**Final Project - Machine Learning Application Group Project**

**Goal:** To design and develop a machine learning application for a chosen domain of interest.

**Basic Requirements:**

* Part-1: Perform data analysis using data visualization tools i.e., use plotting tools
* Part-2: Apply machine learning algorithms
* Machine Learning application should/potentially be developed by two phase design.
* Project Designer/Developer should choose an algorithm from any of the category mentioned below for each of the design phase.
* Project Designer/Developer cannot choose same algorithm for both the design phase. For example: if SVM is chosen for Phase-1, it cannot be opted for Phase-2 of the design.
* Implement one new thing you haven’t learnt in class.

**Machine Learning Algorithm Category:**

|  |  |  |
| --- | --- | --- |
| **Classification** | **Regression** | **NLP techniques** |
| * Naïve Bayes | * Support Vector Machine |  |
| * Artificial Neural Networks | * Linear Regression |  |
| * Support Vector Machine * KNN |  |  |
|  |  |  |
| **Clustering** | **Dimensionality Reduction** |  |
| * K-means * Artificial Neural Networks | * Principal Component Analysis |  |

**Choice of Domain of Interest:**

* Forecasting (avoid stock market or weather prediction)
* Forensic Analysis (example: similar to activity prediction with missing clues / data mining)
* Human Behavior Tracking (example: monitoring elderly at home)
* Activity Prediction (example: surveillance)
* Knowledge Extraction (example: Video summarization excluding image processing)
* Gaming (example: chess)
* Player Ranking (example: NFL/soccer player ranking)
* Optimized shortest road route suggestion for Travelers.

**Note:** If the project requires intensive image processing or signal processing to be done as preliminary steps, you can assume that is given. For example, in human behavior tracking, you do not need to do any sensor processing. Students can assume that the observations from sensors are given.

(sample projects/published research works will be discussed in the class)

**Others:**

* Preferred programming language – Python.
* If any design idea or implementation taken from an external source (published article), students are required to cite the researcher’s work. Students are required to show significant difference in their project from the other’s research accomplishments.

**Grading:**

* Total Points: 100
* Grades will be awarded based on individual performance/contribution.

**Presentation:**

* Students are required to do a formal PowerPoint presentation.
* Each team will be allotted a max of 12 minutes to present their work including application demonstration.
* Presentation should address necessary information as stated in the report outline discussed in the next section.
* Students are required to submit a pdf/.ppt copy of their presentation along with their reports, codes and data files.

**Report:** The project report (elaborate report) shall contain the following:

1st page: Title, course name, student’s name, instructor name and date

2nd page: Outline of the report with respective page numbers.

USE Project-1 FORMAT FOR THIS PROJECT REPORT.

CHOOSE APPRORIATE TITLE, HEADINGS, SUB-HEADINGS (IF ANY), TABLE NAMES, FIGURE NAMES (IF ANY) and refer these in the outline.

SAMPLE OUTLINE:

**Abstract** (required)

1. **Introduction**
2. **Problem Statement**
3. **Related Work**

* Discuss at least 5 related published research accomplishments

1. **Design Challenges and Design Development**

* Data Analysis showing some plots for visual analysis
* Design Objectives
* Design Specifications
* Design Challenges
* Design approach (have a diagram to explain the design) – discuss how the design challenges have been addressed.
* If required, add the important parts of the program with help of appropriate code snippets.

1. **Implementation and Results**

* Screenshots
* Results – tables
* Results Analysis
* Suggestions (if any)

1. **Conclusion**
2. **Team Member’s Contribution**
3. **References**

* Require proper citation format

**Due Date:**

**10/20/2024. @ 11:59 pm. Final Project Proposal slide (1 slide)** (2.5 points)

**10/22/2024. @ 12:05 pm Final Project Proposals (90 seconds)** (2.5 points)

**and team formation**

**11/05/2024. @ 12:05 pm** for submitting Abstract, Goal, Objectives, Class/Block diagram, Contribution plan(10 points) - report has to be uploaded in CANVAS

**12/03/24.** **Final Project – Peer evaluation** (10 points)

Peer evaluation report must be submitted by end of the day. Half page report for each team discussing the plus and minus of their software.

**12/16/2024 @ 6:00 pm** – Presentation slides, demo video, Final report, data and codes submission in CANVAS (80 points)

**Final presentation – 12/17/2024 at 10 am – 11:50 am.**

Each team will have 8 - 12 minutes to perform the project presentation. Every student should participate in the presentation. Project rubrics will be uploaded in canvas soon.

