Compare Support Vector Machines (SVM) to a 3 Layered Neural Network (NN) with Titanic dataset

FAREEN KHAN
Machine Learning Engineer Intern
at
AI Technology and Systems
fareenkhan4rmjozi@gmail.com
https://ai-techsystems.com/

***Abstract*—The Titanic, one of the biggest well known tragedies in history. The horrific accident, on April 15th, 1912, was a catastrophe that impacted the world. Devastating news that this colossal ship everyone thought was indestructible had sunk from a collision with an iceberg, taking 1,500 people with it. Since then, researchers and analysts have been understanding the aspects impacting individual's survival or death. This research proposes to apply two different machine learning techniques, including Support Vector Machines (SVM) and a 3 Layered Neural Network to Titanic dataset, to analyse the survival likelihood of the passengers and crew. Also, obtained accuracy score from both the machine learning techniques are compared with each other.**

Keywords*— *machine learning, support vector machine, neural network, titanic, python, classification, algorithms.

I. INTRODUCTION

The revolution of big data, accompanied by the development and deployment of devices and applications, has enabled the community to apply artificial intelligence and machine learning techniques to vast amounts of data. One of the advantages brought by the innovation is that a wide scope of information can be acquired effectively when mentioned. In this way one of the most famous datasets in information science, Titanic is utilized. White Star Line's imperial mail ship Titanic was the largest English luxury ship to be built. At the time, Titanic was highlighted to be resilient and 'unsinkable' due to the manner in which she was made; however this idea was immediately refuted on April 15th, 1912 when Titanic sunk the base of the Atlantic Sea on her first venture on the way to New York City from Southampton, Britain. The disaster caused around 1500 deaths, making the Titanic's sinking one of the most popular oceanic catastrophes ever.This dataset records different highlights of travelers on the Titanic, including who survived and who didn't. It is understood that some absent and uncorrelated highlights diminished the presentation of expectation. For a definite information examination, the impact of the highlights has been researched. Along these lines some new highlights are added to the dataset and some current highlights are expelled from the dataset. This dataset records different highlights of travelers on the Titanic, including who survive and who didn't. For a definite information examination, the impact of the highlights has been researched. Along these lines some new features are added to the dataset and some current features are expelled from the dataset.

In this research paper, we implement support vector machines and a 3 layered neural network to check whether a passenger survived or not. The performance metrics used is accuracy score for the machine learning techniques

II. METHODOLOGY

A. *Support Vector Machines (SVM)*

SVM, which was developed by Vapnik in 1995, is based on principle of structural risk minimization that exhibits good generalization performance. With SVM, finding an optimal separating hyper plane between classes by focusing on the support vectors is proposed [7]. This hyper plane separates the training data by a maximal margin. SVM solves nonlinear problems by mapping the data points into a high-dimensional space.

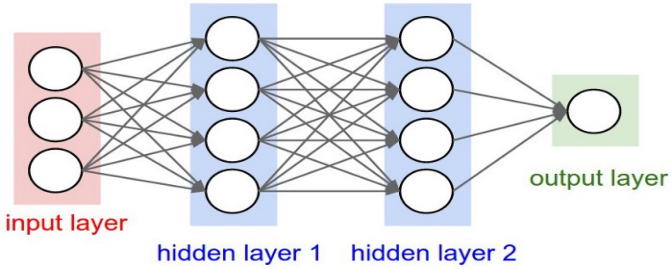
B. *Artificial Neural Networks (ANN)*

Artificial Neural Network(ANN) uses the processing of the brain as a basis to develop algorithms that can be used to model complex patterns and prediction problems. As a result, we can say that ANNs are composed of multiple nodes. That imitate biological neurons of the human brain. Although, we connect these neurons by links. Also, they interact with each other. Although, nodes are used to take input data. Further, perform simple operations on the data. As a result, these operations are passed to other neurons. Also, output at each node is called its activation or node value.
As each link is associated with weight. Also, they are capable of learning. That takes place by altering weight values.

The training process consists of the following steps:

1. **Forward Propagation:**
   Take the inputs, multiply by the weights (just use random numbers as weights)
   Let $Y = W_i I_i = W_1 I_1 + W_2 I_2 + W_3 I_3$
   Pass the result through a sigmoid formula to calculate the neuron's output. The Sigmoid function is used to normalise the result between 0 and 1:
   $1/1 + e^{-y}$

2. **Back Propagation**
   Calculate the error i.e the difference between the actual output and the expected output. Depending on the error, adjust the weights by multiplying the error with the input and again with the gradient of the Sigmoid curve:
   Weight += Error Input Output (1-Output) ,here Output (1-Output) is derivative of sigmoid curve. A 3-layers neural network (we don't count input layer):



III. DATASET

Titanic: Machine Learning from Disaster competition
dataset was provided by Kaggle. The Titanic dataset consist of a training set that includes 891 passengers and a test set that includes 418 passengers which are different from the passengers in training set. A description of the features is given in Table I. While the features such as PassengerId, Survived, Pclass, Age, SibSp, Parch and Fare are numeric values, Name, Sex and Embarked can take nominal values. For a detailed feature engineering we first analysed the features.

