Chaitanya Desai

Tyler Smith

CSE 574

Machine Learning

Assignment 2

March 20, 2022

Part 1: Data Analysis

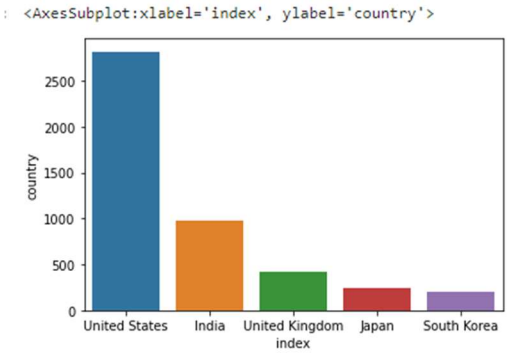
1) **Dataset name: Netflix Dataset**

Netflix dataset consists of Information related to the shows and movies it contains. It gives an in-depth information about the Director of that movie/ TV show, duration, which country it belongs to etc. We encounter the same data including rating and description etc. It also provides the information about the release date and the date it has been uploaded by Netflix on its platform. It contains 12 features and 8807 entries. The dataset comprises of inter/numerical information as well as in form of string that is descriptive information.

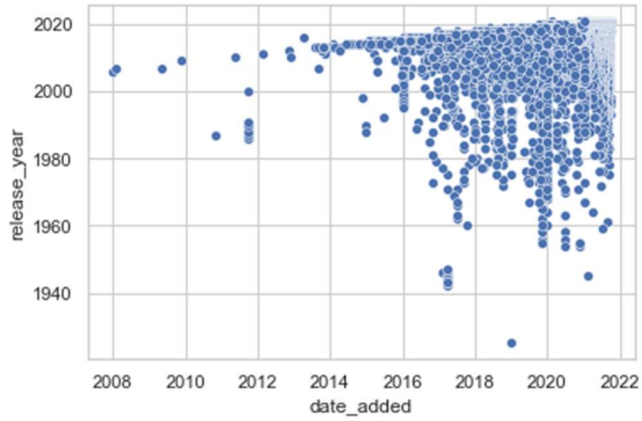
The main statistics about the dataset are:

1) **Mean** = 2014.180198
2) **Std** = 8.819312
3) **Min** = 1925 (Movies staring from the year 1925 have been described)
4) **Max** = 2021 (Movies up to 2021 are described in given dataset)

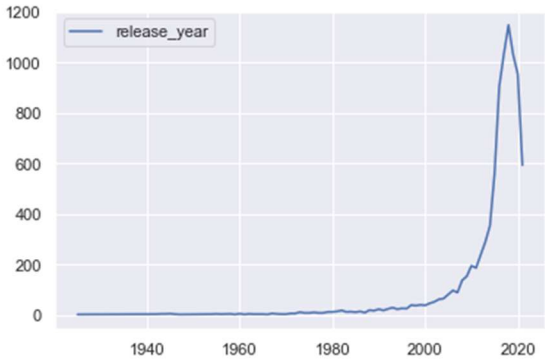
Graphical information:



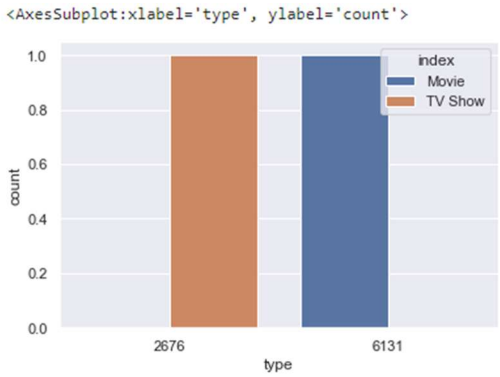
The above graph contains information about how many movies or TV shows are made in following countries. Maximum movies and TV shows are made in USA and least movies are made in South Korea.



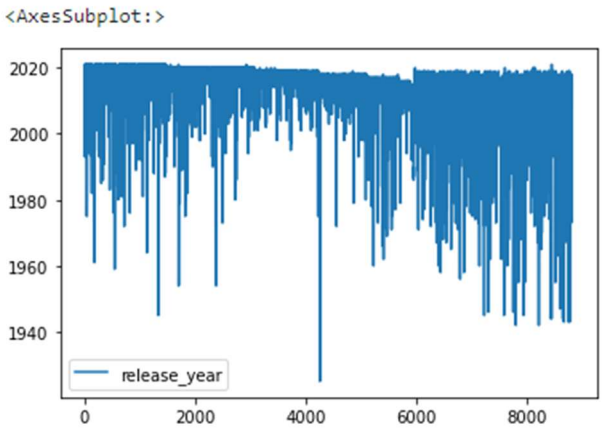
Above scatter plot graph corresponds to each movie/TV show and teels us when was particular movie/ TV show was released and when it was uploaded by netflix on its platform. Most of the movies/ TV shows are uploaded from year 2015.



