**A PRELIMENERY REPORT ON MACHINE LEARNING FROM DISASTER ANALYSIS**

**MACHINE LEARNING FROM DISASTER ANALYSIS**

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY , PUNE IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS

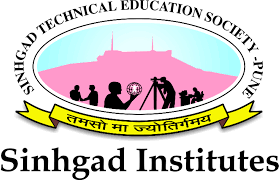
OF

## THIRD YEAR COMPUTER ENGINEERING

## SUBMITTED BY

**CHAITANAYA WAGH** **ROLL NO:139**

**DHANASHRI BELOTE ROLL NO:105**



## DEPARTMENT OF COMPUTER ENGINEERING

## STES’S

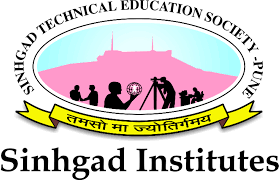
**SMT. KASHIBAINAVALECOLLEGE OF ENGINEERING**

**VADGAON BK, OFF SINHGAD ROAD, PUNE 411041**

## SAVITRIBAI PHULE PUNE UNIVERSITY

## 2019 -2020

## [P-2]



**CERTIFICATE**

This Is To Certify That The Project Report Entitles

**“ MACHINE LEARNING FROM DISASTER ANALYSIS”**

Submitted by

**CHAITANAYA WAGH** **ROLL NO:139**

**DHANASHRI BELOTE ROLL NO:105**

This is a bonafide work carried out by him/her under the supervision of **Prof. Shailesh Patil and Prof. D. H. Kulkarni** and it is approved for the partial fulfillment of the requirement of third year computer engineering.

(12, Sentence case)

## (Prof. Shailesh Patil)

## ( **Prof. D. H. Kulkarni** ) (Dr. P. N. Mahalle)

Guide Head,

Department of Computer Engineering Department of Computer Engineering

## (Dr. A. V. Deshpande)

Principal, Smt.Kashibai Navale College of Engineering Pune – 41

Place : Pune Date : 5/10/2019

## [P3]

**CONTENTS**

**LIST OF ABBREVATIONS I**

**LIST OF FIGURES II**

**LIST OF TABLES III**

|  |  |  |
| --- | --- | --- |
| **Sr.No.** | **Chapters** | **Page No.** |
| 01 | **Abstract** |  |
| 02 | **Acknowledge** |  |
| 03 | **Problem Definition** |  |
| 04 | **Contents** |  |
| 4.1 | Introduction |  |
| 4.2 | Data Set |  |
| 4.3 | Data Analysis |  |
| 4.4 | Approach/Method  4.4.1 SVM  4.4..2 Decision Tree |  |
| 05 | **Results and Discussion** |  |
|  | * 1. Result   2. Conclusion |  |

**ABSTRACT**

**“Machine learning from disaster analysis”**

**Titanic disaster occurred 107 years ago on April 15, 1912, killing about 1500 passengers and crew members. The fateful incident still compel the researchers and analysts to understand what can have led to the survival of some passengers and demise of the others. With the use of machine learning methods and a dataset consisting of 891 rows in the train set and 418 rows in the test set, the research attempts to determine the correlation between factors such as age, sex, passenger class, fare etc. to the chance of survival of the passengers. These factors may or may not have impacted the survival rates of the passengers. In this project, machine learning algorithm namely Support Vector Machine has been implemented to predict the survival of passengers. In particular, this project is based on the basis to calculate higher percentage of accuracy on a test dataset.**

# ACKNOWLEDGEMENT

**We would like to express our sincere gratitude to our supervisor prof. Shailesh Patil & prof. for providing their valuable guidance, comments, and suggestions throughout the course of project. We would specially thank our HOD DR. Parikshit Mahalle sir and principal A. V. Deshpande sir for constantly motivating us. Lastly we would like to thank all the fellow classmates who helped us during difficulties in project.**

# NAME OF THE STUDENTS

**CHAITANAYA WAGH**

**DHANASHRI BELOTE**

# Problem Definition

We are given information about a subset of the Titanic population and asked to build a predictive model that tells us whether or not a given passenger survived the shipwreck. We are given 10 basic explanatory variables, including passenger gender, age, and price of fare, among others. More details about the competition can be found on the Kaggle site, [here](https://www.kaggle.com/c/titanic). This is a classic binary classification problem, and we will be implementing a random forest classifer.

# INTRODUCTION

# Using data provided by www.kaggle.com, our goal is to apply machine-learning techniques to successfully predict which passengers survived the sinking of the Titanic.

# Features like ticket price, age, sex, and class will be used to make the predictions. We take several approaches to this problem in order to compare and contrast the different machine learning techniques. By looking at the results of each technique we can make some insights about the problem. The methods used in the project include Naïve Bayes, SVM, and decision tree. Using these methods, we try to predict the survival of passengers using different combinations of features. The challenge boils down to a classification problem given a set of features. One way to make predictions would be to use SVM [2] on our data to see if we can achieve better results. Lastly we use decision tree analysis [3] and find the optimal decision boundaries.

# DATA SET

# The data we used for our project was provided on the Kaggle website. We were given 891 passenger samples for our training set and their associated labels of whether or not the passenger survived. For each passenger, we were given his/her passenger class, name, sex, age, number of siblings/spouses aboard, number of parents/children aboard, ticket number, fare, cabin embarked, and port of embarkation. For the test data, we had 418 samples in the same format. The dataset is not complete, meaning that for several samples, one or many of fields were not available and marked empty (especially in the latter fields – age, fare, cabin, and port). However, all sample points contained at least information about gender and passenger class. To normalize the data, we replace missing values with the mean of the remaining data set or other values.

