Article

Using CRISP-DM with Acute Liver Failure in mind

André Coutinho [1], César Magalhães [1], Francisco Saraiva [1], Rui Santos [1], Cristiana Neto [2] and José Machado[2]*

[1] University of Minho
[2] Algoritmi Research Center, University of Minho
* Correspondence: jmac@di.uminho.pt;
Tel.: +351 253 604 430;
Fax: +351 253 604 471.

Version February 7, 2021

Abstract: The healthcare sector is highly dependent on data analysis to help prevent the appearance of health complication and disease in people. Having the ability to process data and derive results and conclusions from analysis is a great tool to help in decision making of health specialists everywhere. With techniques and tools for Data Mining it is possible to improve the process of analysing large amounts of data and from them extract conclusions. This works describes the process of taking data from a population, prepare it and analyze it and from it extract possible cases of acute liver failure based on the weight of data and attributes. The data present in this project contains information on indian citizens with various attributes which have acute liver failure. The Rapidminer tool was used for the preparation and analysis of this project. Achieved results were satisfactory with models presenting values of sensitivity up to 94.68% and values of accuracy, specificity and precision above 80%.

Keywords: Data Mining; CRISP-DM; Acute Liver Failure; Machine Learning; Classification

1. Introduction

Currently, the flux of data generated daily by the health sector, as for the different pathologies and clinical situations, and as the health status of the patients, is in such a way complex and meaningful, that the treatment and analysis of data is an added value in helping the process of decision making by health professionals.

One of the diseases that can benefit from data analysis, extracting valuable knowledge, is Acute Liver Failure. This disease causes the scar tissue to replace the healthy liver tissue, preventing the liver from functioning normally. Unlike Chronic Liver Failure, in which the health deterioration of patients is prolonged in time, the aggravation of Acute Liver Failure and clinical state of patients can happen in a matter of days. Due to its rapid acceleration, acute liver failure implies a quick and intensive treatment, many of the times in a specialized care unit. The need for a quick treatment, allied with a resource that allows health professionals to hypothesize different therapeutics and solutions, will lead, in a short and long term, to the development of medicine. [1]

For such results, in the era of data, exists practices and techniques, such as Data Mining that complement the process of research and help in the advancement of a modern medicine. Data Mining is the process by which someone is able to discover patterns and/or interesting relations for the problem at hands with big volumes of data. It's a process commonly used to detect anomalies, patterns, clustering, and, especially in the present investigation, predictive modeling. [2]

The present work will focus in a preventive diagnostics politic, with a data group as support to predict the probability of a certain patient, with certain factors, is prone to suffer from Acute Liver Failure. We'll be using CRISP-DM, a framework financed by the european community, for the process of Data Mining, which helps in the planning and management of data.

The data set used comes from the JPAC Center for Health Diagnosis and Control, which conducts research in India. This dataset consists of selected information from 8785 adults aged 20 years and over, which contains 29 attributes considered relevant in the identification of acute liver failure.

For the purpose of this study/work, we've utilised RapidMiner as the development tool.

2. Methodologies, Materials and Methods

To study the provided dataset and find a solution for the problem at hand the group utilized CRISP-DM, short for standard intersectoral process for data mining, is a process that maintains a structured approach to planning a data mining project. This process consists of an idealized sequence of events, which can be carried out in a predefined order or in a more appropriate order for the project. In this process, there is often a need to return to previous events and repeat the necessary actions. [3]

The process consists of six parts, which are: **Business Understanding**, which aims to understand the objectives and requirements of the project as well as to determine the objective of Data Mining, **Data Understanding**, where the collection, exploration and familiarization with data is made and possible problems in the quality of the data are identified, **Data Preparation**, in which the data is cleaned and the selected according to inclusion / exclusion criteria, **Modeling**, where the Data Mining models to be used are chosen and the models are constructed and evaluated. Finally, we have the **Evaluation** part, where the results from Modeling are evaluated and compared with the initial goals, and the **Deployment** part, where the final models are put into practice. [4]

2.1. Business Understanding

The main objective of this work was simple and clear: to predict acute liver failure through demographic variables, considering the patient's characteristics both in terms of health and social parameters, and also health status of members of patients family. This forecast should seek to achieve high evaluation metrics performance, in order for the disease to be detected as early as possible so that specialists in the field are able to act more quickly and effectively.