A. Preprocessing Steps:

In this study data cleaning, data integration, data transformation and data visualization are applied as pre-processing steps.
-The missing values of Embarked, Age and Fare features are filled by median values of these features.

-The PassengerId, Name, Ticket and Cabin (contained less than 75% of data) features are removed from the feature set.

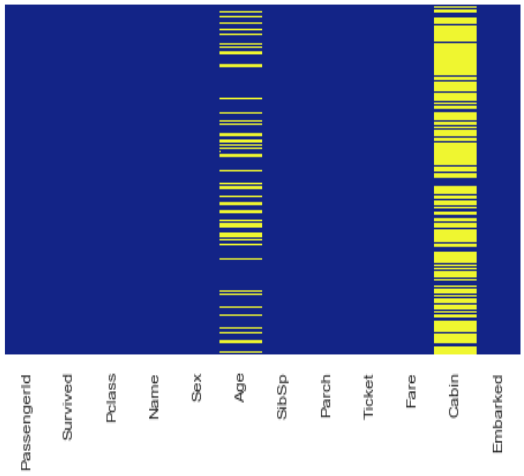
B. DATA ATTRIBUTES TABLE (Table 1)

Attribute	Description	Factors
Survival	Survival of passenger	0 = No, 1 = Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	Male/Female
Age	Age of passengers in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
Embarked	Port from where passenger embarked. C for Cherbourg, Q for Queenstown, S for Southampton	C, Q, S

C. DATA VISUALIZATION

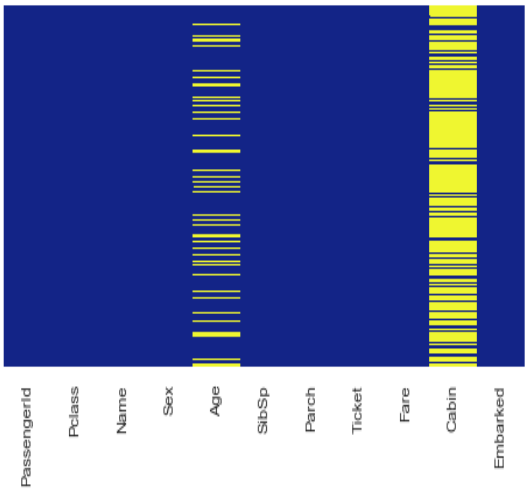
1.)Missing Values: The missing values in both the train and test datasets are checked as: Train.csv

```
Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x118ee8690>
```



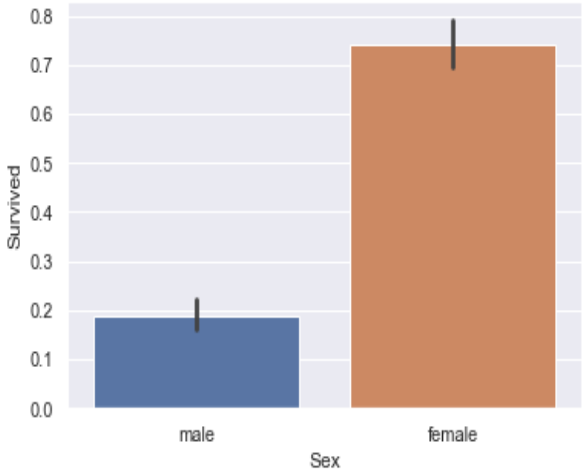
Test.csv:

```
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1af92ed0>
```



2.) Sex Feature: When we consider the distribution of the “Sex” feature, there are 314 female and 577 male passengers. 74.20% of female passengers have been rescued and others have lost their lives. On the other hand, 18.89% of male passengers have been rescued and others have lost their lives. If we analyse these distributions it is realized that the survival rate of women is higher than that of men. It has been concluded that the effect of this feature on predicting the class label is significant.

