Data Analytics with the Fisher Iris Dataset

Vishawam Datta

Introduction

The Fisher Iris dataset is a multivariate dataset introduced by the British statistician and biologist Ronald Fisher in a 1936 paper of his. The dataset contains a collection of 50 samples each of three different species of iris namely setosa, versicolor and virginica.The feature variables in this dataset are the length of petal ,width of petal, length of sepal, width of sepal,where all the dimension are given in centimetres, the major task here is to classify each type of iris flower into its corresponding species given the dimensions of its petals and sepals.The inferences and the analysis of the observations is mentioned in each section along with the tasks performed. A brief summary of all the main and important conclusions is provided at the very end of the report. Note: Throughout the document the species are termed as 'type' and numbers are assigned to each species in the following manner setosa-0 , versicolor-1 , virginica-2.

1. Data Exploration

1.1 Plotting Histogram with Sepal width(y) and Sepal length(x)

Given is the histogram plot with sepal width (cm) plotted on the y axis and the sepal length (cm) plotted on the x axis. The histogram is plotted by dividing the range in ten bins each.

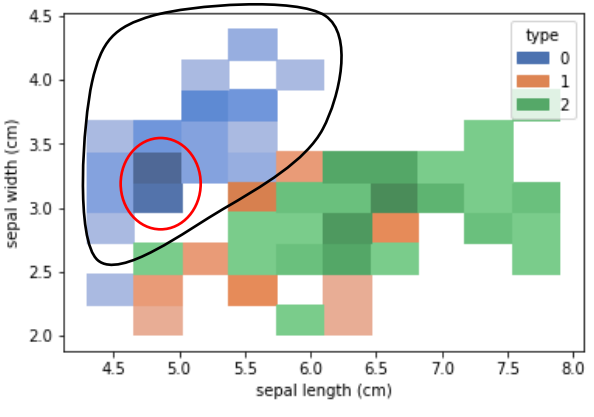

Fig 1.1

Studying the above plot will allow us to get some deep insights into how the data is distributed, here the data is grouped(coloured) based on the type of the species of the flower.As mentioned above 0 stands for setosa ,1 for versicolor , 2 for virginica.The darker a shade of colour in a region, higher is the concentration of samples with the corresponding sepal width and sepal length.For example if we see the cell 4.5-5,3-3.5 (highlighted by the red circle) we can see a darker shade of blue there indicating that there are a lot of samples in specie setosa or type 0 which have sepal length in between 4.5-5cm and petal length between 3 -3.5cm.Furthermore looking at the data we can see that the dimensions of the setosa specie are separated from the other specie (highlighted by the black boundry). However we cannot really see a separation between the specie 1 and 2 namely versicolor and virginica . This gives us an indication that certain classification algorithms may give us satisfactory results when separating the setosa specie from others.However we are unable to draw some perfect conclusions just by studying this plot,hence moving on with data exploration to study a certain scatter plot.

1.2 Scatter plot with Petal width(y) and Petal length(x)

Given below is the scatter plot with the petal length(cm) on the x- axis and the petal width(cm) on the y axis. The data is separated/grouped based on the species ,blue for 0(i.e setosa),orange for 1(i.e versicolor),green for 2(i.e virginica).

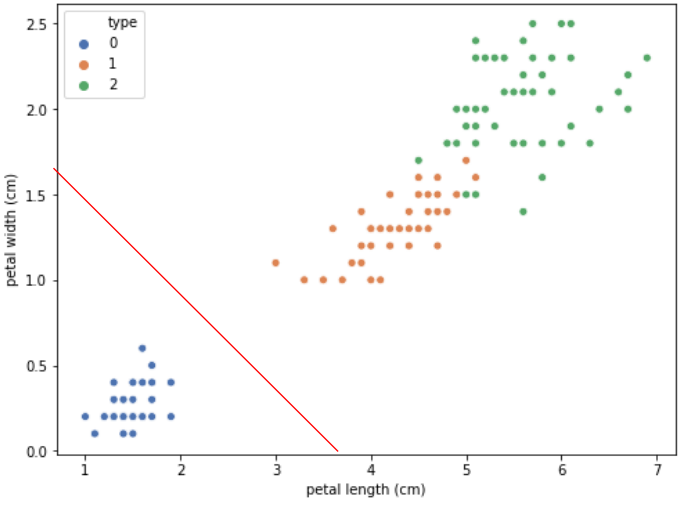

Fig 1.2

Studying the above scatter plot provides us some great insights, here we can see that the specie 0 or setosa is easily separable from the rest of the two species(as clearly indicated by the simple red line drawn on the plot) as the value of both petal length and petal width are quite less (in comparison with the other two species).Thus this gives us a strong

