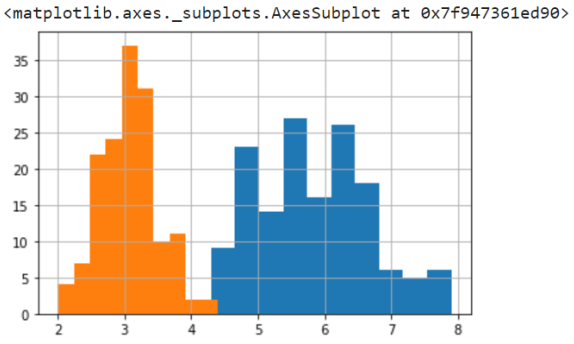
**Task 1: Analysis of Iris dataset using Logistic Regression**

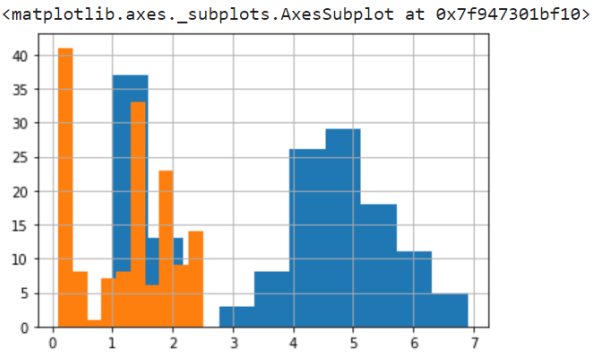
**Overview:** based on a given Iris dataset, we want to deduce, based on the features such as petal length, petal width, sepal length and sepal width, the probability of particular values belonging to a specific species (types) of the iris plant; namely either setosa, versicolor, or virginica.

**The data and the classification process:**

Data visual for sepal:




Data visual for petal:




We can see that the histogram plot of all the features follows the normal distribution. If the size of the data set is large enough, the normal distribution is a good approximation.

We then ran the correlation matrix:

| | sepal length (cm) | sepal width (cm) | petal length (cm) | petal width (cm) | target |
|---|---|---|---|---|---|
| **sepal length (cm)** | 1.000000 | -0.117570 | 0.871754 | 0.817941 | 0.782561 |
| **sepal width (cm)** | -0.117570 | 1.000000 | -0.428440 | -0.366126 | -0.426658 |
| **petal length (cm)** | 0.871754 | -0.428440 | 1.000000 | 0.962865 | 0.949035 |
| **petal width (cm)** | 0.817941 | -0.366126 | 0.962865 | 1.000000 | 0.956547 |
| **target** | 0.782561 | -0.426658 | 0.949035 | 0.956547 | 1.000000 |

The correlation matrix below shows the **correlation coefficient** between petal length and petal width is 0.962865, which is much closer to 1, so we can remove any of the two variables.

We trained the model and print the accuracy:

```
[1 2 2 0 2 1 0 2 0 1 1 2 2 2 0 0 2 2 0 0 1 2 0 1 1 2 1 1 1 2]
82     1
134    2
114    2
42     0
109    2
57     1
1      0
70     1
25     0
84     1
66     1
133    2
102    2
107    2
26     0
23     0
123    2
130    2
21     0
12     0
71     1
128    2
48     0
72     1
88     1
148    2
74     1
96     1
63     1
132    2
Name: target, dtype: int64
Accuracy - 96.66666666666667
```

**Analysis:**

After we trained the model, we are getting an accuracy of 96.67%, which is very decent. The accuracy increases with the increase in the number of data points. From the above plot, we can see that the model is doing a fairly good prediction for plant species.

**References:**


"Exploratory Data Analysis on Iris Dataset". Retrieved from https://www.geeksforgeeks.org/exploratory-data-analysis-on-iris-dataset/

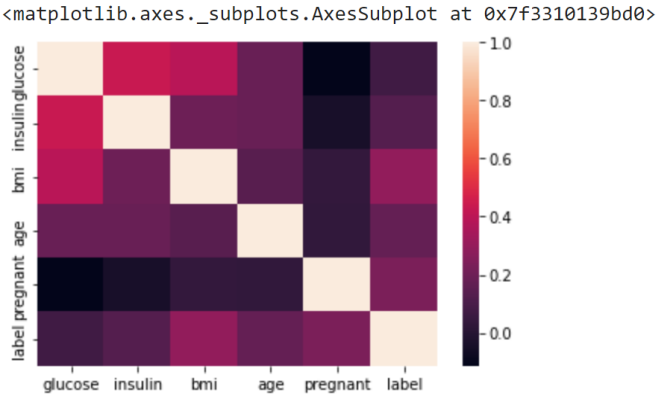
Srivani, K.D. (2022). "IRIS Flowers Classification Using Machine Learning". Retrieved from https://www.analyticsvidhya.com/blog/2022/06/iris-flowers-classification-using-machine-learning/

