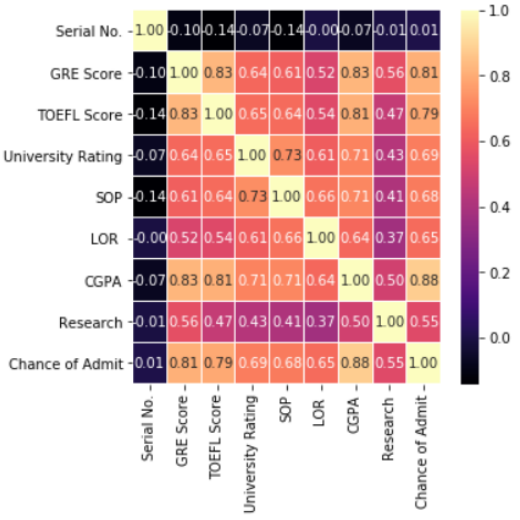
Chances of Admit in Master's and Iris dataset classfication

Ayush Chaurasia
IIT DELHI

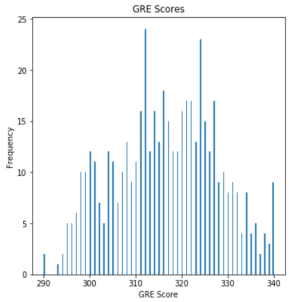
1. Problem and Data set Analysis

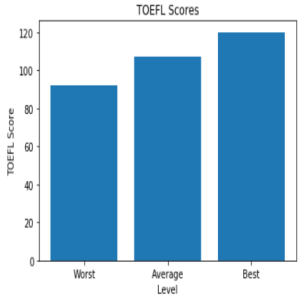
Objective is to predict a student's chances of getting admitted for a master's degree. Data set contains 500 examples. Features in the data set:

- GRE Scores (290 to 340)
- TOEFL Scores (92 to 120)
- University Rating (1 to 5)
- Statement of Purpose (1 to 5)
- Letter of Recommendation Strength (1 to 5)
- Undergraduate CGPA (6.8 to 9.92)
- Research Experience (0 or 1)
- Chance of Admit (0.34 to 0.97)

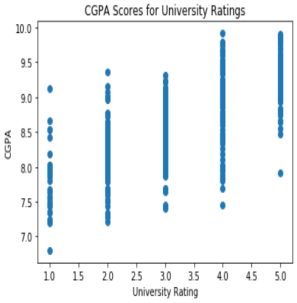


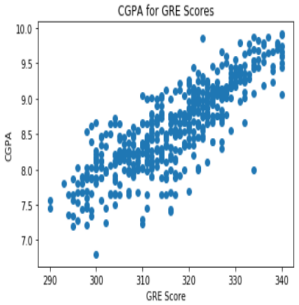
From the above correlation matrix we can be see that correlation between serial no. and chance of admit is very less which had to be since serial number is just a ordering. Also, research is showing less correlation with chance of admit compared to other features hence will be a less important factor(also will be supported in further analysis).



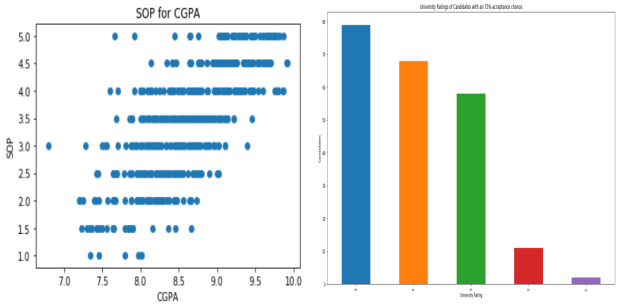


Above are the GRE and TOEFL scores of candidates. Most GRE scores are in 310-330 and average TOEFL score is 107. This implies candidate has high chance of succeeding given his GRE score is above 330 or TOEFL is well above 107. These 2 features are also important since they show good correlation with chances of admitting in the correlation table.





From the above left figure it is clear that candidates with higher rating universities on average have better CGPA. Also, CGPA and GRE scores also show a high correlation(from the above right figure).



