A Comparative Study on Machine Learning Techniques using Titanic Dataset

Ekin Ekinci*, Sevinç İlhan Omurca*, Neytullah Acun*

*Kocaeli University, Faculty of Engineering, Computer Engineering Department
Umuttepe Campus, Kocaeli, Turkey
{ekin.ekinci, silhan}@.edu.tr, neytullah.acun@gmail.com

***Abstract*— The Titanic disaster resulting in the sinking of the British passenger ship with the loss of 722 passengers and crew occurred in the North Atlantic on April 15, 1912. Although it has been many years since this maritime disaster took place, research on understanding what impacts individual's survival or death has been attracting researchers' attention. In this study, we propose to apply fourteen different machine learning techniques, including Logistic Regression (LR), k-Nearest Neighbors (kNN), Naïve Bayes (NB), Support Vector Machines, Decision Tree, Bagging, AdaBoost, Extra Trees, Random Forest (RF), Gradient Boosting (GB), Calibrated GB, Artificial Neural Networks (ANN), Voting (GB, ANN, kNN) and, Voting (GB, RF, NB, LR, kNN) to Titanic dataset, which is publicly available, to analyze likelihood of survival and learn what features have a correlation towards survival of passengers and crew. Also, obtained F-measure score from machine learning techniques are compared with each other and the F-measure score which is obtained Kaggle. As a result of this study, more successful F-measure rates have been obtained with GB and Voting than Kaggle.**

***Keywords*— Machine learning, classification, data analysis, Titanic, Kaggle**

I. INTRODUCTION

The inevitable development of technology has both facilitated our life and brought some difficulties with it. One of the benefits brought by the technology is that a wide range of data can be obtained easily when requested. However, it is not always possible to acquire the right information. Raw data that is easily accessed from the internet sources alone does not make sense and it should be processed to serve an information retrieval system. In this regard, feature engineering methods and machine learning algorithms are plays an important role in this process.

The aim of this study is to get as reliable results as possible from the raw and missing data by using machine learning and feature engineering methods. Therefore one of the most popular datasets in data science, Titanic is used. This dataset records various features of passengers on the Titanic, including who survived and who didn't. It is realized that some missing and uncorrelated features decreased the performance of prediction. For a detailed data analysis, the effect of the features has been investigated. Thus some new features are added to the dataset and some existing features are removed from the dataset.

Chatterjee [1] applied multiple logistic regression and logistic regression to check whether a passenger is survived. He reported performance metrics across different cases comparison and concluded that, the maximum accuracy obtained from Multiple Linear Regression is 78.426%; the maximum accuracy obtained from Logistic Regression is 80.756%.

Datla [2] compared the results of Decision tree and Random Forests algorithms for Titanic dataset. Decision tree is resulted 0.84% correctly classified instances, while Random Forests resulted 0.81%. As the feature engineering steps, they created new variables such as "survived", "child", "new_fare", "title", "Familysize", "FamilyIdentity" which are not included in feature list of Titanic dataset and also replaced a missing value by the mean value of a given feature.

There are several studies in the literature that compared different classification algorithms on multiple dataset. Meyer et al., [3] compared SVM implementation to 16 classification algorithms and for titanic dataset they achieved %20.81 and %21.27 error rates with neural networks and SVM respectively as minimum errors. Ratsch et al. [4] compared Adaboost classifiers to SVM and RBF classifiers. For titanic dataset, %22.4 error rate is obtained from SVM as the minimum error rate. Li et al. [5] used SVM as a component classifier for Adaboost. They used titanic dataset as one of the experimental data and the minimum error rate they obtained is %21.8.

The rest of the paper is organized as follows: Section 2 presents the techniques employed in experimental studies. Experimental setup and results are given in sections 3. Section 4 concludes the paper with a discussion.

