Prediction of the PBC Patient’s Situations

*Abstract*— This paper presents the factors which influence the situation of patients who have PBC (Primary Biliary Cirrhosis) that is a rare but fatal chronic liver disease of unknown cause and thereby, predict the situation of the patients with different machine learning methods such as Support Vector Machine, Neural Network, Xgboost, Random Forest. In this paper, we try to find a better representation to explain our interests with appropriate questions. Then, preprocessed data is applied. The outlier analysis, data cleaning for explicit error and min-max normalization are conducted before making the prediction. After having the clean dataset, the models are conducted. The predictive performances of constructed models are evaluated by classification error rate (CER) accuracy, specificity and sensitivity. All procedures are executed on R-Studio.

Keywords—Logistic Regression, SVM, Artificial Neural Network, Random Forest, PBC

# INTRODUCTION

Primary biliary cholangitis (PBC) is an autoimmune chronic cholestatic liver disease characterized by biliary destruction and progressive intrahepatic cholestasis. It is a very serious disease and may cause the death of the patient. However, if the factors caused the death are investigated earlier, the patients can save. From this point of view, that kind of a problem might be solved by some of the machine learning algorithms.

The provided data has quality problem that should be handled to make more accurate predictions. In this work, the situation of patients, either death or alive, are predicted by using several regression model such as Support Vector Machine, Neural Network, Xgboost, Random Forest. The classification error rate (CER) accuracy, specificity and sensitivity of each method are calculated and compared to identify their performance.

# LITERATURE REVIEW

There are lots of researches on prediction the health situation of PBC patients. A logistic regression model is suggested to predict the health situation of PBC patients in Norway in [1] and, due to narrow data suggested data is not applicable for general. Another research about prediction the health situation of PBC patients in [2] is with CART. The house pricing determinants are researched for Seoul using the quantile regression approach in [3].Similarly, the health situation of PBC patients are examined for Turkey with comparison of hedonic regression and artificial neural network in [4]. Another approach is introduced in [5].

The approach is the long short term memory (LSTM) to predict the PBC patient’s situation. This method is a type of recurrent neural network whose success in classification was proven several times.

# METHODOLOGY

1. *Dataset*

The data set is taken from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. Although there were 424 patients who were suitable for this trial, only 312 of them were participated in this trial since the approval of both patients and physicians, has been asked. The data set includes 20 variables including the patient’s registration number and 312 observations. It has 50 missing values. The information related with the missing observation are going to be given in the following part of the analysis. The variable description is given below.

•N Case number. –discrete variable

•X The number of days between registration and the earlier of death, liver transplantation, or study analysis time in July, 1986. – continuous

•D 1 if X is time to death, 0 if time to censoring – categorical

•Z1 Treatment Code, 1 = D-penicillamine, 2 = placebo. – categorical

•Z2 Age in years. For the first 312 cases, age was calculated by dividing the number of days between birth and study registration by 365. – continious

•Z3 Sex, 0 = male, 1 = female. – categorical

•Z4 Presence of ascites, 0 = no, 1 = yes. – categorical

•Z5 Presence of hepatomegaly, 0 = no, 1 = yes. – categorical

•Z6 Presence of spiders 0 = no, 1 = Yes. – categorical

•Z7 Presence of edema, 0 = no edema and no diuretic therapy for edema; 0.5 = edema present for which no diuretic therapy was given, or edema resolved with diuretic therapy; 1 = edema despite diuretic therapy – categorical

•Z8 Serum bilirubin, in mg/dl. – continious

•Z9 Serum cholesterol, in mg/dl. – continious

•Z10 Albumin, in gm/dl. – continious

•Z11 Urine copper, in mg/day. – continious

•Z12 Alkaline phosphatase, in U/liter. – continious

•Z13 SGOT, in U/ml. – continious

•Z14 Triglycerides, in mg/dl. – continious

•Z15 Platelet count; coded value is number of platelets per-cubic-milliliter of blood divided by 1000. –discrete

•Z16 Prothrombin time, in seconds –discrete

•Z17 Histologic stage of disease, graded 1, 2, 3, or 4. – categorical

1. *Descriptive Statistics*

Descriptive statistics table are shown below. It has been obtained first, since it gives an insight into the data set at the beginning of the exploration. Besides, these values help to create research questions in the next steps of the analysis.