motivation to use supervised learning algorithms like the logistic regression and the naive bayes algorithm to separate type 0 or setosa from the others just by passing two features namely petal width and petal length as predictors. Here we can observe that the specie 1 (namely versicolor) and specie 2 (namely virginica) are somewhat separately distributed but there doesn't exist a distinct or distinguishable boundary between these two species ,hence if we employ naeve bayes or logistic regression classification techniques we will get results but those will not be perfect with some mis classification between the specie 1 and 2.However the specie 0 or setosa can be expected to be completely classified because of the existence of a distinct boundary between setosa and the other species.Moving forward with the data exploration and on to the box plots.

1.3 Box plots of the different features:-

Below are given the box plots of the different attributes together and then separately grouped by the type of specie(i.e 0,1 or 2).Blue for setos(0) ,orange for versicolor(1),green for virginica(2).

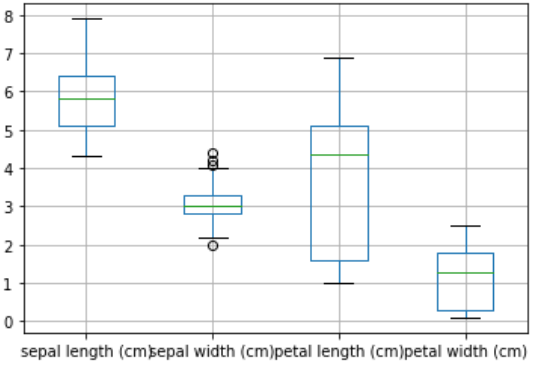

Fig 1.3 a)

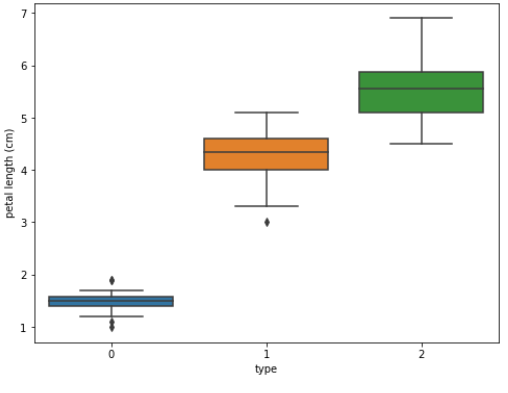

Fig 1.3 b)

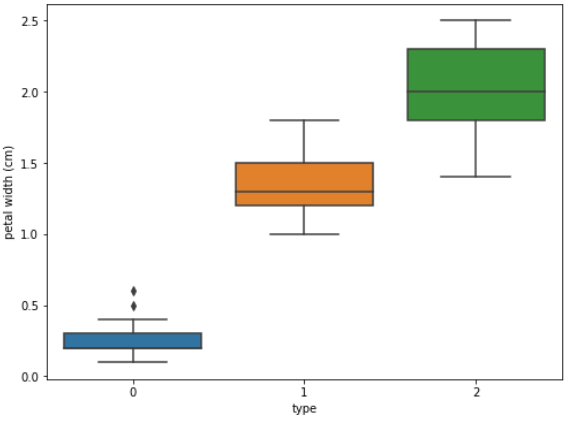

Fig 1.3 c)

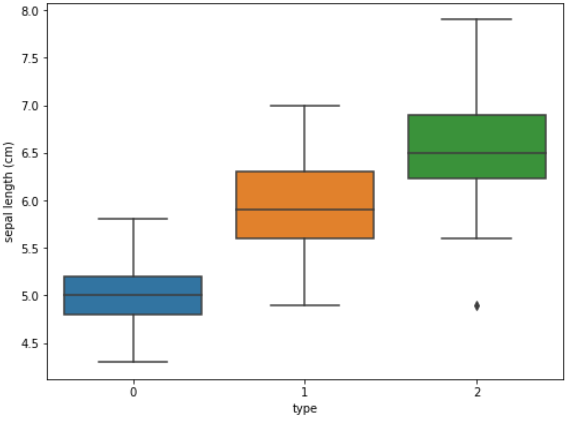

Fig 1.3 d)