II. METHODOLOGY

A. Logistic Regression

LR is one of the most popular methods used to classify binary data. LR is based on the assumption that the value of dependent variable is predicted by using independent variables. In the model, Y is the dependent variable we are trying to predict by observing X which is the input or set of the independent variables $(x_i, \dots, x_k)$. The value of Y that corresponds to the people as either survived (Y=1) or not survived (Y=-1) and is summarized by (X=x). From this definition, the conditional probability follows a logistic distribution given by $P(Y = 1|X = x_i)$. This function called as regression function we need to predict Y.

B. K Nearest Neighbors

kNN is one of the most common, simplest and non-parametric classification algorithms when there is little or no prior knowledge about the distribution of the data. Using the distance metrics to measure the closeness between training samples and the test sample, kNN assigns the test sample with class of its k nearest training samples. In terms of closeness, the kNN is mostly based on the Euclidean distance. The Euclidean distance between training sample $X_1 = (x_{11}, x_{12}, \dots, x_{1N})$ with N features, test sample $X_2 = (x_{21}, x_{22}, \dots, x_{2N})$ with N features and $m = 2$ is

$$distance(X_1, X_2) = (\sum_{i=1}^{N}(x_{1i} - x_{2i})^m)^{1/m}. \quad (1)$$

When m $= 1$ the distance is called as Manhattan and $m >$ 2 the distance called as Minkowski.

C. Naïve Bayes

NB, which is known as effective inductive learning algorithm, achieves efficient and fast classification in machine learning applications. The algorithm is based on Bayes theorem assuming all features are independent given the value of the class variable [6]. This is conditional independence assumption and true in real world applications. Due to this assumption NB performs well on high dimensional and complex datasets.

D. Support Vector Machines

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 hyperplane between classes by focusing on the support vectors is proposed [7]. This hyperplane separates the training data by a maximal margin. SVM solves nonlinear problems by mapping the data points into a high-dimensional space.

E. Decision Tree

Decision trees with their fairly simple structure to create are one of the most used classifiers. A decision tree is a tree structured model with decision nodes and prediction nodes. Decision nodes are used to branch and prediction nodes specify class labels. C4.5 is a kind of decision tree algorithm builds a decision tree from training data by using the information gain. When building decision trees C4.5 uses divide and conquer approach.

F. Bagging

Bagging is one of the oldest and easiest techniques for creating an ensemble of classifiers, improves accuracy by resampling of the training set [8]. The fundamental assumption behind the bagging is to use multiple training sets instead of using a single one to prevent results which depend on a training set. A base single classifier is applied in parallel to generated training sets then generated classification models are combined according to majority voting. With bagging, high accuracy, good generalization performance and reduce in variance and bias are achieved.

G. AdaBoost

AdaBoost is one of the most used and effective ensemble learning methods. The base notion behind AdaBoost is that a strong classifier can be created by linearly combining a number of weak classifiers [9]. In the training process AdaBoost increases the weights of misclassified data points while is decreasing weights of correctly classified data points. That is, AdaBoost reweights all training data in its every iteration. Weak classifiers are applied in serially then generated classification models are combined according to weighted majority voting.

H. Extra Trees

Extra tree (The Extremely Randomized Decision Tree) is a decision tree ensemble classification method. Extra Trees are based on the randomization. For each node of the tree splitting rules are randomly drawn then the best performing rule based on a score is associated with that node [10]. For each tree that composed extra trees whole dataset is used for training.

İ. Random Forest

RF is a classification algorithm developed by Breiman and Cutler that uses an ensemble of tree predictors [11]. It is one of the most accurate learning algorithms and for many datasets; it achieves a highly accurate classifier. In RF, each tree is constructed by bootstrapping the training data and for each split randomly selected subset of features are used [12]. Splitting is made based on purity measure. This classification method estimates missing data and large proportion of the data are missing it still maintains accuracy.

J. Gradient Boosting

GB was developed by Friedman (2001) is a powerful machine learning algorithm that has shown considerable success in a wide range of real world applications. GB handles boosting as a method for function estimation, in terms of numerical optimization in function space [13].

