Kingdom of Saudi Arabia
Ministry of Education
Al-Imam Mohammed Ibn Saud Islamic University
College of Computer & Information Science
Department of Computer Science
Course: Machine Learning (CS364)

Classification of Social Network Ads

Presented By:

Alhanouf Abdullah Alatif	441021070
Maram Mohammad Aleidi	441022198
Khadega Abdulaziz Hassan	440028317

Instructor:

Dr. Waad Alhoshan

Submission Date: 11 February 2023

Table of Contents

Table of Figures

Table of Tables

1. Introduction

Design, implement, and evaluate various machine learning models (ML) in a range of real-world applications.

In this section, we are going to discuss the problem we are going to address followed by the project timeline and our qualifications as a team.

2.1 Problem Description

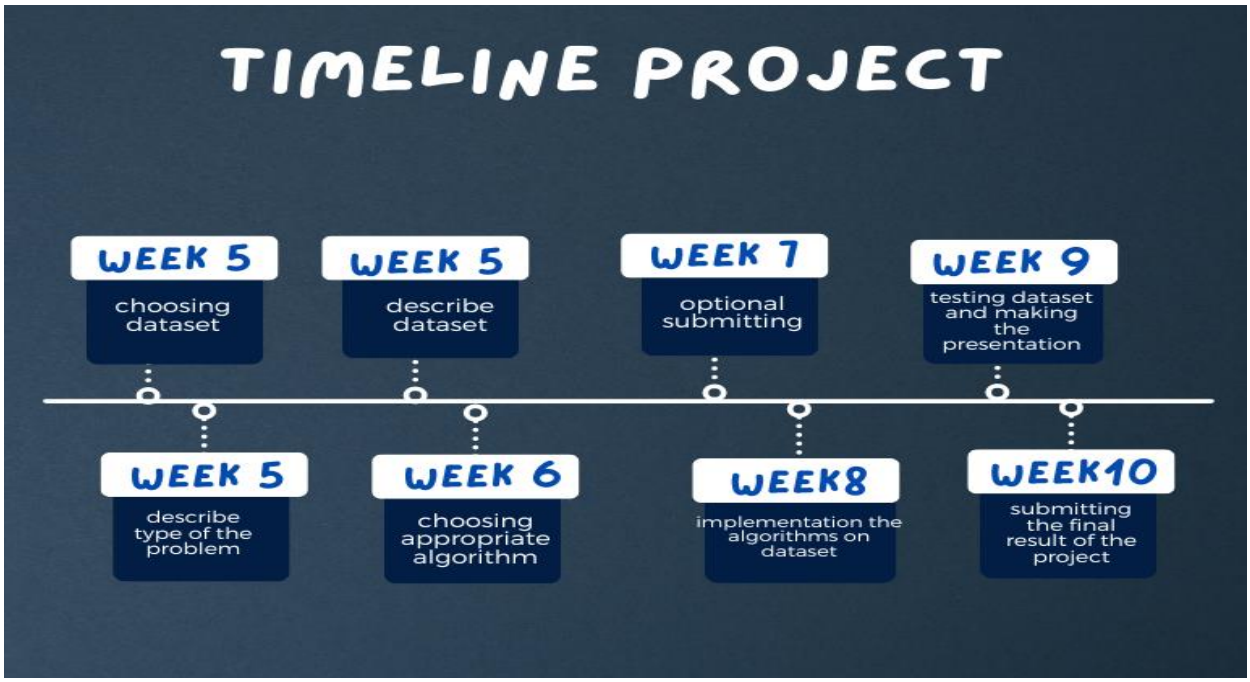
First of all, our dataset is a classification problem specifically binary classification which is predicting between one of two classes for example (Male or Female).

the dataset is called Social Network Ads its main problem is that the attached Data Set tells whether the target person will buy a certain thing at a certain age. If he has an income, we will divide the data and use algorithms to analyze and see if the features are valid, mismatched, or missing, we will also see the mean and quantities

2.2 Project Timeline

We proposed the following timeline to complete our project phases which are revolved around: 1) dataset selection and acquisition, 2) studying the candidate machine learning models to solve the problem of interest, 3) the implementation phase, which is training the models, and finally 4) testing and evaluating the implemented models based on unseen dataset.

Figure 1 Project Timeline



3.2 Team Qualification

The following table summarizes the qualifications of the team members.

