CSDS 391 P2 Extra Credit

Name: Rohan Singh
Case id: rxs1182@case.edu

Topic that I'm exploring:

For this extra credit assignment I will be exploring the classification of NFL Quarterbacks. One of the biggest questions that has been bugging my generation is "Is Joe Flacco an Elite Quarterback?". I will try to answer this question by using K-Means clustering and a Neural Network.

Note: I was a little confused about the meaning of trivial example, so I chose not to use the iris dataset, because that is the most common example on the internet, and instead chose NFL quarterbacks.
Note: The other part of extra credit is at the bottom (asnwering questions)

Approach:

The variables that I will be using for the classification of NFL Quarterbacks will be:

- Net Yards per attempt (Will use this in 2-D version for plot)
- Passer Rating (Will use this in 2-D version for plot)
- Total Yards per game: Passing + Rushing
- Touchdown to Interception ratio

I will be retrieving this data from a CSV file and reading that by using the pandas library and then using my algorithms to classify them as "Elite" or "Not-Elite".

Algorithm:

I started off by pre-processing the data to make sure that they were in the correct format to be used in the algorithms.

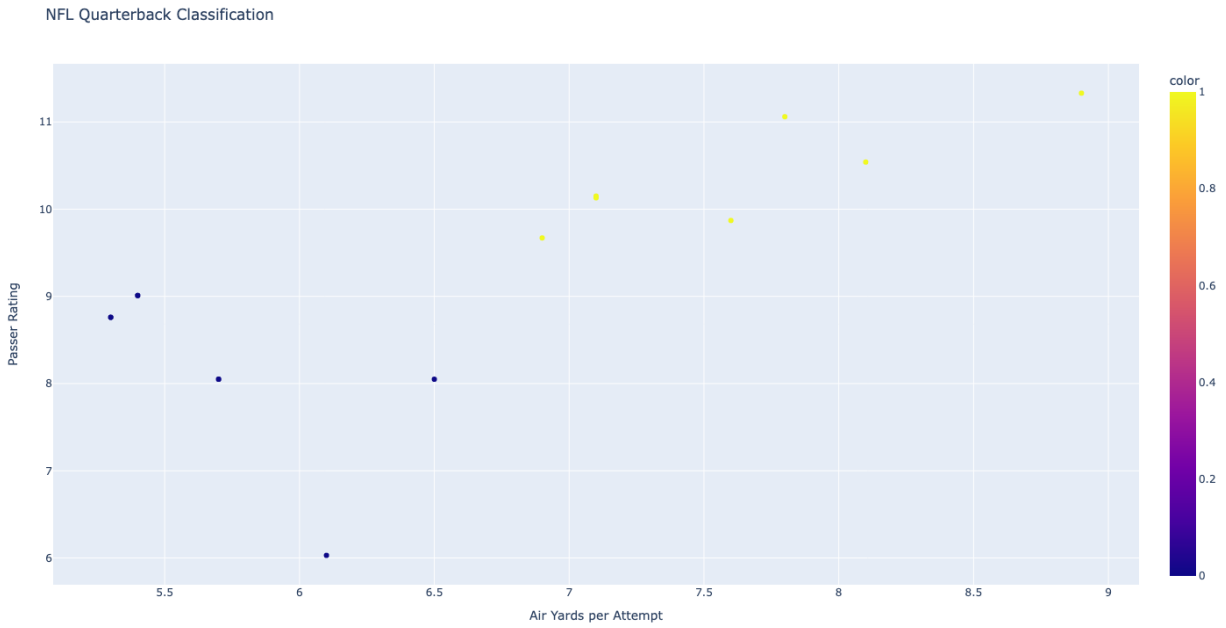
For K-means clustering, I will be using the Scikit-Learn KMeans clustering algorithm. The mean squared error for K Means clustering with Scikit-Learn was less than that of my own code, because they used multiple iterations and not just a single iteration, so I decided to use their function. For the regular classification I used all 4 parameters.
I ran it once more with just 2 parameters (Net yards per attempt and Passer Rating), so that I could plot the results.

For the Neural Network, I used my own algorithm, because the error using my gradient descent optimization algorithm, the mean squared error was very low with iris dataset, whereas the mse was higher in the case of Scikit-Learn's Multiple-Layer Perceptron.