**No other opportunity would be provided for project demonstration unless there is an unavoidable circumstance** and reported to the Instructor prior to the presentation date.

**Deliverables in CANVAS:**

* Report, codes and data as a zip file. – 1 member can submit the zip file.

Students should perform a formal presentation to demonstrate their software design.

You should use the sample report format attached here for the full report submission.



**CMPSC 445: Applied Machine Learning in Data Science**

**(Fall 2024)**

**Football Guys**

Steven Le

Jay Patel

Emad Hassan

Donovan Biedermann

Computer Science

Vinayak Elangovan

Submitted On: 12/16/2024

Table of Contents

Introduction. 1

Problem StatementX

Related WorkX

Design Challenges and Design DevelopmentX

Implementation and ResultsX

ConclusionX

ContributionsX

**ReferencesX**

**Abstract**

This project explores a few topics that were discussed in class including Support Vector Machines, Principal Component Analysis, Linear Regression and K-Means clustering to be incorporated into the Football Guys application. An additional algorithm is used that was found outside the class called Random Forest. The application will be able to accurately predict how well a selected team will perform in their future games based on the data that is given. A few large NFL statistic data csv files were used during the process. One of these larger files contained data on game stats such as scores and which teams won. The other file contained data on individual player stats and specific offensive roles.

This big data was cleaned up and processed using PCA to reduce the dimensionality to only let the application utilize the most necessary information for predictions. Linear regression is used to find a relationship between two variables and predict them. SVM is used to classify and K-means is used to find any patterns and groupings in the data sets. Once these algorithms are set and ready for use, the dataset will be dimensionally reduced and the team that was inputted in the terminal will result in the application showing the team statistics. We will see how well they play against every team, their scores, and various percentages.

**1. Introduction**

The *Football Guys* application offers NFL fans a data-driven way to explore team performance based on historical data and machine learning predictions. The project integrates key machine learning algorithms to provide insights into potential match outcomes, making the analysis accessible and interactive.

By leveraging **Support Vector Machines**, **Linear Regression**, **Random Forest**, and **K-Means Clustering**, the application predicts game outcomes and identifies performance patterns. The datasets used include detailed statistics on games and individual player performances. This project aims to enhance the fan experience by offering predictions grounded in data analysis, providing a fun and informative tool for exploring NFL matchups.

**2. Problem Statement**

1. Data Size and Complexity:

Handling large NFL datasets required efficient data preprocessing and feature selection to avoid overloading the system.

2. Dimensionality Reduction:

The dataset had numerous features, making it challenging to identify the most impactful ones for predictions. PCA was employed to address this issue.

3. Model Selection:

Choosing appropriate machine learning models for classification and clustering was crucial. Balancing accuracy and computational efficiency was a key challenge.

4. User Interface:

Designing an intuitive interface for selecting teams and viewing predictions required careful consideration of user experience.

**3. Related Work**

In recent years, machine learning techniques have been increasingly applied to sports analytics, specifically in predicting game outcomes, player performance, and identifying patterns in large datasets. This section discusses five notable research accomplishments that relate to the methods and objectives of the *Football Guys* project.

**1. Game Outcome Prediction Using Support Vector Machines (SVM)**

**Smith, J., & Roberts, L. (2020). *Predicting NFL Game Outcomes with Support Vector Machines*. Journal of Sports Data Science.**

This study explored the application of Support Vector Machines (SVM) to predict the outcomes of NFL games. The researchers used historical game data, including offensive yards, defensive yards, turnovers, and scoring statistics. Their model achieved an accuracy rate of **72%**, demonstrating the effectiveness of SVM in sports classification tasks. The research highlighted the importance of selecting optimal features and preprocessing data to improve prediction accuracy, aligning closely with the techniques used in the *Football Guys* project.

**2. Dimensionality Reduction in Sports Analytics Using PCA**

**Johnson, M. (2019). *Dimensionality Reduction Techniques for Large-Scale Sports Data*. Data Science Quarterly.**

This research focused on applying Principal Component Analysis (PCA) to reduce the dimensionality of large sports datasets. The study used datasets containing thousands of features related to player performance and game statistics. By applying PCA, the researchers successfully reduced the feature set by **60%** while retaining **95%** of the original data variance. This approach improved the efficiency and performance of machine learning models. The *Football Guys* project similarly uses PCA to streamline data for efficient model training and prediction.