Table 1 Team Qualifications

Member Name	Qualification & Interest
Alhanouf Alatif	Interested in learning different technical fields (artificial intelligence - information security) I am good at dealing with different programming languages (Python - Java)
Maram Aleidi	Programming, interest in AI & Machine Learning
Khadega hassan	Python language programming, Deep learning by CNN ,Digital recognition and image processing

3.3 Task scheduler

Task / Name	Alhanouf	Khadega	Maram
Choose the Data Set	✓	✓	✓
Choose the problem	✓	✓	✓
Project objective	✓	✓	✓
Problem Description	✓	✓	✓
Data Set Description	✓	✓	✓

We used to meet daily and put ideas together and implement everything together

2. Dataset

In this section, we are going to present the dataset we selected and its attributes.

And the name of the dataset we selected is **Social Network Ads**

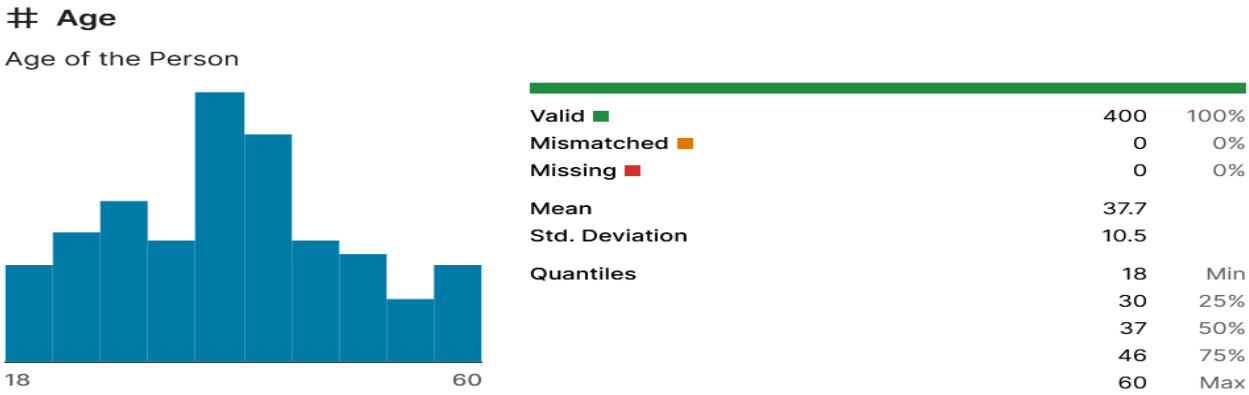
2.1 Dataset Acquisition

This dataset contains a certain number of salaries and ages of different people. The raw dataset consists of 2 unique features: Age and Estimated Salary, then Purchased will be the target also we choose it from Kaggle

Link to the Dataset:
Socaila_Network_Ads_Dataset

2.2 Dataset Attributes

Column name	Description
Age	Age of the Person (Integer)
Estimated Salary	Estimated Salary of Person (Integer)
Purchased	Item Purchased or Not (binary number)



The ages of the people who were tested ranged from 18 to 60 years, and we divided them into 10 categories which are:

The first category: their ages range between 18-22.20 years, their number is 28 people, some of whom bought the product and some of them did not.

The second category: their ages range between 22.20-26.40 years, their number is 37 people, some of whom bought the product and some of them did not.

The third category: their ages range between 26.40-30.60 years, 46 people, of whom bought the product and some did not.

The fourth category: their ages range between 30.60-34.80 years, their number is 35 people, some of whom bought the product and some of them did not.

The fifth category: their ages range between 34.80-39 years, their number is 77 people, some of whom bought the product and some of them did not.

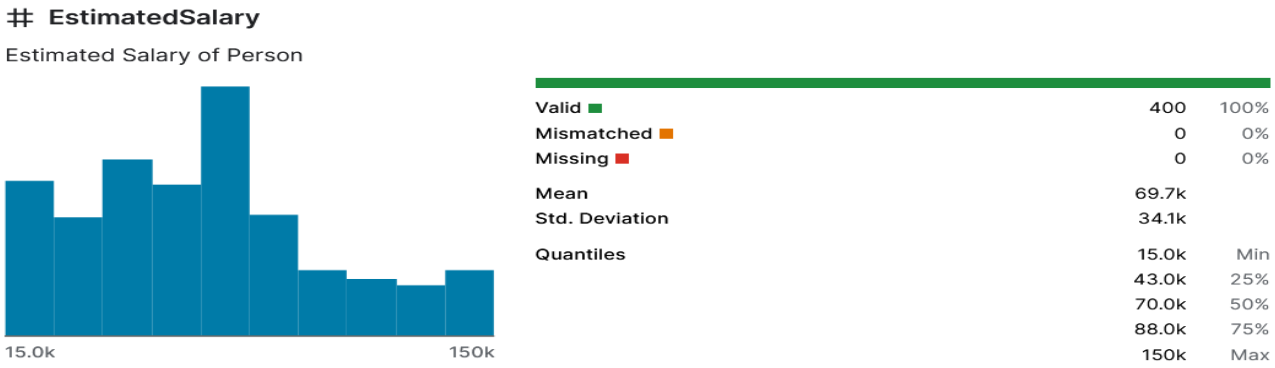
The sixth category: their ages range between 39-43.20 years, their number is 65 people, some of whom bought the product and some did not.