**Task 2: Analysis of Diabetes dataset using Random Forest**

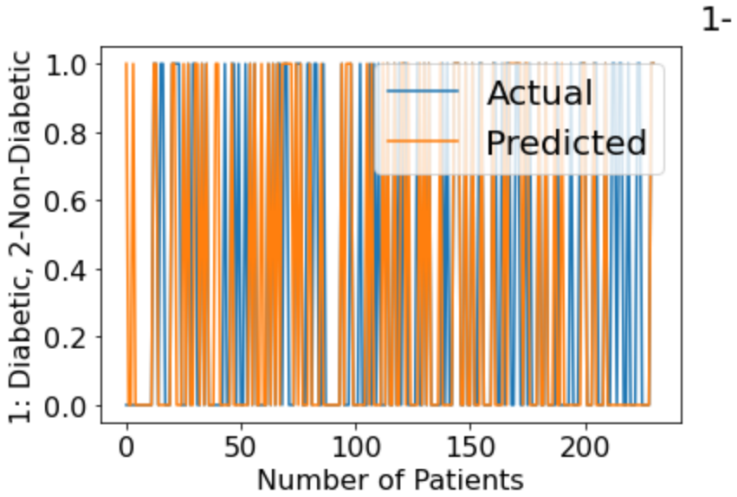
**Overview:** We want to build a machine-learning model to predict whether or not a patient has diabetes

**The data and the classification process:**

There are 5 features that we use: ‘glucose’, ‘insulin’, ‘bmi’, ‘age’, and ‘pregnant’. The correlation between each columns are visualized using heatmap (image below). From the output, the lighter colors indicate more correlation. We notice the correlation between pairs of features, like glucose and pregnancies, or insulin and glucose, etc.

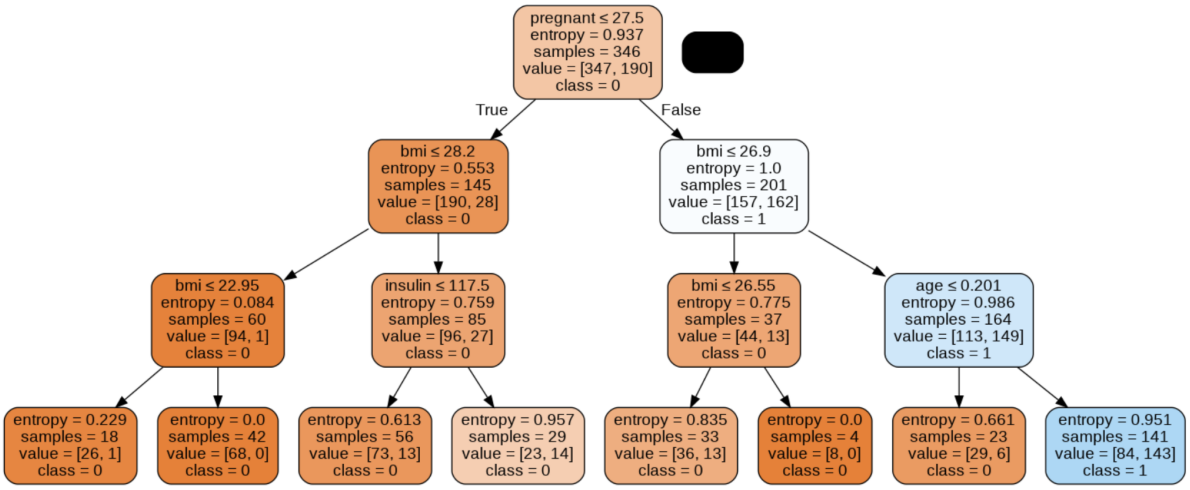



We trained the model by splitting the training and the testing data: 70-30. When printing the accuracy, we found that the accuracy score is over 74 percent, which is quite decent.




```
Accuracy: 74.45887445887446
```

**Visualisation with random forest:**




**Analysis:**

Looking at the accuracy score and the plotting, the machine-learning model performs decently by correctly identifying which patients have diabetes based on given features. I we observe the random tree visualisation, we also notice values like, for example: [347, 190], which means that before the True/False feature split, 347 samples were of the "patient-has-no-diabetes" (Negative) class and 190 were of "patient-has-diabetes" (Positive) class.

