HCC Data Analysis Project

Reza Amini

IDS 572

FALL 2021

TABLE OF CONTENTS

Introduction:

The liver is a vital organ that performs a variety of functions. Filters all blood from the stomach, intestines, and spleen; produces bile, which assists fat absorption; removes toxins from the blood, including medicine; and eventually aids digestion. Liver cancer occurs when cancer cells multiply in the liver. The most common type of liver cancer is hepatocellular carcinoma, which accounts for four out of every five occurrences. Hepatocellular Carcinoma (HCC) accounts for more than 90% of all primary liver malignancies, making it the sixth most often diagnosed malignancy. Malignancy of the liver was the sixth most commonly diagnosed malignancy and the second leading cause of death. Alcohol is responsible for 9.1% of all fatalities [1,2]. In men, more than 90% of primary liver cancers represent a severe global problem. [3]. In order to define the breadth of this illness, some research was conducted in Portugal. [4]. Hospital admissions at HCC doubled between 1993 and 2005, with total admission costs increasing in lockstep. By the end of 2015, the American Society of Hepatology (ASH) predicted a 70% increase in liver cases, demanding a greater national awareness of liver diseases [5]. One of the most difficult difficulties facing medical researchers is predicting survival [6–10].

The HCC has a small dataset size (165 patients), a heterogeneous set of predictive variables (49 clinical variables), a high percentage of missing values (an overall MD rate of 10.22 percent with only eight patients having complete information) and expected heterogeneity between patients due to the range of values in the considered values and the class imbalance for the HCC data. This study aims to look at the existing literature on the use of computational tools for HCC illness and see how far they can be adapted to HCC datasets with complicated features. This study looks at a genuine Hepatocellular Carcinoma database with various clinical variables.

Method

1. Preprocessing Data

Before proceeding with any more steps, we began by cleaning the data before utilizing any methods for this project. After reviewing the dataset, the first step is to analyze each feature and determine whether the data for each patient is complete. The dataset utilized in this research had a large amount of missing data for each characteristic, which had to be deleted or replaced with new values. The data was separated into two categories: continuous and discrete. The discrete data are the classifications that we want to separate from the original data, while the continuous data are the data with continuous variables. To manage them as categorial data, discrete data were converted to factors. For each group's missing data, we adopted a different strategy. To address the missing data in the categorial group, we deleted the missing data under one condition: if the number of missing data was greater than the threshold (5%) of the total number of data, the column was removed; if the total number was less than threshold, only that row of data was removed. For the continuous group, we eliminated the attribute column if the number of missing data exceeded 20% of the total amount of data, unless we replaced the NA with the average of that attribute.

Then, for the continues group, we got boxplots of each object mentioned in the results to see which ones were outliers. We delete the whole data set of that item based on the frequency of outliers. We chose a 0.1 percent threshold for outliers, which indicates that if the number of outliers is fewer than the threshold, we replace them with the attribute's average.

1.1.Distribution Analysis

A distribution is formed by a sample of data, and the Gaussian distribution (Normal distribution) is the most famous distribution. The distribution gives a parameterized mathematical function that may be used to compute the probability of each observation in the sample space. The probability density function is a distribution that describes the group or density of the data. We can show how the percentage of data or the likelihood of the proportion of observations changes over the distribution's range with Density functions. Probability density functions (PDF) and cumulative density functions (CDF) are the two forms of density functions. CDF can calculate the probability of the observation in a distribution. Also, the probability of the observations of the dataset can be summarized by using PDF. CDF can find the value less or equal of each observation.

1.2.Correlation analysis:

The association of two variables and the strength of a linear relationship between them can be determined with correlation analysis. The correlation analysis is a statistical tool which is calculate the change of each variable because of the change of another variable. The stronger relationship between two variables causes the higher correlation and the weaker relationship shows lower correlation. In statistical and research area, analyzing quantitative data collected through research methods is using the correlation analysis.

In this project, we are using two machine learning methods: Supervised and Unsupervised method. We used two clustering methods as unsupervised learning.

2. Clustering Method

Clustering method is one of the machine learning methods which is group the datasets. Clustering is a method of the unsupervised learning which is a common tool for statistical data analysis. It can be used in a wide variety of fields. For a specific dataset, each data point can be classified in a specific group by utilizing the clustering algorithm. So, each group has its own data points that they have similar properties. And if they are not in the same group, they have different features.

To have an insight about dataset and check the features of each group of data, clustering analysis is applied. In this project, we will use two different clustering method to make groups for the given dataset. The first method is Hierarchical clustering and the second one is K-means clustering method.

