IRIS (Aayush Singh 2020CHB1036)

1. Problem Statement..?

We have data about different types of IRIS flowers, and we want to create a system that can tell us if a flower is an iris-setosa, iris-versicolor, or iris-virginica. To do this, we'll look at four things: sepel_length, sepel_width, petal_length, and petal_width. We plan to use two methods to make predictions: a Logistic Regression model and a K-nearest neighbor algorithm with different values of k (from 2 to 14). We'll check how accurate each method is and also use ROC and AUC to compare them.

2. Data and Data Description

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Figure 01 : Above Table Shows Top 5 Observations from Data Set

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.000000	-0.109369	0.871754	0.817954
sepal_width	-0.109369	1.000000	-0.420516	-0.356544
petal_length	0.871754	-0.420516	1.000000	0.962757
petal_width	0.817954	-0.356544	0.962757	1.000000

Figure 02 : data.corr()

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Figure 03 : data.describe()

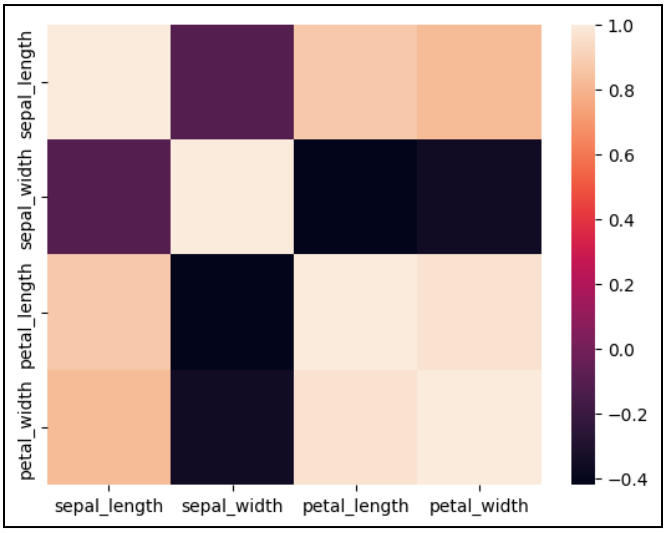

Figure 04 : The corresponding heat map so as to visualize the correlation between the parameters in a better way.

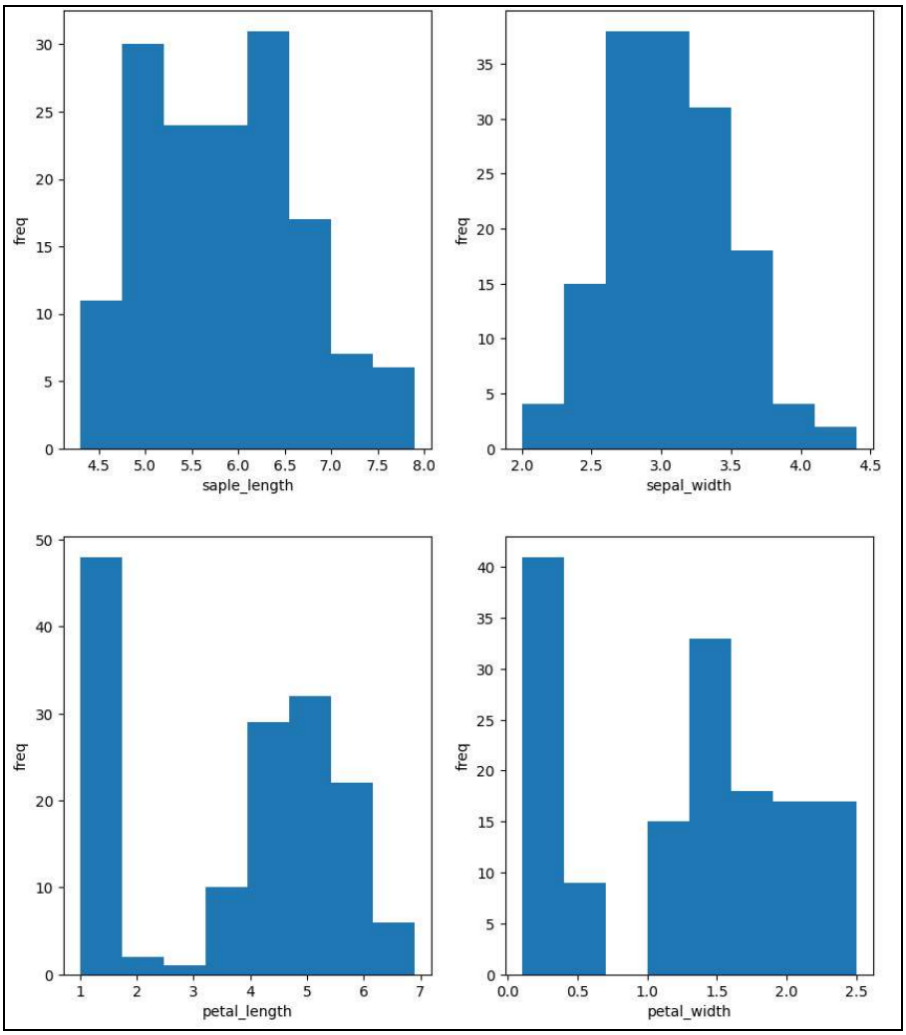

Figure 05 : Histogram of all the Parameters.