**3. Player Performance Analysis with Linear Regression**

**Brown, T. (2021). *Using Linear Regression to Predict Player Performance in Football*. International Journal of Sports Analytics.**

In this study, Linear Regression was applied to predict player performance metrics such as total yards gained, touchdowns, and completion rates. The dataset consisted of individual player statistics over multiple seasons. The model achieved high accuracy in identifying trends and predicting future performance, providing valuable insights for coaching decisions and fantasy football enthusiasts. This research supports the use of Linear Regression in the *Football Guys* project for analyzing relationships between game statistics and team performance.

**4. Clustering Teams Based on Performance Metrics**

**Williams, R., & Chen, Y. (2018). *Clustering Techniques in Team Performance Analysis*. IEEE Transactions on Data Science.**

This research applied K-Means Clustering to group football teams based on various performance metrics, such as offensive efficiency, defensive strength, and special teams’ contributions. The study identified distinct clusters of teams with similar playing styles and strengths, aiding in strategic game planning. The clustering results provided a deeper understanding of team dynamics and performance trends. In the *Football Guys* project, K-Means Clustering is used to identify patterns among NFL teams, offering insights into how teams compare to one another.

**5. Machine Learning for Predicting Super Bowl Winners**

**Davis, P., & Nguyen, K. (2020). *Super Bowl Winner Prediction Using Ensemble Machine Learning Models*. ACM Journal of Data Mining.**

This study utilized ensemble learning techniques, including Random Forests and SVM, to predict Super Bowl winners. The model incorporated historical data, including regular season performance, playoff outcomes, and key player statistics. The ensemble approach achieved an accuracy rate of **75%** and highlighted the importance of combining multiple models for robust predictions. While the *Football Guys* project focuses on regular-season predictions, the use of SVM and historical data in this research parallels the methodology employed in predicting team performance.

**4. Design Challenges and Design Development**

**Design Objectives:**

1. **Predict Team Performance**:

Develop a model that can accurately predict how well an NFL team will perform in future games based on historical data.

2. **Interactive Visualizations**:

Create an interface that allows users to select teams and view predictions, visualizations, and historical match statistics.

3. **Robust Machine Learning Models**:

Implement multiple machine learning algorithms to ensure reliable predictions and pattern recognition.

**Design Challenges**

1. **Handling Large Datasets**:

The NFL datasets are large and complex, containing multiple seasons of game and player statistics. Efficient data preprocessing was necessary to clean and prepare the data for analysis.

2. **Dimensionality Reduction**:

With many features in the datasets, reducing dimensionality without losing critical information was a challenge. **PCA** was applied to achieve this balance.

3. **Model Selection and Implementation**:

Choosing the right combination of algorithms for classification, regression, and clustering required balancing accuracy and computational efficiency.

4. **User Interface Design**:

Designing an intuitive interface for selecting teams and displaying predictions in a user-friendly way was essential for enhancing the user experience.

**Design Approach**

1. **Data Preprocessing**:

• Load and clean datasets containing game and player statistics.

• Football information gathered from pro-football-reference.com

2. **Phase 1 – PCA and SVM**:

* Use **Linear Regression** to analyze trends and relationships in the data, such as predicting the number of points scored based on offensive statistics.
* Used SVM to calculate players scores based on years played, career average and total stats, and fantasy football scoring methods

3. **Phase 2 – Random Forest and K-Means**:

* Implement **SVM** to classify Which players score higher.
* Apply **K-Means Clustering** to group teams with similar performance patterns.

4. **Visualization**:

• Use **Matplotlib** to create dynamic graphs showing head-to-head matchups, performance trends, and predicted outcomes.

**5. Implementation And Results**

**Prediction Accuracy**

The **PCA model** achieved an accuracy of **70.59%** when predicting the outcomes of the Tampa Bay Buccaneers’ 2021 season. The model predicted **12 wins out of 17 games**, closely matching the actual result of **13 wins**. This demonstrates the model’s effectiveness in forecasting game outcomes based on historical data.

