Machine-learning analysis for the Iris dataset

Article · May 2020

CITATIONS
0

READS
3,511

1 author:

Ahmed Ali
University of the Cumberlands
44 PUBLICATIONS **2** CITATIONS

SEE PROFILE

Ahmed Ali

Ph.D. Student

University of Cumberlands

The Iris Dataset Analysis

Introduction

Source code
https://colab.research.google.com/drive/1buACBYRuLvyv0IsAXMToI5W3EHM6CkNo?usp=sharing

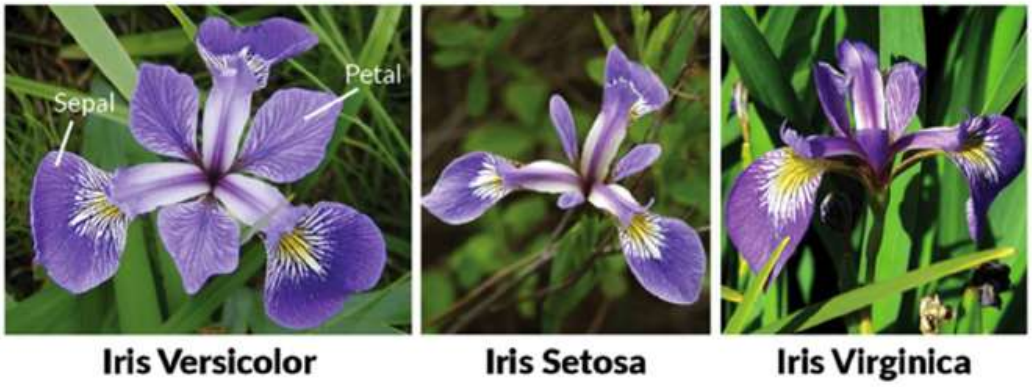


Figure 1 retrieved from Ojha (2019)

The Iris data set used in a specific set of information compiled by Ronald Fisher, a biologist in the 1930s. The data set contains three classes of 50 instances each, where each class refers to a type of iris plant (Ojha, 2019).

The following describes the dataset attributes:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   sepal_length  150 non-null    float64
 1   sepal_width   150 non-null    float64
 2   petal_length  150 non-null    float64
 3   petal_width   150 non-null    float64
 4   species       150 non-null    object
```

The following describes the heads for the dataset:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa

Exploratory Data Analysis

The first step in this section is importing the data into the google Colab exploratory data analysis and find the relationship between attributes. Figure 1 shows the snapshot code of plotting the data from Iris data imported through the Panada library. First, we explore the dependents variables, and plotting the Iris data set shows the plot for all of the length and width distribution dependent variables. Figure 2 shows the histogram of petal and sepal length and width.

