Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)
Department of Computer Science and Engineering (Data Science)

Subject: Machine Learning – I (DJ19DSC402)

AY: 2022-23

Experiment 10

(Mini Project)

Aim: Design a classifier to solve a specific problem in the given domain.

Tasks to be completed by the students:
Select a specific problem from any of the given domain areas, such as: Banking, Education, Insurance, Government, Media, Entertainment, Retail, Supply chain, Transportation, Logistics, Energy and Utility.
Task 1: Select appropriate dataset, describe the problem and justify the suitability of your dataset.
Task 2: Perform exploratory data analysis and pre-processing (if required).
Task 3: Apply appropriate machine learning algorithm to build a classify. Perform appropriate testing of your model.
Task 4: Submit a report in the given format.

- Introduction
- Data Description
- Data Analysis
- Reason to select machine learning model
- Algorithm
- Result Analysis
- Conclusion and Future Scope.
- Python notebook

Task5: Presentation

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)
Department of Computer Science and Engineering (Data Science)

Report on Mini Project

Machine Learning -I (DJ19DSC402)

AY: 2022-23

Football Match Winner Prediction

NAME: Vedanta Yadav

NAME: 60009210185

Guided By
Dr. Kriti Srinivasan

CHAPTER 1: INTRODUCTION

The task at hand is to analyze a dataset consisting of match scores from the Premier League spanning from 2001 to 2016 and to determine the most effective machine learning model for predicting the match winner.

To accomplish this, various prediction models will be applied to the dataset, and the performance of each model will be evaluated and compared. Given the complexity of the problem and the large number of variables that can impact the outcome of a football match, selecting the best model is critical for achieving accurate predictions.

The analysis will require careful data preprocessing, feature selection, and model training and evaluation, with the goal of identifying the model that produces the highest accuracy and precision in predicting the match winner.

This task is of significant importance for sports analysts and enthusiasts, as accurate predictions can provide valuable insights and help inform strategic decisions for both teams and individual bettors.

CHAPTER 2: DATA DESCRIPTION

The dataset in the provided link is a collection of football match data from the English Premier League for the seasons 2001-2016. It contains 6,703 records with 23 variables/features for each record.The variables/features include the date of the match, the teams playing, the goals scored by each team, and various match statistics such as the number of shots on target, corners, fouls, and bookings.

Additionally, the dataset contains a binary variable indicating the outcome of the match, with a value of 0 indicating a loss or draw for the home team and a value of 1 indicating a win for the home team. This variable serves as the target variable for the machine learning models. The dataset has been merged from multiple sources, including Kaggle and football-data.co.uk.

It has been preprocessed to remove irrelevant data, handle missing values, and encode categorical variables.

CHAPTER 3: DATA ANALYSIS

The first visualization in the notebook shows the distribution of the target variable (i.e., home team win or not) in the dataset using a bar chart. This is a useful way to quickly understand the distribution of the target variable and to identify any class imbalance that may affect model performance.

The second visualization is a heat map that shows the correlation between all the features in the dataset. This is a useful way to identify which features are most strongly correlated with the target variable and with each other. Highly correlated features may be redundant and can be removed to simplify the model without losing predictive power.

The third visualization is a scatter plot matrix that shows the pairwise relationships between all the features in the dataset. This is a useful way to identify any linear relationships between features and to identify potential outliers or unusual patterns in the data.

The fourth visualization is a box plot that shows the distribution of the goal difference variable for each value of the target variable (i.e., home team win or not). This is a useful way to visualize the distribution of the goal difference variable and to identify any differences in the distribution between the two classes.

The fifth visualization is a scatter plot that shows the relationship between the goal difference variable and the number of shots on target for the home team. This is a useful way to visualize the relationship between two specific features and to identify any patterns or trends in the data.

Overall, these visualizations provide valuable insights into the relationships between the features in the dataset and the target variable, and can help guide the selection and tuning of machine learning models for predicting football match outcomes.

CHAPTER 4: DATA MODELLING

Here are the different prediction models used in the notebook:

1. Logistic Regression: This is a simple linear model that models the relationship between the input features and the binary target variable (i.e., home team win or not) using logistic function. Logistic regression is a common and interpretable model for binary classification tasks.

2. Gaussian Naive Bayes- It calculates the probability of each class given the observed features, using Bayes' theorem and the estimated mean and standard deviation of each feature for each class. The class with the highest probability is then assigned as the predicted class.

3. Decision Tree: This is a non-linear model that uses a tree-based structure to make decisions about the target variable based on the input features. Decision trees are easy to interpret and can handle both categorical and numerical data, but can suffer from overfitting.

4. Random Forest: This is an ensemble model that combines multiple decision trees to improve performance and reduce overfitting. Random forest is a popular and powerful machine learning model that can handle both classification and regression tasks.

CHAPTER 4: CONCLUSION

In our project of predicting football match winner using machine learning models, we evaluated multiple algorithms including Logistic Regression, Decision Tree, Gaussian Naive Bayes, Random Forest. After training and testing these models on our dataset, we observed that the Random Forest model outperformed the others in terms of accuracy, achieving an accuracy of 98.46% on the test set.

Based on these results, we conclude that the Random Forest model is the best option to use for predicting football match winners in our dataset. Other models did not achieve the same level of accuracy, and therefore we chose the Random Forest model for our prediction task.