K. Artificial Neural Networks

Multilayer perceptron (MLP) is a kind of ANN has ability to solve nonlinear classification problems with high accuracy and good generalization performance. The MLP has been applied to a wide variety of tasks such as feature selection, pattern recognition, optimization and so on. A MLP can be considered as a directed graph in which artificial neurons are presented with nodes and directed and weighted edges connects nodes to each other [14]. Nodes are organized into layers: an input layer, one or more hidden layers and an output layer. MLP uses backpropagation to classify data points and by using backpropagation error is propagated in backward direction to adjust weights.

L. Voting

For obtaining accurate classification results, a bunch of classifiers are assembled for artificial and real-world datasets. Voting is the simplest method which combines predictions from multiple classifiers and made a single contribution. [15]. While majority voting is resulting with class with the most votes, weighted voting makes a weighted linear combination of classifiers and decides class with the highest aggregate.

III. Experiments

A. Dataset

Titanic: Machine Learning from Disaster competition dataset [16] 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.

TABLE I
NUMBER OF FEATURES IN THE DATASET

Feature	Value of Feature	Feature Characteristic
PassengerId	1-891	Integer
Survived	0,1	Integer
Pclass	1-3	Integer
Name	Name of passengers	Object
Sex	Male, female	Object
Age	0-80	Real
SibSp	0-8	Integer
Parch	0-6	Integer
Ticket	Ticket number	Object
Fare	0-512	Real
Cabin	Cabin number	Object
Embarked	S, C, Q	Object

While the features such as PassengerId, Survived, Pclass, Age, SibSp, Parch and Fare are numeric values, Name, Sex and Embarked can take nominal values; the features such as Ticket, Cabin can take numeric and nominal values.

For a detailed feature engineering we first analyzed the features.

1) Sex: When we consider the distribution of the "Sex" feature, there are 314 female and 577 male passengers. 233 of female passengers have been rescued and others have lost their lives. On the other hand, 109 of male passengers have been rescued and others have lost their lives. If we analyze 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.

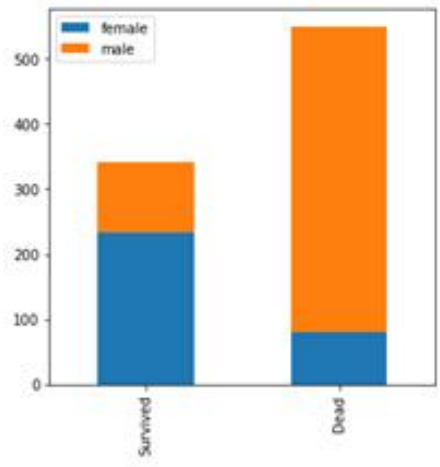

Fig. 1 Distribution of sex feature.

2) Embarked: When we consider the distribution of the "Embarked" feature, there are 644, 168, 77 passengers boarding from the port "S", "C" and "Q" on the ship respectively. The survival rates of passengers boarding from these ports are given in Fig. 2. When this figure is analyzed, C is the port with the highest survival rate of 55%. Thus, this can be interpreted like "embarked" feature gives important clues about survival.

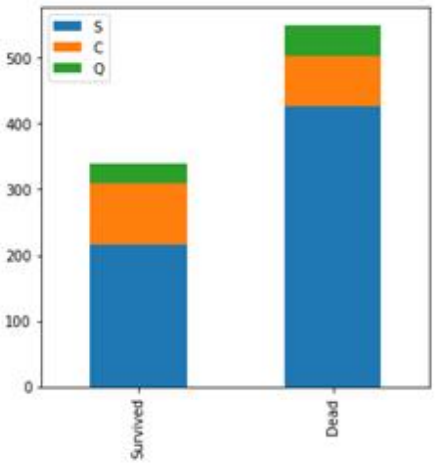

Fig. 2 Distribution of Embarked feature.