```
iris.hist(edgecolor='black', linewidth=1.2)
fig = plt.gcf()
fig.set_size_inches(14,8)
plt.show()
```

Figure 1

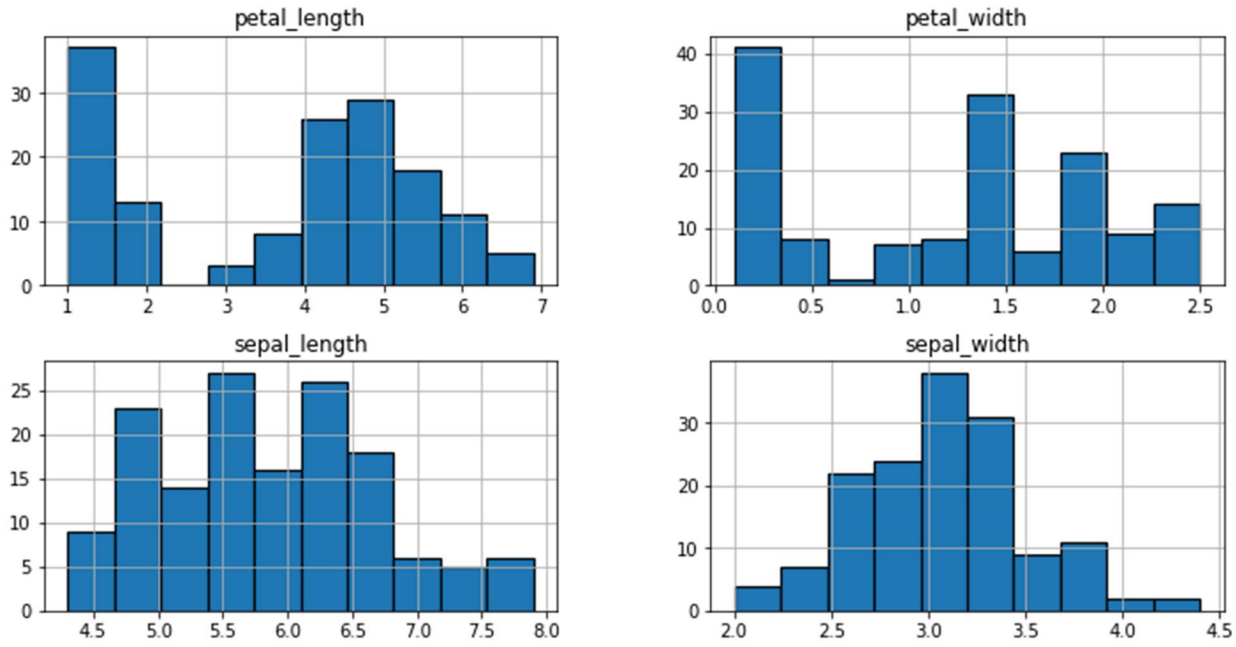


Figure 2 histogram of petal and sepal length and width.

Logistic Regression

Fitting Logistic Regression to the Training set, probability of a particular class or event existing such

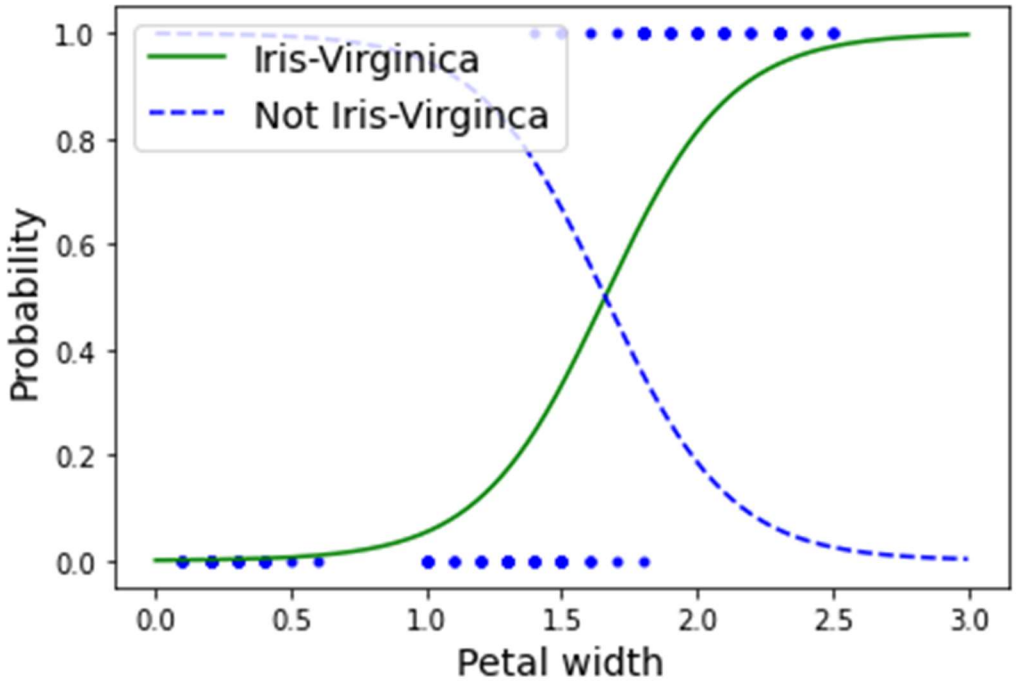


Figure 3

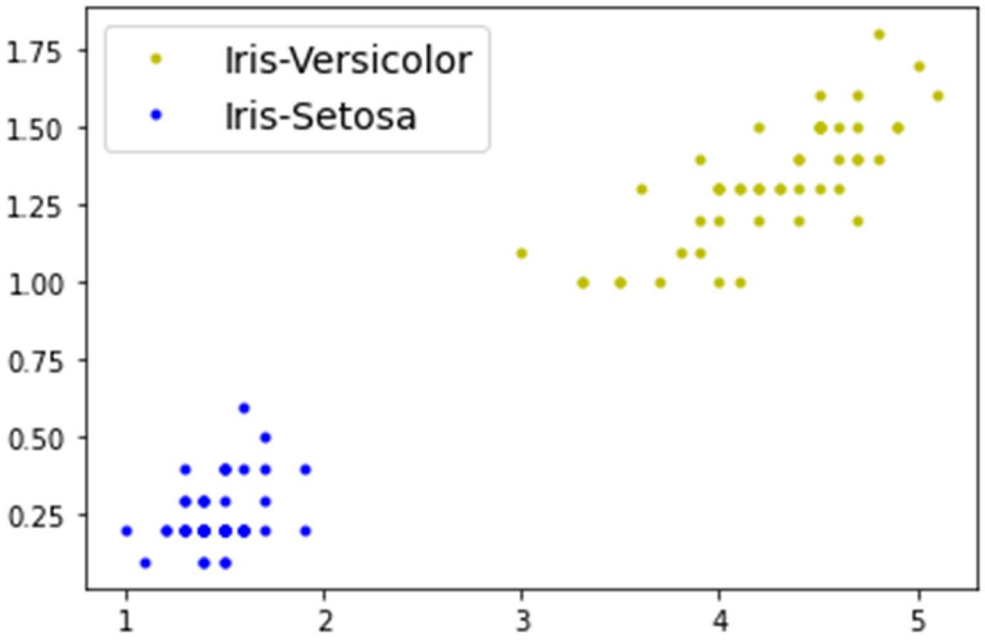


Figure 4

Naive Bayes