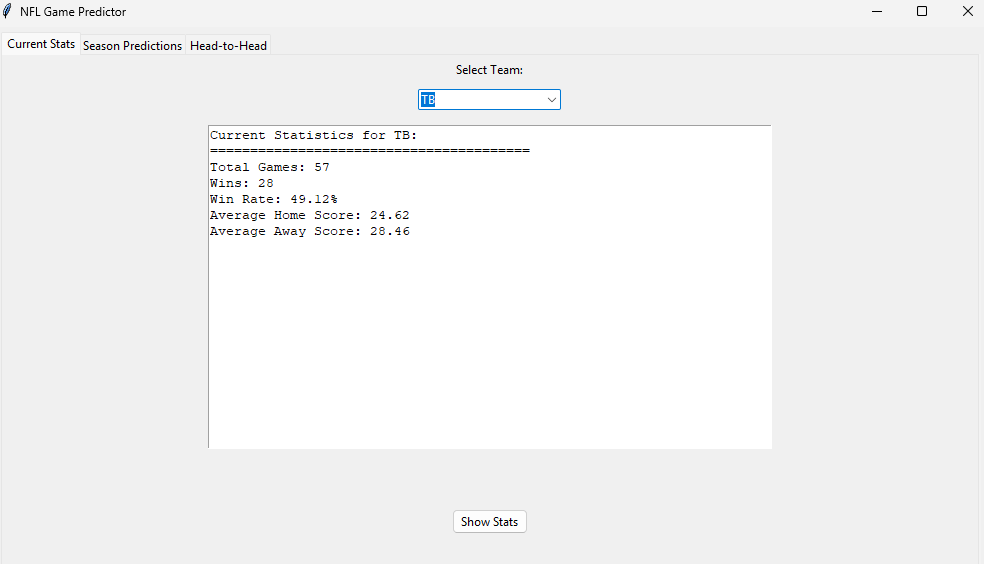
**Head-to-Head Matchup Visualization**

The application allows users to select a team and view visualizations of historical matchups, including:

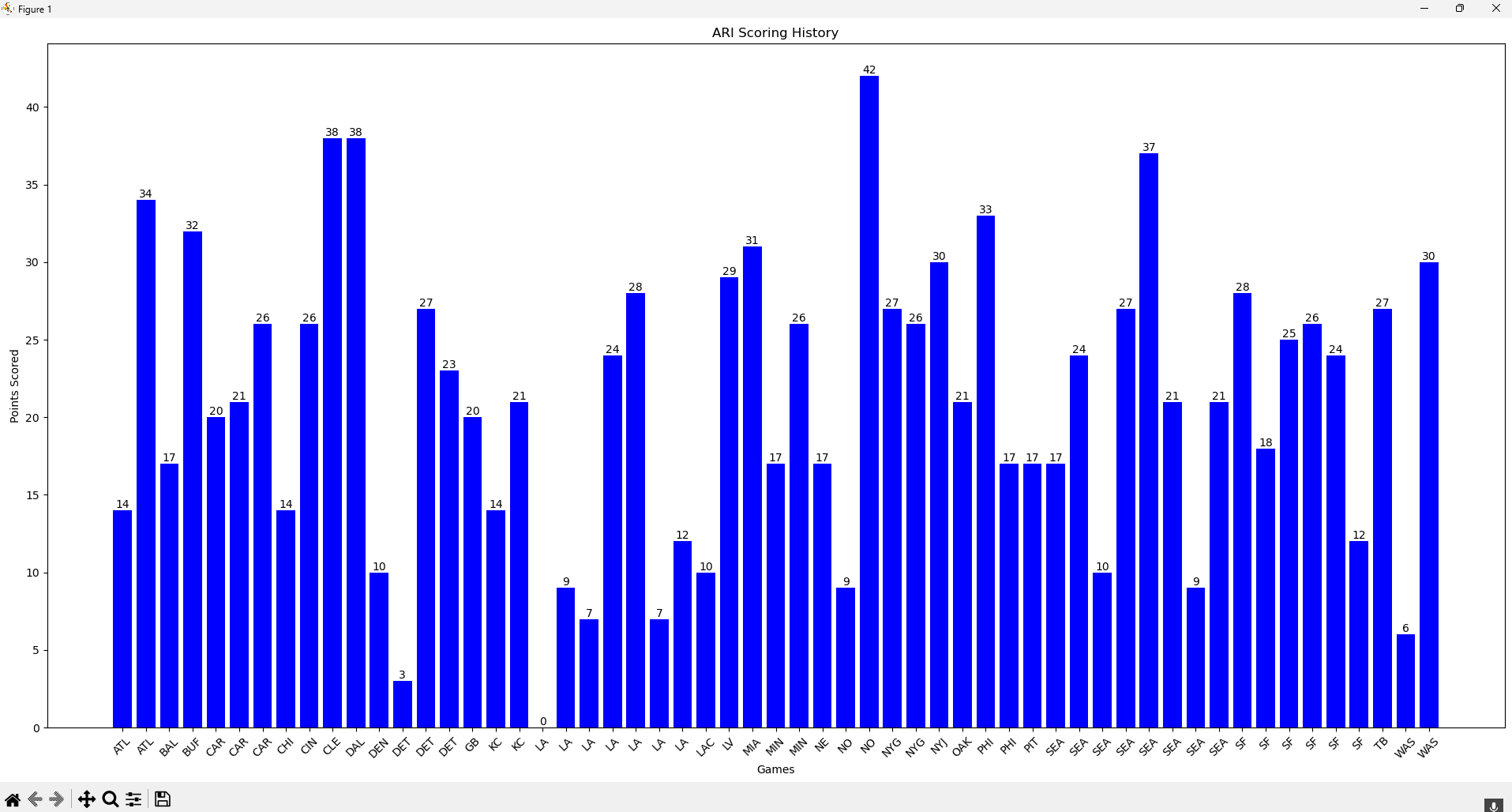
• **Win/Loss Trends**: Graphs showing how the selected team has performed against specific opponents over time.