2.1.Hierarchical Clustering

This method has two different categories called bottom-up and top-down. The bottom-up hierarchical clustering method is also called *hierarchical agglomerative clustering* or *HAC* that represents as a tree. This method counts each data point as a single cluster at the outset and connects each pair of these cluster to merge all the clusters. The root of this tree is the specific cluster that keep all the samples connected and the leaves of this tree are the clusters that has only one sample. There is no requirement to indicate the total number of the clusters and the method is not sensitive to the distance between each cluster or leave of the tree. If the data has a hierarchical structure, and we want to keep the hierarchy property, this method would be the best choice.

2.2.K-means Clustering:

Another clustering method in machine learning is K-mean clustering. This method starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. This method will create and optimize the clusters if the specified number of parameters has been accomplished and we cannot find any changes in the clusters value which means the centroids have been stabilized and the clustering method was successful.

3. Classification:

Four different classification methods were performed in this project.

3.1.**k-nearest neighbor**

It is a supervised machine learning method which is a basic, easy-to-implement technique that may address classification and regression issues. The k closest training examples in a data collection is used as input. Whether k-NN is used for classification or regression determines the outcome. The outcome of KNN classification is class membership. The object is allocated to the most common class among its k closest neighbors (k is a positive integer, typically small). The function is only approximated locally in k-NN classification, and all computation is postponed until the function is evaluated. Because this method depends on distance for classification, normalizing the training data can significantly increase its performance if the features represent various physical units or arrive in wildly different sizes. Assigning weights to the contributions of the neighbors, such that the closer neighbors contribute more to the average than those who are farther away, can be a valuable strategy. The training phase of the method consists solely of storing the training samples' feature vectors and class labels. K is a user-defined constant in the classification phase. An unlabeled vector (a query or test point) is categorized by assigning the most frequently label among the k training examples closest to that query point. Euclidean distance is a widely used distance measure for continuous variables. Another step, such as the overlap metric, can be used for discrete variables, such as text categorization.

3.2.**Naive- Bayes classifier**

It is a probabilistic machine learning model that performs classification tasks. The Bayes theorem lies at the heart of the classifier. We can calculate the likelihood of A occurring if B has already occurred using Bayes' theorem. The evidence is B, and the hypothesis is A. The predictors/features are assumed to be independent in this case. That is, the presence of one attribute has no bearing on the other. The number of parameters required by Naive Bayes classifiers is linear in the number of variables (features/predictors) in a learning task. The decoupling of the class conditional feature distributions means that each distribution may be estimated as a one-dimensional distribution separately, which helps to solve difficulties caused by the curse of dimensionality, such as the necessity for data sets that grow exponentially in size as the number of features increases.

3.3.**Logistic regression**

This method calculates the likelihood of an event occurring. Logistic regression is utilized when the dependent variable (target) is categorical. There are two forms of logistic regression: 1) Binary Logistic Regression: There are only two possible outcomes for a categorical answer. 2) Multinomial Logistic Regression: A regression model with three or more categories that are not ordered.

4. Feature Selection

The process of selecting a subset of relevant variables to using in the machine learning model called feature or variable selection. We used them to prevent the curse of dimensionality and simplify the models to interpret easier by the data scientists or researchers. Also, the redundant or irrelevant data could be removed without missing any information by using feature selection methods.

Results

In this section, the results from R code are described. Figure 1 shows the boxplots for each attribute after removing the outliers. The Figure 1 shows that the threshold of removing the outliers from each column was appropriate, and we can see the summary of the data clearly.