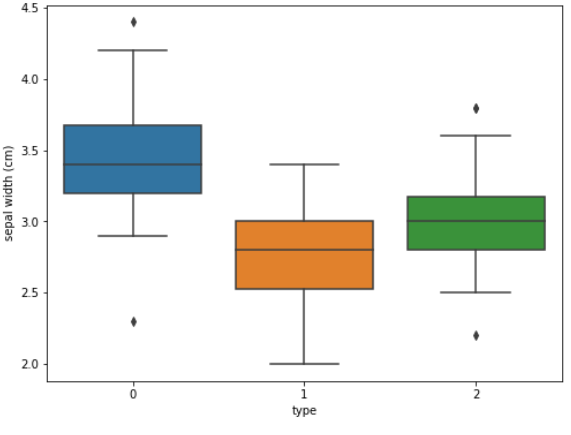

Fig 1.3 e)

Studying the plot of fig 1.3 a) we can observe that the data doesn't have any massive outliers and therefore unsupervised learning algorithms which are sensitive to outliers could be employed for classification of the flowers into different species.By studying the fig 1.3 b) and fig 1.3 c) further consolidates the analysis made in section 1.2 ,i.e the petal width and petal length value of specie 0 or setosa (blue) are much less than the other two species and the max and min value of these two attributes don't overlap with the other species at all. When we study the rest of the plots we get a really interesting observation that the average value of all the features (i.e sepal width and length,petal width and length) of the specie 1 or versicolor (orange) are less than the average values of the corresponding features of the specie 2 or virginica (green).However despite the average value of specie 1 being lesser there is overlap between the features of the species 1 and 2(namely versicolor and virginica).

Summarizing the data exploration process we can say that it looks quite promising to use supervised learning algorithms like logistic regression and naeve bayes to separate the species from each other (classify them) , these algorithms would do an excellent job in classifying sample into setosa ,these would also classify the other species as well however they may face some discomfort because of absence of a distinct boundary between versicolor and virginica despite the average value of features being less for versicolor (type 1).Due to absence of outliers we can also think of using the unsupervised clustering algorithm K-means.

2. Data Classification

2.1 Gaussian Naïve Bayes

Gaussian naïve bayes is a supervised learning algorithm which uses a probabilistic model in the backend to predict that the given data point is most likely to fit which class. It is a generative model because it doesn't directly calculate the probability that a particular point belongs to a particular class but uses the baye's theorem to calculate the class which the data most likely belongs to.It does so by assuming that the different features arise from gaussians of different classes(for example the probability that a setosa flower would have a certain sepal width is given by a normal/gaussian distribution) (assumption).It seems quite encouraging to use this model in the given case as analysed in the exploration step as well as the fact that all the features are a result of natural/biological selection process (petal and sepal dimensions) it is encouraging to use gaussian distribution to calculate their probabilities. The gaussian naïve bayes assumes no correlation between its feature variables as well and it is required to calculate a prior probability that if a flower is picked randomly then what is the probability that it belongs to a particular class ,in this case it is 1/3.The gaussian naïve bayes performs better than the one vs rest logistic classifier when the data set given is small, however fails to do so in larger data sets.

2.1.1 The confusion matrix: -

This matrix helps to evaluate the accuracy of a particular classification model.By definition the confusion matrix is such that the (i,j)th entry is equal to the number of observations known to be in group i and predicted to be in group j.Given below is the confusion matrix table.The data that is tabulated is from the testing data.(which was obtained after a random 80:20 split)

	Setosa(0)	Versicolor (1)	Virginica(2)
Setosa(0)	13	0	0
Versicolor(1)	0	5	0
Virginica(2)	0	1	11

As predicted during data exploration this algorithm doesn't have any problem when classifying points into type 0 or setosa specie.However it has misclassified a data point of virginica specie into a versicolor specie which may be due to the inseperable nature of the specie 1 and 2.More on this is explained in the misclassification analysis section 3.