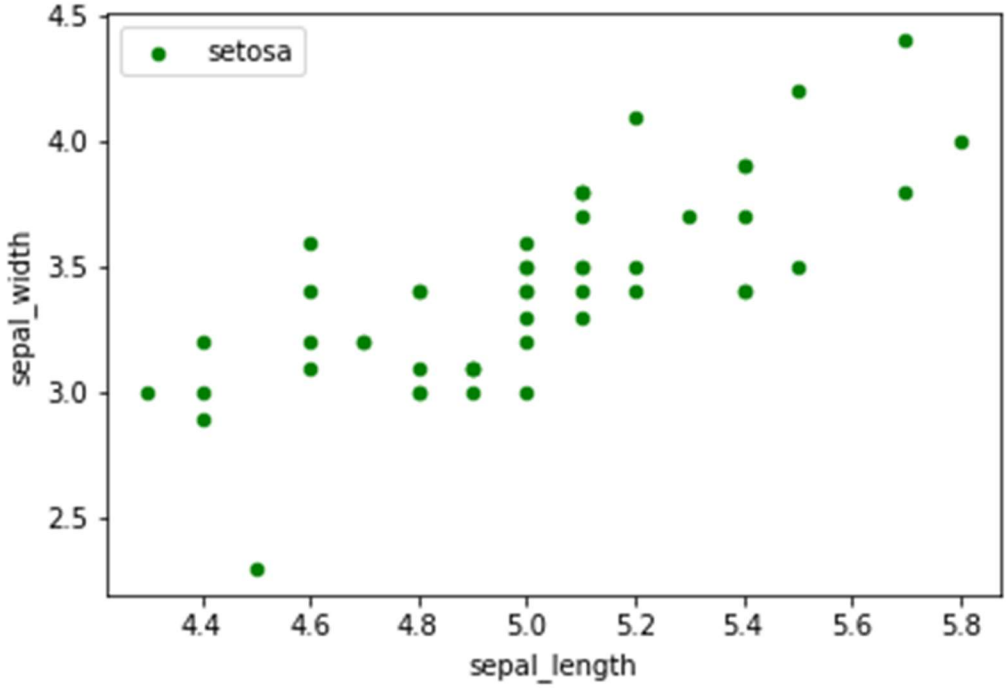


Figure 5

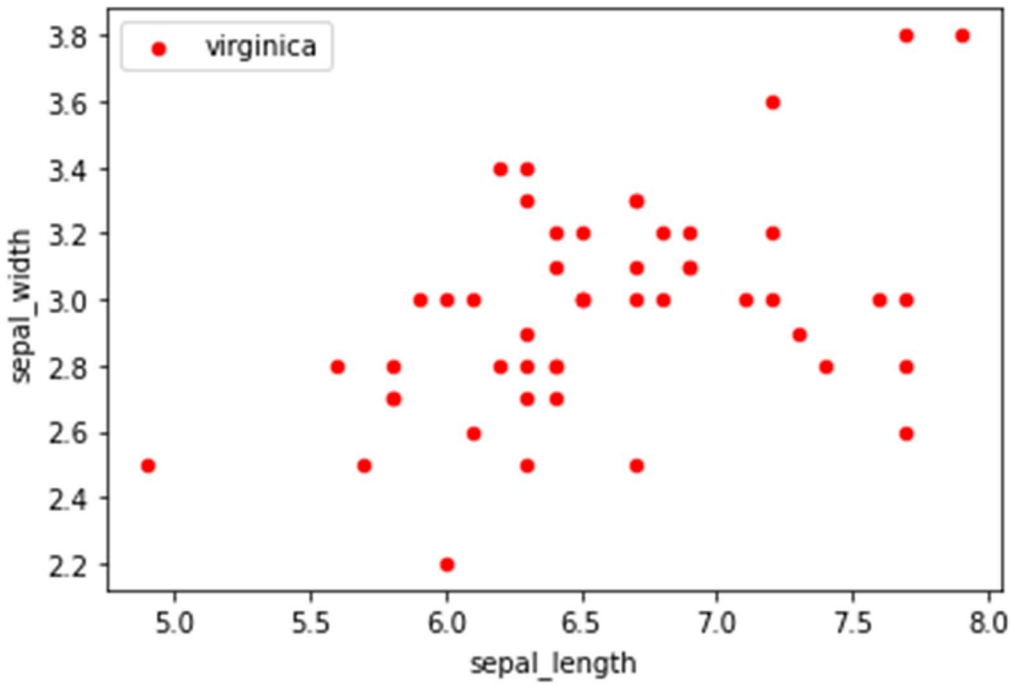

virginica
sepal_width
sepal_length
3.8
3.6
3.4
3.2
3.0
2.8
2.6
2.4
2.2
5.0
5.5
6.0
6.5
7.0
7.5
8.0

Figure 6

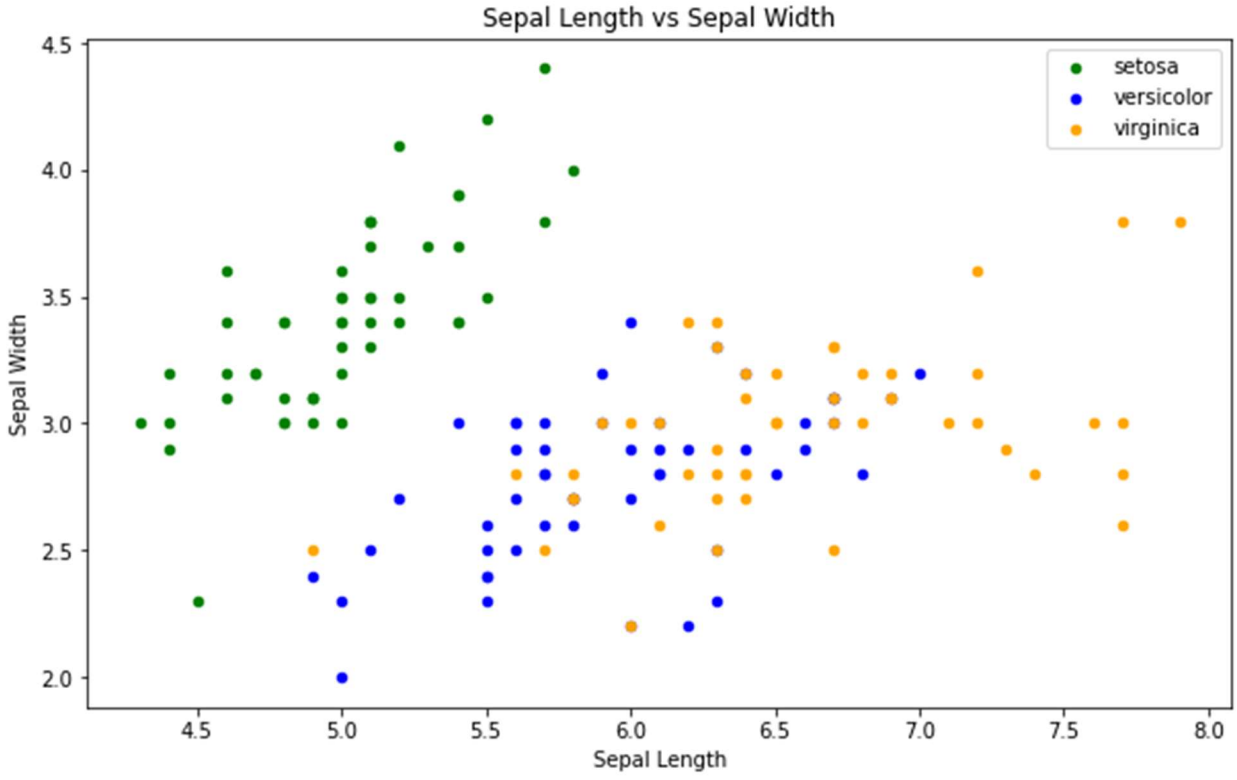

Sepal Length vs Sepal Width
setosa
versicolor
virginica
Sepal Width
Sepal Length
4.5
4.0
3.5
3.0
2.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

Figure 7

KNN

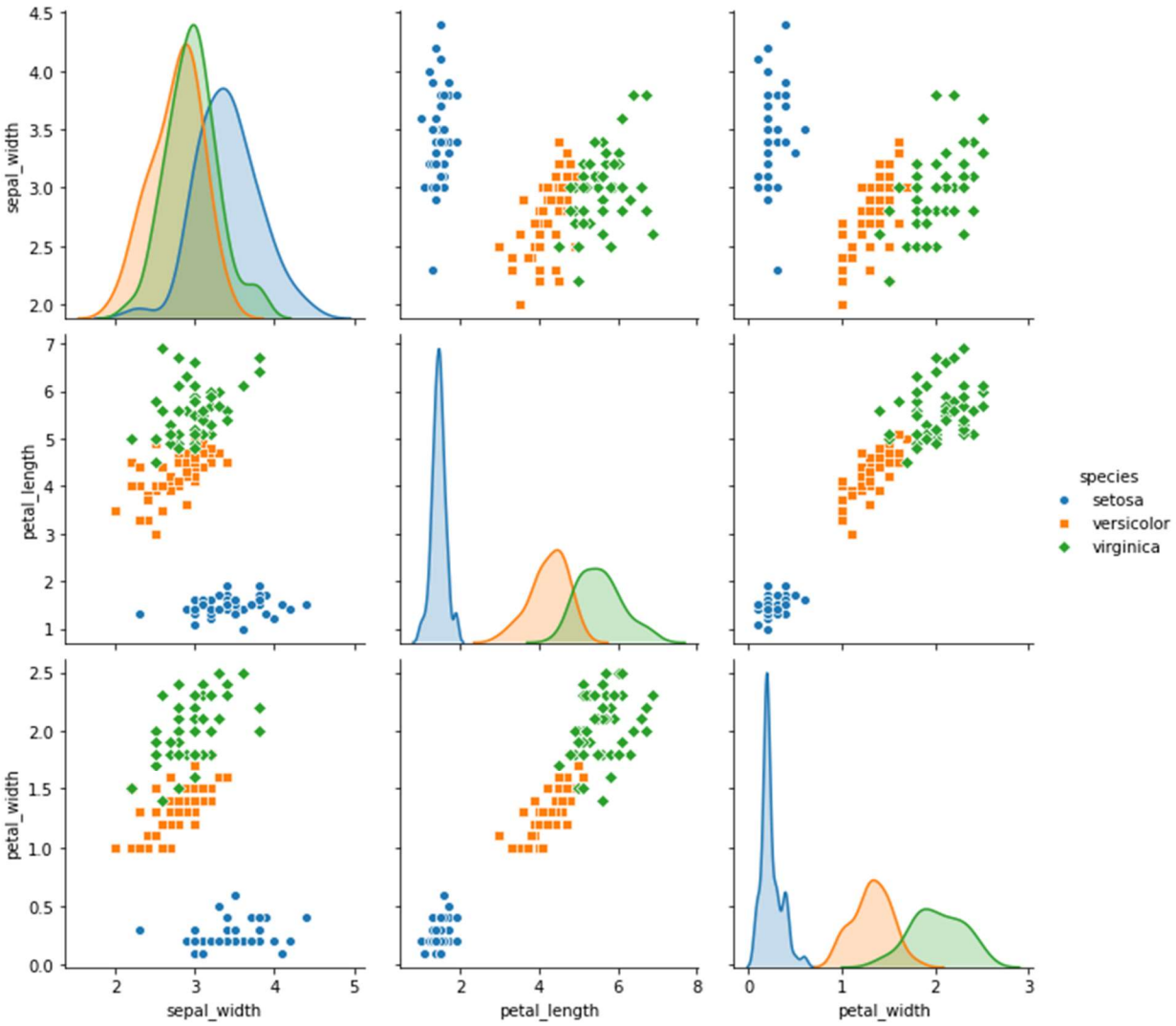
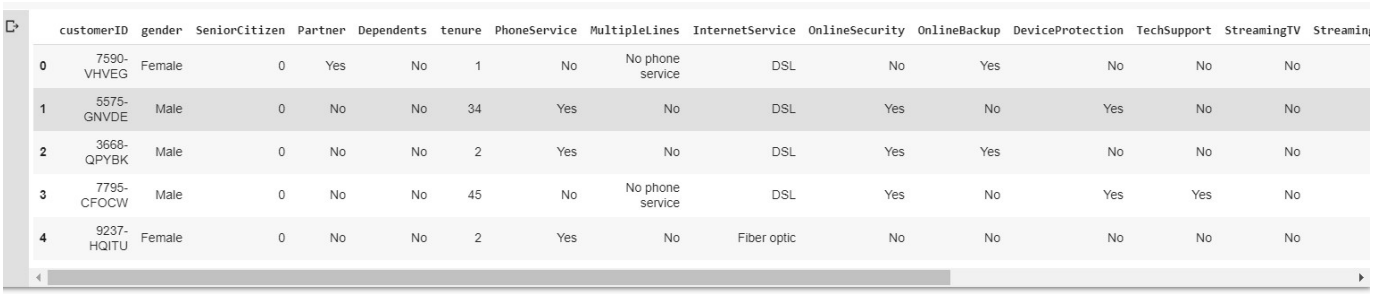


Figure 8

Churn Dataset

The dataset telco churn is concerning about losing customers. Big companies' services such as telephone, insurance as TV cables often use the analysis from customer Attrition To measure business sustainability. They are recovering customers back after long-term separation is more complicated than newly recruited clients. An important distinction should be taken into account between voluntary churn and involuntary churn. When customers decided to move to a different company, this is voluntary. On the other hand, involuntary churn occurs due to unexpected circumstances such as a customer moving to a different city. The dataset analysts are focus on voluntary churn because the reason is more likely related to the company, such as policy, price, and relationship.



	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	Streamin
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	

Figure 9