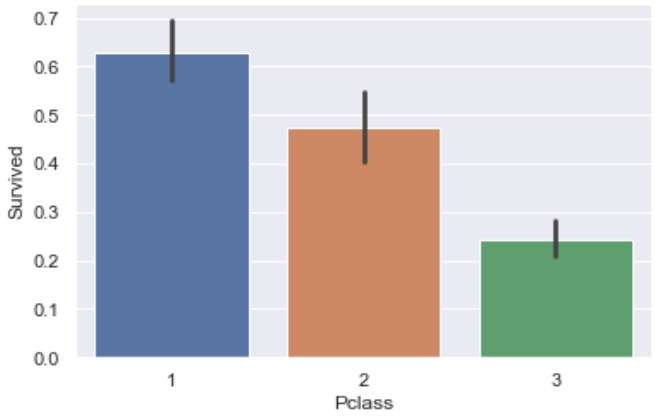
```
Percentage of females who survived: 74.2038216560509
Percentage of males who survived: 18.890814558058924
```



3.) Pclass Feature: Pclass1 has 216 passengers, Pclass2 has 184 passengers and Pclass3 has 491 passengers.

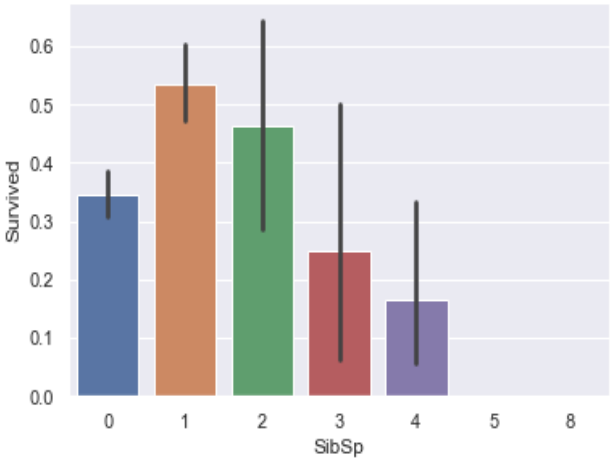
```
Percentage of Pclass = 1 who survived:
 62.96296296296296
Percentage of Pclass = 2 who survived:
 47.28260869565217
Percentage of Pclass = 3 who survived:
 24.236252545824847
```

Pclass

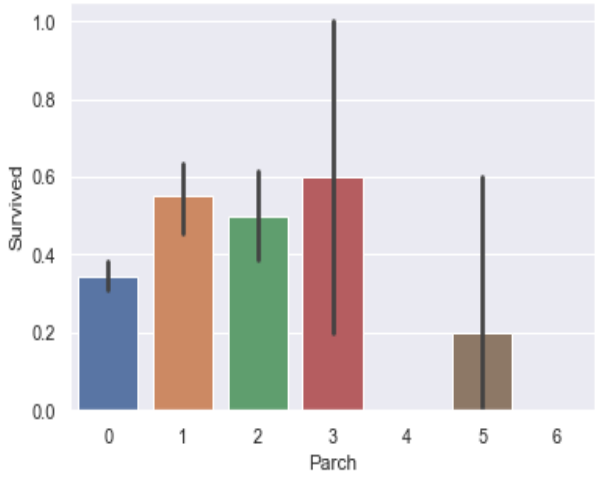


4.) SibSp Feature: stands for Sibling-Spouse.

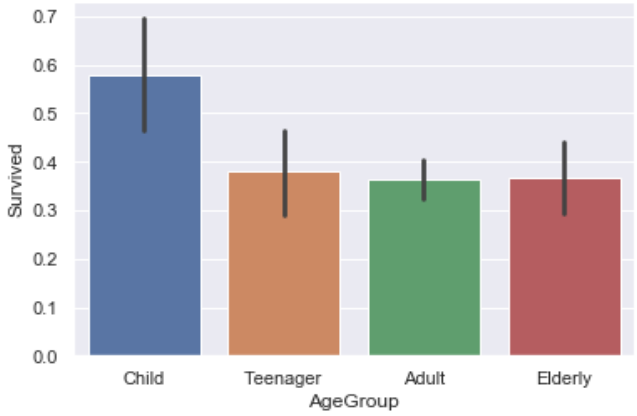
```
Percentage of SibSp = 0 who survived:
34.53947368421053
Percentage of SibSp = 1 who survived:
53.588516746411486
Percentage of SibSp = 2 who survived:
46.42857142857143
```