2.1.2 The classification report: -

This is a text document or report that mentions most of the common classification performance metrics such as precision , recall of all features,the overall accuracy of the model used for classification etc.

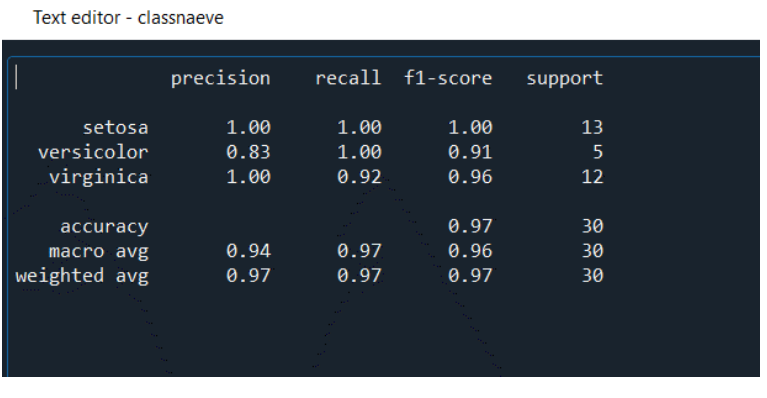

Fig 2.1

It is clear from the above classification report that the algorithm has a high value of recall and precision(1 to be precise) when considering the case of setosa,however these values are slightly less for the case of the other species because of the misclassification of one data point.The overall accuracy of the model is also high at 97%, hence overall the algorithm seems to perform pretty well.

2.2 One vs Rest Logistic Regression

This is a supervised learning algorithm and uses a discriminative model to classify the data points into separate species, it work by applying binary logistic regression three times in this case , every time to find the probability function of what is the probability that a certain data point belongs to a certain class. It doesn't assume anything about the probability distribution function unlike the GNB which assumes a gaussian distribution of the features.However there are certain assumptions used by logistic regression , the model assumes that the given data points are linearly seperable.Also it works by finding a distinct boundary/plane/hyperplane between the data points to be classified.Given this information about the algorithm it is evident that it will be able to separate setosa (type 0) easily from others but again may face discomfort for the remaining two species most probably due to their inseparable nature.

2.2.1 The confusion matrix: -

This matrix helps to evaluate the accuracy of a particular classification model.By definition the confusion matrix is such that the (i,j)th entry is equal to the number of observations known to be in group i and predicted to be in group j.Given below is the confusion matrix table.The data that is tabulated is from the testing data.(which was obtained after a random 80:20 split).

	Setosa(0)	Versicolor (1)	Virginica(2)
Setosa(0)	13	0	0
Versicolor(1)	0	4	1
Virginica(2)	0	1	11

As predicted during data exploration this algorithm doesn't have any problem when classifying points into type 0 or setosa specie. However it has misclassified a data point of virginica specie into a versicolor specie and a data point of the versicolor specie into a virginica specie which was also evident in the data exploration step.Its performance is slightly lower than the naïve bayes algorithm as predicted because of the size of the data

set and also the features are assumed to come from a gaussian distribution in the gaussian naïve bayes which seems to be a good assumption. More on why these algorithms are unable to distinguish between versicolor and virginica in the misclassification analysis in section 3.

2.2.2 The classification report: -

This is a text document or report that mentions most of the common classification performance metrics such as precision , recall of all features, the overall accuracy of the model used for classification etc.