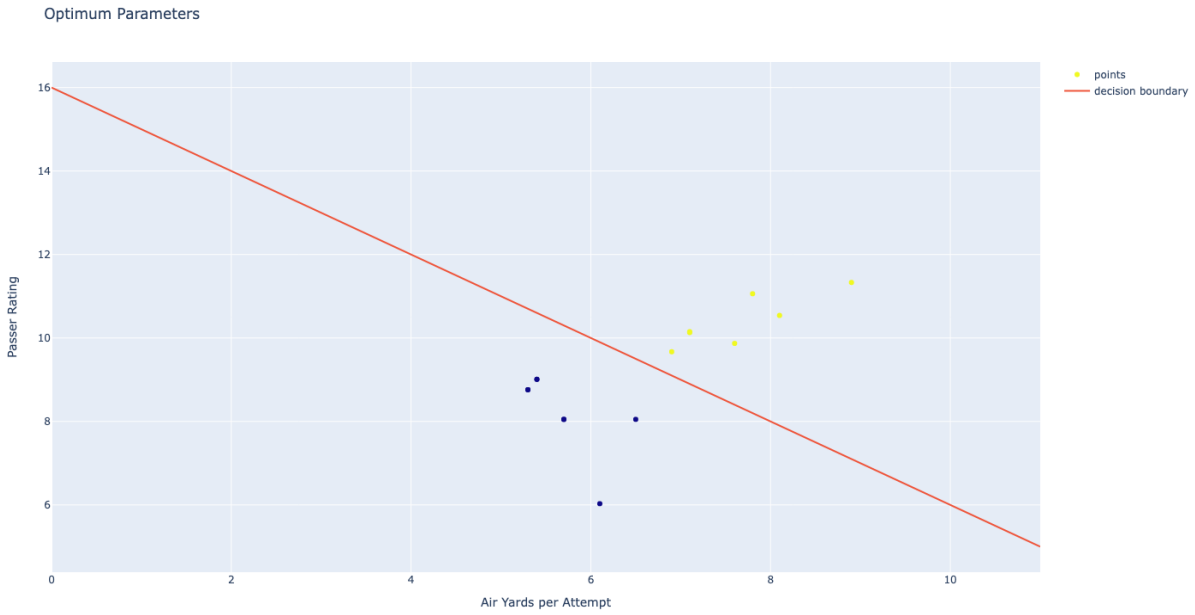
Conclusion:

After this analysis I was able to find out which NFL Quarterbacks are elite and which ones were not.
Here are the plots to justify the reasoning that I have been giving my friends about Elite Quarterbacks:
*Note: Yellow means Elite, Blue means scrubs
*Sorry but yellow is kind of hard to see

Using KMeans:



Using a Neural Network:



Using this I found out which Quarterbacks in the NFL are Elite using Machine Learnign, here is the result in terms of a pandas dataframe, in case it isn't very clear in the plot (again sorry for the terrible color choice ;-;

	Player	Air Yards per Attempt	Passer Rating	Total yards per game	Touchdown-Interception ratio	Classification
0	Lamar Jackson	8.9	11.33	290.1	5.9	1
1	Patrick Mahomes	8.1	10.54	305.6	4.1	1
2	Josh Allen	7.6	9.87	317.8	3.7	1
3	Justin Herbert	7.1	10.13	325.1	3.1	1
4	Tom Brady	6.9	9.67	301.3	4.5	1
5	Aaron Rodgers	7.8	11.06	250.6	5.8	1
6	Kyler Murray	6.5	8.05	234.1	2.7	0
7	Sam Darnold	6.1	6.03	201.7	0.8	0
8	Joe Burrow	5.4	9.01	254.3	2.8	0
9	Derek Carr	5.7	8.05	287.6	1.9	0
10	Jacoby Brissett	5.3	8.76	216.7	0.9	0
11	Matt Ryan	7.1	10.15	287.8	7.1	1
12	Ben Roethlisberger	5.4	9.01	254.3	2.8	0
13	Baker Mayfield	5.7	8.05	287.6	1.9	0
14	Kirk Cousins	5.3	8.76	216.7	0.9	0

Extra Credit Iris dataset:

I used Scikit-Learn's Multiple Layer Perceptron for classification of the IRIS dataset, because I had used Scikit-Learn before. The MLP is a multi layer neural network.

I have attached my source code for the same where I use all 4 dimensions of the IRIS dataset to get the results.

Answers for the Extra Credit Assignment:

- Multi-Layer Neural Networks make decisions by implementing hidden layers where the inout vectors are converted into an intermediate state and then relayed forward to another layer, using activation steps. Sigmoid functions are an example of activation functions.
- Some non-lineairties used in neural network classifications activation functions are sigmoid, tanh (hyperbolic tan), Relu and Softmax.
- In the aforementioned example about NFL Quarterbacks, the KMeans classifier worked much better, since there were relatively few datapoints and not the neural network.