3) Pclass: "Pclass" feature describes three different classes of passengers. There are 216 passengers belong to the class 1, 184 passengers in class2 and finally 491 passengers in class 3. The survival rates of passengers due to "Pclass" feature are given in Fig. 3. The passengers with the highest survival rates are the first class passengers with 63%. This ratio also shows that wealthy people are alive.

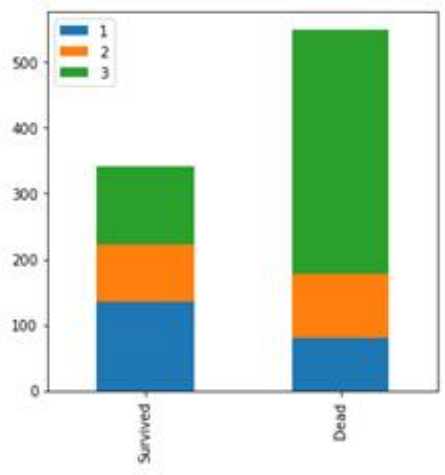

Fig. 3 Distribution of Pclass feature.

4) Age: When the "age" feature" is considered it is seen that, the age of passengers are range from 0 to 80. If we group the passengers by specific age ranges such as 0-13, 14-60 and 61-80 then we realized that most of the passengers in the 0-13 age group are survived and a large majority of passengers in the age group 61-80 lost their lives. This statistical information proves that the first children were rescued when the ship started to sink.

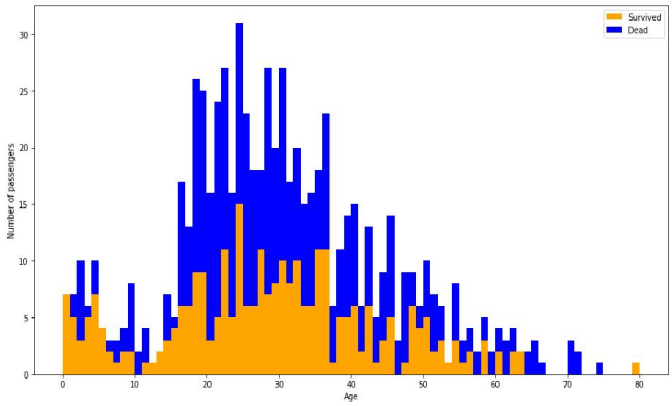

Fig. 4 Distribution of Age feature.

5) Fare: The "fare" feature specifies the fare paid by the passenger and it changes between 0-512. If we distinguish this feature with two groups as 0-90 and 91-512, then it is seen that most of the passengers who paid between 0 and 90 lost their lives and the majority of the passengers who paid between 91and 512 survived.

6) Correlation between data: In classification task of machine learning, the correlation which is often used as a preliminary technique to discover relationships between variables can be a key to improve the accuracy of a prediction model. In classification models, the positive or negative correlation between feature values can be used to discover which ways the independent features influence intuitive forecasting. The correlation between features of Titanic dataset is shown in Fig. 5. Due to the embarked and sex features have nominal values they are not included in this figure. When the correlation scores are evaluated it is observed that the correlation between "Survived" and "Sex" is highest while the correlation between "Survived" and "Age" is the minimum. Apart from that, "Sibsp" and "Parch" features are correlated by 0.41. Accordingly by combining these two features a new feature can be created.

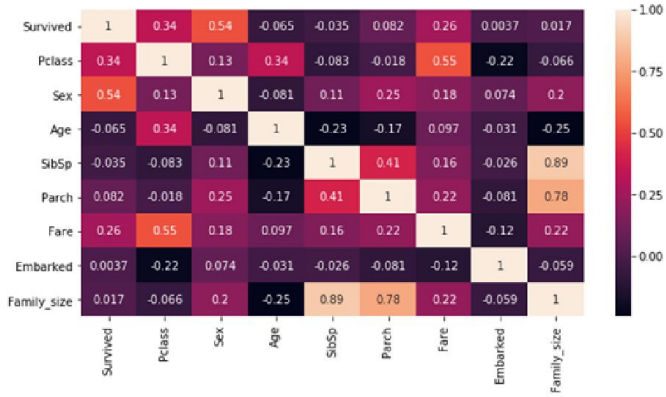

Fig. 5 Correlation between data.

