Compare Support Vector Machine to a 3 layer Neural Network on the titanic dataset.

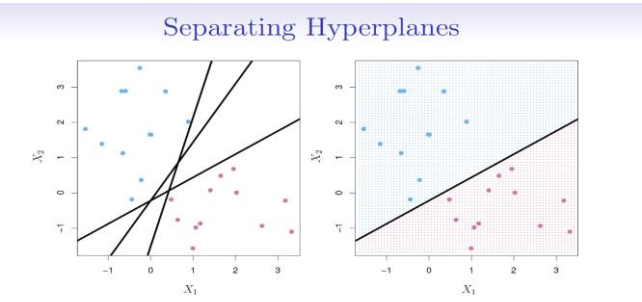
Jalaj Doraiburu
Machine Learning Intern
AI Technology and Systems
jalajdorai@gmail.com
www.ai-techsystems.com

***Abstract* -Titanic dataset is one of the basic datasets. It contains data of various details of passengers of titanic cruise. Name, Sex, Age, Sibling, Parents, ticket, Fare, Cabin, Embarkment and survival of passenger are the data contained in this data set.**
The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.
It's the Assignment-3 given to me as a Machine Learning Intern. I have done the Exploratory data Analysis, Preprocessing, modelled two algorithm -Support Vector Machine and a 3-layer neural network

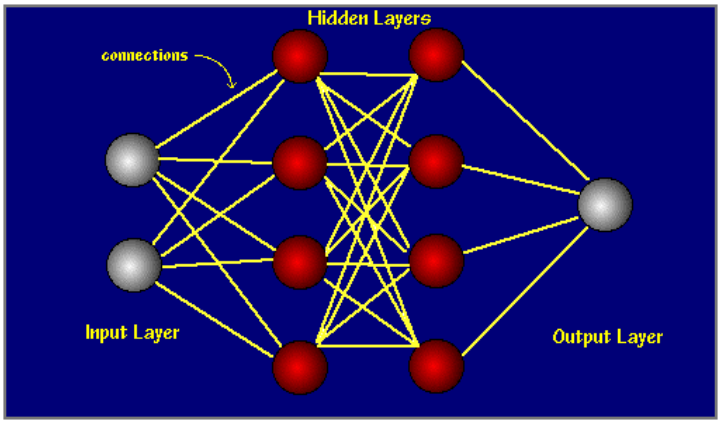
Keywords—support vector machine, neural network, prediction, python, classification, Confusion matrix, Machine learning.

I. INTRODUCTION

Support vector machines attempt to pass a linearly separable hyperplane through a dataset in order to classify the data into two groups. This hyperplane is a linear separator for any dimension; it could be a line (2D), plane (3D), and hyperplane (4D+)



Neural networks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer' where the answer is output as shown in the graphic below.



II. METRIC OVERVIEW

A. *Confusion Matrix*

Confusion Matrix is used to compare the Performance between the two Algorithms.

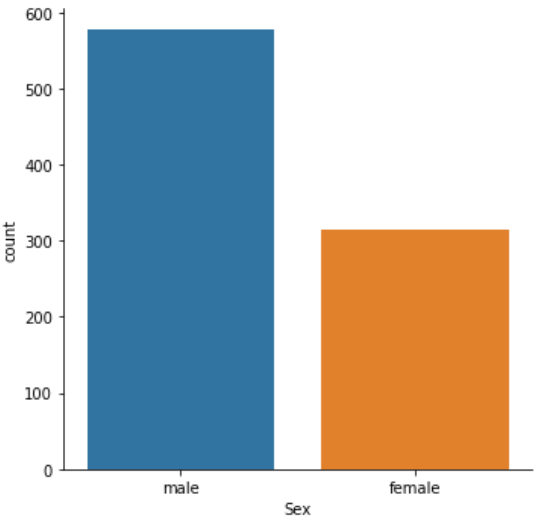
	Positive	**Negative**
Positive	True Positive	False Positive
Negative	False Negative	True Negative

III. EXPLORATORY ANALYSIS

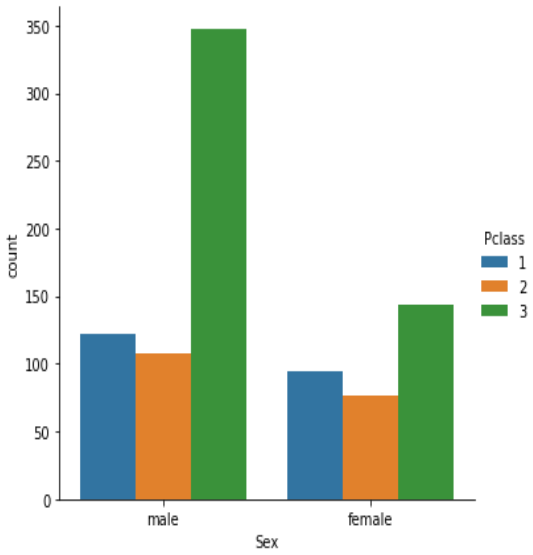
I have done some Exploratory Data Analysis on the Titanic Data Set to gain some insight on the data and get as much as information from the data in form of data visualization.

