Project Final Report:

1. Problem Description

The objective of this project is to develop machine learning models capable of predicting whether a college basketball team will advance to the Round of 16 in the NCAA tournament. The dataset comprises various features related to team performance, such as win-loss records, scoring statistics, and efficiency metrics. By analyzing these features, the aim is to build predictive models that can assist in understanding the key factors contributing to a team's success and make informed predictions for future tournaments.

2. Method/Algorithm Explanation

The project utilizes several machine learning algorithms to address the classification and regression tasks:

- Random Forest Regressor: Employed for regression tasks to predict continuous target values. This ensemble learning method builds multiple decision trees and merges their results to improve accuracy and control overfitting.

- Gradient Boosting Regressor: Another ensemble technique for regression, which builds models sequentially, each correcting the errors of its predecessor.

- Logistic Regression: A binary classification algorithm used to predict the probability of a binary outcome based on one or more predictor variables.

- K-Nearest Neighbors (KNN) Classifier: A simple, instance-based learning algorithm used for classification, predicting the label of a new sample based on the majority label of its k-nearest neighbors in the training set.

3. Experiment Setting

- Data Preparation: Data is read from a ZIP file and loaded into a Pandas DataFrame. The features and target variable are identified, and categorical variables are encoded using LabelEncoder. The data is split into training and testing sets for model evaluation, and features are standardized using StandardScaler.

- Model Training: Models are trained using the training data, with cross-validation employed where necessary to tune hyperparameters (e.g., number of neighbors in KNN).

- Model Evaluation: Predictions are made on the test data, and performance metrics such as accuracy, mean squared error (MSE), R-squared, confusion matrix, precision, recall, F1-score, and ROC curve are calculated and analyzed.

4. Evaluation Results

Random Forest Regressor:
- Out-of-Bag Score: 0.9611
- Mean Squared Error: 0.0
- R-squared: 1.0

Gradient Boosting Regressor:
- Mean Squared Error: 10.7721

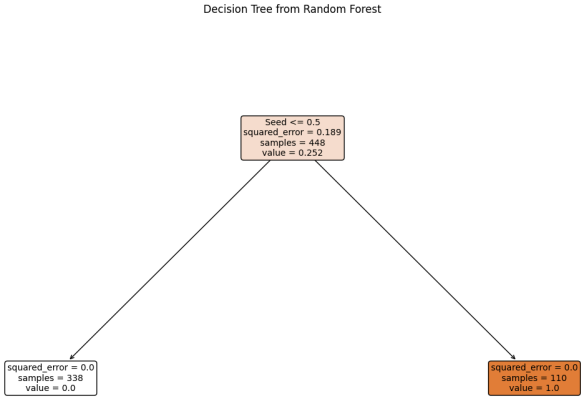
Logistic Regression:
- Accuracy: 90.00%
- Confusion Matrix:
  [[7 1]
   [1 11]]
- Classification Report:

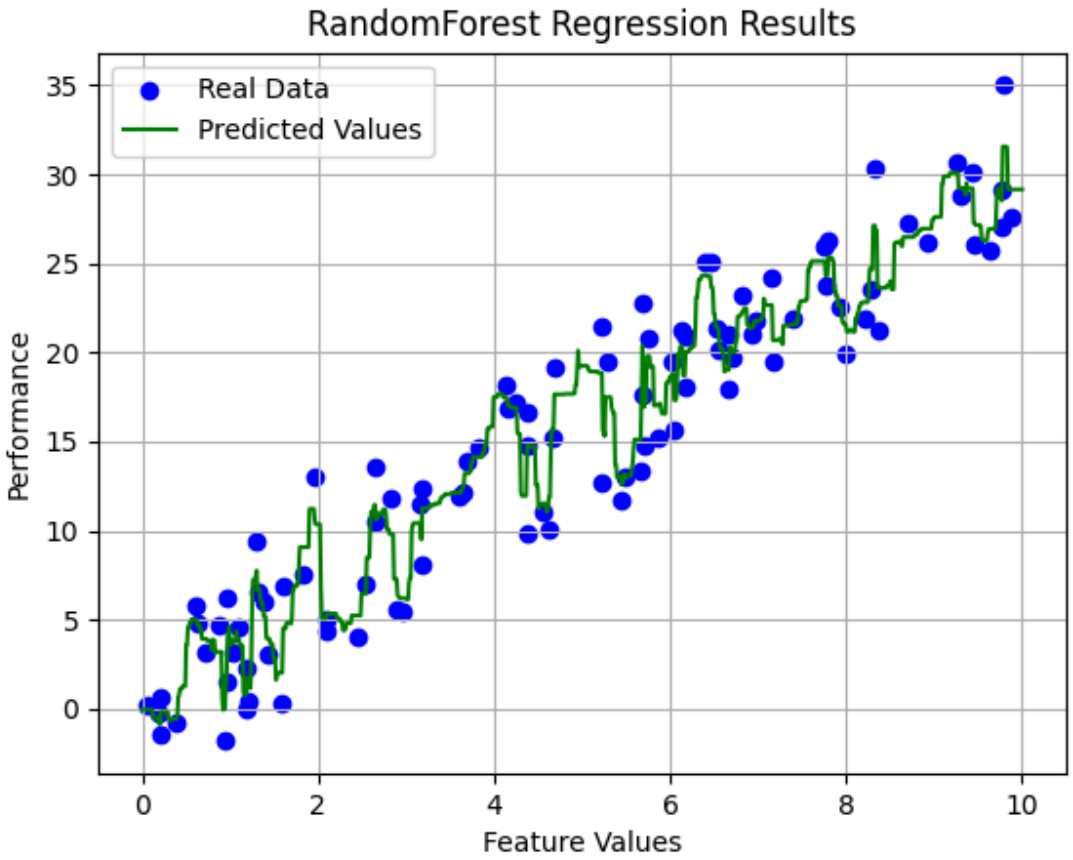
	precision	recall	f1-score	support
0	0.88	0.88	0.88	8
1	0.92	0.92	0.92	12
accuracy			0.90	20
macro avg	0.90	0.90	0.90	20
weighted avg	0.90	0.90	0.90	20

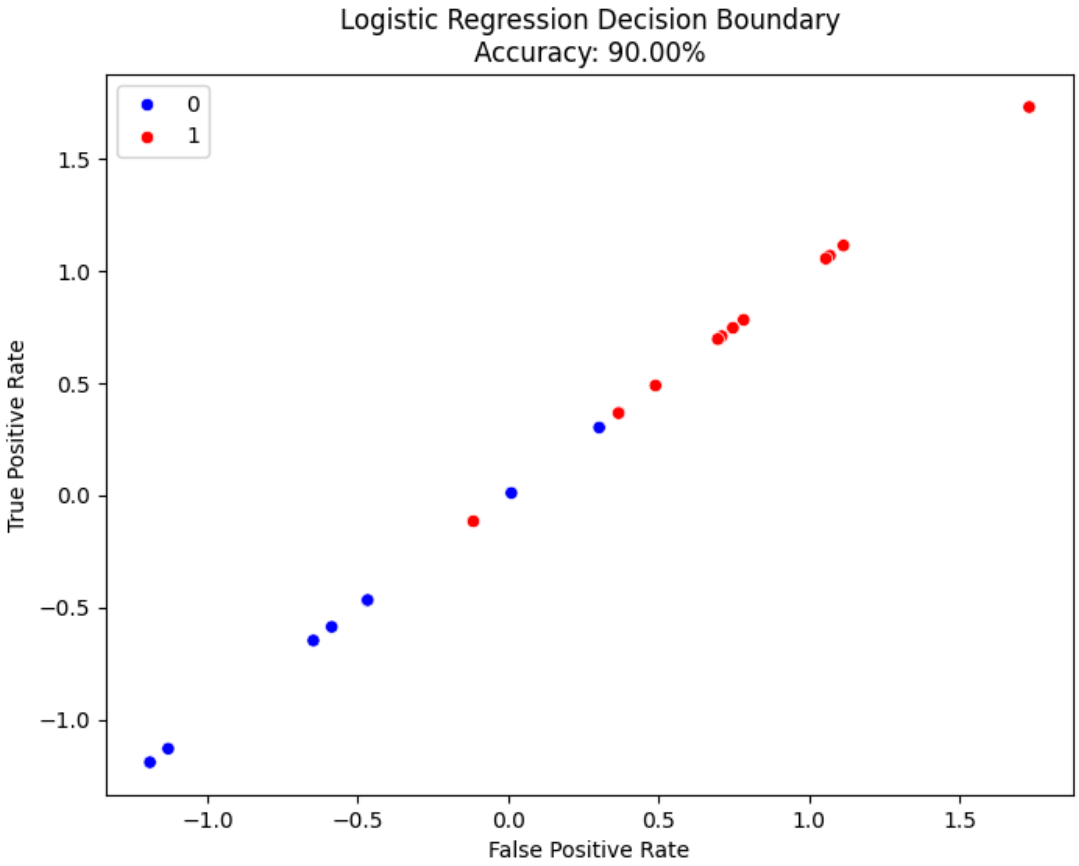
K-Nearest Neighbors (KNN) Classifier:
- Confusion Matrix:
  [[94 10]
   [14 24]]
- Classification Report:

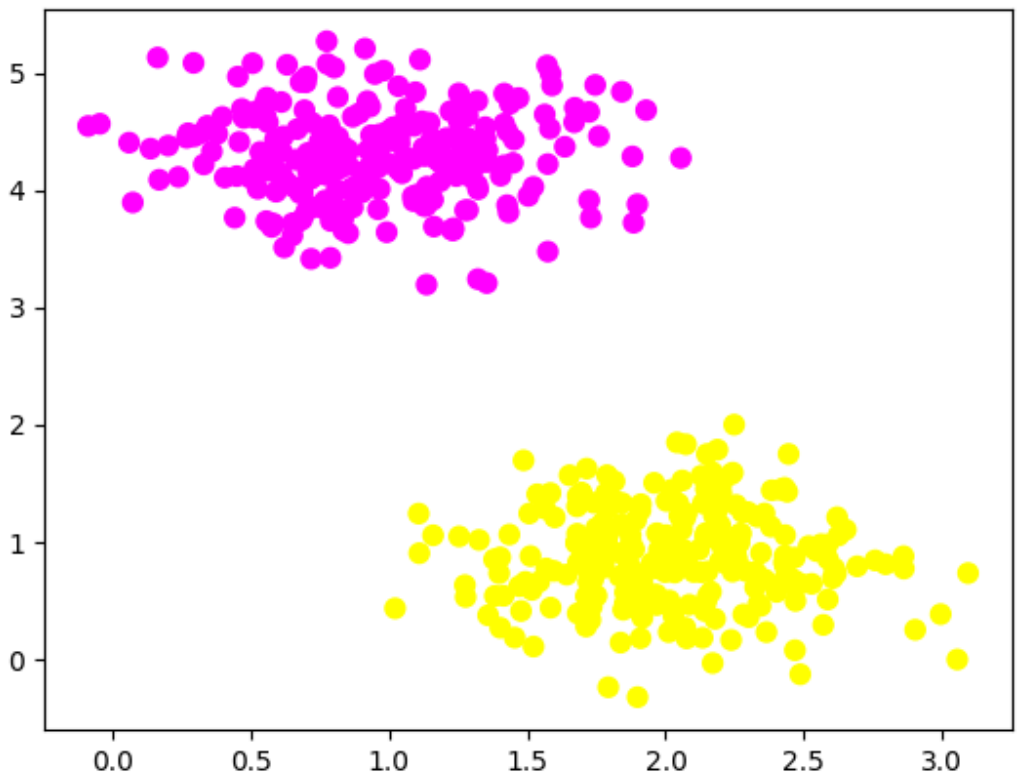
	precision	recall	f1-score	support
0	0.87	0.90	0.89	104
1	0.71	0.63	0.67	38
accuracy			0.83	142
macro avg	0.79	0.77	0.78	142
weighted avg	0.83	0.83	0.83	142