```
Features :
 ['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSup

Missing values :   0

Unique values :
 customerID          7043
gender                 2
SeniorCitizen          2
Partner                2
Dependents             2
tenure                73
PhoneService           2
MultipleLines          3
InternetService        3
OnlineSecurity         3
OnlineBackup           3
DeviceProtection       3
TechSupport            3
StreamingTV            3
StreamingMovies        3
Contract               3
PaperlessBilling       2
PaymentMethod          4
MonthlyCharges      1585
TotalCharges        6531
Churn                  2
dtype: int64
```

Exploratory Data Analysis

Figure 10 shows the gender distribution of non-churn customers, where it shows 50.7% of customers are males while the females are 49.3. for the churn customers, the males are 50.2% while the female percentage is 49.8.

Figure 10

Figure 11 shows the customers' attrition distribution which indicates 73.4% have attrited Figure 12 shows a comparison of attrition distribution between non-churn and churn customers. Figure 13 shows the senior citizen attrition percent.

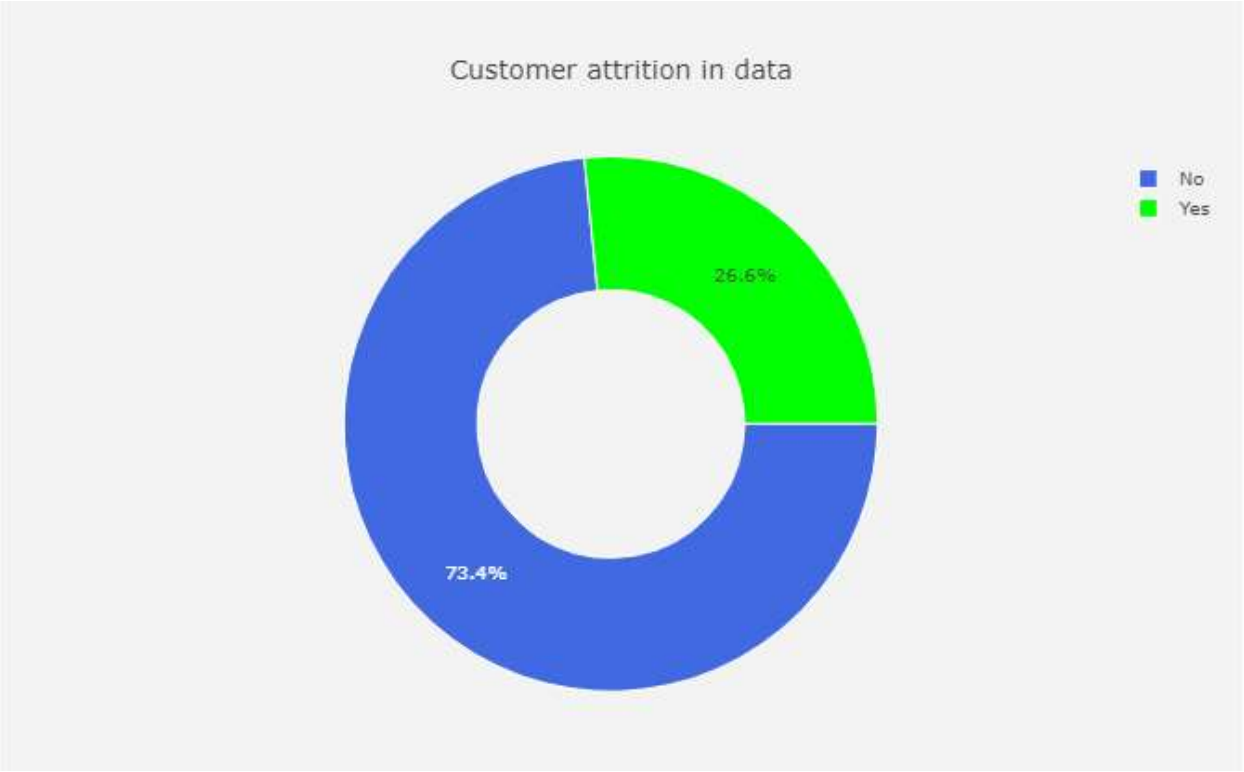


Figure 11

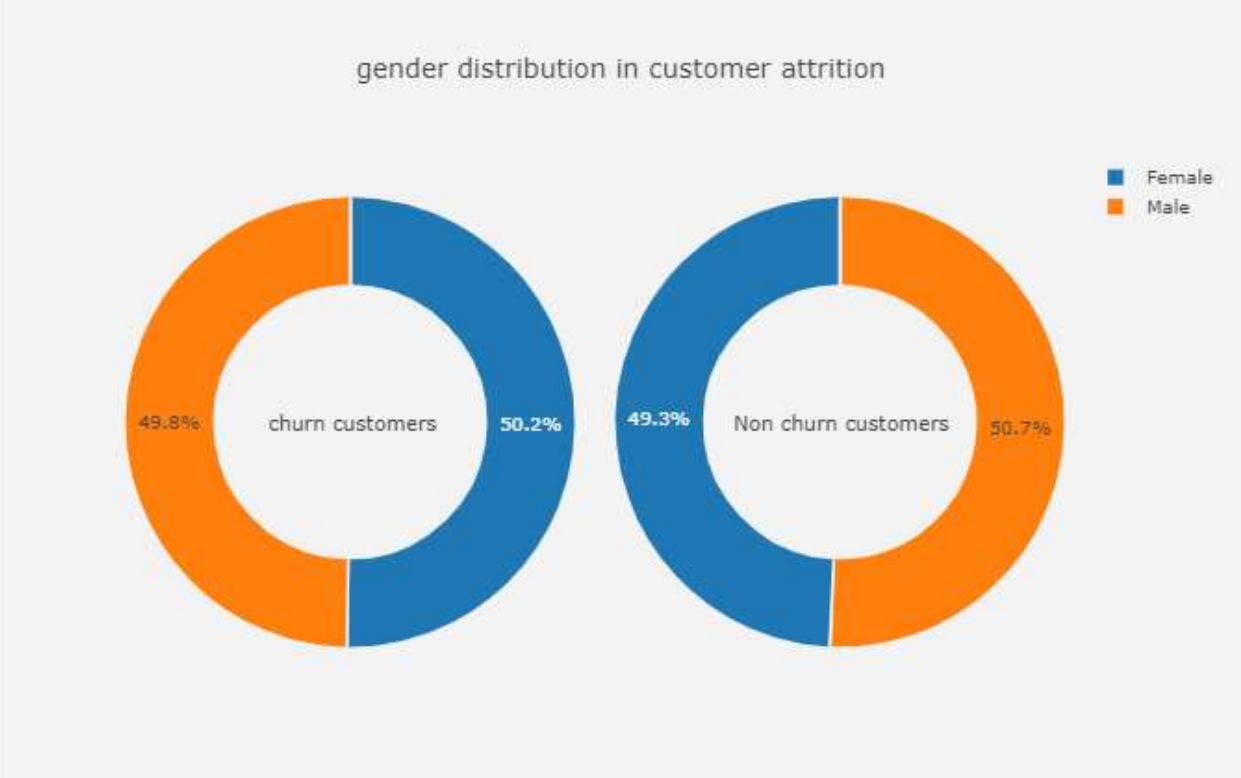

gender distribution in customer attrition
Female
Male
49.8%
churn customers
50.2%
49.3%
Non churn customers
50.7%

Figure 12

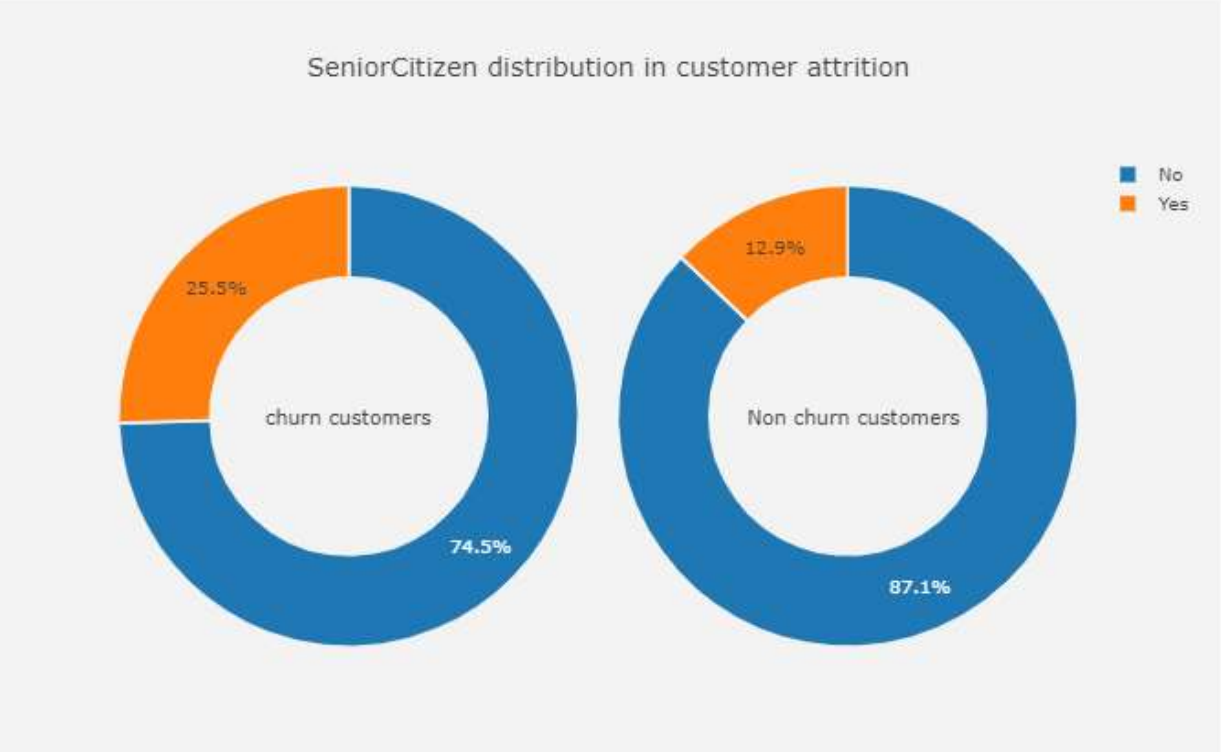


Figure 13

Decision Tree

Figure 14 shows the decision tree classification based on condition tenure <= 16.5 which will result in classifying the dataset into two main groups.