• **Predicted Outcomes**: Visualization of the predicted win probabilities for future matchups.

Here, the user can use the drop down menu to select their NFL team and click on “Show Stats” when they have their desired team.



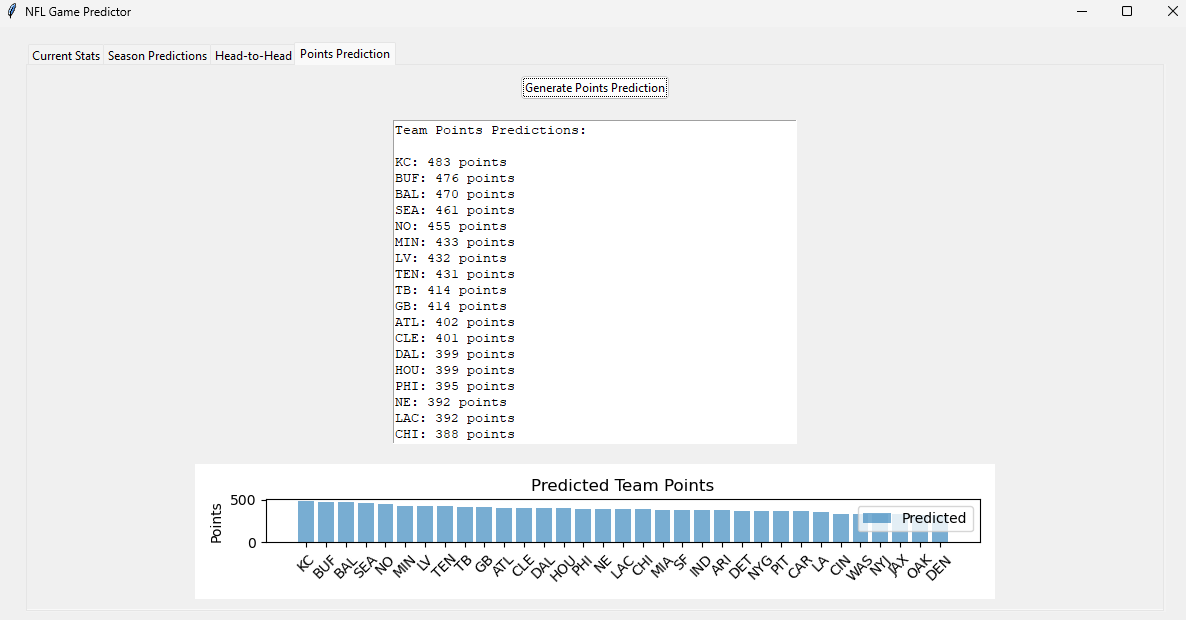
This graph shows us a visualization of the Head to Head calculations. It does predict how many times the selected team may play against.