A descriptive summary is attached in Table 1 for numerical attributes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Days | Age | Serum Bil. | Serum Chol. | Albumin |
| Min. | 41 | 26.28 | 0.300 | 1.00 | 1.960 |
| 1st Qu. | 1191 | 42.34 | 0.800 | 44.00 | 3.310 |
| Median | 1840 | 49.76 | 1.300 | 86.50 | 3.555 |
| Mean | 2005 | 49.99 | 3.233 | 90.11 | 3.520 |
| 3rd Qu. | 2708 | 56.70 | 3.400 | 139.75 | 3.800 |
| Max. | 4556 | 78.44 | 28 | 200.00 | 4.640 |
| NA's | 2 | 3 | 4 | 2 | 2 |

Table Descriptive Statistical Summary of Numerical Data

This table belongs to the continuous variables in the data set. It represents the Tukey’s five number summary statistics which are the minimum value, 1st quantile, Median, 3rd quantile and the maximum value of the variables. Given table also provides the mean value of the given variables and the number of missing values for each variable separately. Interpretation are below by using these values.

As an example, the minimum and the maximum values of the produced albumin (gm/dl) level in the liver are 1.960 and 4.640 respectively. Both the mean of 3.520 and median of 3.555 indicate where the albumin is located. Thus, the average grams of albumin in the liver is between 3.5 and 3.55. Besides, the variability is calculated by using the interquartile range which is difference between 1st quartile and 3rd quartile. It is equal to 0.49 gm. In addition, this variable has 2 missing values.

A descriptive summary for categorical variables is attached in Table 2 for numerical attributes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Patient | Treatment | Sex | Ascites | Hepato |
| 0 | 187 | 156 | 35 | 283 | 151 |
| 1 | 124 | 154 | 274 | 23 | 158 |
| NA | 1 | 2 | 3 | 6 | 3 |

Table 2 Descriptive Statistical Summary of Caterogical Data

The table represents the categorical variables in the data set and since they are categorical variables, frequency representation has been chosen. In addition to the number of frequencies, the total number of missing values are observed simply for each variable. It is clearly seen from the table, for example, there is only one missing value in the situation of patients variable.

1. *Explaratory Data Analysis*

In this part, six research questions has been stated and the solutions according to the questions has been given by using suitable statistical methods. As it has been mentioned previously, these statistics gives a general information about the variables and it enables us to go one step further in the analysis.

*C.1* *Does the histologic stage of disease experience a change when the patient gets older?*

In the data set, when the histologic stage of disease on the liver increases, PBC gets to the critical levels. In this question, we expect the histologic stage of disease to increase when the patient’s age increases.

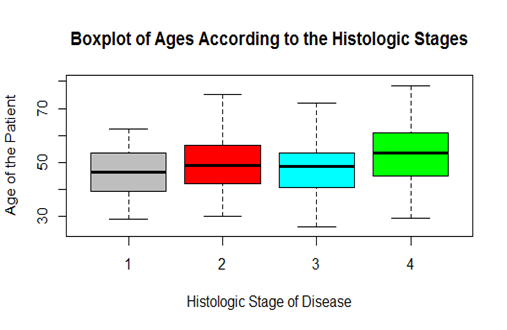


Figure Box Plot

The box plot shows that the expectation given above is almost satisfied. As it is observed from the plot, the median of each stage that represents age increases when the stage of disease rises.