The above graph tells us how many movies was released in following years. We can see a significant portion was released after 2005.



Following graph tells us the number of TV shows and number of movies that are present on Netflix.



Following graph also gives information about release date and gives us general information about it.

2) Dataset name: Insurance Dataset

Insurance dataset consists of the information that are considered while deciding the cost of insurance premium like age, bmi, smoker. It also gives information about the region and sex of the individual. The dataset consists of 7 features and 1338 entries. Features like bmi has integer entries while features like region has string entries.

The main statistics about the dataset are:

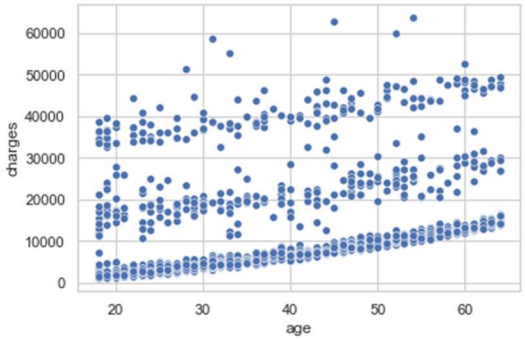
1) Mean
Age = 39.207025
Bmi = 30.663397
Children = 1.094918
Charges = 13270.422265

2) Std
Age = 14.049960
Bmi = 6.098187
Children = 1.205493
Charges = 12110.011237

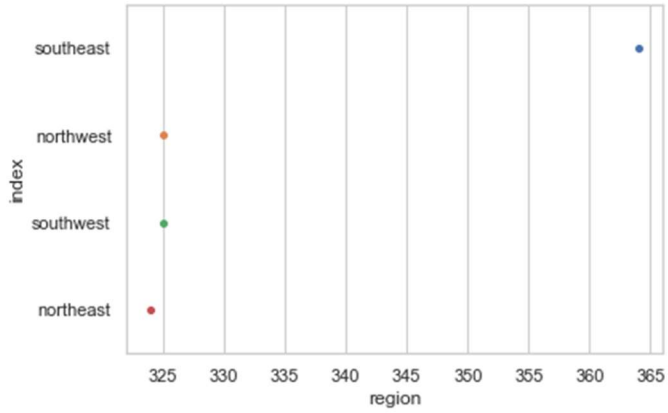
3) Min
Age = 18
Bmi = 15.96
Children = 0
Charges = 1121.873900

4) Max
Age = 64
Bmi = 53.13
Children = 5
Charges = 63770.428010

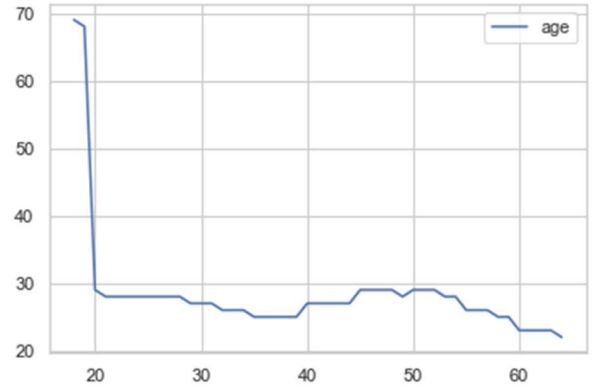
Graphical information:



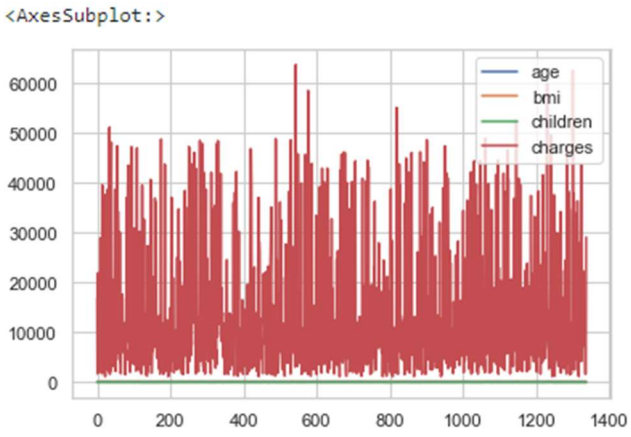
Above graph gives information about the age and the charges



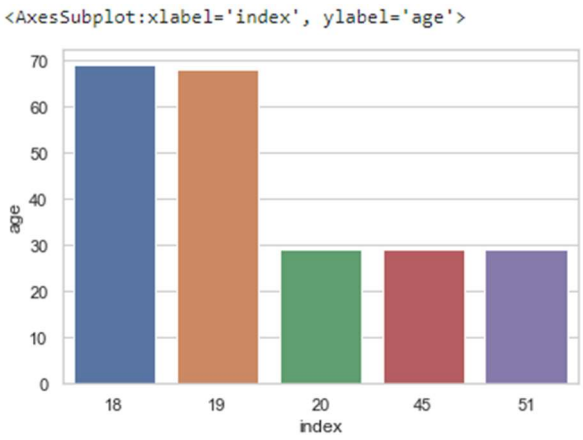
Above graph gives information about the number of people in given region



Above graph gives information about the age and count of people that is their count w.r.t their age.