3. Some necessary transforms before applying the Algorithms

- We start by separating the data into inputs and outputs. We create a matrix (n*4) and put all the input observations in X, where n is the number of observations.
- Then, we place the actual values in a matrix Y. Since we can't work with non-numerical data, we convert the output data to numerical form. For example, we change iris-setosa to 0, iris-versicolor to 1, and iris-virginica to 2.
- After that, we scale the data appropriately, making the variance of the input data equal to 1.
- Finally, we divide the data into training and testing sets. The training data is used to teach the model, and the testing data is used to evaluate the model's performance.

4. Logistic Regression

$$\pi(\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}$$

$$\tilde{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}y$$

- The above mentioned formula is for Logistic regression. X_1, X_2, X_3……X_n are the parameters, in our case we only need 4 of them.
- B_0, B_1,…….B_K are Multiple Regression Coefficients. We evaluate those by using the formula given in the right.
- We predict the class of the object using π(X).
- Depending on this value we put the observations into different classes.

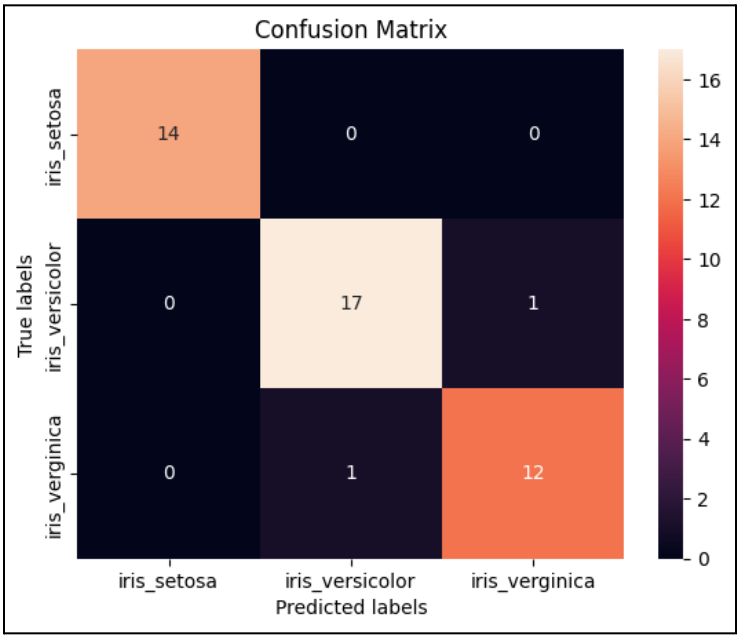

Figure 06 : Confusion Matrix that we obtained using the logistic regression model.

The accuracy of this model is nearly 95.5%.

5. KNN Algorithm

- Basically in this algorithm we calculate the distance between the particular observation we need to predict the class of and all other observations.
- We take the K nearest observation from all this distance. That is we pick the K nearest neighbors of our corresponding observation.
- Among these K nearest neighbors we check in which class most of these K neighbors lies
- We predict the same class for our observation.
- K can be varied in algorithm along with how we measure distance.
- For this following project I am evaluating the model for K = 2 to 14 and we are using euclidean distance for calculating distance between the observations.