7) Family size: In machine learning applications, features extension methods as well as feature reduction methods can also improve the classification performance. In this study, a feature named "Family_size" is created in addition to the existing features. This feature is calculated by adding the value of Sibsp feature to the value of Parch feature. After that, we have distinguished this feature with two groups. In the first group consist of passengers whose family_size is 0, 4, 5, 6, 7 or 10 and in the second group there are passengers whose family_size is 1, 2 or 3. It is observed that most of the first group lost their lives and the majority of the second group is survived. These results show that the number of family members strengthens the possibility of survival.

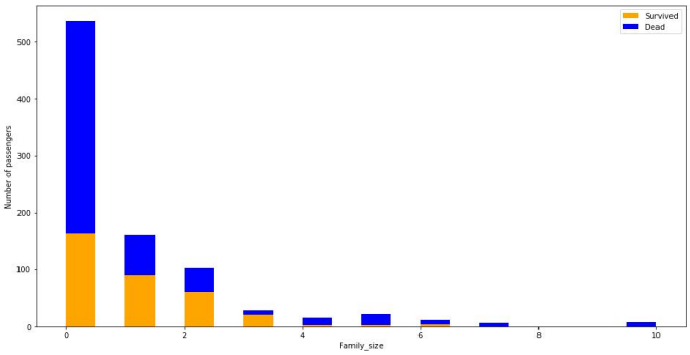

Fig. 6 Distribution of Family_size feature

B. Preprocessing Steps

In this study data cleaning, data integration, data transformation are applied as preprocessing steps. The missing values of Age and Fare features are filled by meadian values of these features. The missing values of "Embarked" feature are filled by "C" value. The PassengerId, Name, Ticket and Cabin features are removed from the feature set.

C. Experimental Results

All algorithms are run in order to analyze 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.

Logistic regression is applied with a penalty term which is decided as "l2". For kNN, number of neighbors is selected as 8 and Minkowski is selected as distance measure. Naïve Bayes algorithm is used based on Bernoulli distribution. In SVM, the penalty parameter is important to control level of misclassification and determined as 3. Gini is used as impurity measure in C4.5 decision tree. Apart from that maximum depth is a parameter that makes the search space finite and also prevents decision tree from growing to an extremely large size and decided as 25. In Bagging, maximum bag size is determined as 2.6%. Decision tree is used as the base estimator in bagging and adaboost. Number of trees in the forest is selected 15 and 200 for Extra Trees and Random Forest classifiers respectively. In Gradient Boosting, Logistic Regression is used as loss function to be optimized. Gradient Boosting is calibrated with sigmoid function in Calibrated Gradient Boosting. MLP with backpropogation is used as Artificial Neural Networks. In the first voting algorithm, GB, ANN and kNN are voted, in the second GB, RF, NB, LR and KNN are voted.

Algorithms are evaluated according to accuracy and F-measure. We compare our F-measure scores with F-measure scores obtained from Kaggle. The performances of the algorithms are listed in Table II. It is observed that the best performance is provided with Voting (GB, ANN, kNN) with F-measure score of 0.82. Compared with Kaggle, more successful F-measure rates have been obtained with GB and Voting. With Calibration we expect to see better results but our calibration doesn't yield increase in F-measure score while Kaggle yields.