2.2. Data Understanding

Data Understanding of the CRISP-DM process consists of understanding the data presented in a study, identifying any possible issues and exploring the quality of the present information.

For this study, a dataset was supplied with various attributes to study their impact on predicting acute liver failure. Within this dataset the following attributes and information of the patients were given: **Age**, **Gender**, **Region**, **Weight** in kg, **Height** in cm, **Body Mass Index** (BMI), **Obesity**, **Waist** size in cm, **Maximum Blood Pressure** and **Minimum Blood Pressure** of the patient in mm Hg, **Good**, **Bad** and **Total Cholesterol** levels in mg/dL, the presence of **Dyslipidemia**, **Peripheral Vascular Disease** (PVD), **Physical Activity**, **Education**, **Unmarried**, **Income** level, **Source of Care** or health care, **Poor Vision**, **Alcohol Consumption**, **Hypertension**, **Family HyperTension**, **Diabetes**, **Family Diabetes**, **Hepatitis**, **Family Hepatitis**, **Chronic Fatigue** and lastly **Acute Liver Failure** (ALF).

Tables 1 and 2 further describe some of these attributes.

Table 1. Non True or False discrete attributes from the dataset

Discrete Attribute	Description	Variable	Values
Gender	Patient's gender	Male Female	M F
Physical Activity	Patient's level of physical activity	low moderate high very high	1 2 3 4
Region	Patient's region within India	North South East West	north south east west
Source of Care	—-	Clinic Government Hospital Private Hospital Never Consulted	clinic = = =

Table 2. Continuous attributes from the dataset

Continuous Attribute	Description	Min	Max
Age	Patient's age in years	20	85
Weight	Patient's weight in kg	25.6	193.3
Height	Patient's height in cm	130.4	200.1
BMI	Patient's Body Mass Index	12.04	66.44
Waist	Patient's waist in cm	58.5	173.4
Minimum Blood Pressure	Patient's min blood pressure	10	132
Maximum Blood Pressure	Patient's max blood pressure	72	233
Good Cholesterol	Patient's good cholesterol in mm/dL	8	160
Bad Cholesterol	Patient's bad cholesterol in mm/dL	27	684
Total Cholesterol	Patient's total cholesterol in mm/dL	72	727

