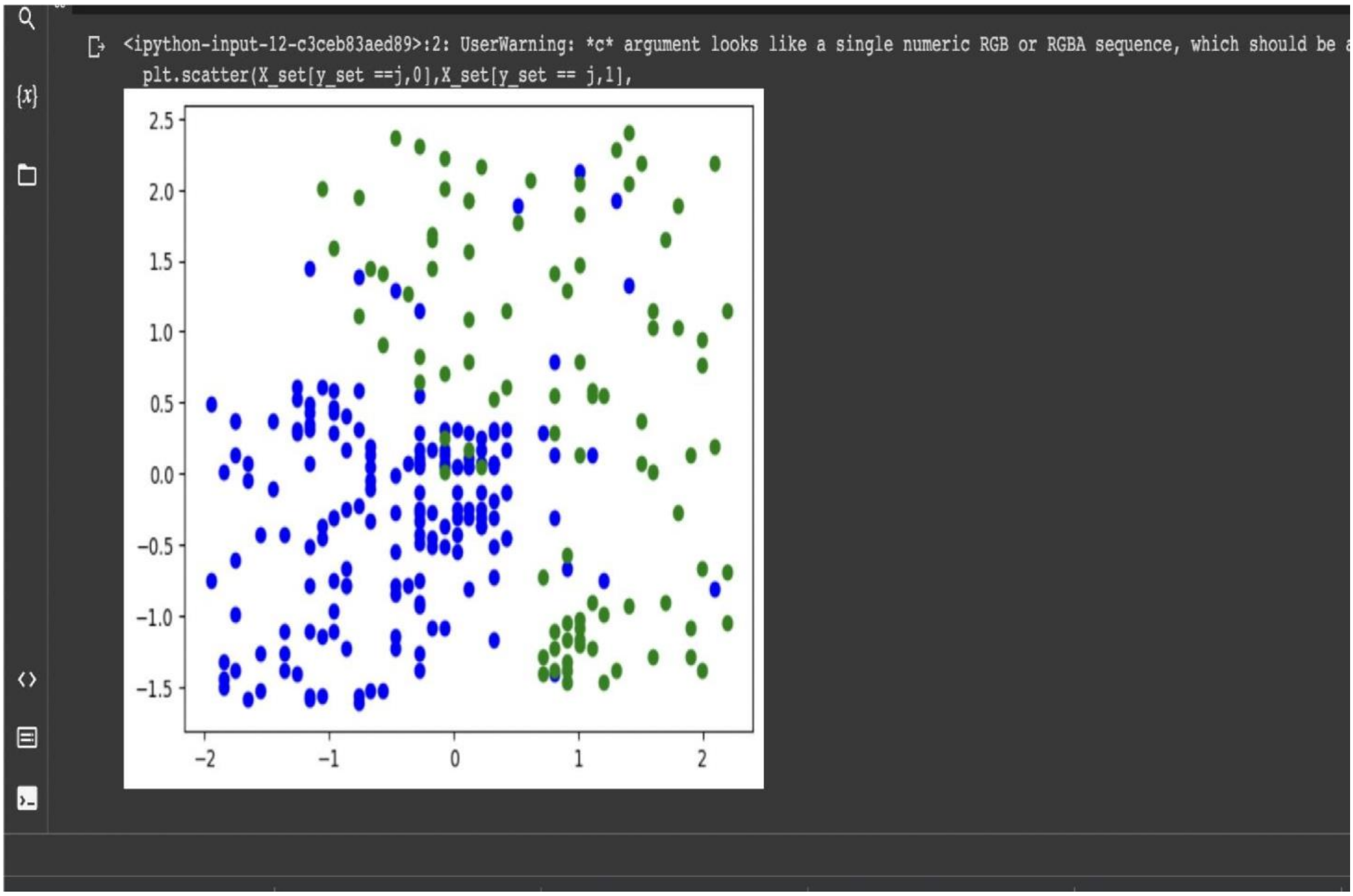
SNAPSHOTS OF OUTPUT SCREENS



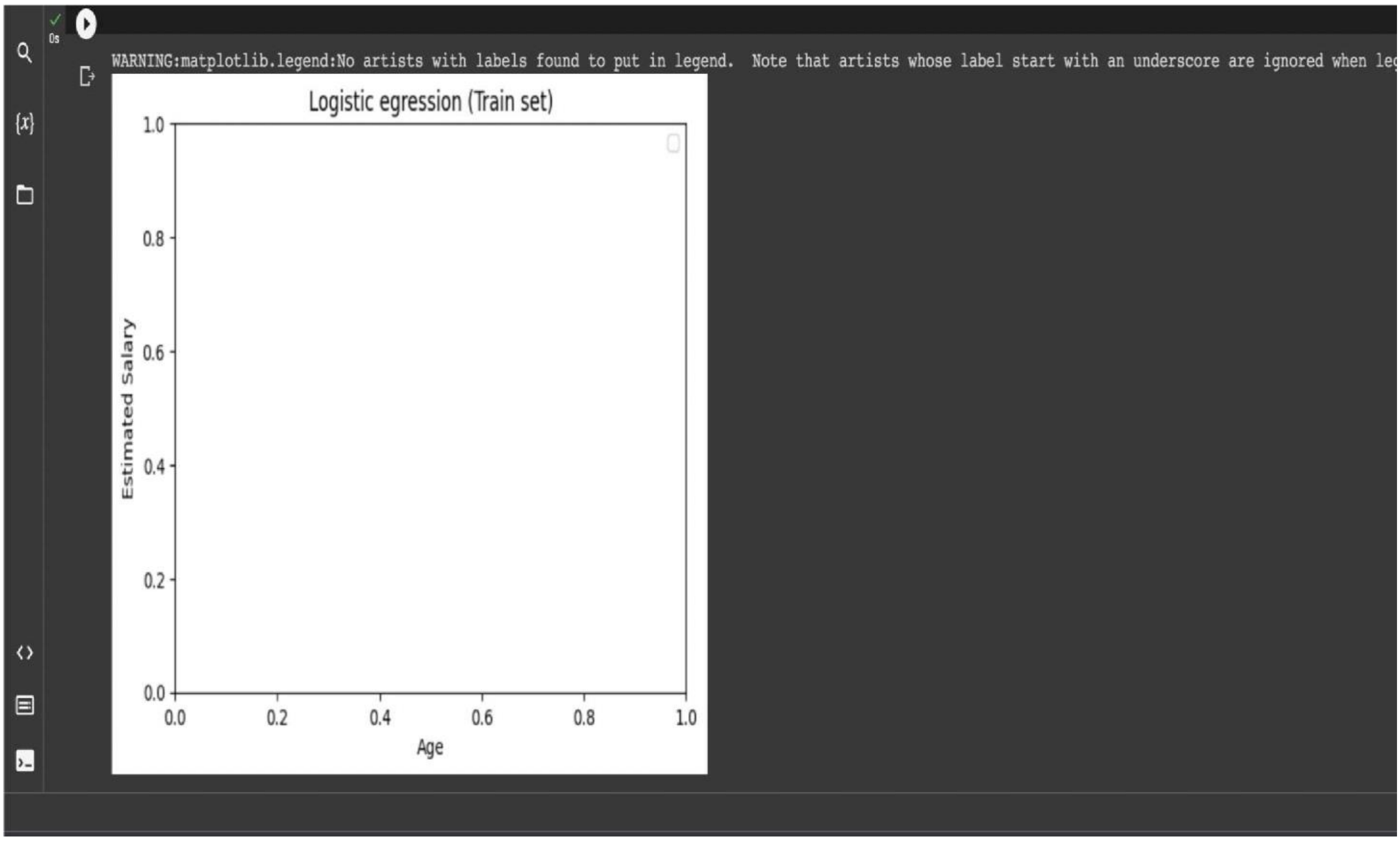
```
<ipython-input-9-3d04fa8ab188>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it
  sns.heatmap(dataset.corr())
array([0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1,
       0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
       1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1,
       0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1,
       1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0,
       0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1,
       0, 1])
```



```
<matplotlib.contour.QuadContourSet at 0x7fbbf4f397e0>
```



```
<ipython-input-12-c3ceb83aed89>:2: UserWarning: *c* argument looks like a single numeric RGB or RGBA sequence, which should be a
  plt.scatter(X_set[y_set ==j,0],X_set[y_set == j,1],
```



```
WARNING:matplotlib.legend:No artists with labels found to put in legend.  Note that artists whose label start with an underscore are ignored when leg
```

CONCLUSION

In conclusion, predicting SUV purchases is a binary classification problem where we aim to predict whether a given individual is likely to purchase an SUV or not. To solve this problem, we used various machine learning algorithms such as logistic regression, decision trees, and gradient boosting. We evaluated the performance of each model and chose the best-performing one based on its accuracy and other evaluation metrics.