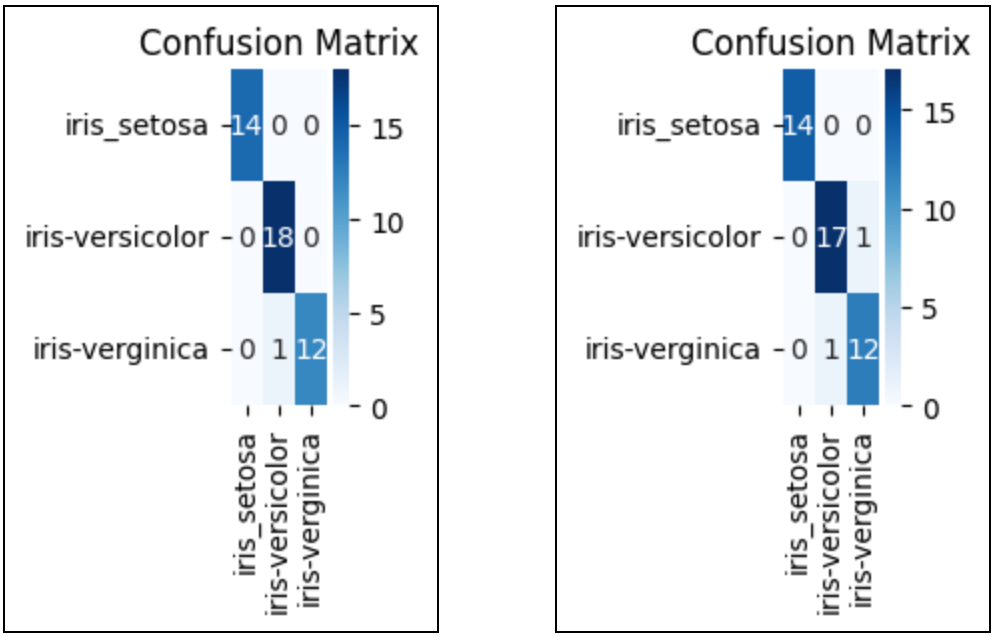

Figure 07 : Confusion Matrix for K = 3 and K = 14

6. Performance of Models Using ROC

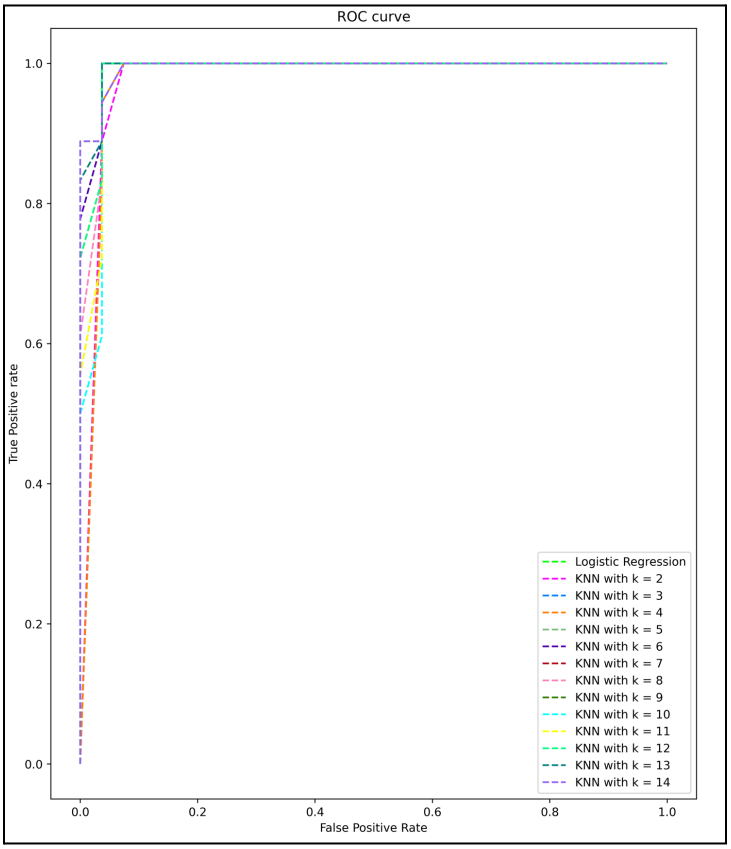

Figure 08 : ROC Curve for Logistic Regression and KNN Model

- The figure on the left the shows the ROC for all the prediction models, namely our logistic regression model and our KNN model for K = 2 to 14
- From the ROC we can know which model is the best by knowing the area under the ROC curve covered by each of the models.
- Whichever model has the highest area under the curve or AUC is the best model.
- Next we will evaluate the AUC for each model.

7. Performance of Models Using AUC

```
The AUC for the Logistic Regression is 0.9978269758362351
The AUC for the KNN algorithm for k = 2 is 0.9780323414582673
The AUC for the KNN algorithm for k = 3 is 0.9787759180120291
The AUC for the KNN algorithm for k = 4 is 0.9780323414582673
The AUC for the KNN algorithm for k = 5 is 0.9933078110161443
The AUC for the KNN algorithm for k = 6 is 0.9955385406774296
The AUC for the KNN algorithm for k = 7 is 0.9940513875699061
The AUC for the KNN algorithm for k = 8 is 0.9925642344623826
The AUC for the KNN algorithm for k = 9 is 0.9903335048010975
The AUC for the KNN algorithm for k = 10 is 0.9881027751398123
The AUC for the KNN algorithm for k = 11 is 0.9903335048010975
The AUC for the KNN algorithm for k = 12 is 0.9940513875699061
The AUC for the KNN algorithm for k = 13 is 0.9962821172311913
The AUC for the KNN algorithm for k = 14 is 0.9962821172311913
The Best Model is Logistic Regression
```

Figure 09 : AUC Values for Logistic Regression and KNN Model

- The figure on the left shows the AUC for each model that we used.
- Based on these AUC scores we can judge which model is performing the best.
- The higher the AUC the better the model.
- Hence the best model for our prediction data is the Logistic regression model.

8. Conclusions

- We applied Logistic Regression and KNN Classification algorithm model for prediction of the variety of flowers.
- Using ROC and AUC analysis we found out that the best model for the prediction among the two is the Logistic Regression model.
- By this project I got to learn how KNN and Logistic Regression works. I learnt how to compare models so as to know which is better.
- Thus by this project I learnt the importance of data science concepts and how important data is, using data and analyzing the data can help us predict the future outcomes and hence it can become a very important part of our life.