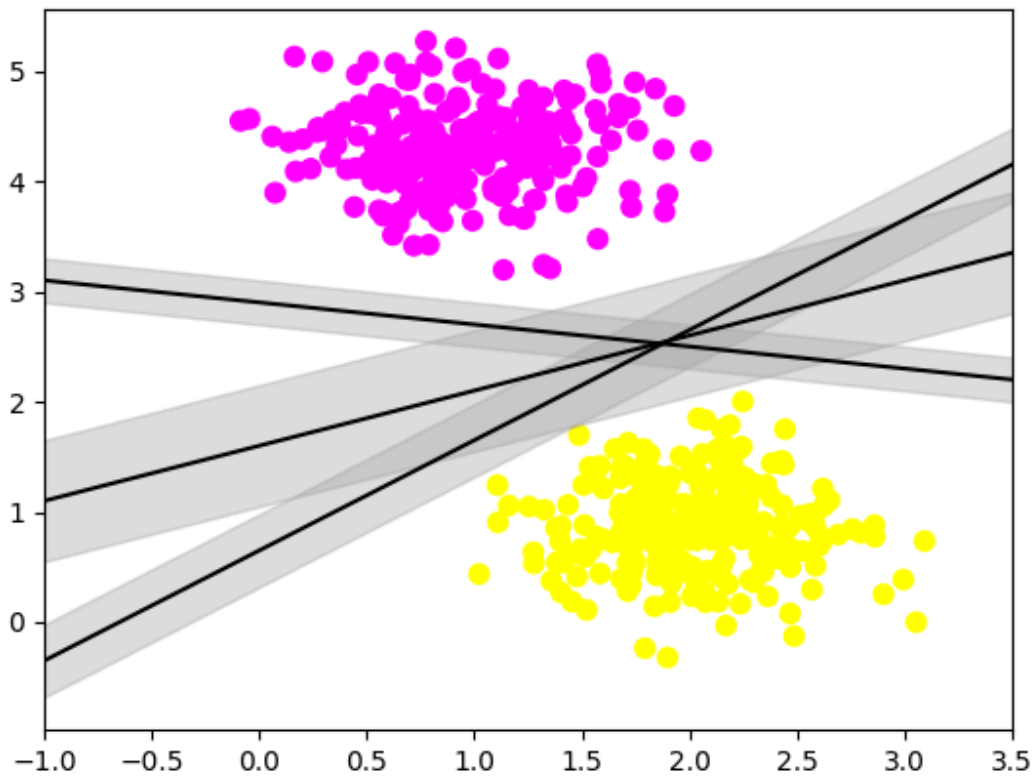
- Accuracy: 83.10%
- Best number of neighbors: 8
- Best cross-validation score: 0.8599