Above graph gives information about all the features of the graph but also tells us limitation of using it. As charges are more in magnitude they dominate other values and reflect as prominent feature.



Above graph gives information about the index and age

3) Dataset name: Titanic Dataset

Titanic dataset gives us the information about the people which were on titanic. It describes their name, sex, age, siblings/ spouse with them or parents or children with them. It also gives us the information about their fare and how many of them survived. Features like age consists of integers and features like name includes string. It includes 887 entries within 8 features.

The main statistics about the dataset are:

1) Mean

Survived = 0.385569
Pclass = 2.305524
Age = 29.471443
Siblings/ Spouses Aboard = 0.525366
Parents/ Children Aboard = 0.383315
Fare = 32.30524

2) Std

Survived = 0.487004
Pclass = 0.836662
Age = 14.121908
Siblings/ Spouses Aboard = 1.104669
Parents/ Children Aboard = 0.807466
Fare = 49.78204

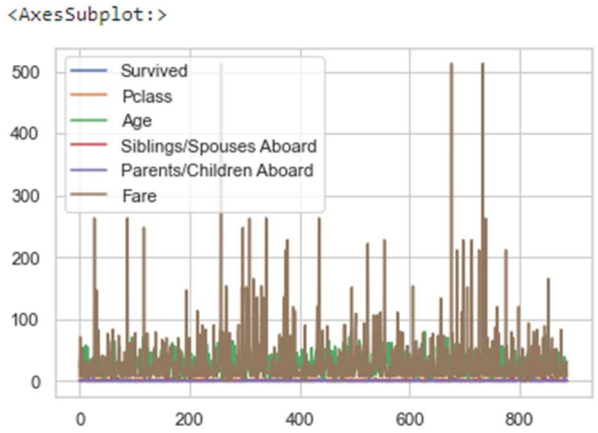
3) Min

Survived = 0
Pclass = 1
Age = 0.42
Siblings/ Spouses Aboard = 0
Parents/ Children Aboard = 0
Fare = 0

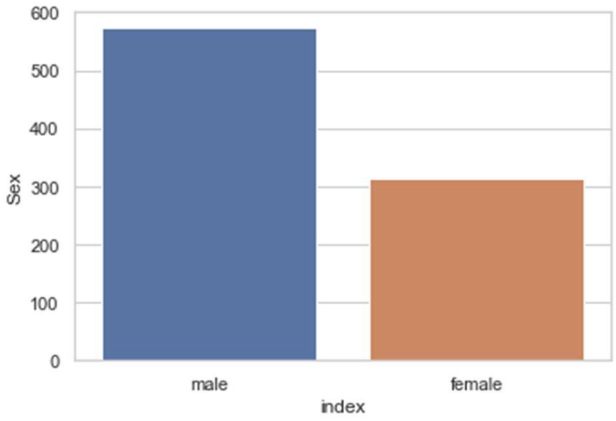
4) Max

Survived = 1
Pclass = 3
Age = 80
Siblings/ Spouses Aboard = 8
Parents/ Children Aboard = 6
Fare = 512.32920

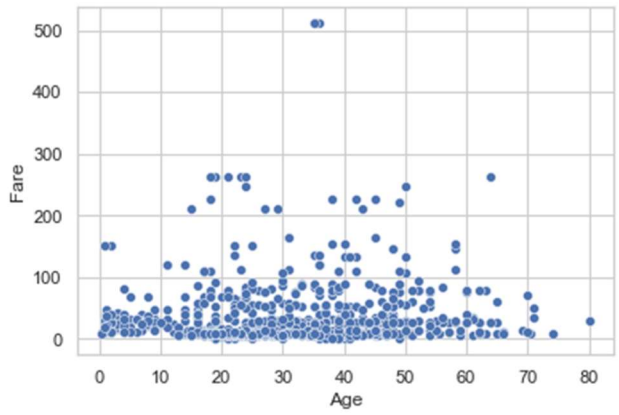
Graphical information:



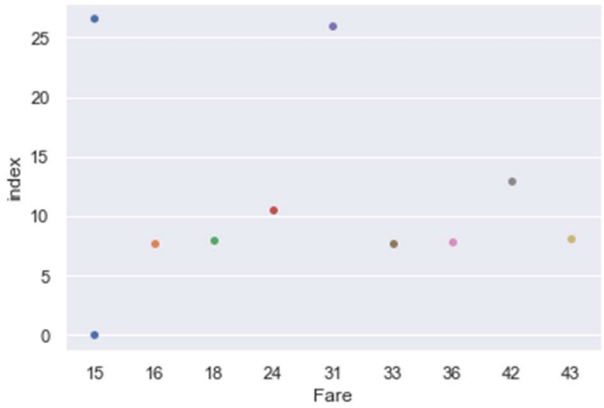
This graph gives information about all the parameters but also shows the limitation as prominent feature that is fare dominates and hides all other parameters.