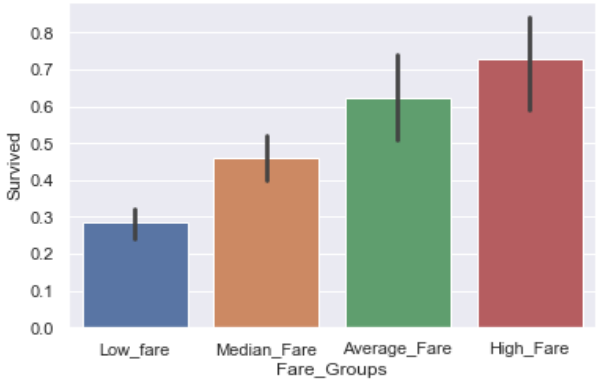
5.) Parch Feature: stands for Parents-Children



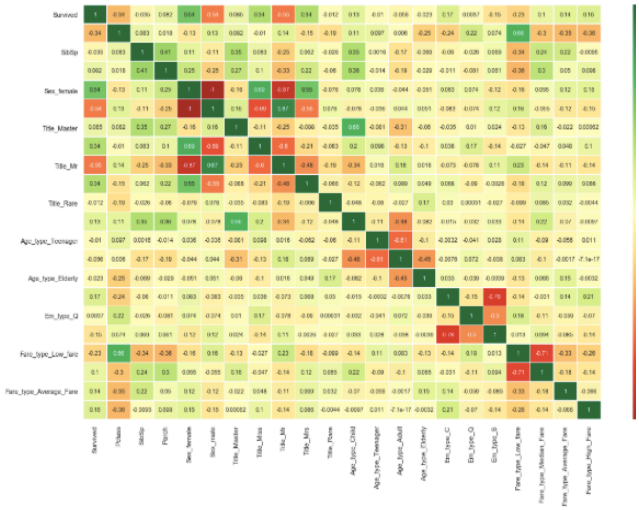
6.) Age Feature:Age of passengers ranges from 0-90 years. The age groups are 0-12, 13-19, 20-49 and 50-90 years.



7.) Fare Feature: Fare groups are 0-20, 20-60, 60-100and 100-250. The passengers who paidthe highest fare, survived the most. Thus, Fare feature is plays a significant role.



8.) Correlation between data: A vital aspect of machine learning, the correlation is a method that discovers relationships between variables. In classification, the positivity and negativity of features helps us determine the influence of independent features on intuitive forecasting



We notice that in the figure below the correlation between 'Survived' and 'Sex'(females) is high.

IV. EXPERIMENTAL RESULTS

All algorithms are run in order to analyse likelihood of survival and learn what features have a correlation towards survival of passengers and crew. When applying algorithms to titanic dataset, we have seen that to make the algorithm accurate, some more adjustments on some model parameters are required.

1.) Support Vector Machine:

Support Vector Machine (SVM) algorithm is first applied on the processed dataset. The model is trained on the X_train, y_train datasets. The SVM model gives an accuracy of 83.5% on the training dataset. This trained model is used for predicting 'Survived' values, i.e, 0 or 1. The obtained values are then stored in the submission_SVM.csv file.

2.) Artificial Neural Network:

After this, a 3 layered neural network is applied n the training dataset using keras. Folloowing are the points to be considered:

Model=Sequential()
Layers=3(Input+2Hidden+Output)

InputDimensions=21
ActivationFunctions:Relu,Sigmoid.
KernelInitializer=Uniform
Optimizer=Adam

Loss=BinaryCrossentropy
BatchSize=32
Epochs=200

Upon training the model on the training dataset, the accuracy achieved is 84.28%. This trained model is used for predicting 'Survived' values, i.e, 0 or 1. The obtained values are then stored in the submission_ANN.csv file.

	Titanic Dataset		
	ML Algorithms Used	***SVM***	***ANN***
	Metrics Used: Accuracy	83.5%	84.28%

Conclusions

Obtaining valuable results from the raw and missing data by using machine learning and feature engineering methods is very important for knowledge-based world. In this paper, we have proposed models for predicting whether a person survived the Titanic disaster or not. First, a detailed data analysis is conducted to investigate features that have correlation or are non-informative. Some of them are excluded such as name, ticket and cabin. Secondly, in classification step 2 different machine learning algorithms are used for classifying the dataset formed in pre-processing step.

The proposed model can predict the survival of passengers and crew with 84.28% accuracy score: ANN. As conclusion, this paper presents a comparative study on machine learning techniques to analyse Titanic dataset to learn what features effect the classification results and which techniques are robust.

References

[1] Kaggle, Titanic: Machine Learning from Disaster. Online: https://www.kaggle.com/c/titanic/

[2] Towards Data Science: Fundamental Techniques of Feature Engineering with Machine Learning. Online: https://towardsdatascience.com

[3] Towards Data Science: understanding Neural Networks. Online: https://towardsdatascience.com

[4] Medium: Exploratory Data Analysis of Titanic Dataset with Python. Online: https://medium.com