- Missing MAP: Cabin and Age has the maximum number of missing data.
- This plot shows that the potential data that I can feature engineer.
- According to this plot cabin has the maximum, number of missing data,so it is the feature that we can drop even though cabin and deck is an essential data for prediction.
- But I don't have the complete data of cabin so it will be more relevant if I drop cabin.

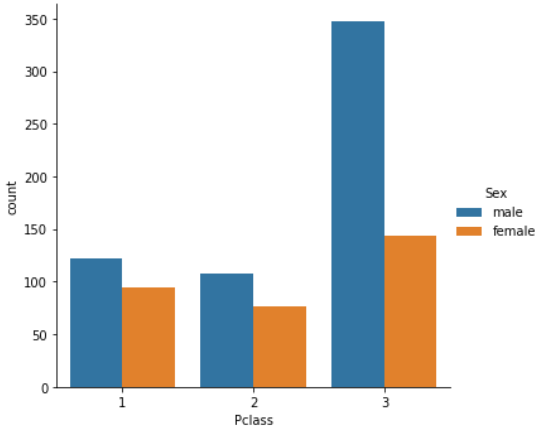
- The factor plot shows number of male and female abroad on the cruise.
- There are more male than female.
- Its probable male survival rate may be less than female



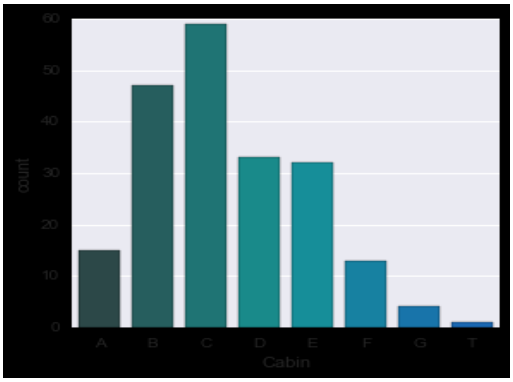
- The factor plot shows number of male and female abroad on the cruise according to Ticket class.



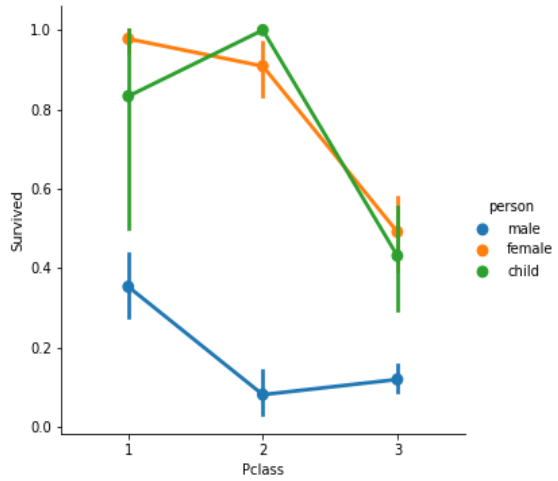
- The factor plot shows number of male and female in different Ticket class.



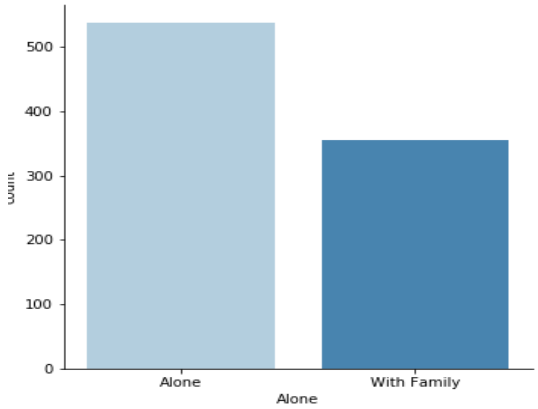
- This plot gives the relationship between Cabin and Ticket class.