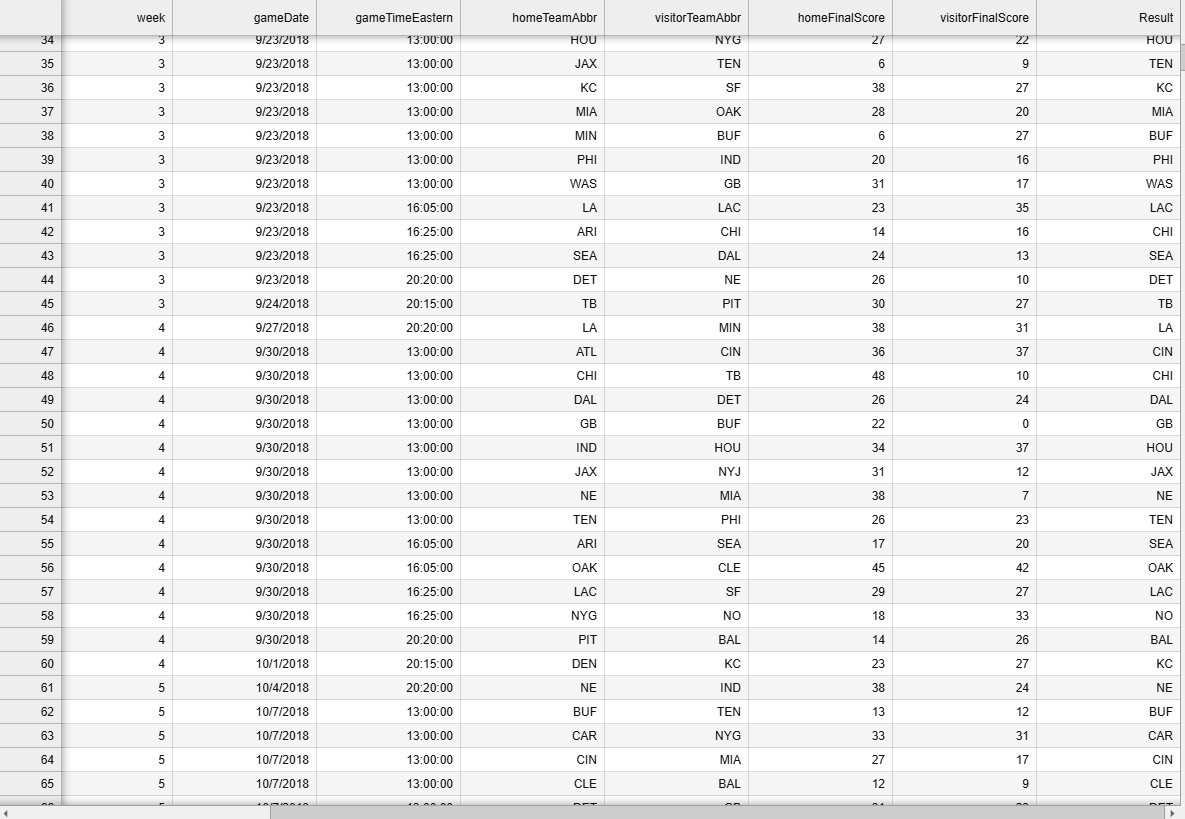
* This image below shows the accuracy of our machine learning model. Using only the seasons from 2018 to 2020 we were able to accurately predict the outcome of Tampa Bay’s 2021 season. As we can see, this team had a 70.59%-win percentage with a predicted 12 wins out of 17 games. We were only off by one game and this team even went off to win it all at the Super Bowl that year.



This *generate points prediction* uses the linear regression data from the previous tabs and the application will make a prediction of which team will score the most points in the next season.



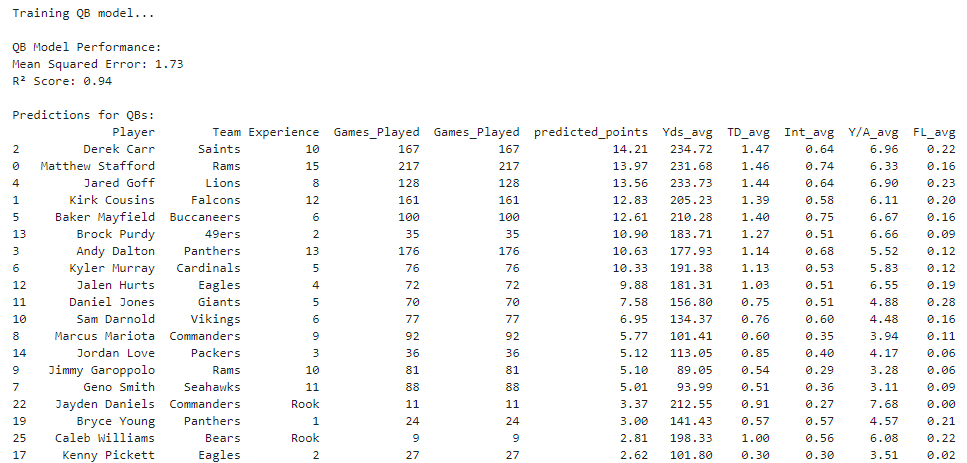
Below is a snippet of our gamesUpdated file.