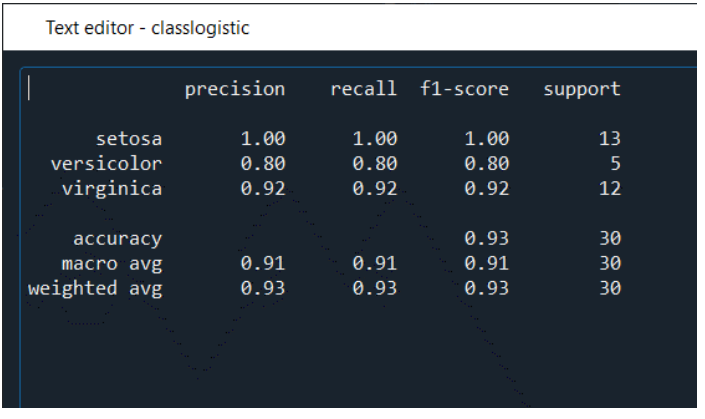

Fig 2.2

It is clear from the above classification report that the algorithm has a high value of recall and precision(1 to be precise) when considering the case of setosa,however these values are slightly less for the case of the other species because of the misclassification of two data points.The overall accuracy is 93% which is slightly less than the naïve bayes classification but we can safely say that we can use logistic regression to get promising results in classifying this dataset.Moving on to another classification technique.

2.3 K means

K means is an unsupervised learning technique which attempts to cluster the data points into certain spherical boundaries based on the Euclidian distance of data points from the centre of such clusters. The accuracy of the model depends on the choice of the initial cluster centres and if not chosen wisely then the algorithm would lead to high variance, however this can be avoided by choosing different cluster centres randomly in different iterations and those centres be chosen where the variation of data points is least after classification. This algorithm doesn't need the target values to learn but instead just takes in a feature data set and spits out the classification of the data points.The problem in the implementation of this algorithm in the scikit learn library is that the labels that this algorithm gives may not be consistent with the original labels provided in the data set , that

is to say that it may give setosa specie a label 1 when the label given in the training data is 0,hence a manual change is required (interchanging certain label values like make all labels that are given to be 0 to 1 and so on).The pros of this algorithm is that it is fast and converges to a minima. However the cons are also serious, that is it assumes equal covariance in all direction which may not always be the case and furthermore it detects spherical clusters instead of a distinct boundary of other shapes which may lead to misclassification in the given data set.

2.3.1 The confusion matrix: -

This matrix helps to evaluate the accuracy of a particular classification model.By definition the confusion matrix is such that the (i,j)th entry is equal to the number of observations known to be in group i and predicted to be in group j.Given below is the confusion matrix table.The data that is tabulated is for the entire data since there is no need to split the data into separate training and testing data in the case of K-means algorithm.

	Setosa(0)	Versicolor (1)	Virginica(2)
Setosa(0)	50	0	0
Versicolor(1)	0	48	2
Virginica(2)	0	14	36

In this case also the algorithm is easily able to classify the data points into setosa specie or type 0 but we can see there are not one or two but several misclassifications(16 to be precise) in the case of the species versicolor and virginica.This may be due to the fact that this algorithm detects spherical clusters and hence leads to high misclassification for virginica and versicolor. More analysis on this is done in section 3.

2.3.2 The classification report: -

This is a text document or report that mentions most of the common classification performance metrics such as precision, recall of all features ,the overall accuracy of the model used for classification etc.

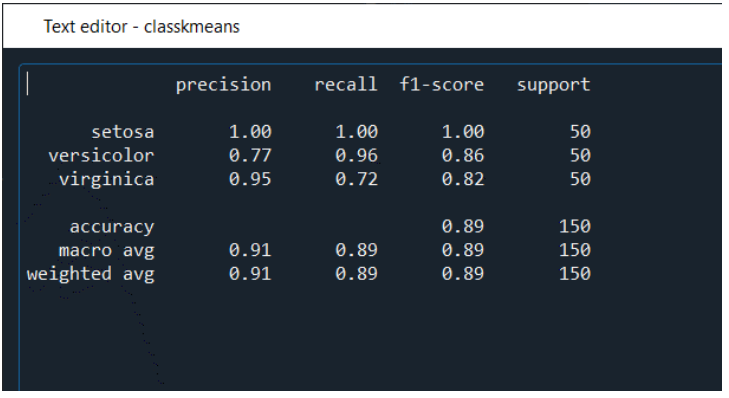

Fig 2.3

It is clear from the above classification report that the algorithm has a high value of recall and precision(1 to be precise) when considering the case of setosa,however the values of precision for versicolor and recall for virginica see a significant drop because of the misclassification of 16 data points as observed in the confusion matrix, We observe a significant drop in the accuracy of the model which is 89% and lower than the naïve bayes and logistic regression models. The misclassification and the low comparative performance of this model is explained and analysed in the upcoming section on misclassification analysis.