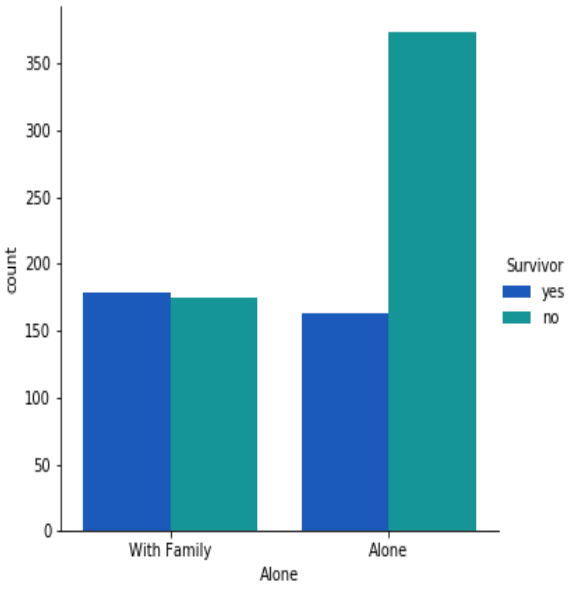
- This Plot shows that most of the survivor where children and women.



- This Plot Shows people who were alone and who were with family.

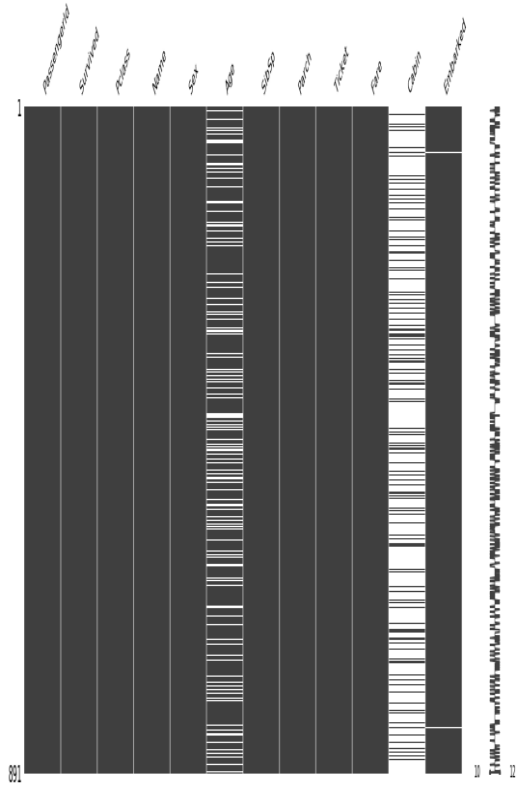


- This plot shows whether being alone was a factor in survival of the person.



IV. PREPROCESSING

I have an idea of the data I am working with. So far I have 12 variables and 1309 observations. 891 observations from the train set and 418 observations from the test set.
From the missmap it is clear that most of the missing values are from cabin and age.So those values are put nan.



- Missing Map

I have created a function that regroups the passengers into male, female and children

This is because children and females where moved out of the sinking ship first. And from the graph it is clear that the survival rate of child and female is actually more than male.

I have than created a function which separates people who have family members and who are alone. I think it is a necessary feature engineering as the survival rate of alone people are less than a family person.

I have the data preprocessing into two parts first I will take some important parameter such as age, gender, ticket class.

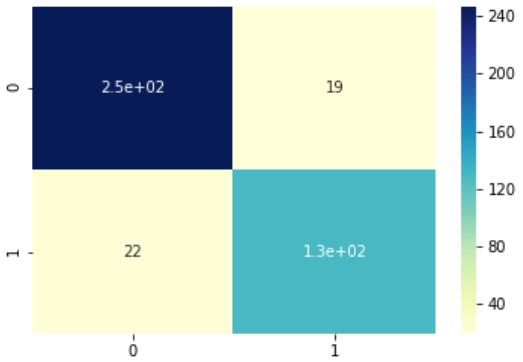
After that I took a parameter alone which specifies whether the person is with family or is, he alone.

It is important to feature engineer children. As the survival rate of female and children is more.

V. MODELS

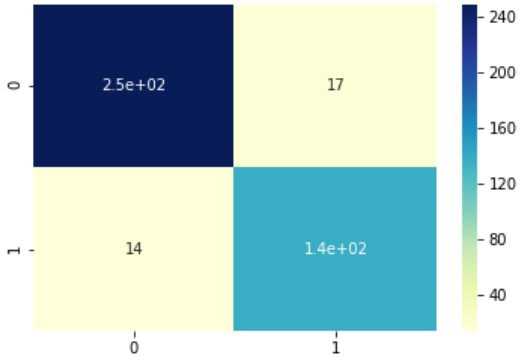
I have used support vector machine model and a 3-layer neural network and compared the performance.