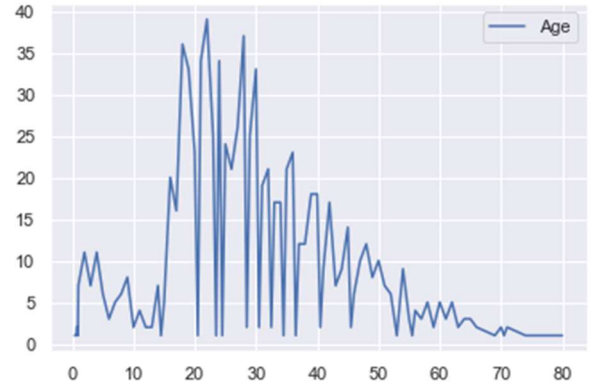
Above graph gives the information about number of males and females in the titanic



Above graph information about the relation between age and fare of the people on titanic



Above graph gives the information about the fare and their count



Above graph gives the information about the age of people and their count

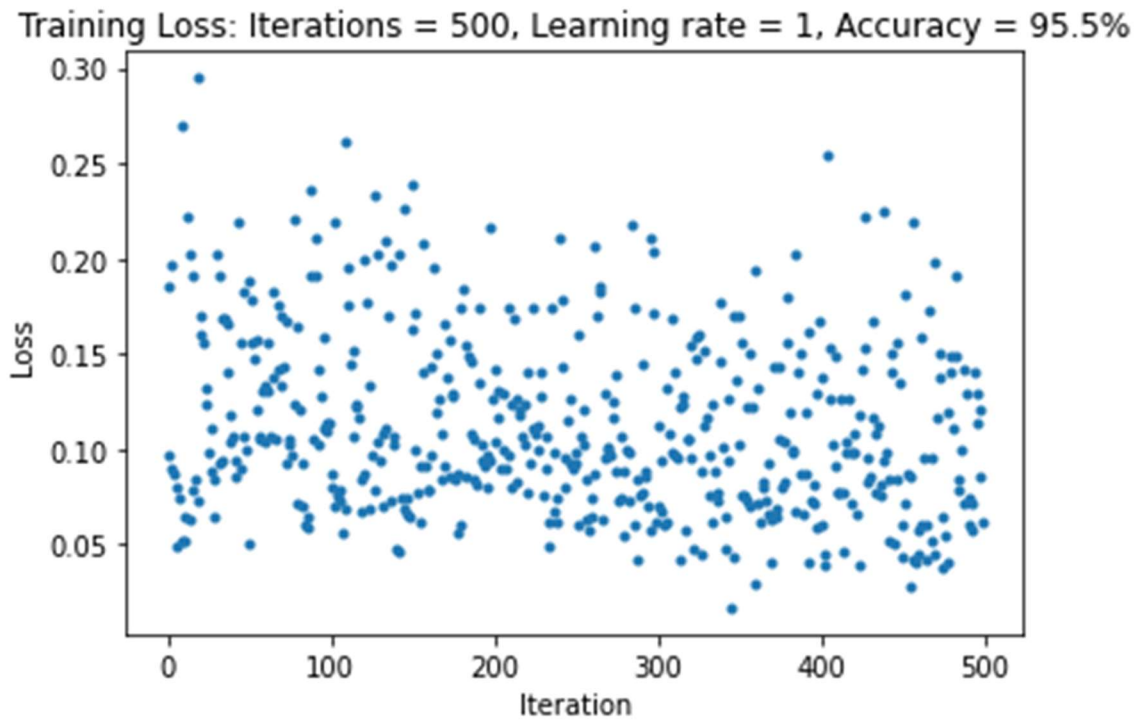
Part 2: Logistic Regression

Logistic regression was used to calculate updated weight vectors based on four input variables to predict an output variable. The data input included the penguin's culmen length, culmen depth, flipper length, and body mass. The output was the penguin's sex. After these weight vectors were calculated with a training set of data, a test set of data was used to check the accuracy of the weights. The best accuracy tested was with 500 iterations, a learning rate of 1, and an accuracy of 95.5%. Below are the associated weights and training loss graph.