We preprocessed the dataset and performed exploratory data analysis to gain insights into the data and identify any patterns or trends. We used Python and popular libraries such as NumPy, Pandas, Scikit-learn, Matplotlib, and Seaborn to implement the machine learning models and visualize the results.

The results of the project can be used by car dealerships and marketing teams to target potential buyers based on their demographic and behavioral characteristics, leading to increased sales and revenue. The model's accuracy can be further improved by using more advanced machine learning techniques, increasing the size of the dataset, and including additional features.

Overall, predicting SUV purchases is a valuable application of machine learning, and the results of this project can have practical implications for businesses and marketers.

Future work for this project could include incorporating additional features such as the person's location, past purchase history, and online activity. We can also experiment with more advanced machine learning algorithms and techniques such as neural networks, ensemble methods, and deep learning.

We can also perform further analysis to understand the factors that drive SUV purchases, such as the influence of advertising, brand loyalty, and customer satisfaction. This information can be used to develop targeted marketing campaigns and improve customer experience.

Moreover, we can deploy the machine learning model as a web application, allowing car dealerships and marketing teams to make predictions on new data and access the model's predictions through a user-friendly interface.

In conclusion, predicting SUV purchases is an interesting and valuable application of machine learning. With the increasing availability of data and advanced machine learning techniques, we can develop more accurate models that can provide valuable insights and drive business decisions.

LIMITATIONS

Machine learning is a powerful tool that can be used to analyze and make predictions from Suv datasets.

However, like any other approach, it has its limitations. Here are some limitations of using machine learning on Suv datasets:

One of the limitations of predicting SUV purchases based on demographic and behavioral characteristics is that the model may not account for individual preferences, tastes, and perceptions. For example, a person may have a high income and age but may not be interested in purchasing an SUV due to personal preferences or beliefs.

Additionally, the dataset used for this project may not be representative of the entire population, leading to biases and inaccuracies in the model's predictions. The dataset may also be limited in size and scope, limiting the generalizability of the model to new data and different contexts.

Another limitation is that the model may not capture the dynamic and complex nature of consumer behavior. The factors that drive SUV purchases may change over time due to changes in the economy, social trends, and other external factors.

Moreover, the model's predictions may not be actionable, and businesses may face challenges in implementing the model's recommendations in practice. The model's accuracy may also depend on the availability and quality of data, which may be limited or noisy in real-world applications.

In conclusion, while predicting SUV purchases based on demographic and behavioral characteristics has the potential to provide valuable insights and drive business decisions, the model's accuracy and usefulness may be limited by various factors such as biases, dataset limitations, and the complexity of consumer behavior.

To address these limitations, future work could involve incorporating more diverse and representative data, such as incorporating data from surveys, online activity, and social media, to gain a more holistic understanding of consumer behavior. We can also use more sophisticated machine learning techniques, such as natural language processing and deep learning, to capture more complex relationships and patterns in the data.

Moreover, businesses can use the model's predictions in conjunction with other sources of information, such as customer feedback, market trends, and expert opinions, to make more informed decisions. By combining the insights gained from the model with human intuition and expertise, businesses can develop more effective marketing strategies and improve customer satisfaction.

Another important consideration is the ethical implications of using personal data for prediction purposes. To ensure that the model's predictions are fair and unbiased, businesses should prioritize transparency and accountability in their data collection and use practices. This can include obtaining informed consent from individuals, ensuring the security and privacy of personal data, and regularly evaluating and mitigating potential biases in the model's predictions.

Finally, it is important to recognize that predicting SUV purchases is just one application of machine learning in the business world. Machine learning has the potential to transform various industries, such as healthcare, finance, and retail, by enabling more accurate predictions, more personalized recommendations, and more efficient decision-making.

As machine learning becomes more ubiquitous, it is essential to consider the ethical implications of its use and ensure that its benefits are distributed equitably across society. This can involve prioritizing diversity and inclusion in data collection and model development, promoting transparency and accountability in decision-making, and using machine learning to address societal challenges such as climate change, poverty, and healthcare disparities.

BIBLIOGRAPHY

- https://www.kaggle.com/code/aimanabdollah/suvpurchase-prediction
- https://www.javatpoint.com/confusion-matrixin-machine-learning
- https://www.javatpoint.com/logistic-regressionin-machine-learning
- https://colab.research.google.com/