3. Misclassification Analysis using PCA

PCA or principal component analysis is a type of feature extraction procedure where in the number of dimensions/variables are reduced by making the new variables a linear combination of the initial features/variables, this helps us to eliminate dimensions without totally eliminating the less significant variables. The dimensions that are left after applying pca are those which explain the most variability in the given data points and hence applying PCA greatly helps us by allowing easy visualization ,for example when we reduce 4 dimensions to 3 we can visualize a 3d plot which is the best 3d plot capable of explaining the most variation in the data. If we reduce the 4 dimensions to 2 we get a 2d plot / plane which is the plane capable of explaining the most variation (on a plane surface) in the data. However there is a drawback of PCA, if we reduce the dimensions we aren't able to exactly interpret the new variables that are formed.

In this example of the iris dataset we apply the PCA to reduce the dimension to 3 and 2 to understand the misclassification occurring b/w the type 1 and type 2 flowers namely the versicolor and virginica specie. This can be done as we will get the plots which explain the most variability in the data and we can see whether we can get spherical clusters within the data or some distinct boundary between the two species.Given below are the plots.Note again:- type 0- setosa , type 1-versicolor ,type 2- virginica.The 3d plot is captured from several different angles for easy visualization and interpretation.

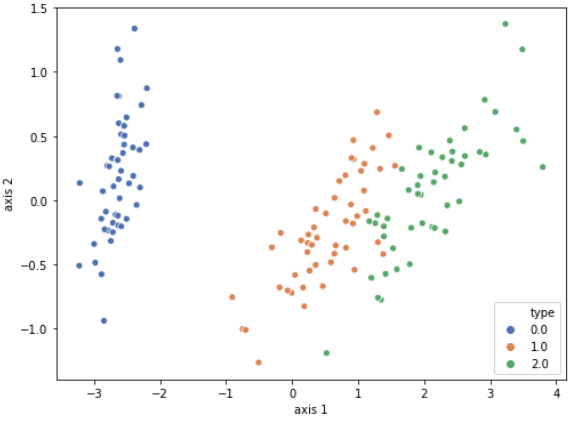

Fig 3.1(Reducing to 2 dimensions)

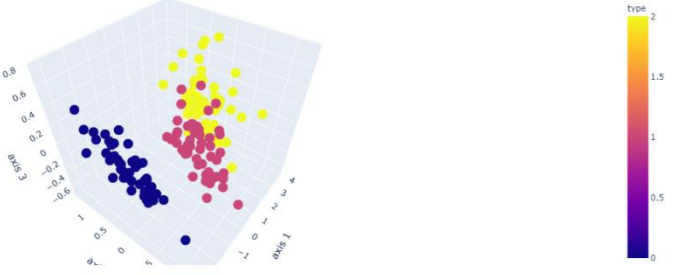

Fig 3.2(Reducing to 3 dimensions)

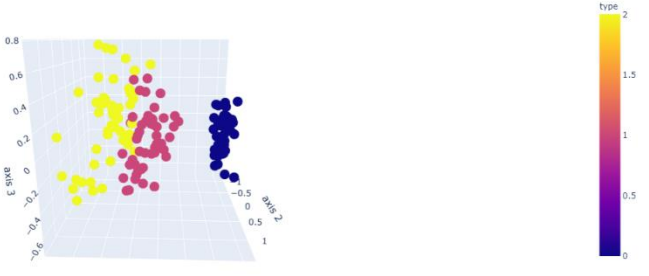

Fig 3.3 (Reducing to 3 dimensions)

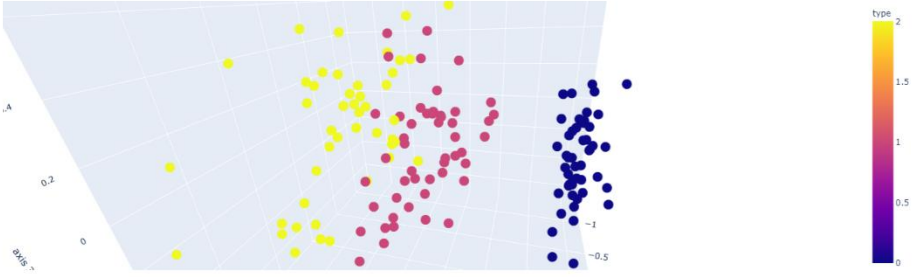

Fig 3.4 (Reducing to 3 dimensions)