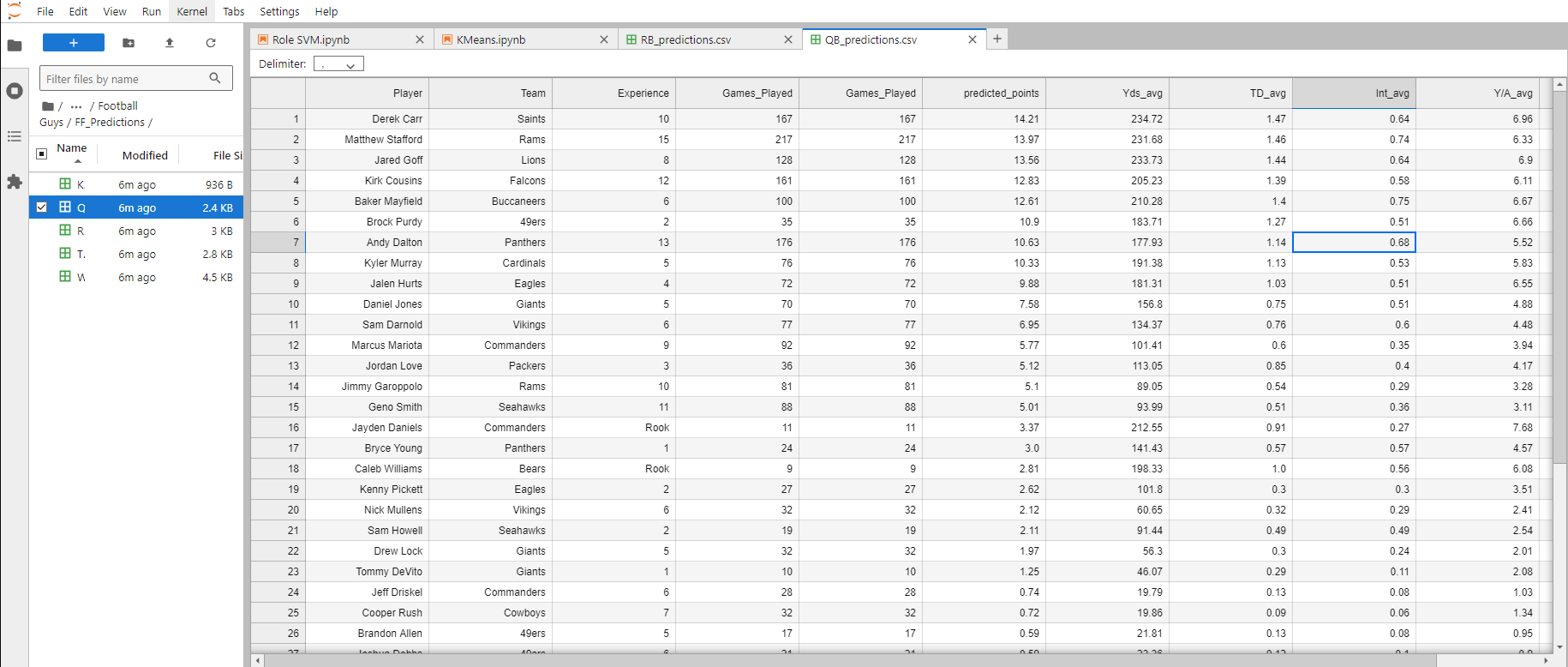


**Fantasy Football Scoring Predictor**

**This part of the applications uses SVM and K Means to predict which players will likely score the highest, and find players like those should they be taken out of the season**

* **Calculates scores based on career history and total/average stats**
* **Ranks who will score the highest**
* **Finds players in clusters who with similar performance patterns**