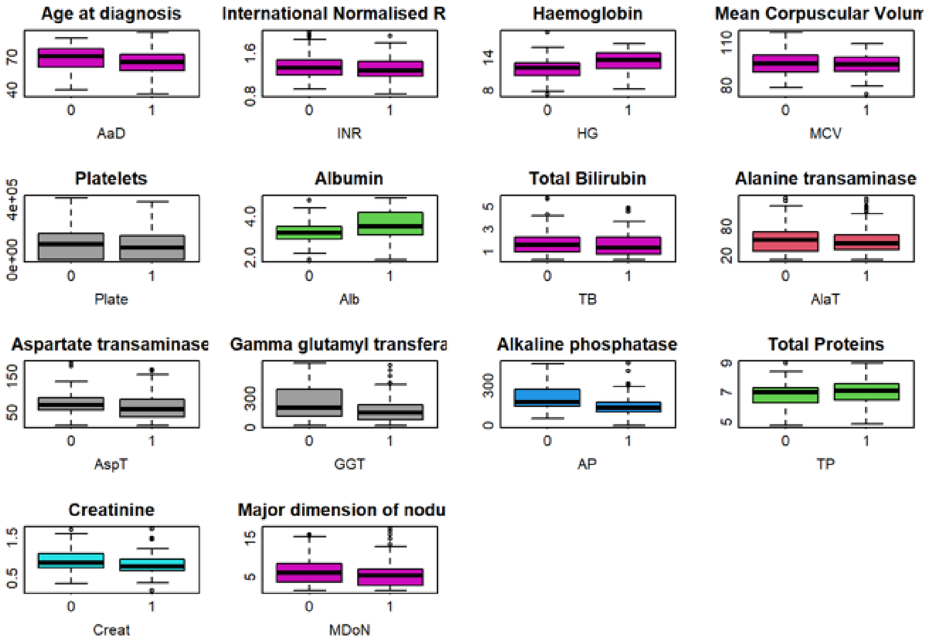

Figure 1: Boxplots after removing the outliers

The distribution of each selected variable and the QQ plot are plotted in Figure 2 and 3.

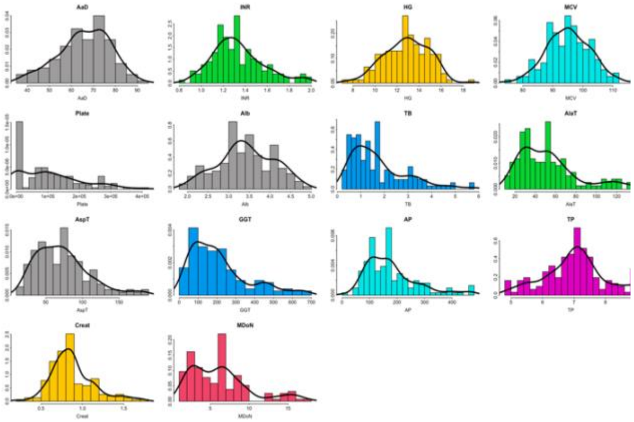

Figure 2: Distributions Plot

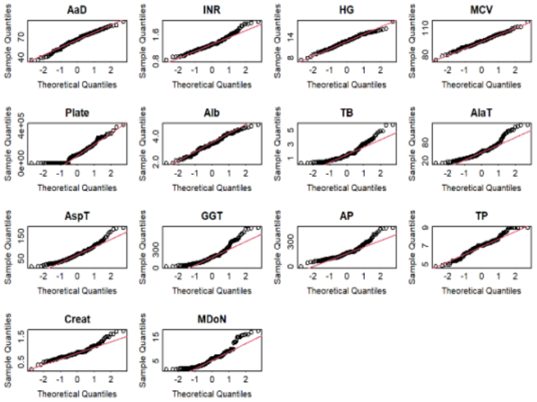

Figure 3: QQ-Plot

The correlation between each attribute is shown in Figure 4. The figure shows that there is not a strong correlation between each column of data. But between Alanine transaminase and Aspartate transaminase, there is a 0.63 correlation which is more than other columns.

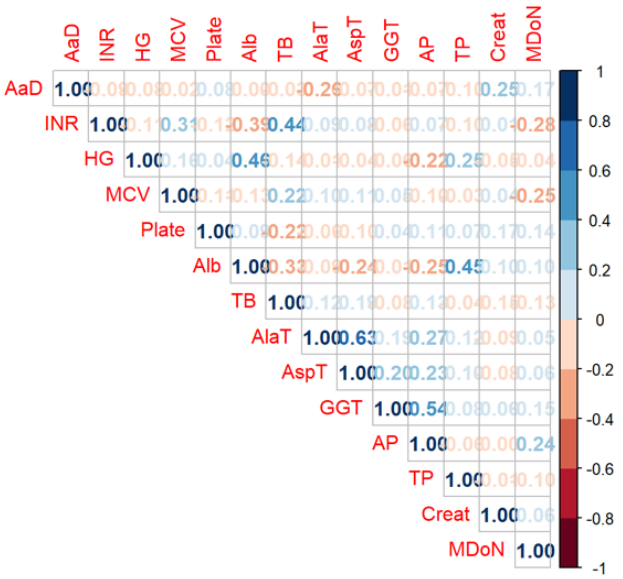

Figure 4: Correlation between attributes

Figure 5 shows the variance of principal components for the first ten columns with the highest variance. The plot shows that the variance for the VPC decreased from 4 to 1, which means …