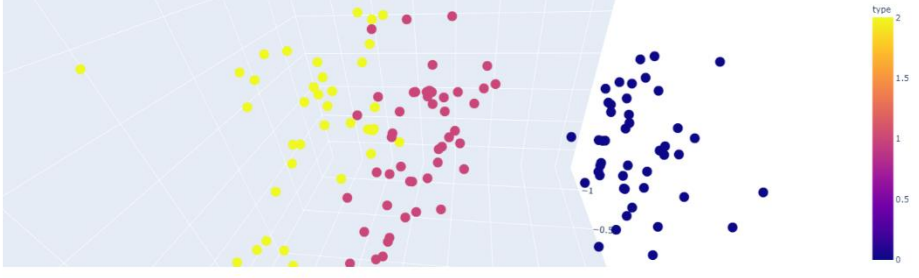

Fig 3.5 (Reducing to 3 dimensions)

If it is unclear from the legends in the plots , for 2d plot blue – setosa (type 0),orange- versicolor (type 1),green- virginica(type 2), for 3d plot blue – setosa (type 0),red- versicolor (type 1),yellow- virginica(type 2).

First analyzing fig 3.1 (the 2d plot) ,it is clear after looking at the plot that setosa is easily seperable from the other species because of a large gap between the blue dots and the green and orange dots , which explains why all the three algorithms work well to do so , even k means which makes spherical clusters doesn't face difficulty classifying the data points into setosa.However we can see some overlap between the green and orange data points which explains why all the algorithms do some misclassification while classifying versicolor and virginica because of the inseparable nature of the two species , however the k means algorithm performs particularly bad while separating type 1 and 2 which will become clearer when we study the 3d plots.

Coming on to the 3d plots, as seen in Fig 3.2 and Fig 3.3 , setosa(blue dots) is clearly well separated from the species virginica and versicolor, explaining the good performance of all

the classification algorithms in seperating setosa, and there seems to be overlap between the red and yellow points.Coming on to fig 3.4 and fig 3.5 which are the zoomed in version of the plots, we can clearly see significant overlap between the yellow and red points which clearly explains the slight performance drop faced by logistic regression and naïve bayes algorithms while classifying species into type 1 and type 2.However K means suffers a drastic performance drop while classifying species into type 1 and type 2, this happens because it detects spherical clusters in the data points and as is clear form fig 3.4 and fig 3.5 if we make spheres around the yellow and red data points we will have massive overlap between the spheres which will result in bad performance and high misclassification while separating versicolor and virginca. Naïve bayes and logistic regression don't face such performance drops as they don't use spherical boundaries to separate the species , logistic regression tries to find a distinct surface for a boundary (which will not lead to as massive an overlap as spherical surfaces) and naïve bayes assumes gaussian distribution of the features which seems to work the best.

Summary of conclusions and learnings: -

1. The Setosa specie is easily separable from rest of the species because of its features which are in a different range than the features of the other species virginica and versicolor.All the algorithms perform well in separating setosa.

2. There is some amount of overlap in versicolor and virginica despite the average values of features of versicolor being less than virginica which leads to some amount of misclassification by the algorithms while classifying them.

3. The gaussian naïve bayes seems to perform slightly better than logistic regression ,however the performance of these supervised learning algorithms are more or less the same.K means performs the worst among the three algorithms and is not suggested to be used to separate versicolor and virginica.

4. After PCA it is evident why K means doesn't have good performance while classifying virginica and versicolor because of significant overlap between the data points of these species, it doesn't work well because it detects/forms spherical clusters which leads to massive misclassification in the case of versicolor and virginica.

Bibliography

[1] https://pandas.pydata.org/docs/

[2] https://numpy.org/doc/

[3] https://scikit-learn.org/stable/

[4] https://matplotlib.org/

[5] https://seaborn.pydata.org/

[6] https://plotly.com/python/3d-scatter-plots/

[7] Stack overflow for some syntax.