Complete this document as part of your project plan about your final project. Your answers need to be short and brief. About one to three sentences for each question is fine. Write your answers under each question below. Make sure to maintain the question numbering.

Maximum Grade Points = 50 points

1. What is purpose of your project? (This should include the algorithms you are comparing, the output variable and the input variables captured in a general way).

   Sports analytics is nothing new, but what is relatively new in the industry of Professional Athletics would be E-sports, bringing a wide variety of data. Every video game will have data in common, but there's also a lot of data unique to each individual game (*e.g. baseball vs basketball*). I will be comparing various classification algorithms to see which one is most appropriate to the data

2. What is the significance of your project? That is, why is this project important? What potential positive impact does your project have?

   As seen on T.V. sports analytics are a big part of what makes up the "experience". Insights from game data can be advantageous to all those involved, from players to officials, and even game developers.

3. What is the research question? (This should include the algorithms you are comparing, the output variable and the input variables captured in a general way. A research question is the purpose of your project flipped into a question).

- What Machine Learning models would be appropriate for predicting whether a player is mvp or not?

- Can an equation be derived to calculate what will be called "game performance" to predict whether a player is mvp or not?

4. What algorithms will you be comparing to answer your research question?

   I will be comparing logistic classification, random forest, and decision tree algorithms

5. What is your output variable?

   y = all_players_df['mvp']

6. What are the input variables for your analysis? You should list at least five input variables in the dataset to be used for analysis/modeling.

   ```
   X = all_players_df[
       ['shots', 'shots_against', 'goals', 'goals_against', 'saves', 'assists',
        'score', 'avg_distance_to_ball', 'avg_distance_to_ball_possession',
        'avg_distance_to_ball_no_possession', 'avg_distance_to_mates',
       'time_defensive_third', 'time_neutral_third', 'time_offensive_third',
        'time_defensive_half', 'time_offensive_half', 'time_behind_ball',
        'time_infront_ball', 'time_most_back', 'time_most_forward',
       'goals_against_while_last_defender', 'time_closest_to_ball',
        'time_farthest_from_ball', 'percent_defensive_third',
        'percent_offensive_third', 'percent_neutral_third',
        'percent_defensive_half', 'percent_offensive_half', 'percent_behind_ball',
        'percent_infront_ball', 'percent_most_back', 'percent_most_forward',
        'percent_closest_to_ball', 'percent_farthest_from_ball']
   ]
   ```

7. What kinds of data preprocessing are you anticipating performing on the dataset before modeling?

   - Conversion of HTML data

   - Unflattening of lists of dictionaries

   - Removal or replacement of null values

8. What exploratory data analyses (on which variables) do you plan to do before modeling?

   - To have tables of each game with stats of each team's performance w/ complementary visuals

   - Descriptive statistics

   - Dataframe of player data constructed from earlier exploratory analysis steps

9. What visualizations (on which variables) do you plan to do before modeling?

   - To have the input variables against the output variable on a scatter plot

   - Bar plot of team statistics by game

10. What is the link to your data source? If your dataset is not from an internet data source, you will need to upload it to GitHub and share the link here. Avoid using common internet datasets such as Iris dataset, mushroom dataset, etc. Your dataset should be something interesting to you and uncommon). Feel free to use a dataset from this link: https://vincentarelbundock.github.io/Rdatasets/datasets.html or from some other source

such as Kaggle competition but don't present any copy and paste internet solutions for your final project. You will use the tools you have learned from this course to do your own project independently. Please, provide the link to your data source here:

https://ballchasing.com