1. Support Vector Machine
   We can separate the different objects with an infinite number of hyperplanes. But which hyperplane is best for the model. Well, the best hyperplane is the one that maximizes the margin. The margin is the distance between the hyperplane and a few close points. These close points are the support vectors because they control the hyperplane.
   I have used age, pclass and sex as the initial feature.



This is the confusion matrix. As it's pretty clear that SVM has predicted

- 19 False Positives
- 22 False Negatives.



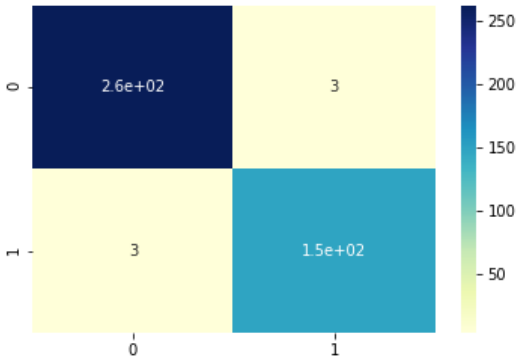
This is the confusion matrix. As it's pretty clear that SVM has predicted

- 17 False Positives
- 14 False Negatives.

.

2. 3-layer neural network
   I have used a 3-layer neural network.
   With the delta rule, as with other types of backpropagation, 'learning' is a supervised process that occurs with each cycle or 'epoch' (i.e. each time the network is presented with a new input pattern) through a forward activation flow of outputs, and the backwards error propagation of weight adjustments. More simply, when a neural network is initially presented with a pattern it makes a random 'guess' as to what it might be. It then sees how far its answer was from the actual one and makes an appropriate adjustment to its connection weights.

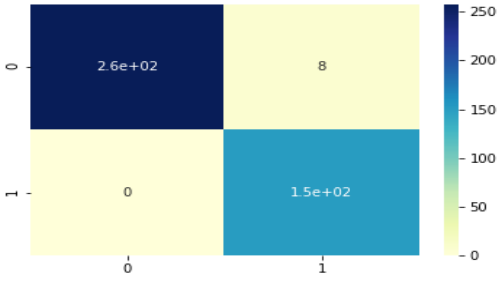
I have used keras tensor flow to build this neural network.
I have used age, pclass and sex as the initial feature.



This is the confusion matrix. As it's pretty clear that neural network has predicted

- 3 False Positives
- 3 False Negatives

I have then added alone/ family person as a featured parameter. After this I can see that the neural network has improved a lot from the previous result.



- This is the confusion matrix. As its pretty clear that neural network has predicted
  - 8 False Positives
  - 0 False Negatives

So, After adding a featured data the neural network has improved a lot.

VI. Result

According to my observation,
Accuracy of SVM classifier = 90.1 %

```
from sklearn.metrics import accuracy_score
Accuracy = accuracy_score(y_test,y_pred)
print("Accuracy",Accuracy)
```

```
Accuracy 0.9019138755980861
```

Accuracy of SVM classifier = 92.5%
After adding alone parameter its efficiency increased.

```
from sklearn.metrics import accuracy_score
Accuracy = accuracy_score(y_test,y_pred)
print("Accuracy",Accuracy)
```

```
Accuracy 0.9258373205741627
```

Accuracy of 3-layer neural network =98.56%

```
from sklearn.metrics import accuracy_score
Accuracy = accuracy_score(y_test,y_pred)
print("Accuracy",Accuracy)
```

```
Accuracy 0.9856459330143541
```

Accuracy of 3-layer neural network = 98.0%
After adding alone parameter its efficiency decreased so this suggest alone is not a good parameter to judge survival.

```
from sklearn.metrics import accuracy_score
Accuracy = accuracy_score(y_test,y_pred)
print("Accuracy",Accuracy)
```

```
Accuracy 0.9808612440191388
```

VII. Conclusion

I have performed the Exploratory Data Analysis on the Titanic Data set and Preprocessed it for the modelling. Then I successfully Modelled SVM Classifier Algorithms and compared it with a 3-layer neural network. I found out that 3-layer network outperformed SVM classifier on the titanic data set with accuracy of 98.56%
This was very much expected as I thought neural network will perform far better than support vector machine

VIII. References

[1] Kaggle, Titanic: Machine Learning Form Disaster [Online].Available: http://www.kaggle.com/

[2] https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html