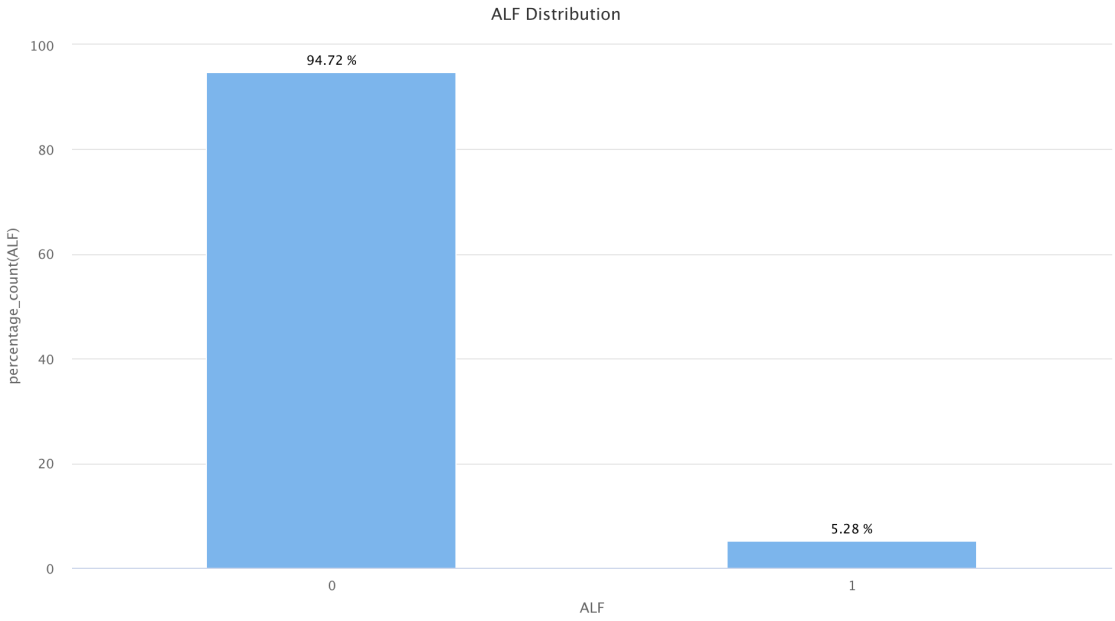

Figure 1. Distribution of the target variable ALF in the original dataset

For a grand total of 8785 instances and 30 attributes, looking through the data a lot of the attributes are seen to be binomial as in being made up of either true or false, with only a few being numeric values or scaled values like Physical Activity. The dataset presented some missing values for different attributes and an unbalanced set of patients regarding the ALF diagnosis. These were errors the group tackled in the next phase.

2.3. Data Preparation

The Data Preparation phase of CRISP-DM is based on selecting the data and cleaning, constructing, integrating and formatting for use. After understanding the dataset provided for the study the group started working on preparing the data for working and analysis. The group looked to remove duplicate values from the dataset as well as missing values with averages of neighbouring values, all this while remapping the attributes of 0 and 1 into false and true respectively, with a technique called binomial remapping. This left the dataset in an improved state for the modelling and analysis phase to better judge the efficiency and prediction of the algorithms. [5]

To tackle the analysis and prepare further the group decided to split the dataset into three different scenarios:

- **Scenario 1:** The base dataset with cleaned data of duplicates and empty values and with all attributes into account.
- **Scenario 2:** The same cleaned dataset but with the top 10 attributes by weight of their value in prediction ALF in patients.
- **Scenario 3:** Another cleaned dataset but now with only the top 5 attributes by weight.

For all three of these scenarios the group took into account three approaches, one by oversampling the dataset, which consisted of generating more rows for the minority of ALF positive patients, another by undersampling the dataset of the majority of negative ALF cases to the same amount of positive ALF patients and the last by not sampling the dataset. Making sure the dataset is balanced is extremely important for algorithms to more accurately predict models and reduce predictions errors. [6]

2.4. Modeling

Data Modeling is the phase of the CRISP-DM process that succeeds the Data Preparation. It consists of selecting the modeling technique, generating test designs, build the model and assess the model. After analysing the results of the Data Preparation it was decided to use three different Data Mining Techniques: Naïve Bayes, k-Nearest Neighbours and Decision Trees.

The dataset had multiple different types of variables, with binomial, polynomial, integer and real values. With the resulting datasets from the previous stage being divided into three scenarios: S1 All attributes ; S2 Top 10 attributes: Age, Height, Max Blood Pressure, Total Cholesterol, Hepatitis, Bad Cholesterol, PVD, HyperTension, Diabetes, Chronic Fatigue ; S3 Top 5 attributes: Age, Height, Max Blood Pressure, Total Cholesterol, Hepatitis; and only one binomial target variable that was ALF.

In addition, there were two different sampling methods being put to use: 10-fold Cross Validation and Split Validation (in which 70% of the data was used for training and the rest for testing) ; each one being tested with three different Data Approaches: Undersampling, Oversampling and No Sampling.

After the final datasets and the corresponding classifiers had been selected, the sampling methods, more specifically cross-validation, percentage split, and supplied testset, were evaluated.

2.5. Evaluation

In Data Mining, the handling of the data can be quite error prone while a variety of techniques and methods act on it. Therefore, the evaluation of the results needs to be both minute and detailed. This can be achieved through the use of performance metrics: numerical measures that allow the evaluation of the strenghts and weaknesses of a given classifier. With the help of RapidMiner's *Performance Operator*, a confusion matrix was obtained for each desired metric: Accuracy (1), Kappa (2), Precision (3), Sensitivity (4) and Specificity (5). A confusion matrix generates values that help calculate these mentioned metrics. The number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) are all values presented in it.

	true 0		true 1	
pred. 0	6757	TN	773	FN
pred. 1	1564	FP	7548	TP

Figure 2. Example Confusion Matrix (with "pred." being "predicted")