Similarly, SOP rating is correlated with CGPA(from the above left figure). A similar pattern is observed for SOP rating with GRE score. Another thing to observe from the above right figure is that chance of admit significantly reduces with decrements in university rating and henceforth university rating also plays important role in a candidate's chances of admit.

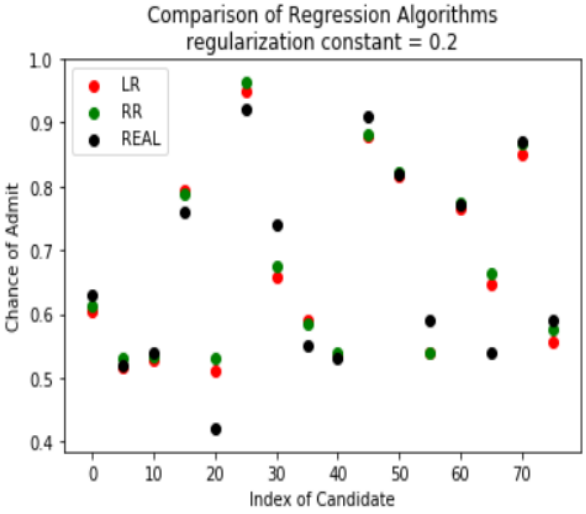
1.1. Principal Component Analysis

On applying PCA to the above data set the following order of importance was observed in the features(topmost being most important and bottommost being least important):

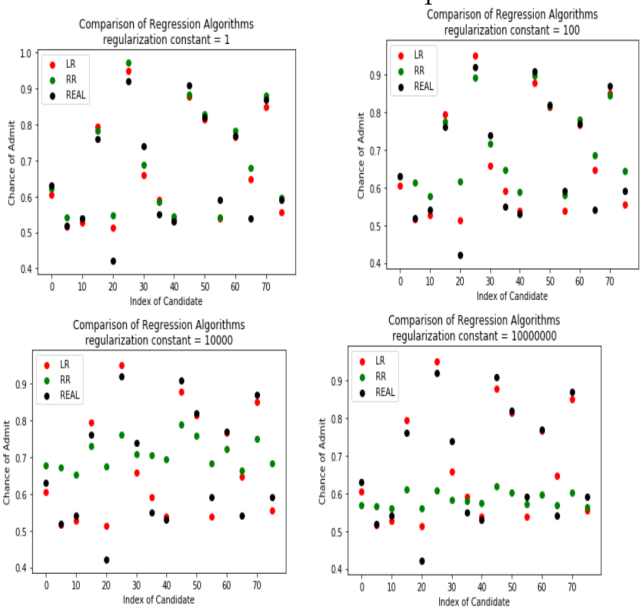
1. GRE Score
2. TOEFL Score
3. University Rating
4. CGPA
5. Research
6. LOR
7. SOP

2. Regression Analysis

Data was divided into 80:20(train:test) i.e. 400 training examples and 100 test examples.



Above graph shows chances of admit for each candidate in the test data. The black dots show the actual chances whereas red dots and green dots show the predictions from linear regression(LR) and ridge regression(RR) respectively (MSE LR = 0.0018 MSE RR = 0.0331). Regularization constant is 0.2 as indicated in the image title. Another point to note is that since most of the candidates in the data have over 0.7 chances of admit, many candidates with less chance of admit are not well predicted.



Above 4 figures show the analysis for 4 different values of regularization constant. As can be seen increasing the value to 1 and then to 100 decreases the over fitting or better say dampens the green dots(ridge regression predictions). to some base horizontal line. On increasing the regularization constant all the predictions lie nearly on horizontal line i.e. weights are completely damped and a highly

under-fit model is produced. Also, the error kept on increasing as the regularization constant was increased showing that the model kept on under fitting.