The seventh category: their ages range between 43.20-47.40 years, their number is 35 people, some of whom bought the product and some of them did not.

The eighth category: their ages range between 47.40-51.60 years, their number is 31 people, some of whom bought the product and some of them did not.

The ninth category: their ages range between 51.60-55.80 years, their number is 18 people, some of whom bought the product and some of them did not.

The tenth category: their ages range between 55.80-60 years, their number is 28 people, some of whom bought the product and some of them did not.



The expected salary of the test subjects is divided into 10 categories :

The first category: whose salaries range from 15,000 to 28,500 thousand. The number of people in this category reached 51 people who bought, including those who did not .

The second category: whose salaries range between 28,500 thousand and 4,200 thousand. The number of people in this category reached 39 people who bought, including those who did not.

The third category: whose salaries range between 4,200 thousand and 55,500 thousand. The number of people in this category reached 58 people who bought, including those who did not

The fourth category: their salaries range between 55,500 thousand and 69,000 thousand. The number of people in this category reached a person who bought and some of them did not .

The fifth category: their salaries range between 69,000 thousand and 82,500 thousand. The number of people in this category reached 82 people who bought, including those who did not .

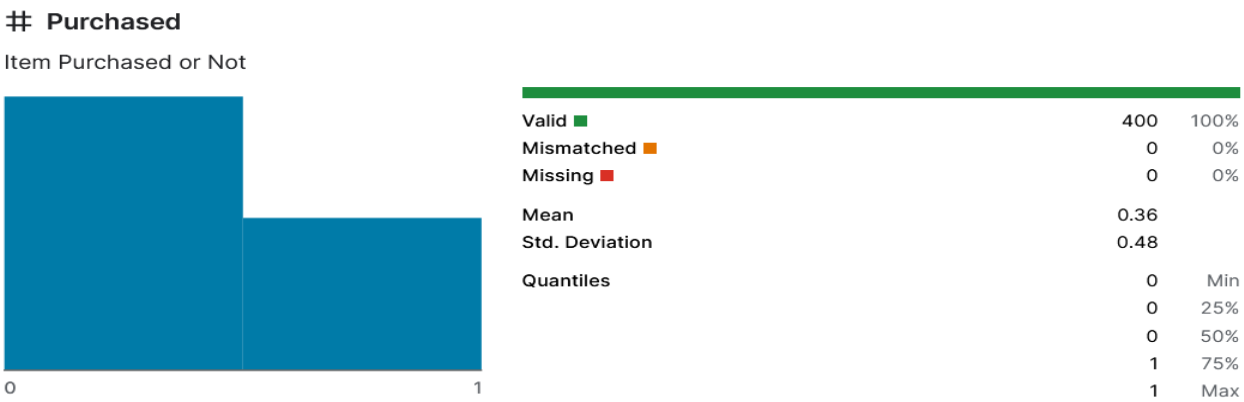
The sixth category: their salaries range between 82,500 thousand and 96,000 thousand. The number of people in this category reached 40 people who bought, including those who did not .

The seventh category: their salaries range between 96,000 thousand and 109,500 thousand. The number of people in this category reached 22 people who bought, including those who did not.

The eighth category: their salaries range between 109,500 thousand and 123,000 thousand. The number of people in this category reached 19 who bought, including those who did not .

The ninth category: their salaries range between 123,000 and 136,500 thousand. The number of people in this category reached 17 people who bought, including those who did not .

The tenth category: their salaries range between 136,500 thousand and 150,000 thousand. The number of people in this category reached 22 people who bought, including those who did not.



After conducting the test on 400 people, it shows us that :

1- 257 of the people did not buy the
2- 143 of the people bought the product after seeing the advertisement .