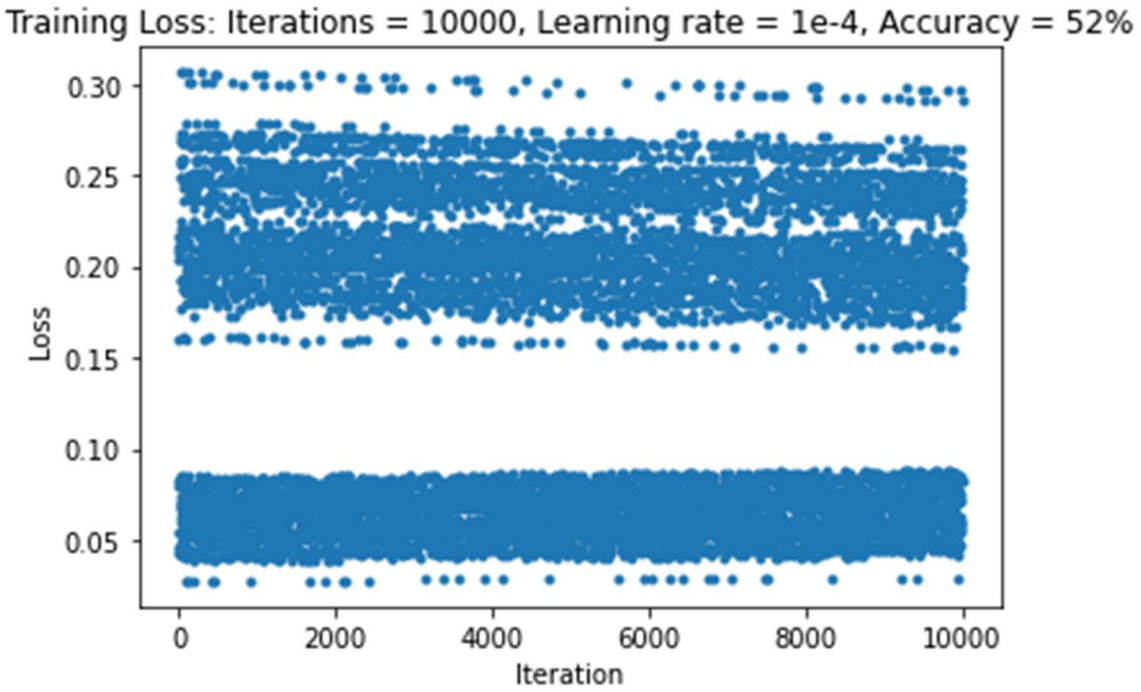
Best accuracy: 95.5%
Weights: ['w0': -3.261766, 'w1': 1.063380, 'w2': 2.429360, 'w3': 0.555320, 'w4': 2.461572])

Weight 0: Bias
Weight 1: Culmen length weight
Weight 2: Culmen depth weight
Weight 3: Flipper Length weight
Weight 4: Body Mass weight

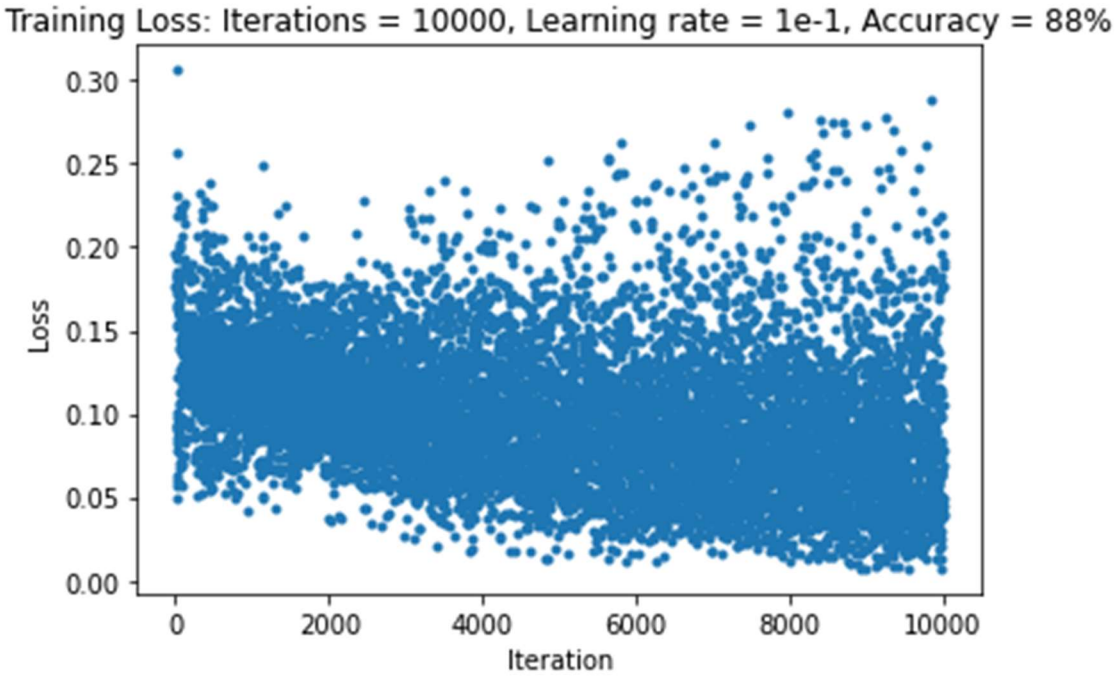
Include loss graph and provide a short description 3.