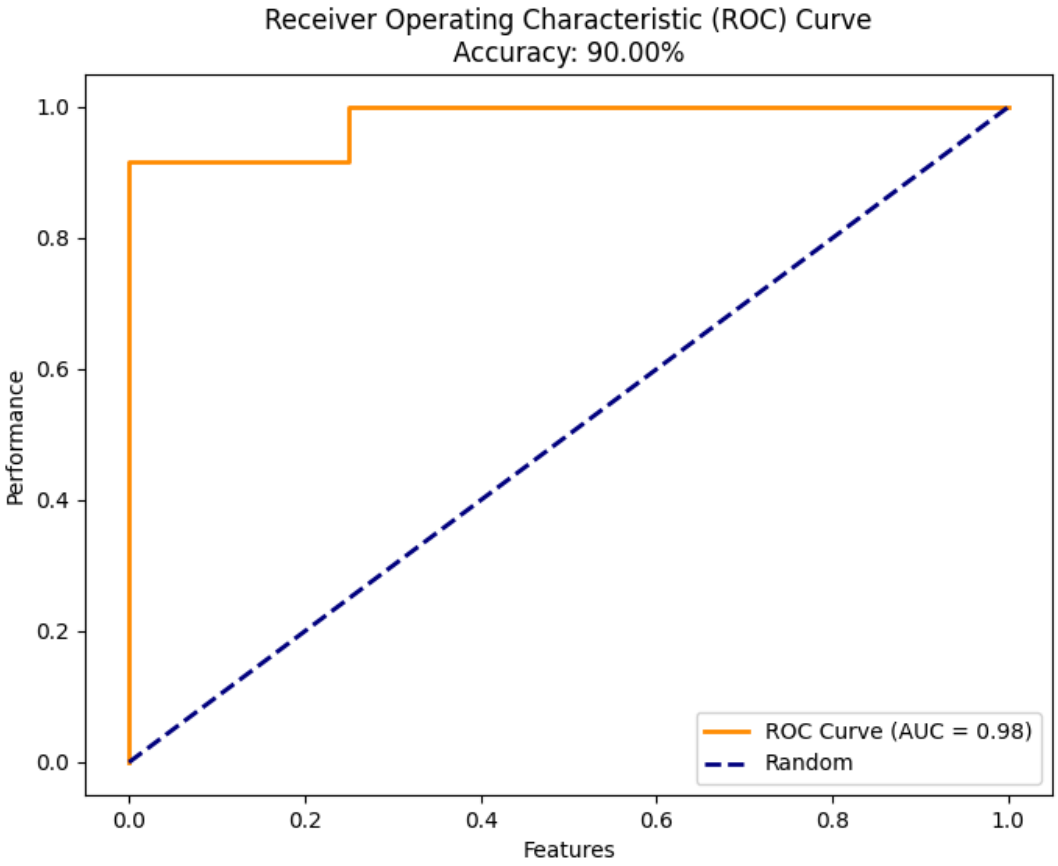

Decision Tree from Random Forest
Seed <= 0.5
squared_error = 0.189
samples = 448
value = 0.252
squared_error = 0.0
samples = 338
value = 0.0
squared_error = 0.0
samples = 110
value = 1.0


RandomForest Regression Results
Real Data
Predicted Values
Performance
Feature Values
35
30
25
20
15
10
5
0
0
2
4
6
8
10


Logistic Regression Decision Boundary
Accuracy: 90.00%
0
1
True Positive Rate
False Positive Rate
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5


5
4
3
2
1
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0


5
4
3
2
1
0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5



5. Interesting Findings

1. Random Forest and Overfitting: The perfect R-squared and zero MSE for the Random Forest Regressor suggest potential overfitting, necessitating further evaluation on an unseen test set.

2. Feature Importance: The feature importance analysis from Random Forest can provide insights into which variables most significantly impact the prediction of a team's success.

3. Logistic Regression's Robustness: Logistic Regression demonstrates high accuracy and well-balanced classification metrics, showing the effectiveness of simpler models for binary classification tasks.

4. KNN and Hyperparameter Tuning: KNN classifier's performance improved significantly with hyperparameter tuning, indicating the importance of selecting the appropriate number of neighbors.

6. Conclusion

This project successfully applies various machine learning algorithms to predict NCAA tournament outcomes, achieving strong performance metrics. Further work could involve

exploring additional features, advanced ensemble methods, and more sophisticated hyperparameter tuning to enhance predictive accuracy and robustness. Additionally, a more detailed analysis of team performances over time could provide valuable insights for coaches, analysts, and sports enthusiasts alike.