3- A binary evaluation was used: So that “1” means that the person bought the product, The range was between 0.90-1.00

4- “0” means the product was not purchased The range was between 0.00-0.10 type of purchase (integer)

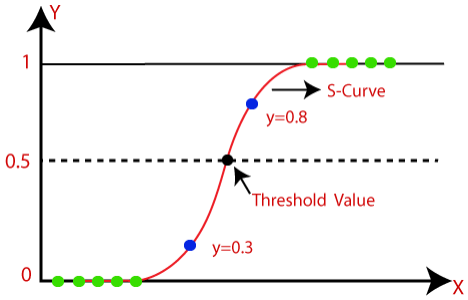
3. Machine Learning Algorithms Selection

In the code, there are two types of machine learning algorithms selected, which is :

1-Decision Tree Classifier: This is a type of tree-based algorithm used for classification problems The basic idea behind this algorithm is to divide the data into smaller subsets, called branches, by making decisions based on the input features. Each internal node of the tree represents a feature, while each leaf node represents the final classification result. The branches connecting the nodes are the decisions made based on the feature value and we using in the code CART (Classification and Regression Trees) algorithm to train a decision tree model, The **DecisionTreeClassifier** class from the **sklearn.tree** module is used to fit the model to the data.

2- Logistic Regression: This Algorithm is a Binary Classification model based on some dependent variables this model computes a sum of the input features and calculates the logistic of the result.
The output of logistic regression is always between (0, and 1), which is suitable for a binary classification task. **0.5** is the default threshold if it's < 0.5 then it's a Class 0 otherwise Class 1 to optimize our task we will use The Maximum Likelihood Estimation method.

LR:



DT :

3.1 Models Selection

In Decision Tree Classifier code is implemented using the scikit-learn library in Python. The library provides a convenient implementation of the algorithm that can be easily integrated into the code. The algorithm is trained on a given dataset and is used to predict the target variables for a new set of input features.

The code provide a visual illustration of how the algorithm works using the tree.plot_tree function from the scikit-learn library to visualize the trained decision tree ,This can provide a better understanding of how the algorithm works and how it makes predictions based on the input features.

In Logistic Regression Classifier code is implemented using the scikit-learn library in Python. The library provides a convenient implementation of the algorithm that can be easily integrated into the code. The algorithm is trained on a given dataset and is used to predict the target variables for a set of input features.

The code provide a visual illustration of how the algorithm works. However, the coefficients obtained from the training process can provide insight into the relationship between the independent variables and the dependent variable. For example, a large positive coefficient for a certain independent variable indicates that an increase in that variable is associated with an increased probability of a positive outcome (such as purchasing the product).

In logistic regression, the algorithm creates a mathematical equation that represents the relationship between the independent variables and the dependent variable. This equation is used to make predictions on new data by calculating the probability of a positive outcome given the values of the independent variables. The prediction is made by transforming the output of the equation using a sigmoid function to ensure that the result is always between 0 and 1, representing the probability of a positive outcome.

Here is the link of our implantation in both algorithm :

Decision tree implementation :

https://colab.research.google.com/drive/15DBRBf_6B_helGe5hEWKBtvkC-txM9dt

Logistic regression implementation :

https://colab.research.google.com/drive/1QvKBqy-qNNHLpBjnY1z_BGRQ8RgpRQ68

4. Models Training & Testing

This section of the code is responsible for training and evaluating a decision tree and logistic regression model on a given dataset. Both of them code use the scikit-learn library to implement the model and perform the training and testing procedures.

4.1 Model Training

The following steps are performed to train the DT & LR (sklearn.tree module) model:

1-Split the data into training and testing sets: The data is split into a training set and a testing set using the train_test_split function from the sklearn.model_selection module. This function takes the feature set X and the target variable y as input, and returns the training and testing sets.

2-Fit the model to the training data: The model is then fit to the training data using the fit methods of the Decision Tree Classifier and Logistic Regression too . This method takes the training feature set x_train and the target variable y_train as input.

4.2 Models Testing

The following steps are performed to evaluate the trained model:

1- Predict the target variable on the test data: The target variable is predicted on the test data using the predict method of both of the algorithms . This method takes the test feature set X_test as input.