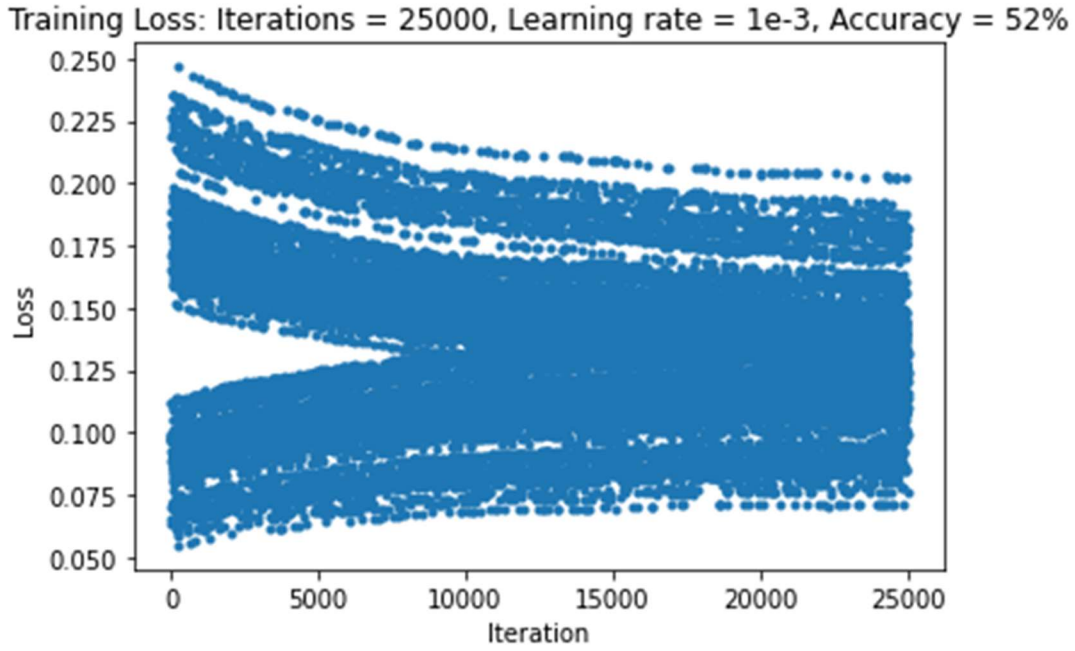
If the learning rate is too small, there is a tendency for the weights to not adjust quickly enough. Given 100,000 iterations and a learning rate of 1e-4, the accuracy is only 52% since the weights are unable to be updated in a significant way.



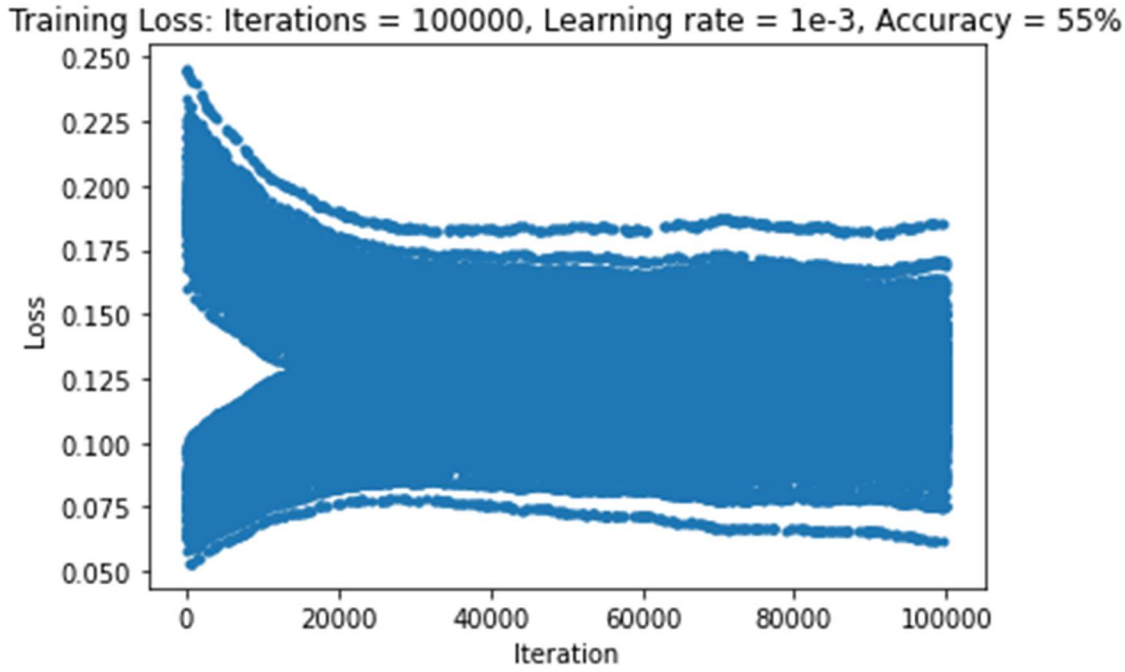
If there is a lower number of iterations, but a higher learning rate, this normally yielded better accuracy results. Below is an instance where 88% accuracy was tested with a learning rate of 1e-1 with 10,000 iterations.