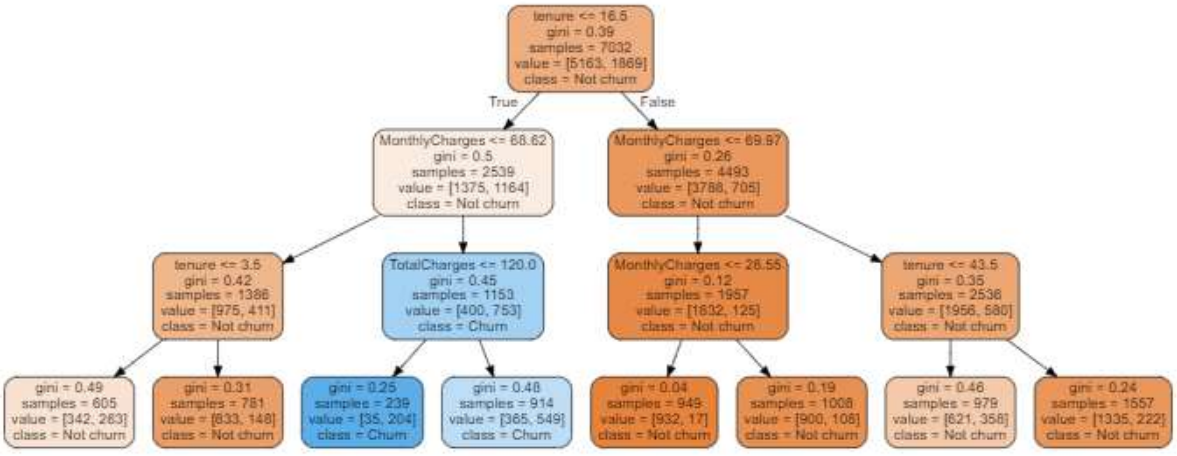


Figure 14

KNN Classifier

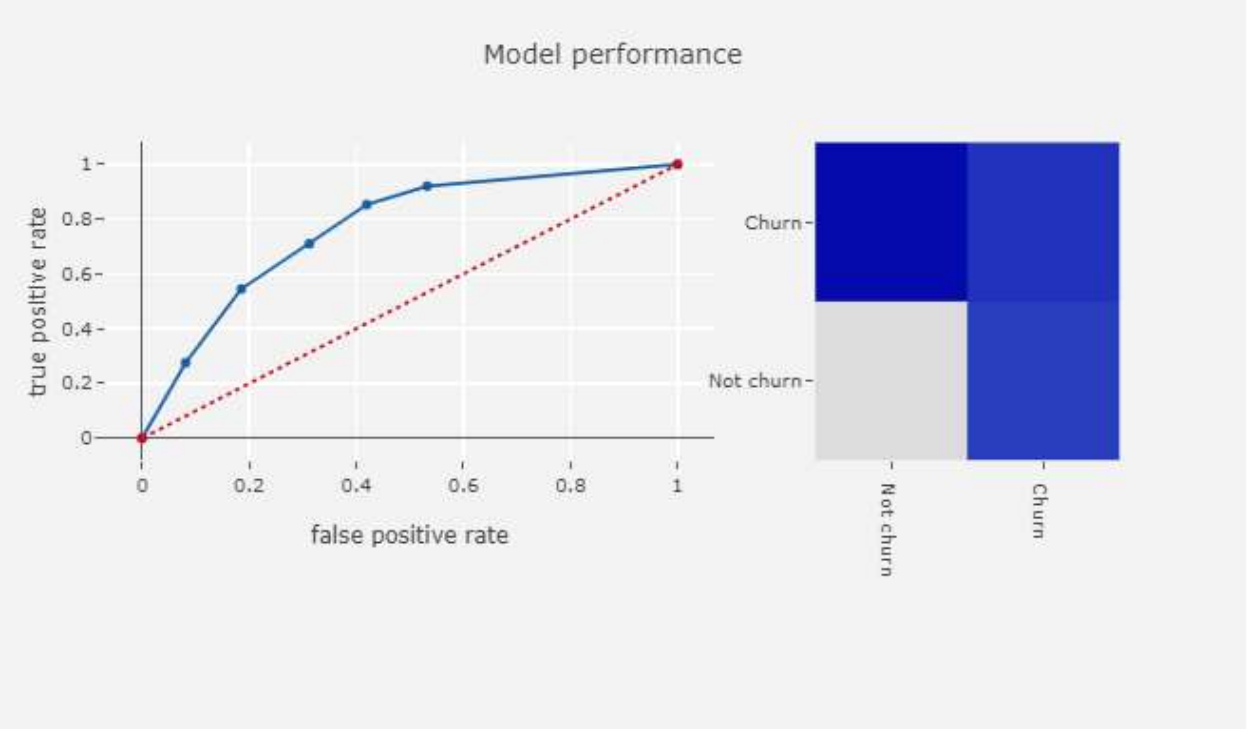
```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=1, n_neighbors=5, p=2,
           weights='uniform')

 Classification report :
              precision    recall  f1-score   support

          0       0.86      0.69      0.76      1268
          1       0.47      0.71      0.56       490

avg / total       0.75      0.69      0.71      1758

Accuracy Score   :  0.6939704209328783
Area under curve :  0.6989506212579669
```



References

Ojha, S. (2019). Exploratory Data Analysis On IRIS DATASET https://www.kaggle.com/uciml/iris

Jamesdhope (2017). Logistic Regression for Iris Classification Retrieved from https://www.kaggle.com/jamesdhope/logistic-regression-for-iris-classification