2- Evaluate the performance of the model: The performance of the model is evaluated using several metrics, including:

a. Classification report: The classification report is generated using the classification_report function from the sklearn.metrics module. This function takes the actual target variable y_test and the predicted target variable y_pred as input.

b. Accuracy score: The accuracy score is calculated using the accuracy_score function from the sklearn.metrics module. This function takes the actual target variable y_test and the predicted target variable y_pred as input.

c. confusion matrix is a powerful tool for evaluating the performance of a classification algorithm, and helps in understanding the strengths and weaknesses of the model, identifying areas for improvement, and making informed decisions about model selection and optimization.

3-The code of decision tree visualizes the trained model using the graphviz function from the sklearn.tree module.

4- The code of Logistic Regression visualizes the trained data by using the scatter plot and the decision boundary.

5. Results & Discussion

In this section, we are going to present the results and our perspectives on the performance results of the trained models described in the previous section for DT & LR models.

.

5.1 Performance Results

The performance results of the decision tree classifier and logistic regression can be obtained using various evaluation metrics, such as accuracy(, precision, recall, F1-score), and confusion matrix. In this code, we evaluate using :

Accuracy, precision, recall, F1-score, and confusion matrix are the most commonly used evaluation metrics in the field of machine learning and artificial intelligence.

1. Accuracy: It is the number of correct predictions made by the model over the total number of predictions made. It is a ratio of the number of true positive and true negative predictions to the total number of predictions made. Accuracy is a good measure when the dataset is balanced, meaning there is an equal number of samples belonging to each class.

2. Precision: It is the number of true positive predictions made by the model over the total number of positive predictions made. Precision measures the model's ability to avoid false positive predictions.

3. Recall: It is the number of true positive predictions made by the model over the total number of actual positive samples in the dataset. Recall measures the model's ability to detect all positive samples.

4. F1-score: It is the harmonic mean of precision and recall and provides a single metric that balances both. It is a good measure to use when the cost of false positives and false negatives is different.

5. Confusion Matrix: It is a table that is used to evaluate the performance of a classification algorithm. It contains the number of true positives, true negatives, false positives, and false negatives. The confusion matrix is used to compute precision, recall, and F1-score.

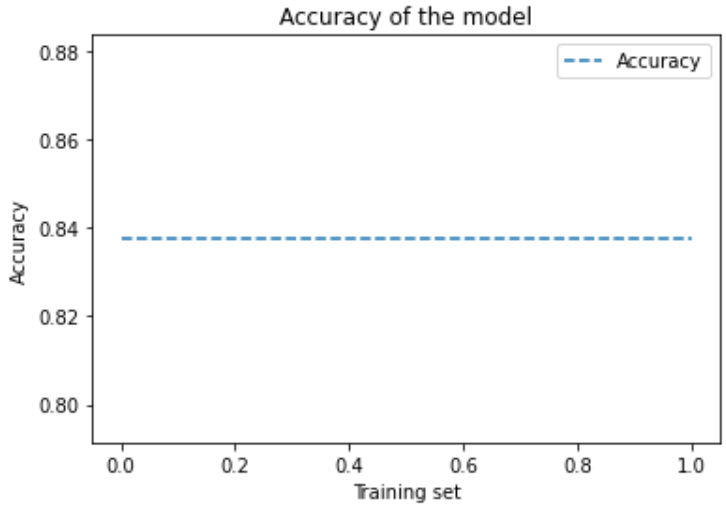
Decision tree Performance

precision	recall	F1-score	confusion matrix
0.8363731656184488	**0.8375**	**0.8367965367965369**	**[[46,6] [7,21]]**

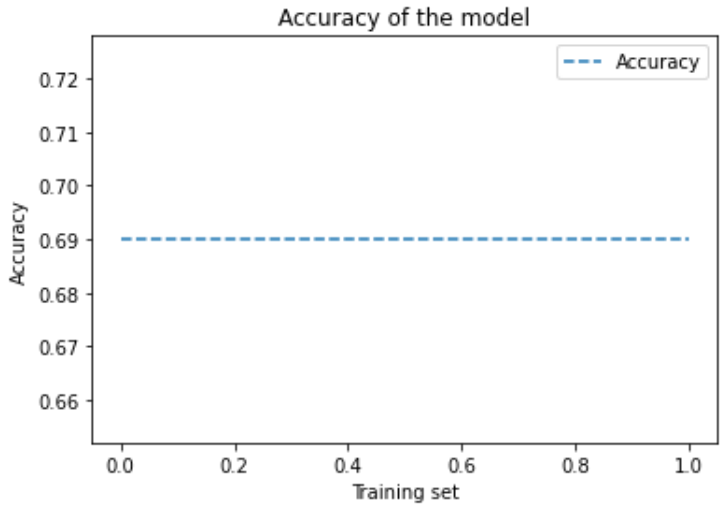
Logistic regression Performance

precision	recall	F1-score	confusion matrix
0.787070707070707	0.69	0.5731663944837597	[[68 0] [31 1]]