# 

# DATA ANALYSIS

# In order to prepare our data for training in our SVM, we do not need to bin values together. Instead we simply turn all the values into numerical values. In order to do this, we’ll interpret the bit representation of strings and characters as float represented numbers.

# 

# In figure 1. of the previous page, we see the breakdown of the data to get a better sense of what features might be good indicators of our classification problem. First, we notice that out of all the passengers in the test data, 36.38% survived. If we breakdown the group into sex, we see that a significant difference in survival between females (74.20%) and males (18.89%). This is a strong indicator that sex would probably be a good feature to use. Continuing our analysis, we see that of the females in first class and second class (first class can be thought of as upper class, second class as middle class, and third class as lower class), more than 90% survived. Third class fared much worse with a 50% survival rate. Of the males, first class had a much higher survival rate (36.89%) than second (15.74%) or third (13.54%). Interestingly, there was not significant variation in survival given a person’s age in any subgroup except for youths in second class.

# APPROACH/METHOD

# Basic Naïve Bayes classification in [1] is used as a baseline to see what is achievable. More sophisticated techniques like SVM in [2] and decision tree analysis is used [3] to see if improvements can be made in the classification test. We experimented with using different feature sets of each method and found the optimal feature combination on the test group.

# SVM

# To improve our classification, we used support vector machines [2]. We considered the following features: 1) passenger class, 2) sex, 3) age, 4) number of siblings, 5) patriarchal status, 6) fare, and 7) place of embarkation. We used a Gaussian radial basis function as our kernel and set the tolerance to 𝜀 = .001.

# 

# Unlike Naïve Bayes, no extra data cleaning was needed. Iterating through all possible feature combinations, we were able to achieve an accuracy rate of 77.99% on the test data set using only three features. The three features that achieve this rate were class, sex and place of embarkation. Using age, fare, and place of embarkation resulted in the worst accuracy of 58.13%. It is interesting to note that this accuracy would be less if we had just guessed that all test points died (accuracy of 63.23%). This suggests that perhaps class and sex are strong indicators of survival whereas age and fare are weaker indicators of survival. In figure 2, we see the SVM learning curve using the features class, sex and place of embarkation. At around 400 samples, the training curve has reached its asymptotic value of 77.99% and any additional sample does not improve the accuracy.

# 

# Decision Tree

# We built our decision using the following features – gender, passenger class, age, and fare. We first split the data into males and females because it was most correlated with the chance of survival. From just using a single feature, we achieved an accuracy of 76.79%, which is the same percentage as Naïve Bayes with just the gender feature. This is expected since with just the gender feature, both classifiers are labeling test samples the same way (which is marking all females as survived and all males as died). Then we split both males and females into passenger classes. Even after splitting the data into passenger class, males in each class are more likely to die, and passengers in each class, other than class 3, are more likely to survive. If we choose the hard decision that female passengers in class 3 all survive, it will still produce an accuracy of 76.79% because the classifier hasn’t changed from the earlier process of labeling all males as died and females as survived. However, if we choose the hard decision that all females in class 3 will die, our accuracy improves to 77.27% on the test data. Next, we look at the feature age. Since the domain of age is continuous, we have to find a good decision boundary to split our data. After plotting the age and survival of passengers in each gender and passenger class, we decided to use a binary decision because in most cases, older passengers were more likely to die than younger ones. Instead of using the same age boundary for each gender and passenger class, we considered each gender and passenger class, case by case and found different boundary thresholds for each. To find our boundary threshold, we tried to minimize the classification error on our training set. This means that we chose the age boundary for each gender and passenger class such that if we classify all samples below the age boundary as survived and all above as died, we minimize the classification error on the training set. After including age in the decision tree, we achieve a classification error of 78.94%.

# 

# Results

# Of the two methods, SVM performed the worst and decision tree performed the best. However, the best and worst performance only differs by 2.64% so all the methods have roughly the same performance on our data set. This is probably because there was one feature that was strongly correlated with whether a passenger survives. The decision tree does not make this assumption. Even though the decision tree considers correlation between features, it only performs marginally better than SVM. So this shows that assuming that features are independent is not necessarily a bad assumption for our problem. Table 3 offers a summary of the achievable accuracy using SVM and decision tree analysis. Even though we were given many features of passengers in our data, we found that most of the features were not useful in classification. For example, the number of sibling/spouses and the number of parents/children did not help with classification in any of the three models. Knowing the number of relatives aboard did not help with classification, but perhaps, if we were given the links between passengers then we’d be able to infer more about the survival rate. Since family units tend to all die or all survive, knowing the family links would have been useful.

# Table 3. Comparison of Performance

# SVM 84.74 % Accuracy

# Decision Tree 87.88 % Accuracy

# 

# Conclusion

# There were not significant differences in accuracy between the three methods we experimented with. Even using every combination of features, we were still not able to produce an accuracy rate that was much different than simple SVM classifier using only sex as a feature. It appears that the other features were only weakly indicative of survival, as sex seemed to dominate the others in terms of being able to accurately predict survival. Even with more sophisticated algorithms, we were not able to achieve much improvement. This shows the importance of choosing important features and obtaining good data. It would be interesting to continue this analysis with other possible features or with other machine learning algorithms like random forests or KNN.

# REFERENCES

# [1] A. Ng. CS229 Notes. Stanford University, 2012

# [2] Cortes, Corinna; and Vapnik, Vladimir N.; "Support-Vector

# Networks", Machine Learning, 20, 1995.