We can verify the result shown by the previous plot using ANOVA. ANOVA is used because age has normal distribution (p>0.05) and the variance between groups are homoscedastic. (p>0.05) The ANOVA result is given below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | P value |
| Hist | 3 | 142.4 | 142.37 | 28.66 | 0.0009 |
| Res. | 308 | 705.3 | 4.97 |  |  |
| Total | 311 |  |  |  |  |

Table 3 The Results of ANOVA

It is enough evidence to conclude that there is significant difference in the age of patients among their histologic state of the disease. (p<0.05) Then, we apply pairwise comparison as post-hoc test after finding significant difference in ANOVA. Therefore, it is found that the significantly highest age belongs to patients in the stage of 4. (p<0.05)

1. *Missingness*

Missing data arises in almost all serious statistical analyses. When exploring missing data, it is important to come to a conclusion about the mechanism of missingness. It means that finding out the reason for why data are missing. In our data set, we have missing values created by the mechanism of missing completely at random. It means that the probability that missing observations are unrelated to the value of independent variables or to the value of any other variables in the data set. When we deal with such kind of data, we may choose to remove variables that include missing observations or we may try to impute them. In this project, we chose the imputation methods by using “mice” packages that is already available in R.

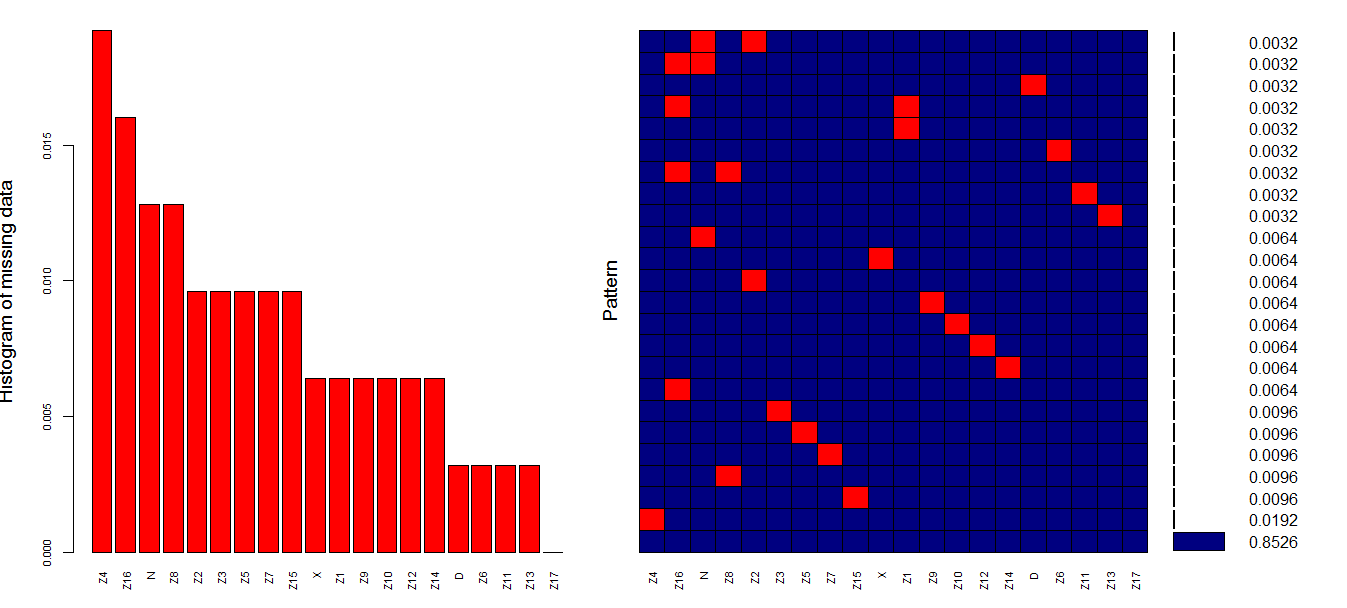


Figure 3 Aggregation Plot of Missing

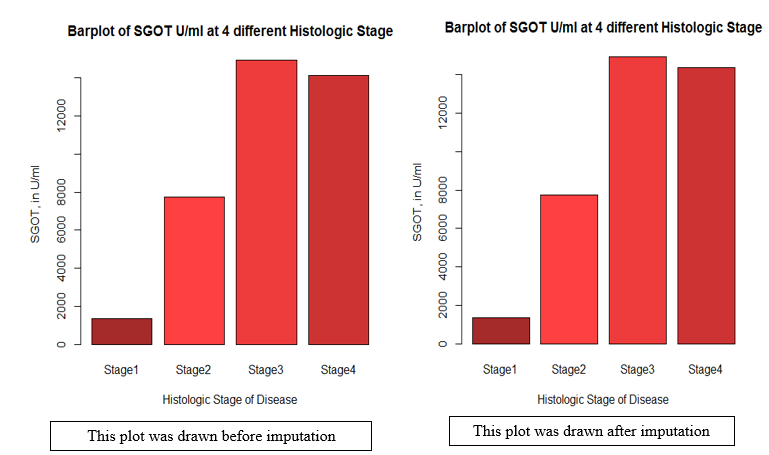
In this work, the aim is to predict land prices. Therefore, the QQ plot is generated to see the distribution in Figure 3.

The bar plot that is the left side of the graphs shows the percentage of the missing values in each variables in the data set. The aggregation plot that is the right side of the graphs shows the all combination of non-missing and missing parts in the observations where red denotes the missing and the blue denotes the non-missing parts. Besides, the aggregation plot also shows the corresponding percentiles of these parts. Therefore, we can say that data set contains the 50 missing values as previously mentioned in the description part by looking at both plots.

