**Student Name:** David Haynes (z1767127)

**Data Source:** The data chosen for my project was obtained through Kaggle which consists of NFL combine data from 2010 through 2023. I did not have to combine any data to complete my analysis since the file was available as a csv download. There are a total of 15 columns/variables and 4,741 observations/rows in the data set. I do not have to keep this data confidential since it is on a public website.

**Python Code:**

[**https://colab.research.google.com/drive/1t6jPUNmFA3NE4s42KU7XvUp7bxIr0rXo?usp=sharing**](https://colab.research.google.com/drive/1t6jPUNmFA3NE4s42KU7XvUp7bxIr0rXo?usp=sharing)

* **Describe the dataset:** NFL combine data from 2010 through 2023 was selected for this project. The data was obtained via Kaggle and consists of 4,741 rows and 15 columns which is a shape of (4,741, 15). The data set consists of 1 integer, 1 boolean, 4 objects, and 9 floats. Upon initial analysis of the data, it was discovered there was a large amount of missing data. The total number of missing entries was 11,500 for 16.17%. Imputation was conducted for all the missing entries. KNN imputation with neighbors = 2 was conducted for all the float values to best uphold the integrity and accuracy of the data. Imputation was only needed for one additional variable, an object ‘height,’ which was imputed with mode to remain consistent in the data set and give realistic findings. After imputation, the total number of missing values is zero.
* **Apply two types of machine learning techniques to analyze the data:** The two types of machine learning techniques performed to analyze the data include K-means clustering and logistic regression.
* **Specify analytic questions for each technique**: K-means clustering is used since two numerical variables are being analyzed. The question to be answered is whether 40-yard dash times have an impact on the pick number in the draft? Logistic regression is used since analysis is conducted on a binary variable “Drafted” and aims to build a model that can confidently determine if a player will be drafted based on player statistics. Can it be determined if a player will be drafted based on NFL combine player numerical values?
* **Elaborate results from both statistics and practical perspectives**: Following imputation, I wanted to run a K-means cluster on 40-yard dash time and the number sequence a player was picked in a draft. I have watched football my entire life and firsthand know faster players shine in the NFL. I was very interested in seeing actual data on the subject and the clusters that form. The elbow curve tapered off after 3 clusters which is why I chose 4 clusters for the analysis. As shown in the attached analysis, we can see in the scatter plot that a player who was picked at or below 100 with a 40-yard dash time of under 4.75 are clustered together. This is to be expected since quicker players tend to hold more value in the NFL. On the other hand, players with a 40-yard dash time of greater than 5 are clustered with picks above 125. This seems accurate since slower players are typically less desirable in the NFL. The remaining clusters include pick numbers greater than 125 with 40-yard dash times of less than 4.75, and picks below 100 with a 40-yard dash time of greater than 5. These seem to be the middle range players that were drafted in respect to 40-yard dash times and are more related to one another than the first two clusters described above since these two clusters are connected. Based on this analysis it can be determined that 40-yard dash times do impact a players pick number in the draft.

Independent variables position, school, height, weight, 40yd, vertical, and bench were used in the logistic regression model to predict the dependent variable drafted. The goal is to create a well-fitting model that can help determine if a player will be drafted based on their personal statistics at the combine. Once assigning categorical variables data values, a confusion matrix can be seen displaying the predicted vs actual values. As shown in the matrix, the model accurately predicted 508 (X) values from the test data set for players who were drafted and 131 (X) values for players who were not drafted. However, there were 310 predictions, either false positives or false negatives which impact the strength of the model. Also, the AUC score is 0.687 which is teetering on acceptable. Caution should be used when relying solely on this model. When looking at the classification report, the accuracy for the model is 0.67 which leads to the conclusion that the model is not good and that it cannot be determined if a player will be drafted based on NFL combine player statistics.

**Summary:** The attached python analysis was conducted on NFL combine data from 2010 through 2023 that was obtained via Kaggle with a shape of (4,741, 15). The data set consists of 1 integer, 1 boolean, 4 objects, and 9 floats. The data set initially consisted of 11,500 missing entries for 16.17%. KNN imputation was conducted for all the float values to best uphold the integrity and accuracy of the data. The remaining data was imputed with mode to remain consistent in the data set and give realistic findings. To give meaningful insights, the two types of machine learning techniques chosen to analyze the data include K-means clustering and logistic regression. These methods were chosen based on the numerical and binary categorical data available in the data set.

Regarding K-means clustering, the question is whether 40-yard dash times have an impact on the pick number in the draft? Logistic regression is optimal for the binary categorical variable “drafted.” Can it be determined if a player will be drafted based on NFL combine player numerical values? The following summary analysis provides answers to these questions, as well as further analysis to the data.

A K-means cluster analysis was performed on 40-yard dash times and the number sequence a player was picked in a draft. The number of clusters chosen was 4 based on the elbow curve. According to the scatter plot, a player who was picked at or below 100 with a 40-yard dash time of under 4.75 are clustered together. Players with a 40-yard dash times greater than 5 are clustered with picks above 125. The remaining clusters include pick numbers greater than 125 with 40-yard dash times of less than 4.75, and picks below 100 with a 40-yard dash time of greater than 5. According to this analysis, faster players in the 40-yard dash are picked higher in the draft than slower players and it can be determined that 40-yard dash times do impact a players pick number in the draft.

A logistic regression model including independent variables position, school, height, weight, 40yd, vertical, and bench was built to predict the dependent variable drafted. A confusion matrix can be seen displaying the model accurately predicted 508 (X) values from the test data set for players who were drafted and 131 (X) values for players who were not drafted. There were also 310 predictions as either false positives or false negatives. The AUC score is 0.687 and the accuracy for the model is 0.67. Based on these findings, it cannot be determined that a player will be drafted based on NFL combine numerical values.

Overall, the analysis performed a form of supervised and unsupervised learning in respects to logistic regression and K-means clustering. Meaningful insights were gathered from the K-means analysis regarding 40-yard dash time impacting draft number. Logistic regression was chosen to determine if a player will be drafted based on the data provided. The model utilized many variations of independent variables with the same outcome of low AUC scores below 0.70. While it is disappointing that the logistic regression model isn’t as accurate as expected, it aligns with the uncertainty of the NFL.