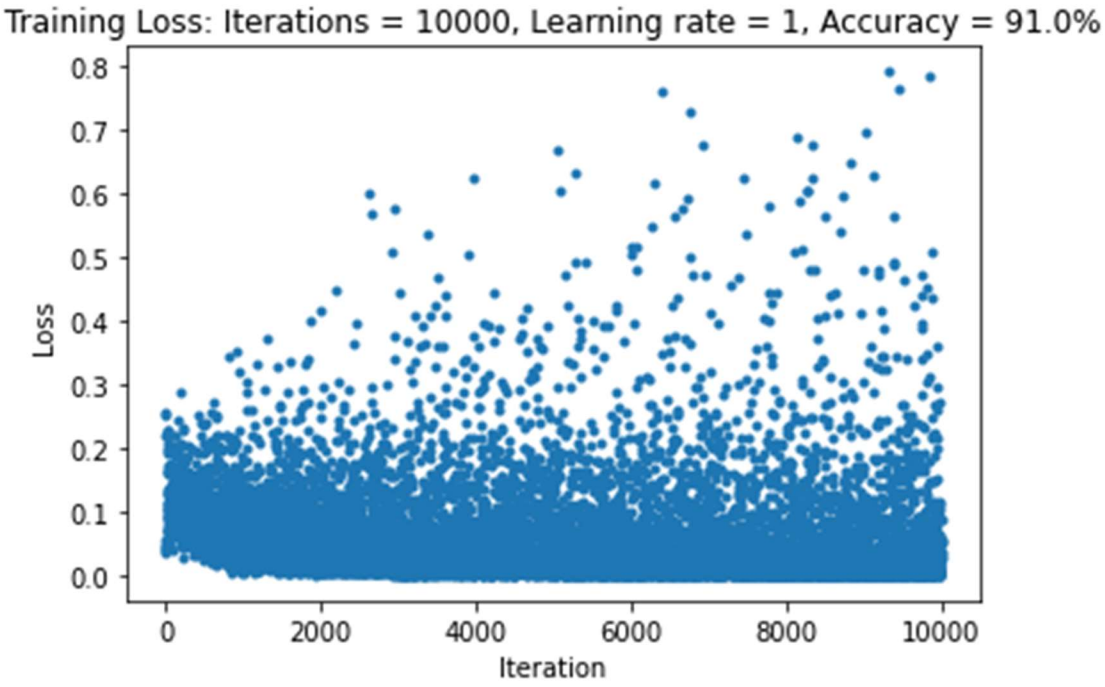
The greater the number of iterations, it can be seen that the overall loss decreases. This can be seen under 25,000 iterations at a learning rate of 1e-3. This still, however, does not yield good accuracy at all, and the time to complete this was approximately 1.25 minutes.



With a learning rate of 1e-3 at 100,000 iterations, we still see a low accuracy of these values, with a process time of approximately 5 minutes.



As the learning rate is increased significantly, there is a trend for higher scaling of the weights, and thus more noise for edge cases. The accuracy, however, during the tests has been better. The required process time is also significantly lower. This is seen for when the hyperparameters are 10,000 iterations with a training rate of 1. The tested accuracy is 91%.



Some of the benefits of logistic regression is that it's easy to implement and efficient to train. Multiple input arguments can be used, and it's ideal for discrete data with linear relationships between the dependent and independent variables. One of the drawbacks of logistic regression is that it can only be used for discrete functions, and it always assumes a linearity between the independent and dependent variables. Data that may be identified with non-discrete inputs or that may be non-linear, may not mesh well with logistic regression analysis. If there is a low number of inputs for training, this may lead to overfitting. With this being said, there are many great applications for logistic regression analysis.

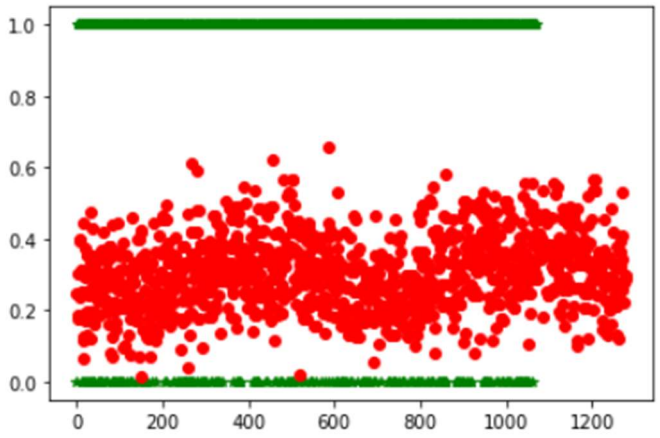
Part 3: Ordinary Least Square

Ordinary Least Squares (OLS) approach was used to analyze a given dataset. The input variables include the person's age, BMI, quantity of children, and charges. The output variable is whether the person is a smoker or non-smoker. Below is the given loss value and weight vectors for a linear regression of the data.