Mice package helps to fill the missing values in the data set with appropriate values. These values are generated from the distribution that especially designed for each missing point in the data set.

Therefore, there are several methods to fill in the missing parts of the data. However, we should not concern the disadvantages of imputation methods along with the advantages of these methods. Besides, at the end of one of these processes we should not get any different distribution or visual presentation than the original data. Since it is important for that data after imputation is valid for analysis.

In order to see that our imputations are suitable to continue analysis, we check our two questions which are in EDA part.



*Figure 3 Bar Plot*

As mentioned below the plots, the left bar plot was drawn before imputation process and the right bar plot was drawn after the imputation process. It can be clearly seen that both graphs have a bimodal shape. In general look they are similar to each other. Limits of stages in plot remains same although new terms are added in the SGOT variable. That’s why, the allocation of amounts of SGOT with respect to stages does not change. In other words, we can make a same interpretation about the results after the imputation. Therefore, we can say that our imputation on the missing values in data set is appropriate for the SGOT values.

After the examining of the raw data, it is seen that the data needs cleaning for meaningless data. Such as, very small land area, or a price that is meaningless are vanished from the data.

After cleaning the data, min-max normalization is used for all the numerical feature.

To check the accuracy of the model, the training data is divided into two part 80% and 20%, training set and validation set respectively.

## Modelling

After missing imputation and controlling the correction of methodology, the data set is suitable to conduct a model to predict the situation of patients. However, before constructing a model, we divide the data into two parts that are train data and test data. This methodology is called Cross-Validation. The reason is that when we have new observations about the variables, we want to test the suitability of our model after adding these new observations.

Therefore, we use 250 observations which is the 80% of the data set as a train data to construct a model. In addition, we accept 62 observations as a test data that are 20% of the data set.

## Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.