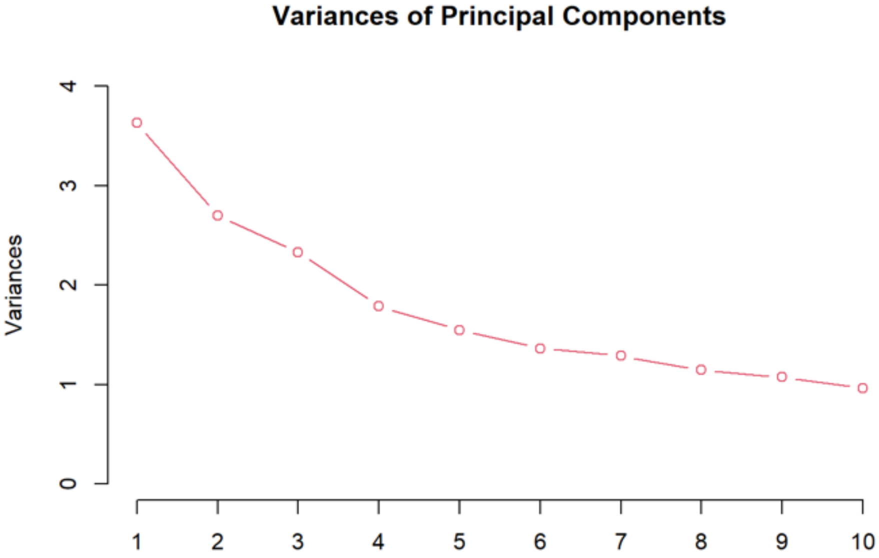

Figure 5: VPC

The first method that we used was hierarchical clustering as an unsupervised model that is showed in Figure 6. After checking different displacement and clustering methods, we found that Pearson displacement with Mcquitty clustering method has the best results.

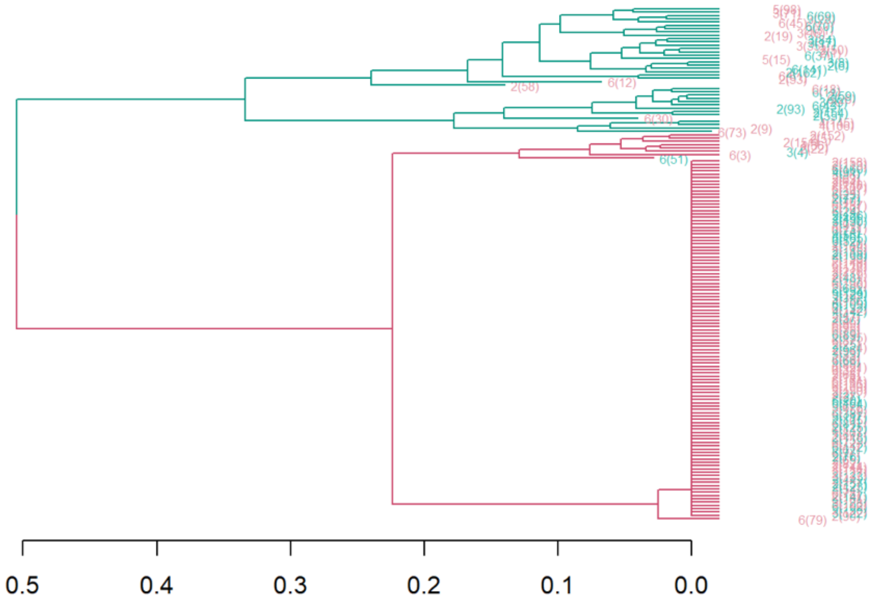

Figure 6: Hierarchical Clustering

For supervised model, we used four different methods such as: KNN classification, Naïve-Bayes classification, logistic regression, and multinomial logistic regression. The multinomial logistic regression has the best results that presented in Table 1.