The values presented in the tables 3 through 8 are made possible via the following formulas [7]:

$$Accuracy = TP + TN/(TP + TN + FP + FN) \tag{1}$$

$$Kappa = (po - pe)/(1 - pe) \tag{2}$$

where:

$$po = observedaccuracy = (TP + TN)/(TP + FP + FN + TN)$$

$$pe = expectedaccuracy = [(TP + FP)(TP + FN) + (FN + TN)(FP + TN)]/[(TP + FP + FN + TN)^2]$$

$$Precision = TP/(TP + FP) \tag{3}$$

$$Sensitivity = TP/(TP + FN) \tag{4}$$

$$Specificity = TN/(TN + FP) \tag{5}$$

Table 3. DM models with the best accuracy results for each DM technique

DM Technique	Scenario	Sampling Method	Data Approach	Accuracy
kNN	S1	Cross Validation	Oversampling	85.96%
NB	S1	Split Validation	Oversampling	82.25%
DT	S1	Cross Validation	Oversampling	79.32%

Table 4. DM models with the best kappa results for each DM technique

DM Technique	Scenario	Sampling Method	Data Approach	Kappa
kNN	S1	Cross Validation	Oversampling	0.719
NB	S1	Split Validation	Oversampling	0.645
DT	S1 S2	Cross Validation Split Validation	Oversampling	0.586

Table 5. DM models with the best precision results for each DM technique

DM Technique	Scenario	Sampling Method	Data Approach	Precision
kNN	S1	Cross Validation	Oversampling	82.84%
NB	S1	Split Validation	Oversampling	79.19%
DT	S3	Split Validation	Oversampling	72.80%

Table 6. DM models with the best sensitivity results for each DM technique

DM Technique	Scenario	Sampling Method	Data Approach	Sensitivity
kNN	S3	Cross Validation	Oversampling	93.01%
NB	S1	Split Validation	Oversampling	87.50%
DT	S1	Cross Validation	Oversampling	94.68%

Table 7. DM models with the best specificity results for each DM technique

DM Technique	Scenario	Sampling Method	Data Approach	Specificity
kNN	S1	Cross Validation	Oversampling	81.20%
NB	S1	Split Validation	Oversampling	77.00%
DT	S3	Split Validation	Oversampling	65.54%

The analysis of the tables above brings out the best overall results accross all metrics. With these results in mind, certain thresholds were established to filter out the best model or models. The defined thresholds were:

- **Accuracy > 80%**
- **Kappa > 0.7**
- **Precision > 80%**
- **Sensitivity > 90%**
- **Specificity > 80%**

Table 8. Best model respecting the established thresholds

DM Model	Accuracy	Kappa	Precision	Sensitivity	Specificity
S1, kNN, OS & CV	85.96%	0.719	82.84%	90.71%	81.20%

As observed, the established thresholds resulted in only one suitable model making use of *Scenario 1*, the *k-Nearest Neighbours* algorithm, *Oversampling* and *Cross Validation*. The *sensitivity* is not the highest of the ones presented, however the model is the most balanced and with the highest values across all performance metrics and still with a satisfactory *sensitivity* result.

Further tools can be used to evaluate an algorithm or model. In this case, three ROC curves were constructed. By plotting the true positive rate against the false positive rate, these curves offer a graphic illustration of the trade-offs between sensitivity and specificity[9].

Along with the curve, there is an area associated: the Area Under the Curve (AUC), the most commonly used global index of diagnostic accuracy[8] due to its discriminatory power[10].

In Figure 3, the curves and corresponding AUC's can be visualized to further substantiate the models adequacy.