In this study, we eliminate the insignificant variables using backward elimination method and the final model is fitted. The output of the final model is given below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | *Coefficients* | | | | | | *Death* | **Estimate** | **Std. Error** | **z value** | **Pr(>|z|)** | | *(Intercept)* | -15.70 | 2.79 | -5.620 | 0.001 | | *Days* | -0.0078 | 0.0016 | -4.730 | 0.0003 | | *Age* | 0.0531 | 0.00017 | 3.019 | 0.002 | | *Serum chol* | 0.008 | 0.0004 | 2.530 | 0.011 | | *Alkaline phos* | 0.0003 | 0.00009 | 3.543 | 0.0003 | | *SGOT* | 0.009 | 0.0003 | 2.894 | 0.0038 | | *Proth* | 1.088 | 0.22 | 4.869 | 0.0001 | | Null deviance: 334.82 on 249 degrees of freedom / AIC: 235.25 | | | | | | Residual deviance: 221.25 on 243 degrees of freedom / Number of Fisher Scoring iterations: 5 | | | | | |

Table 4 The Results of Logistic Regression

The significance of the model is checked considering the difference between null and residual deviance, and it is found that the model is significant. (p<0.05).

It is clearly seen that all variables have a positive effect on the probability that is death except days. When we interpret the effect of the independent variables individually, the following interpretations are made.

The expected decrease in log-odds is 0.0078 for one unit increase in day. Then, we also say that one unit change in Age reflects as the 0.0053 change in the log-odds of death. 0.008 is the estimated rate of death for per unit change in cholesterol.

Besides, the rate of expected change in the odds ratio of death is 1.000296 when alkaline phosphatase increases one unit. Then, for one unit increases in SGOT levels results in 1.008730 rate of change in the odds ratio of the success variable which is death.

Lastly, when prothrombin time increases one unit, the expected change in the odds ratio of death is 2.967573 units. It also indicates that prothrombin time is the most influential variables on the situation of patients. In other words, the risk of death increases when the prothrombin time of the patients increases.

1. *Support Vector Machine*

Support Vector Machine is a supervised learning algorithm that can be used for both classification and regression problem.

In this method, we try to predict the response variable by considering the following features, days, age, serum chol., alkaline phos., SGOT, prothrombin.

After determining the features, SVM is set as C-classification type SVM and radial basis function is used as kernel function. Moreover, cost and gamma parameters are tuned as 10 and 0.001, respectively.

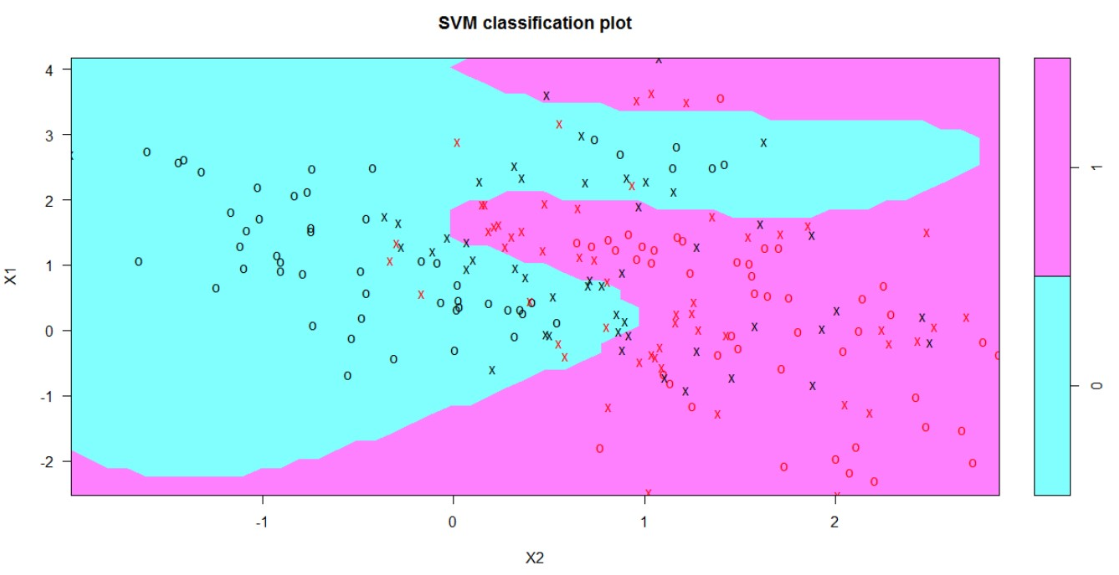


Figure 4 Classification Performance of SVM

1. *Artificial Neural Networks*

Artificial Neural Network is another supervised learning algorithm that can be used for both classification and regression problem.