Table 1: results of multinomial logistic regression

	K-Nearest Neighbors	**Naïve-Bayes Classifier**	**Logistic Regression**	**Multinomial Logistic Regression**
Accuracy	**0.7097**	**0.7097**	**0.6452**	**0.7742**
P-Value	**0.5887**	**0.02281**	**0.1839**	**0.04504**
Sensitivity	**0.6667**	**0.6000**	**0.5294**	**0.7500**
Specificity	**0.7273**	**0.8125**	**0.7857**	**0.7895**
+ Pred Value	**0.5000**	**0.7500**	**0.7500**	**0.6923**
- Pred Value	**0.8421**	**0.6842**	**0.5789**	**0.8333**

Prevalence	**0.2903**	**0.4839**	**0.5484**	**0.3871**

Conclusion

The way we handled the missing and outlier information may not be naïve, but many higher standard alternatives can be suggested, such as Interpolation, imputation, and distance-based methods. Although interpolation can make sense in small dimensions, for a high-dimensional dataset, like the data analyzed in this project, in should not elevate the overall significance of data. The same is applicable for imputation methods which act weakly on large datasets. Distance-based methods like k-nearest neighbors can reproduce quite relevant information but they can affect classification and clustering results. Still, the ratio of missing values was so large in some cases that no known method could save several variables.

In our analysis, clustering algorithms presented better reliability comparing to classification techniques. For the case of clustering analysis, hierarchical clustering with Pearson distance matrix and McQuity algorithm produced the highest cophenetic correlation coefficient, comparing to other distance matrices and algorithms. K-means also demonstrated good results when exposed to two to four first principal components of data. In general, clustering analysis introduced opportunities to further analysis of data.

In conclusion, to improve the response of our analysis, more variables and observations should be included. Genetic information of the patients, more detailed description of symptoms and patient habits that are main risk factors for HCC and more automated procedure to collect records can prepare the dataset for more accurate analysis.

References

[1] W.H. Organization, Globocan 2012: estimated cancer incidence, mortality and prevalence worldwide in 2012. <http://globocan.iarc.fr/>.

[2] W.H. Organization, Cancer fact sheet, 2014. <http://www.who.int/mediacentre/factsheets/fs297>.

[3] Anon., European association for the study of the liver, European organization for research and treatment of cancer, EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma, J. Hepatol. 56 (4) (2012) 908–943.

[4] R.T. Marinho, J. Giria, M.C. Moura, Rising costs and hospital admissions for hepatocellular carcinoma in portugal (1993–2005), World J. Gastroenterol. 13 (10) (2007) 1522–1527.

[5] L.P.C. Cancro, Cancro do fígado pode aumentar 70 por cento até, 2015. <http:// www.ligacontracancro.pt/noticias/detalhes.php?id=115>.

[6] H.B. Burke, P.H. Goodman, D.B. Rosen, D.E. Henson, J.N. Weinstein, F.E. Harrell, J.R. Marks, D.P. Winchester, D.G. Bostwick, Artificial neural networks improve the accuracy of cancer survival prediction, Cancer 79 (4) (1997) 857–862.

[7] J. Thongkam, G. Xu, Y. Zhang, F. Huang, Toward breast cancer survivability prediction models through improving training space, Expert Syst. Appl. 36 (10) (2009) 12200–12209.

[8] N. Esfandiari, M.R. Babavalian, A.-M.E. Moghadam, V.K. Tabar, Knowledge discovery in medicine: current issue and future trend, Expert Syst. Appl. 41 (9) (2014) 4434–4463.

[9] P.H. Abreu, H.A. Amaro, D. Castro-Silva, P. Machado, M.H. Abreu, N. Afonso, A. Dourado, Overall survival prediction for women breast cancer using ensemble methods and incomplete clinical data, in: L.M. Roa Romero (Ed.), XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013, IFMBE Proceedings, vol. 41, Springer, 2014, pp. 1366–1369.

[10] P.H. Abreu, H.A. Amaro, D. Castro-Silva, P. Machado, M.H. Abreu, N. Afonso, A. Dourado, Personalizing breast cancer patients with heterogeneous data, in: Y.-T. Zhang (Ed.), International Conference on Health Informatics, IFMBE Proceedings, vol. 42, Springer, 2014, pp. 39–42.

[11] J. Yuan, T. Fine, Neural-network design for small training sets of high dimension, IEEE Trans. Neural Netw. 9 (2) (1998) 266–280.

[12] R. Andonie, Extreme data mining: Interference from small datasets, Int. J. Comput. Commun. Control 5 (3) (2010) 280–291.

[13] F. Harrell, K. Lee, D. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Stat. Med. 15 (4) (1996) 361 387.

[14] P.J. García-Laencina, J.L. Sancho-Gómez, A. Figueiras-Vidal, Pattern classification with missing data: a review, Neural Comput. Appl. 19 (2010) 263–282.

[15] K. Qi, D. Wu, L. Sheng, D. Henson, A. Schwartz, E. Xu, K. Xing, D. Chen, On an ensemble algorithm for clustering cancer patient data, BMC Syst. Biol. 7 (Suppl. 4) (2013) S9.