Loss Value = 0.391345

Weight vectors = [0.10963443]
[0.09917397]
[0.62098034]
[1.34598381]
[0.27515140]
[-1.88597301]

Below is a plot comparing the predictions vs the actual test data:



One of the main benefits of using Ordinary Least Square (OLS) method is that it is simple, and computation is easy. The method may perform poorly on multi-variate dataset that contains single independent variables set and multiple dependent variables sets. A large dataset is required in order to find the accurate results.

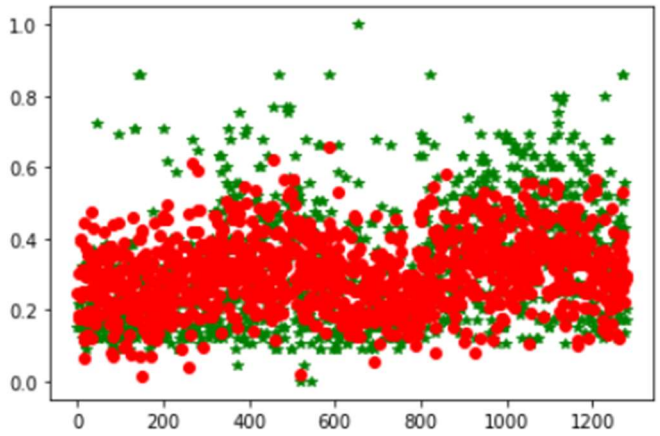
Part 4: Ridge Regression

Ridge Regression was used to analyze a given dataset. The input variables include fixed acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, pH, alcohol, and quality. The output variable is the amount of alcohol in the drink. Below is the given loss value and weight vectors for a ridge regression of the data.

Loss Value = [78.66626323]

Weight Vector = [0.07584017]
[0.28648583]
[-0.61604135]
[-0.08249034]
[0.06362043]
[-0.19793931]
[0.21766678]
[0.09720396]
[-0.29710749]
[0.26932664]
[0.4303553]

Below is a plot comparing the predictions vs the actual test data:



One of the main benefits of linear regression is that it is computationally fast. It gives us the relation between dependent and independent variables using a perfectly fitted line. Ridge regression helps in reducing impact of correlated inputs (Independent variables are highly correlated). If data has potentially many correlated features, ridge regression is most likely the optimal option of the two. The main motivation to use l2 regularization is that it helps in reducing impact of correlated inputs. In l2 regularization weights close to zero have little effect on model complexity, while outlier weights can have a huge impact.

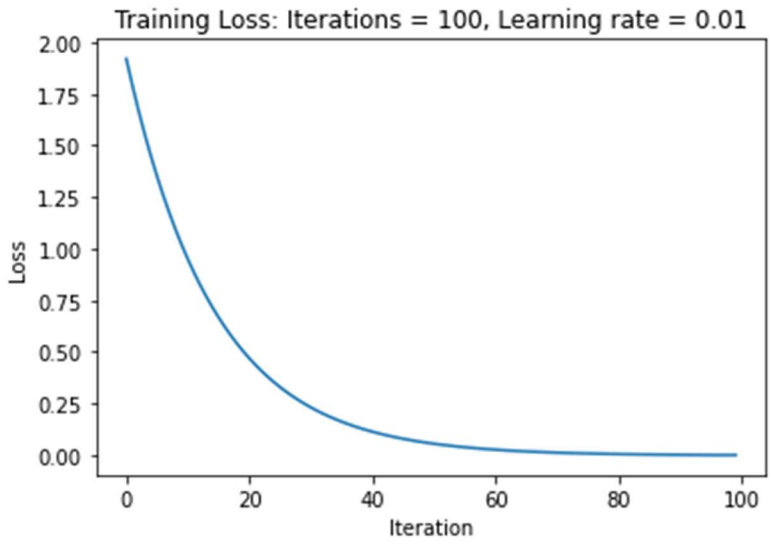
Bonus: Gradient Descent from Scratch

Calculated weights using gradient descent from scratch where w0 is the bias:

Weights (bias, volatile acidity, citric acid, chlorides, density, sulphates, fixed acidity, residual, sugar, free sulfur dioxide, total sulfur dioxide, pH, quality):

[-0.561052]
[0.429944]
[0.314394]
[0.273530]
[0.076994]
[-0.150799]
[0.805846]
[0.326065]
[0.728920]
[0.019957]
[0.299420]
[0.068562]

Below is the loss shown over each iteration:



The results seem to be more consistent than the ridge regression, but the time required to calculate takes much longer. Each single iteration requires a calculation of all of the data sets combined which will add up if larger data sets are used for a high number of iterations.