In this method, we try to predict the response variable by considering the following features, days, age, serum chol., alkaline phos., SGOT, prothrombin..

After determining the features, they are normalized and sigmoid function is used as activation function of the model. The model has one layer with one neuron. The plot of the model is given below.

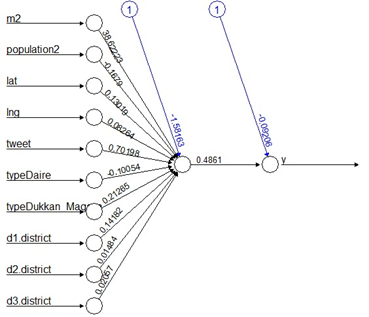


Figure 5 Representation of Neural Network for One Hidden Layer

1. *Random Forests*

Random forests is an another supervised learning based on tree algorithms which is applicaple for both regression and classification problems.

In this method, we try to predict the response variable by considering the following features, days, age, serum chol., alkaline phos., SGOT, prothrombin. Also, mtree number is tuned as 14.

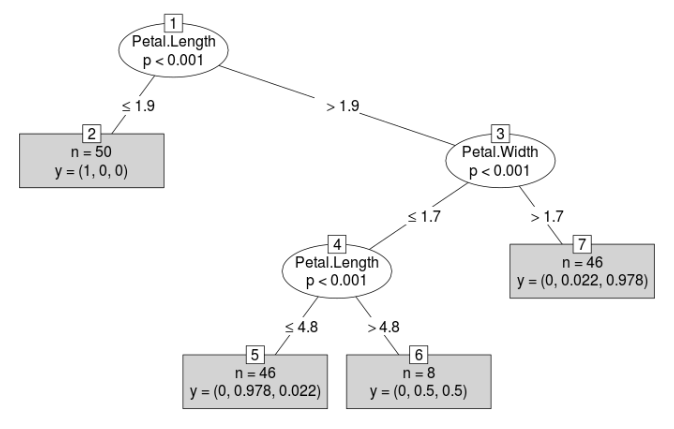


Figure 6 Classification Performance of Random Forest

1. *XgBoost*

Xgboost is a machine learning method based on tree algorithm and can be used for both regression and classification problems.

In this method, we try to predict the response variable by considering the following features, days, age, serum chol., alkaline phos., SGOT, prothrombin.

The eta, gamma and max\_depth parameters are tuned as 0.3, 10 and 6, respectively.

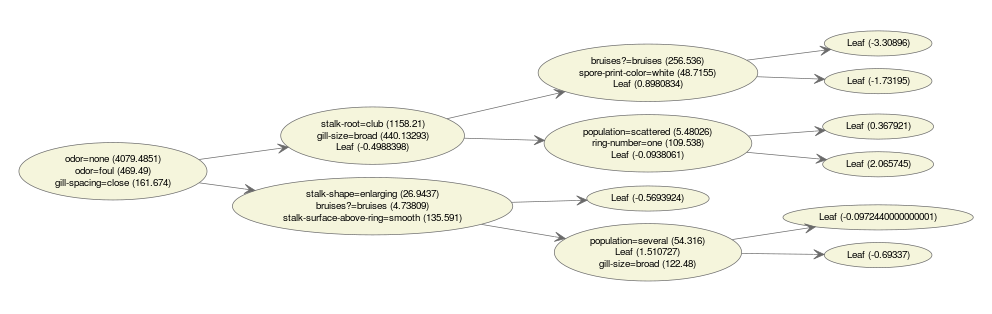


Figure 7 Classification Performance of XgBoost

1. *Perfomance Comparison on Test Data*

In this work, classification error rate, accuracy, sensitivity and specificity are used to identify the performance of the methods.

