DEPARTMENT OF COMPUTER & INFORMATION SYSTEMS ENGINEERING

BACHELORS IN COMPUTER SYSTEMS ENGINEERING

Course Code: CS-324

Course Title: Machine Learning

Open Ended Lab

TE Batch 2021, Spring Semester 2024

Grading Rubric

TERM PROJECT Group

Members:

Student No.	Name	Roll No.
S1	RASIB HASAN	CS-21071
S2	MAAZ BIN MANSOOR	CS-21094

CRITERIA AND SCALES				Marks Obtained	
				S1	S2
Criterion1: Data Collection					
0	1	2	3		
The student has not chosen a suitable dataset for predictive modeling.	The student has chosen a dataset, but it may not be suitable for predictive modeling, or it lacks enough features.	The student has chosen a suitable dataset for predictive modeling, and it has enough features to work with.	The student has chosen an excellent dataset for predictive modeling, which has rich features and is well-suited for the task.		
Criterion 2: Data Preprocessing					
0	1	2	3		
The student has not performed data cleaning, handling missing values, or encoding categorical variables	The student has performed basic data cleaning and handled missing values, but has not encoded categorical variables.	The student has performed data cleaning, handled missing values, and encoded categorical variables.	The student has performed thorough data cleaning, handled missing values effectively, and encoded categorical variables efficiently.		
Criterion 3: Exploratory Data Analysis (EDA)					
0	1	2	3		
The student has not performed exploratory data analysis (EDA) or provided minimal analysis with no meaningful insights.	The student has performed basic exploratory data analysis, but the analysis lacks depth, and insights are limited	The student has performed thorough exploratory data analysis, identifying important variables, correlations, and providing meaningful insights.	The student has performed exceptional exploratory data analysis, providing comprehensive insights, and utilizing a variety of visualization techniques effectively.		

Criterion 4: Feature Engineering					
0	1	2	3		
The student has not performed feature engineering.	The student has performed basic feature engineering, but has not created new features or scaled/normalized existing features.	The student has performed feature engineering, creating new features and scaling/normalizing existing features if required.	The student has performed advanced feature engineering, creating meaningful new features and effectively scaling/normalizing existing features.		
Criterion 5: Model Building					
0	1	2	3		
The student has not built any predictive models.	The student has built models using machine learning algorithms, but the implementation lacks depth, and multiple algorithms were not used.	The student has built models using multiple machine learning algorithms, implementing them using Python packages, and evaluated their performance.	The student has built models using multiple machine learning algorithms, implemented them both using Python packages and without Python packages, and		
			thoroughly evaluated their performance.		
Criterion 6: Model Evaluation					
0	1	2	3		
The student has not evaluated model performance or has done so inadequately.	The student has evaluated model performance but has not used different techniques or compared the performance of different models.	The student has evaluated model performance using different techniques, compared the performance of different models, and selected the best-performing model.	The student has thoroughly evaluated model performance using various techniques, performed a detailed comparison of different models, and selected the best-performing model based on comprehensive evaluation metrics.		
Criterion 7: Conclusion					
0	1	2	3		
The student has not provided a conclusion or has provided a conclusion with minimal insights.	The student has provided a basic conclusion with some insights but has not discussed model limitations or suggested improvements.	The student has provided a detailed conclusion with meaningful insights, discussed model limitations, and suggested improvements.	The student has provided an exceptional conclusion with comprehensive insights, thorough discussion of model limitations, and insightful suggestions for improvements.		
Criterion 8: Report					
0	1	2	3		
The submitted report is unfit to be graded.	The report is partially acceptable.	The report is complete and concise.	The report is exceptionally written.		
			Total Marks:		

PROJECT: PSL Match Outcome Prediction

INTRODUCTION

We always find win predictors in the t20 matches broadcasted in 2nd innings to predict the win percentages of the respective teams, thus, to get an intuition that how it is being performed and to apply the learned machine learning techniques and those apart from the course were experimented throughout this end-to-end machine learning project.

The project involves data collection, Exploratory Data Analysis (EDA), and training three machine learning algorithms: Logistic Regression, Random Forest, and Naïve bayes. The models are then evaluated based on performance metrics to determine their accuracy and effectiveness.