The goal of a decision tree is to make the "best" True/False splits so right at the very bottom of the tree we get these "pure" nodes where there's only one type of class or another. A visual way to see how good a feature split is represented is that the darker the hue of the box is, the purer the node. In the image above, at the top of the tree, nodes are generally light-brown/light-blue/light-orange, and as they go down, the colour of the node becomes darker.

**References:**

Jay, R. (2021). "What's in a "Random Forest"? Predicting Diabetes" in Towards Data Science. Retrieved from https://towardsdatascience.com/whats-in-a-random-forest-predicting-diabetes-18f3707b6343

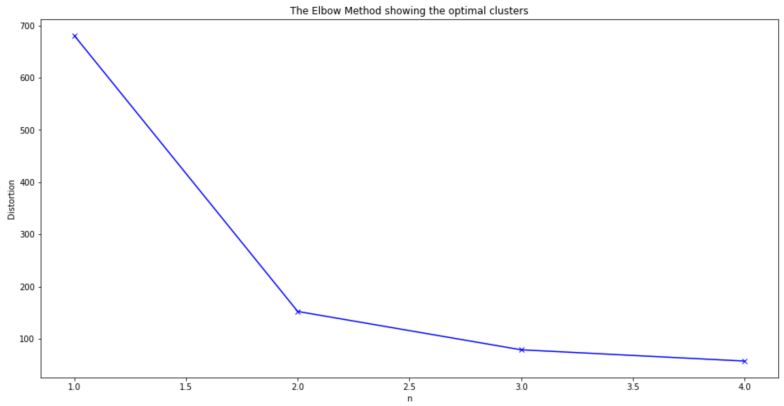
Nayak, L (2022). "Predicting Diabetes with Random Forest Classifier" in Towards Data Science. Retrieved from https://towardsdatascience.com/predicting-diabetes-with-random-forest-classifier-c62f2e319c6e

**Task 3: Analysis of Iris dataset using K-means clustering**

**Overview:** based on a given Iris dataset, and using an unsupervised learning method, we want to deduce, based on the features, the probability of particular values belonging to a specific species (types) of the iris plant.

**The data and the classification process:**

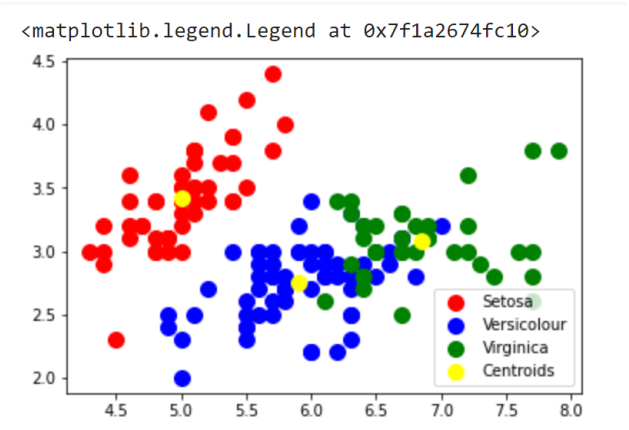
When we use the Elbow Method, we find that the optimal number (centers) of clusters is 3.0.




When we implemented the K-Means, the result was three clusters with the values of 0, 1, and 2. When visualised, we get an insight from the scatterplot below that the model is accurate in determining Iris flower species.

Although we already know the answer is 3 as there are 3 unique class in Iris flowers, we can get an absolute segmentation when we put higher K values. However, if the points within each cluster are very less, then the variation on the real data will be high - leading it into over simplifying the data.

So, with K=3 we have obtained an optimal distortion/inertia with which we can segment the data into 3 different clusters with minimal error in segmentation.



**References:**

Bandgar, S. (2021). “CLUSTERING ON IRIS DATASET IN PYTHON USING K-Means” in Analytics Vidhya. Retrieved from https://medium.com/analytics-vidhya/clustering-on-iris-dataset-in-python-using-k-means-4735b181affe

Khotijah, S. (2022). “K-Means Clustering of Iris Dataset“ in Kaggle. Retrieved from https://www.kaggle.com/code/khotijahs1/k-means-clustering-of-iris-dataset

Saxena, Y. (2021). “Analyzing Decision Tree and K-means Clustering using Iris dataset” in Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/analyzing-decision-tree-and-k-means-clustering-using-iris-dataset/