All of these metrics are based on confusion matrix, that’s why the elements of confusion matrix are identified before the equation of these metrics.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted** |  |  |
|  |  | **Positive** | **Negative** |
| **Observed** | **Positive** | TP | FN |
|  | **Negative** | FP | TN |

Table 5 Confusion Matrix

Where TN and TP are the number of correctly predicted negative and positive values.and FN and FP are the number of wrongly predicted negative and positive values.

1. *Classification Error Rate*

|  |  |
| --- | --- |
|  | (1) |

1. *Accuracy, Sensitivity and Specificity*

|  |  |
| --- | --- |
|  | (2) |
|  | (3) |

|  |  |
| --- | --- |
|  | (4) |

# RESULTS

In this part, the results are expressed for following classification models;

1. Logistic Regression
2. Support Vector Machine
3. Artificial Neural Network
4. Random Forests
5. XgBoost

In Table 6, Classification error rate, accuracy, specificity and sensitivity of each method is compared. According to the table, best prediction for the validation set is provided by artificial neural networks.

Classical regression methods are significantly faster than the Neural networks approach. On the other hand, their error values are quite high.

The result of the artificial neural networks with one hidden layer is quite promising due to its computation and error values. RMSE and MAE is calculated for normalized price data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CER | ACC | SN | SP |
| LR | 0.0461 | 0.756 | 0.77 | 0.15 |
| SVM | 0.0462 | 0.82 | 0.79 | 0.26 |
| ANN | 0.0458 | 0.88 | 0.80 | 0.35 |
| RF | 0.0596 | 0.83 | 0.58 | 0.28 |
| XGBoost | 0.0218 | 0.815 | 0.66 | 0.16 |

Table 6 Performance Comparison

# CONCLUSION

In this work, exploratory data analysis such the graphical techniques and the descriptive statistics are firstly conducted. Then, the outlier analysis, data cleaning for explicit errors and min-max transformation for data are applied in order to make quality of the data better. Finally, the health situation of PBC patients is tried to predict by using several methods. Their results are shown in the previous chapter. According to provided data, it is observed that days, age, serum chol., alkaline phos., SGOT, prothrombin are the most efficient factors on the health situation of PBC patients. In addition to this, it is difficult unseen the success of the artificial neural network approaches in the prediction of the health situation of PBC patients. Their accuracy is significantly better than accuracy of classical approaches.

# REFERENCES

|  |  |
| --- | --- |
| [1] | H. Kuşan, O. Aytekin ve İ. Özdemir, «The Use of Fuzzy Logic in Predicting House Selling Price,» *Expert Systems with Application,* Vol. 37, Issue 3, pp. 1808-1813, 2010. |
| [2] | A. Can, «The Measurement of Neighborhood Dynamics in Urban House Prices,» *Economic Geography,* Vol. 66, Issue 3, pp. 254-272, 1990. |
| [3] | H. Kim, S. W. Park, S. Lee ve X. Xue, «Determinants of House Prices in Seoul: A quantile Regression Approach,» *Pacific Rim Property Research Journal,* Vol. 21, Issue 2, pp. 91-113, 2015. |
| [4] | H. Selim, «Determinants of House Prices in Turkey: Hedonic Regression Versus Artificial Neural Network,» *Expert Systems with Applications,* Vol. 36, Issue 2, pp. 2843-2852, 2009. |
| [5] | S. C. Bourassa, M. Hoesli ve J. Sun, «A Simple Alternative House Price Index Method,» *Journal of Housing Economics,* Vol. 15, Issue 1, pp. 80-97, 2006. |
|  |  |