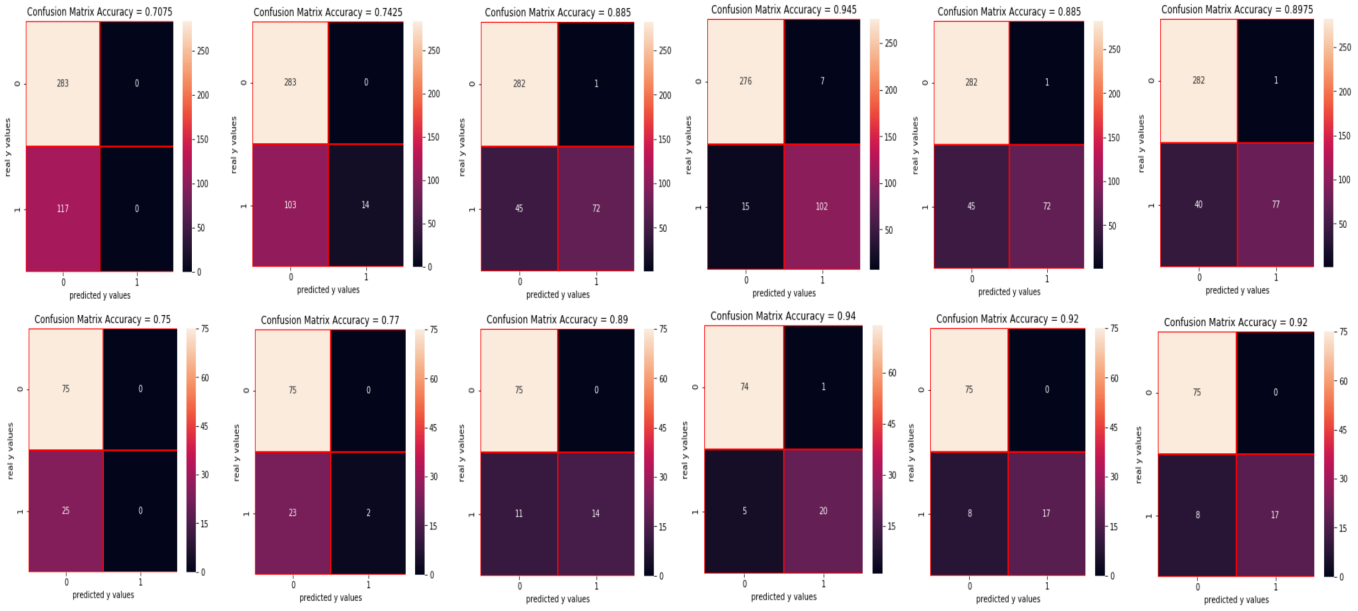
3. Binary Classification

We modified the regression problem to a binary classification problem by the below simple rule:

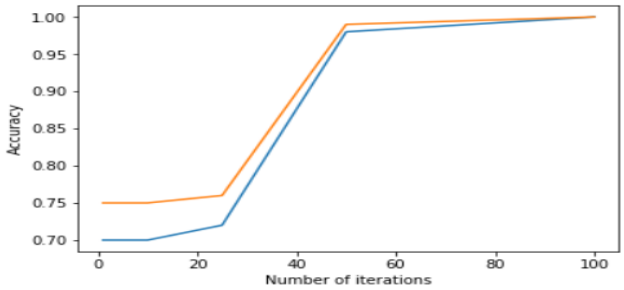
- If a candidate's Chance of Admit is greater than 0.8, the candidate will receive 1 label.
- If a candidate's Chance of Admit is less than or equal to 0.8, the candidate will receive 0 label.

3.1. Logistic Regression

Different combinations of learning rates and number of iterations were tried. Batch Gradient Descent was used. First we will vary the learning rates and keeping the number of iterations fixed at 1000.



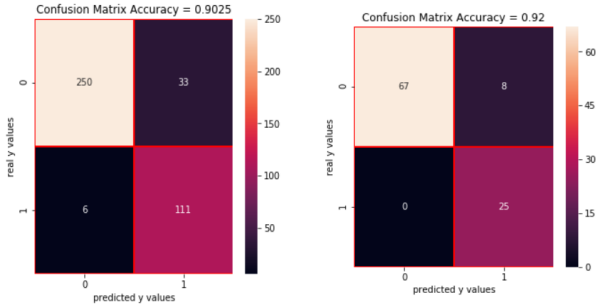
In the above confusion matrices the learning rate from left to right are 0.01,0.1,1,10,100,1000 respectively. Figures in the above row is for the training data and below is for the test data. As seen, both the training and test accuracy keep on increasing till learning rate is 1. After that increasing learning rate decreases the accuracy for both. The reason for this is, when low learning rates are kept than training is vulnerable to noise in the data and henceforth it adapts to it giving low accuracy. Large learning rates may not be able to adapt the data completely or better may not be able to optimize the gradient step since the step size will be too large. Hence 1 is the optimal learning rate here.