TABLE II
COMPARISON OF ACCURACY, F-MEASURE AND KAGGLE SCORES OF ALGORITHMS

Algorithm	Accuracy	F-measure	Kaggle
Voting (GB, ANN, kNN)	0.869	0.82	0.794
Gradient Boosting	0.869	0.815	0.789
Calibrated (GB)	0.866	0.81	0.813
Voting (GB, RF,NB,LR kNN)	0.851	0.79	0.789
Random Forest	0.848	0.781	0.789
Artificial Neural Networks	0.813	0.743	0.766
AdaBoost	0.814	0.741	0.78
Decission Tree	0.817	0.738	0.789
Bagging	0.806	0.731	0.775
Logistic Regression	0.802	0.728	0.766
Naive Bayes	0.789	0.714	0.762
Extra Trees	0.815	0.713	0.785
k Nearest Neighbors	0.802	0.712	0.665
Support Vector Machines	0.787	0.71	0.766

IV. 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. And as a preprocessing step some new features are added to dataset such as family_size and some of them are excluded such as name, ticket and cabin. Secondly, in classification step 14 different machine learning algorithms are used for classifying the dataset formed in preprocessing step.

The proposed model can predict the survival of passengers and crew with 0.82 F-measure score with Voting (GB, ANN, kNN).

As a conclusion, this paper presents a comparative study on machine learning techniques to analyze Titanic dataset to learn what features effect the classification results and which techniques are robust.

REFERENCES

[1] T. Chatterjee, "Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms," *International Journal of Emerging Research in Management &Technology,* vol. 6, pp. 1-5, June 2017.

[2] M. V. Datla, "Bench Marking of Classification Algorithms: Decision Trees and Random Forests – A Case Study using R," *in Proc. I-TACT-15*, 2015, pp. 1-7.

[3] D. Meyer, F. Leisch and K. Hornik, "The support vector machine under test," *Neurocomputing,* vol. 55, pp. 169-186, Sept. 2003.

[4] G. Rätsch, T. Onoda, and K.-R. Müller, "Soft Margins for AdaBoost," *Machine Learning,*vol. 42, pp. 287-320, Mar. 2001.

[5] X. Li, L. Wang, and E. Sung, "AdaBoost with SVM-based component classifiers," *Engineering Applications of Artificial Intelligence*, vol. 21, pp. 785-795, Aug. 2008.

[6] S. İlhan Omurca and E. Ekinci, "An alternative evaluation of post traumatic stress disorder with machine learning methods," *in Proc. INISTA 2015*, 2015, pp. 1-7.

[7] C. Cortes, and V. Vapnik, "Suppoprt-Vector Networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.

[8] G. Liang, X. Zhu, and C. Zhang, "An empirical stıdy of bagging predictors for different learning algorithms," *in Proc. AAAI'11*, 2011, pp. 1802-1803.

[9] Y. Ma, X. Ding, Z. Wang and N. Wang, "Robust prcise eye loation under probabilistic framework," *in Proc. Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 339-344.

[10] C. Desir, C. Petitjean, L. Heutte, M. Salaun and L. Thiberville, "Classification of Endomicroscopic Images of the Lung Based on Random Subwindows and Extra-Trees," *IEEE Transactions on Biomedical Engineering*, vol. 59, pp. 2677-2683, Sep. 2012.

[11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp.5-32, 2001.

[12] R. Diaz-Uriarte and S. Alvadez de Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, p. 3, Jan. 2006.

[13] S. B. Taieb and R. J. Hyndman, "A gradient boosting approach to the Kaggle load forecasting competition," *International Journal of Forecasting,* vol. 30, pp. 382-394, Apr. 2014.

[14] I. Maglogiannis, K. Karpouzis, B. A. Wallace, and J. Soldatos, Eds., *Supervised Machine Learning: A Review of Classification Techniques*,

ser. Emerging Artificial Intelligence Applications in Computer Engineering. IOS Press, 2007, vol. 160.
[15] E. Bauer and R. Kohavi, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants," Machine Learning, vol. 36, pp. 105-139, 1999.
[16] (2018) The Kaggle website. [Online] Available: http://www.kaggle.com/