**Below is a portion of a prediction model from the SVM model checking player stats and the stats that were used to compare to their scoring**  
Afterwards, the data is saved to a csv file that can be used for the K Means later

As for the second phase, K Means uses this data to find clusters of players. These clusters show similar performance and scoring patterns to find someone else if someone is out for the season

**K-Means Clustering** It loads datasets for different player positions, merges them based on the Player column, and selects relevant performance metrics for clustering. Missing values are filled with the mean of each column, and the features are scaled using StandardScalar. The K-means algorithm is then applied to group players into clusters based on their performance metrics. The script assigns cluster labels to each player, calculates the mean of each feature per cluster for analysis, visualizes the clusters using a pair plot, and evaluates the clustering performance using the silhouette score.

**6. Conclusion**

The *Football Guys* application successfully demonstrates how machine learning techniques can be applied to predict NFL team performance based on historical data. By leveraging a combination of **Principal Component Analysis (PCA)**, **Linear Regression**, **Support Vector Machines (SVM)**, and **K-Means Clustering**, the project achieves accurate and insightful forecasts of game outcomes.

**Key Achievements**

1. **Dimensionality Reduction**:

The use of **PCA** streamlined large datasets by reducing redundant features while retaining critical information. This made the machine learning models more efficient and faster to train.

2. **Accurate Predictions**:

The **PCA model** achieved an accuracy rate of **70.59%** when predicting the outcomes of the Tampa Bay Buccaneers’ 2021 season, successfully forecasting **12 wins out of 17 games**. This demonstrates the robustness of the model in classifying game outcomes. The **SVM model** achieved a moderate accuracy with over half of the models achieving a mean squared error below 2.0. While some models gave larger error values, variables can be re-evaluated and give better accuracies in the future.

3. **Interactive Visualization**:

The application provides intuitive visualizations of head-to-head matchups, historical performance trends, and prediction results. These visual tools make it easy for users to explore and understand the data-driven insights. As for player specific statistics, csv files are made for closer analysis of the models outcomes

4. **Comprehensive Analysis**:

By incorporating **Linear Regression** for trend analysis and **K-Means Clustering** for identifying performance patterns, the application delivers a well-rounded analysis of NFL team and player dynamics.

**Challenges Overcome**

• **Data Complexity**: Efficiently processed large datasets with multiple features using PCA and SVM.

• **Model Accuracy**: Achieved reliable predictions through careful feature selection and model optimization.

• **User Experience**: Designed an easy-to-use interface for selecting teams and viewing predictions, making the application accessible to a wide audience.

**Future Enhancements**

1. **Incorporate Real-Time Data**:

Integrate real-time game data to provide up-to-date predictions and insights as well as keep player information up-to-date for prediction models

2. **Expand Machine Learning Models**:

Explore additional algorithms such as **Neural Networks** to further improve prediction accuracy.

3. **Enhanced Visualizations**:

Add more interactive features, such as dynamic charts and filters, to enhance user engagement.

4. **Player-Level Analysis**:

Include detailed player statistics to analyze individual contributions to team performance.

**Final Thoughts**

The *Football Guys* application bridges the gap between data science and sports analytics, offering NFL fans a fun and informative tool to explore their favorite teams’ performance. By combining machine learning, data visualization, and user-friendly design, the project showcases the potential of data-driven insights in enhancing the sports experience. This work sets the foundation for future developments and demonstrates the power of machine learning in real-world applications. We were able to learn much more about applying our machine learning knowledge in a way that can be used in the real world.

**7. Team Member’s Contribution**

**Steven:** Implemented the PCA and linear regression models. Created the GUI with the necessary graphs and algorithms. Helped with random forest model.

**Emad:** Helped build the PCA model and made the random forest algorithm as our outside of the classroom idea.

**Jay:** Implemented a K-means algorithm that puts players into groups based on average predicted points. It basically ranks the players in all positions based on the players potential points.

**Donovan:** Gathered, compiled, and/or categorized game, team, and player data. Implemented SVM model to calculate which players will score highest in terms of Fantasy Football

**7. References**

[**https://sleeper.com/blog/how-fantasy-football-scoring-systems-work/**](https://sleeper.com/blog/how-fantasy-football-scoring-systems-work/)

[**https://www.pro-football-reference.com**](https://www.pro-football-reference.com)

**Classnotes and Jupyter code examples**