METHODOLOGY

Data Collection

The dataset used for this project is sourced from PSL (Pakistan Super League) match data, which includes details such as the year, match number, teams involved, innings, over, ball, runs, wickets, and match outcomes. This data is suitable for predictive modeling due to its rich features that can influence match outcomes.

Data Preprocessing

Data preprocessing steps included:

- **Data Cleaning** :
  Handling missing values, especially in the wicket and wicket_text columns.
  .
- **Data Transformation:**
  Creating new features like total_runs and wickets to capture cumulative statistics within an innings

Exploratory Data Analysis (EDA)

Model Description: EDA was conducted to understand the distribution of variables and identify correlations.

Key insights were:

- **- Distribution of fours and sixes throughout years.**
- **- Correlation between each variables.**
- **- Visualization of match outcomes based on different teams and innings.**
- **- Distribution of types of wickets**

Visualizations used include bar plots and pie plots to explore these patterns

Feature Engineering

Feature engineering steps included creating new features and selecting relevant ones for modeling. Important features considered were:

- **– balls left**
- **– runs left**
- **- is_four**
- **- is_six**

Used corr heatmap to choose best features for model training

Model Building

Multiple machine learning models were constructed to predict match outcomes, utilizing both Python packages custom implementations. These models include:

1. **Random Forest Classifier**
2. **Naive Bayes**
3. **Logistic Regression**

The models were developed both with the aid of Python packages like scikit-learn and through custom-built implementations to ensure flexibility and robustness in the prediction system.

Model Evaluation

The models were evaluated using accuracy, confusion matrix, and classification report. The performance comparison showed:

Random Forest Classifier:

```
Accuracy:

 0.8052511887533595

 Confusion Matrix:

 [[1541  518]
 [ 424 2354]]

 Classification Report:

              precision  recall  f1-score   support

           0       0.78      0.75      0.77      2059
           1       0.82      0.85      0.83      2778

    accuracy                           0.81      4837
   macro avg       0.80      0.80      0.80      4837
weighted avg       0.80      0.81      0.80      4837
```

Logistic Regression:

```
Accuracy:

 0.7868513541451313

 Confusion Matrix:

 [[1503  556]
 [ 475 2303]]

 Classification Report:

              precision   recall  f1-score  support

           0       0.76      0.73      0.74      2059
           1       0.81      0.83      0.82      2778

    accuracy                           0.79      4837
   macro avg       0.78      0.78      0.78      4837
weighted avg       0.79      0.79      0.79      4837
```

Naive Bayes:

```
Accuracy:

 0.7066363448418441

 Confusion Matrix:

 [[ 948 1111]
 [ 308 2470]]

 Classification Report:

              precision  recall  f1-score   support

           0       0.75      0.46      0.57      2059
           1       0.69      0.89      0.78      2778

    accuracy                           0.71      4837
   macro avg       0.72      0.67      0.67      4837
weighted avg       0.72      0.71      0.69      48
```

Conclusion

Random Forest is the Best Model Based on the evaluation metrics for the three models (Random Forest, logistic regression, and Naive Bayes), it is evident that the Random Forest (RF) model performs the best on the cricket training data. Here's why:

Highest Accuracy: Random Forest: 80.53% Logistic Regression Classifier: 78.54% Naive Bayes: 70.66% The Random Forest model achieves the highest accuracy, indicating that it correctly predicts the outcomes more often than the other models.

Confusion Matrix Analysis: The Random Forest model has the highest number of correct predictions for both classes (1541 true positives and 2354 true negatives), with fewer misclassifications compared to logreg and Naive Bayes.

Techniques Used By Random Forest: The combination of ensemble learning, bagging, feature randomness, variance reduction, handling of missing data and outliers, feature importance, and its non-parametric nature makes Random Forest a powerful and effective model. These techniques collectively contribute to its superior performance compared to Logistic Regression Classifier and Naive Bayes on the cricket training data.

Limitations

- Limited dataset size and scope might affect the model's generalizability.
- Models may overfit due to the limited number of matches in PSL history.