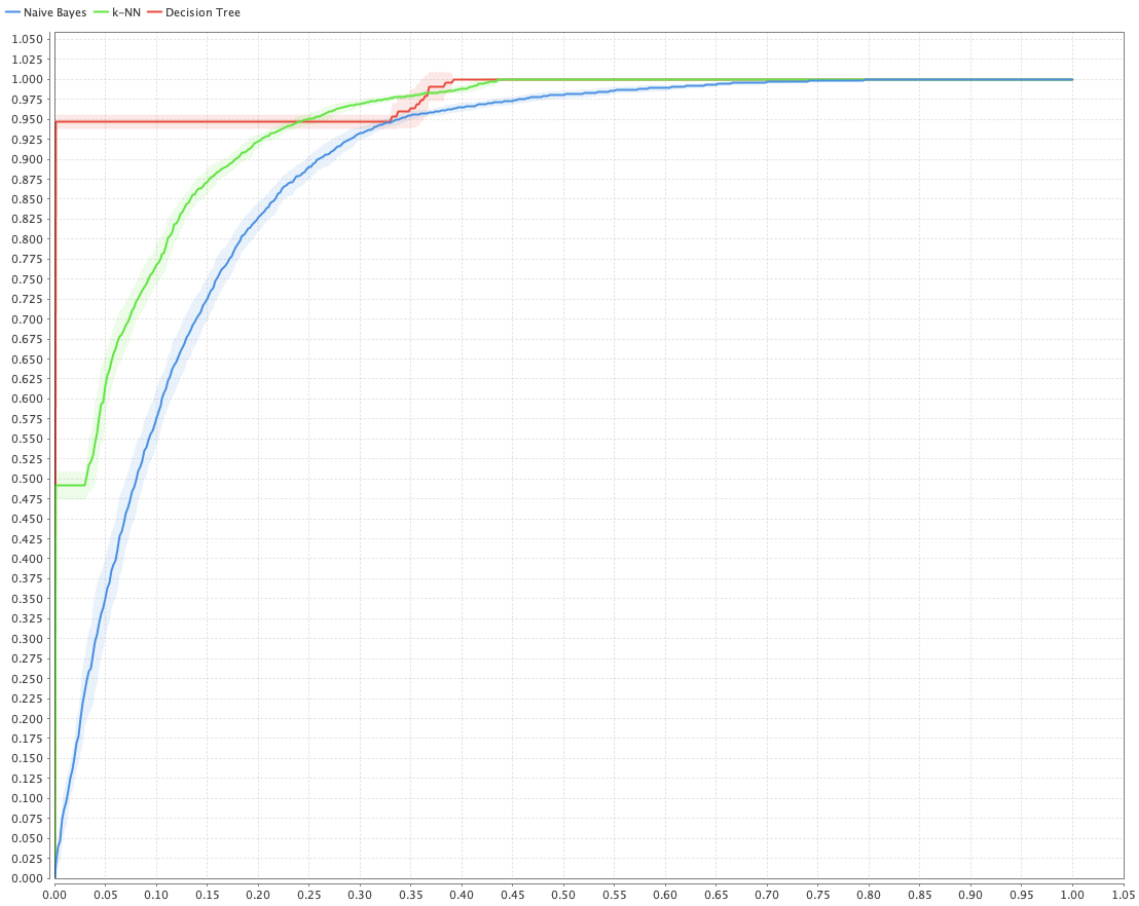

Figure 3. ROC curves of the best models

3. Discussion and Conclusions

After thorough analysis of the 270 rows, resulting from the evaluation of each Scenario/Data Modelling Technique/Data Approach/Metric combination, conclusions were drawn regarding the following metrics of evaluation: Accuracy, kappa, Precision, Sensitivity and Specificity.

Results regarding sensitivity allowed to conclude that while using the oversampling data approach, the highest percentages were obtained among the other data approaches, averaging approximately 89.38% sensitivity across the board.

Models using Cross Validation as sampling method got the best results in all five of the selected metrics in comparison with models using Split Validation. The former uses the totality of the dataset for training purposes while the latter uses only a specified portion. This means the more the data available for training, the more and more effectively algorithms learn.

Between undersampling and oversampling, there was a clear distinction in results. This might have occurred due to the low amount of examples being considered while using undersampling - 928 - versus the 16642 being considered when using oversampling.

Differences in the results when comparing metrics between models using both Cross Validation and Oversampling were not at all accentuated. This might be observed because all attribute selection was made possible by extracting the top k results from the operator Weight by Information Gain Ratio in RapidMiner and like so the top results were always of similar importance to the algorithm when compared to the bottom k.