Above shown is the accuracy with changing the number of iterations keeping the learning rate constant at 1. As the number of iterations were increased the training data is expected to get an accuracy of 1 which is also seen(blue line) but test data accuracy(orange line) also approaches 1 which is not expected. This can be because only 100 examples were there for test data it might not be very varied and hence not behaving as expected(this argument

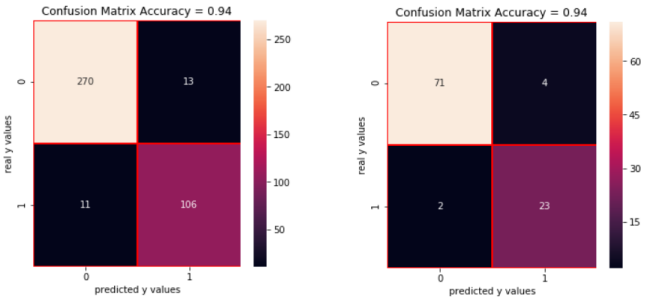
is also supported by Gaussian Naive Bayes method as explained below).

3.2. Gaussian Naive Bayes



This method gave us a train accuracy of 0.90 and a test accuracy of 0.92. This method gave less accuracy than logistic regression because this method is naive in the sense that it assumes independence of features whereas we already did see high correlations among few features like CGPA and GRE score(another reason for less accuracy than logistic regression can be that logistic regression over fits the data and due to poor test data as explained before in logistic regression section). Another point to note is that here the test accuracy is more than the train accuracy again supporting the point raised in logistic regression section that the test accuracy was also increasing with increase in number of iterations. This supports the possible problem in test data explained in the last in logistic regression section.

3.3. K-Nearest Neighbours



Applying KNN gave good accuracy ranging from 0.92-0.96 for different values of k ranging from 1-15. In this method the research data point was ignored primarily because of the reasons below:

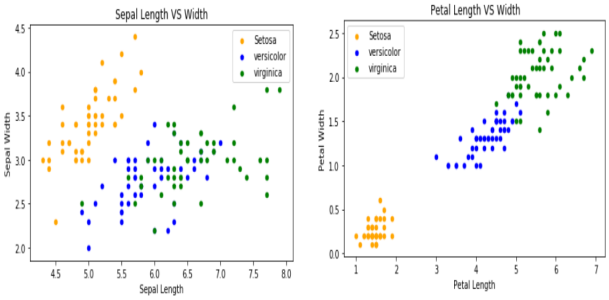
- It is a discrete quantity and simply taking it in the euclidean distance(1 distance value between different research value i.e. 0 and 1 otherwise 0 distance value) is not very intuitive and not representative whereas other features were continuous and could be scaled to 0-1 for good distance representation.
- We saw that research had a very less correlation with chance of getting admit and hence we can take out the feature without worrying about the results(also evident from PCA analysis).

4. Multi Class Classification

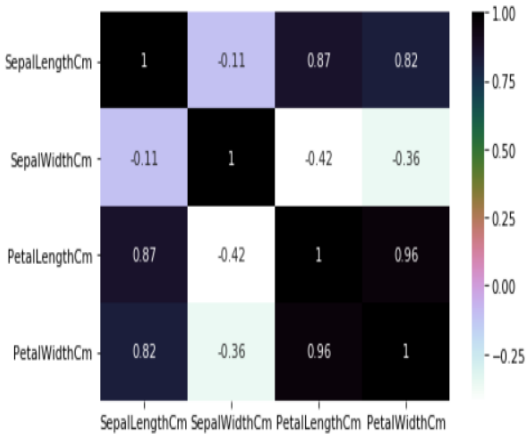
Objective is to predict the species of flower based on its characteristics. Data set contains 150 examples. Features in the data set:

- Sepal Length(in cm)
- Sepal Width(in cm)
- Petal Length(in cm)
- Petal Width(in cm)
- Species:Iris-Setosa,Iris-versicolor,Iris-virginica

4.1. Data Analysis

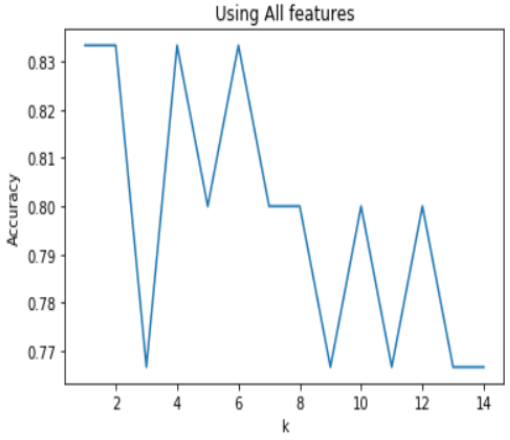


In the above figures it can be seen that sepal length vs width for each specie is not that clearly clustered compared petal length vs width for each specie. But since both the figures(petal and sepal) are still showing clustering indicates that we can apply KNN or K-means algorithm to them to predict the specie. Also, petal clustering is more good(distant) than sepal clustering, one might expect to have better results in clustering algorithms with only petal features taken into account rather than sepal features(as seen further).

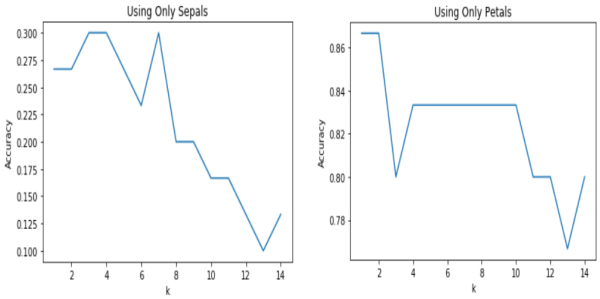


From the above correlation table it is observed that the Sepal Width and Length are not correlated whereas the Petal Width and Length are highly correlated.

4.2. K-Nearest Neighbour



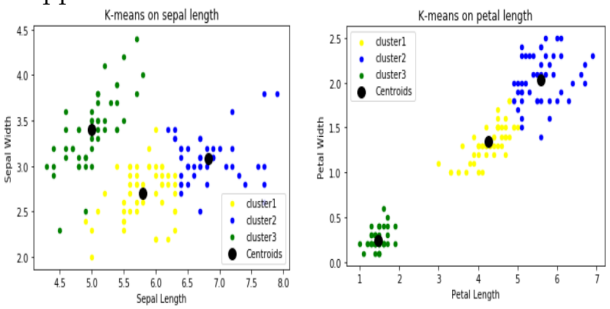
Data is split in 80:20(train:test). In the above figure we used all the features and applied KNN algorithm. Highest accuracy have been obtained for 4 and 6 clusters with accuracy of 0.83 over test data.



In the above figure on left only sepal features were used for KNN i.e. petal width and length were not considered. The highest accuracy achieved is only 0.3 (for 3 clusters). This is also expected as discussed in the data analysis section that clustering of sepal width vs length was not good and also they weren't highly correlated. In the above figure on the right only petal features are used for KNN and a high accuracy is obtained for 2(accuracy = 0.86) or 3 clusters(accuracy = 0.83)(higher accuracy than using all features). This is expected also for the similar reasons discussed in data analysis part.

4.3. K-Means

Since it is a unsupervised we use all the data points i.e. no train and test split. We separately apply K-means for 2 cases: first in which petal features are dropped and second in which sepal features are dropped.



In both the above figures K-means has quite well separated the clusters with each cluster given a particular color and black dots showing the centroid. These figures are quite good replica of the actual species figures shown below which shows that K-means works well and is doing good classification.