Future Improvements

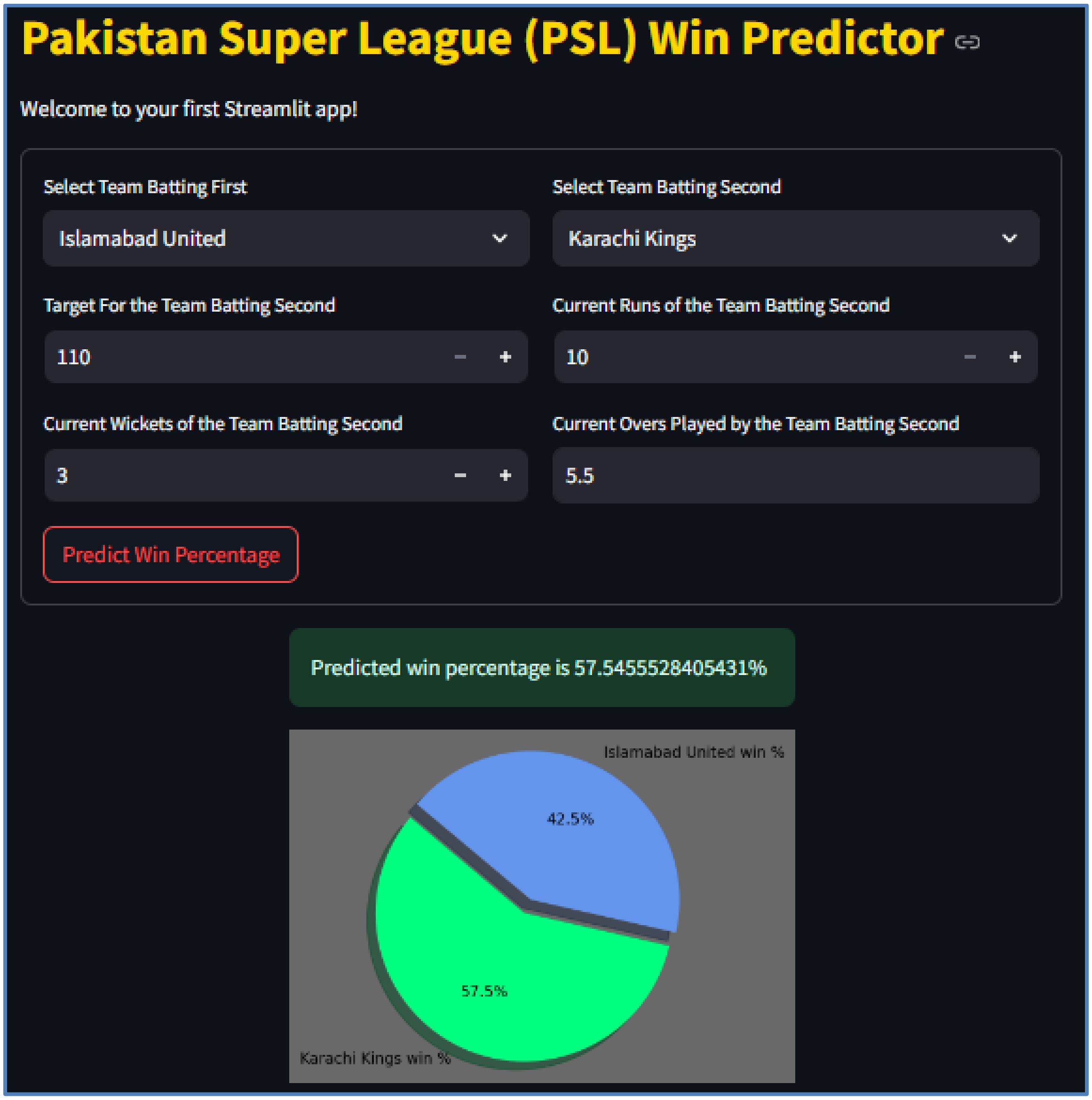
- Gathering more extensive datasets from different cricket leagues.
- Incorporating additional features such as player statistics and weather conditions.

Suggestions for Improvement

- Gathering more extensive datasets from different cricket leagues.
- Incorporating additional features such as player statistics and weather conditions.

Bonus: User-Friendly Interface

A user-friendly interface is developed using Streamlit to allow users to input match details and get predictions in real-time.



Pakistan Super League (PSL) Win Predictor

Welcome to your first Streamlit app!

Select Team Batting First: Islamabad United

Select Team Batting Second: Karachi Kings

Target For the Team Batting Second: 110

Current Runs of the Team Batting Second: 10

Current Wickets of the Team Batting Second: 3

Current Overs Played by the Team Batting Second: 5.5

Predicted win percentage is 57.5455528405431%