TASK 6 – Insights, Conclusion & Future Scope

1. Introduction

This project applied machine learning techniques to predict the most likely finalists for the FIFA World Cup 2026 using historical data from 1930 to 2022. The process involved cleaning data, visualising trends, building models, optimising performance, and finally generating predictions. The aim was to understand the important performance factors that influence how far a team progresses in a World Cup.

2. Key Insights

2.1 Data Insights

During data cleaning, we learned that raw match counts are not very meaningful on their own. Derived metrics like goal difference, win rate, goals per match, and FIFA ranking were much more powerful indicators of team strength.

2.2 EDA Insights

Visualisations clearly showed that teams with strong attacking and defensive balance — especially those with high positive goal difference and consistent performance — have historically reached finals more often. FIFA ranking, which summarises long-term performance, also had strong influence.

2.3 Model Insights

Two models were tested:

- Logistic Regression
- Random Forest

Random Forest performed significantly better because it can capture complex nonlinear patterns in football performance.

2.4 Optimisation Insights

Using GridSearchCV improved the Random Forest model considerably. The optimised model achieved:

- Accuracy: ~87%
- F1 Score: ~83%
- AUC: ~0.90

This means the model is reliable and captures the important features well.

2.5 Prediction Insights

The model predicted that the top finalist-probability teams for FIFA 2026 are:

- Argentina
- France
- England
- Brazil
- Portugal

Argentina ranked higher despite Germany having more match wins because:

- Higher win rate percentage
- Higher goal difference
- Better recent FIFA ranking and form
- Stronger overall performance stability

The model evaluates overall long-term performance, not just raw wins — which is why Argentina scored higher.

3. Final Predictions Summary (Task 6)

Based on the optimised model, the strongest contenders for the 2026 FIFA World Cup finals are:

Argentina, France, England, Brazil, and Portugal.

These predictions align with current world rankings and recent tournament performance patterns.

4. Conclusion

This project shows how machine learning can be meaningfully applied to sports analytics. Using engineered features and a tuned Random Forest model, we can generate realistic and data-driven predictions. The model performed strongly and produced results consistent with real-world expectations.

5. Future Scope

Future extensions can include:

- Adding player-level features such as age, top scorers, and injuries
- Using deep learning models like LSTMs to capture time-based patterns
- Creating a real-time dashboard to update predictions as matches happen
- Predicting not just finalists, but also group winners, semi-finalists, and tournament winners

6. References

- FIFA Official Website
- Kaggle World Cup Datasets
- Wikipedia (1930–2022 World Cup Data)
- Scikit-Learn Documentation