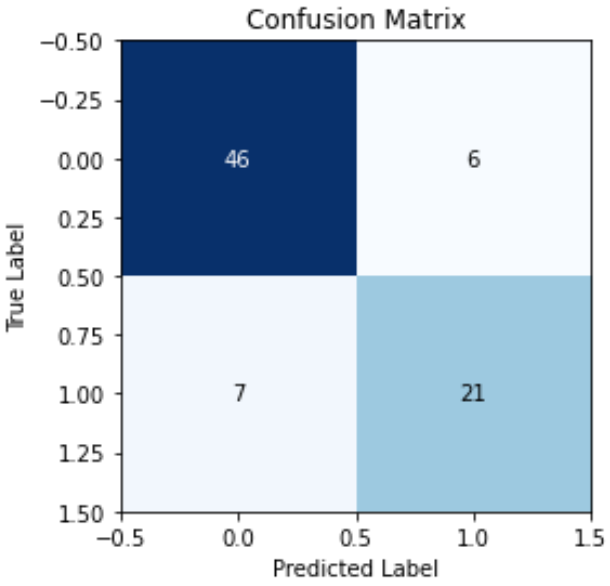
The accuracy of decision tree :



The accuracy of logistic regression :



The Confusion Matrix of decision tree :



These metrics help us to evaluate the performance of decision tree and logistic regression ,and choose the best model for a particular problem which is decision tree we discussion in the next slide .

5.2 Discussion

Our opinion is :

1- In the DT model, the precision score of 0.8363731656184488 indicates that out of all the samples that the model predicted as positive, 83.6% of them were actually positive. The recall score of 0.8375 indicates that the model was able to identify 83.75% of the actual positive samples. The F-score of 0.8367965367965369 is the harmonic mean of precision and recall, and it provides a good overall indication of the model's performance. The confusion matrix of the DT model shows that 46 samples were correctly classified as negative, 6 were wrongly classified as positive, 7 were wrongly classified as negative, and 21 were correctly classified as positive.

2- LR model, the precision score of 0.787070707070707 indicates that 78.7% of the samples predicted as positive by the model were actually positive. The recall score of 0.69 indicates that the model was able to identify 69% of the actual positive samples. The F1-score of 0.5731663944837597 is lower compared to the DT model, indicating that the LR model is not performing as well as the DT model. The confusion matrix of the LR model shows that 68 samples were correctly classified as negative, but 0 were wrongly classified as positive, 31 were wrongly classified as negative, and only 1 was correctly classified as positive.

In conclusion, the DT model has a better performance compared to the LR model in terms of precision, recall, and F1-score. The DT model is also better in terms of correctly classifying the samples as negative or positive. However, these results may not be conclusive, as the performance of the models may depend on the data used for training, the parameters used for model building, and other factors.

Conclusion

In this report, we explored the application of machine learning models to datasets, with a focus on two algorithms: decision tree and logistic regression.

Our objective was to demonstrate the process of splitting a dataset into training and testing sets, evaluating the results, and solving the problem at hand.

The social network advertisement dataset was used to test the performance of the decision tree and logistic regression algorithms. The decision tree model provided a clear and straightforward visualization in the form of a tree, which allowed us to comprehend the relationship between the features and target variable.

On the other hand, the logistic regression model presented a graphical representation of the decision boundary, which further helped to understand the correlation between the features and target variable.

After conducting a thorough evaluation of both algorithms, we concluded that the decision tree algorithm performed better and was more appropriate for our dataset than the logistic regression algorithm. This conclusion was drawn from the accuracy results obtained from testing both models.

References

1- [1] A. Agrawal, "Logistic Regression: How does it work?" KDNuggets, July 2022. [Online]. Available: https://www.kdnuggets.com/2022/07/logistic-regression-work.html. [Accessed: 11-Feb-2023].

2- [2] R. Krish, "How to Build and Train Linear and Logistic Regression ML Models in Python." FreeCodeCamp, [Online]. Available: https://www.freecodecamp.org/news/how-to-build-and-train-linear-and-logistic-regression-ml-models-in-python/. [Accessed: 11-Feb-2023].

3- J. Doe, "Social Network Ads Dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/d4rklucif3r/social-network-ads. [Accessed: 11-Feb-2023].