It was verified that the algorithm with the best performance indicator values for this dataset was the k-Nearest Neighbour, but only considering all the established value thresholds and

Scenario/Sampling Method/Data Approach combination (as per Table 8). Curiously, this was also the algorithm with worst sensitivity values, although this time with undersampling taken into consideration regardless of chosen Scenario, Sampling Method or Data Approach. Taking all this information into consideration, we can safely state that the best model used the k-Nearest Neighbour algorithm along with Cross Validation, Scenario 1 and oversampling. Despite achieving overall good results in all metrics, specificity, for example, can be safely ignored since, in this specific case of diagnosing Acute Liver Failure, it is better to focus on a positive diagnosis rather than a negative one. Specificity focuses on the ability to identify negative results hence the attention put towards sensitivity, the one that focuses on the amount of actual positives identified as positives, also known as, true positives.

The use of different scenarios presented differences in results but overall, it has to be considered that major complications come with increasing age, and all of its underlying complications, making this a big factor to take into consideration, but also manifest themselves through other risk factors. Whether its origin lies in a viral source or autoimmune disease[11], hepatitis is a major risk factor when it comes to relate ALF to a cause.

Between scenario 3 and 1, the first comprises Diabetes, another factor known to increase the risk of acute liver failure[12] and this can be an indicator of the performance of the models making use of the first scenario.

In the end, the Scenario 1 dataset prepared with oversampling, coupled with Cross Validation and making use of the k-Nearest Neighbour model was the one with the overall best metric results, and most importantly, presenting a very satisfactory value of sensitivity, making this the most suitable example to address these data. Moreover, figure 3 shows the ROC curves for the best data modelling techniques under the influence of a Scenario 1, oversampling and Cross Validation. Knowing that the goal when drawing a ROC curve is for it to be as close to the upper-left corner as possible, making it a perfect classifier, and considering what it is now know about the models analysed under this study, our most suitable model presents itself as being a prime example for this situation.

4. Future Work

This study shows the importance of data mining and its techniques when evaluating a life threatening condition. Resorting to different data mining techniques can significantly help health organisations all over the world to diagnose, treat and prevent such conditions and the models constructed in this work can certainly help in some way or another. It is necessary to point out though, that further tests and new approaches are advised to be made and considered before any decisions and conclusions can be drawn when thinking about implementing a helpful system based on the work shown here.

Oversampling was a major determinant in results which might indicate that future work might need to cover a larger number of examples. International joint investigations can be of great help since they are able to cover a large amount of situations and countries. This makes it easier to identify new patterns in new data uncovered in different nations. In addition, several investigations that were already undertaken can be coupled together to draw even more conclusions in an effort to cast more light over the issue at hand.

Author Contributions:

Conceptualization: André Coutinho, César Magalhães, Francisco Saraiva, Rui Santos, Cristiana Neto and José Machado **Supervision:** Cristiana Neto and José Machado **Validation:** André Coutinho, César Magalhães, Francisco Saraiva, Rui Santos, Cristiana Neto and José Machado **Introduction:** Rui Santos **Methodologies, Materials and Methods:** André Coutinho, César Magalhães, Francisco Saraiva and Rui Santos **Business Understanding:** Rui Santos **Data Understanding:** Francisco Saraiva **Data Preparation:** Francisco Saraiva **Modeling:** César Magalhães **Evaluation:** André Coutinho **Discussion and Conclusions:** André Coutinho **Future Work:** André Coutinho

Abbreviations

The following abbreviations are used in this manuscript:

ALF	Acute Liver Failure
AUC	Area Under the Curve
CRISP-DM	Cross-industry standard process for data mining
CV	Cross Validation
DM	Data Mining
DT	Decision Tree
kNN	k Nearest Neighbours
NB	Naïve Bayes
OS	Oversampling
ROC	Receiver Operator Characteristic
S1	Scenario 1
S2	Scenario 2
S3	Scenario 3

References

1. Steven K. Herrine, MD, Sidney Kimmel Medical College at Thomas Jefferson University, *Insuficiência Hepática*, http://tiny.cc/pymlrz, (accessed: 29 of June 2020)
2. Christopher Clifton (2019), *Data mining*, https://www.britannica.com/technology/data-mining, (accessed: 29 of June 2020)
3. Smart Vision, *What is the CRISP-DM methodology?*, https://www.sv-europe.com/crisp-dm-methodology/, (accessed: 29 of June 2020)
4. Cristiana Neto(2020), A: *Descoberta do Conhecimento, Data Mining e CRISP-DM*, (accessed: 29 of June 2020)
5. Rona Sara George(2020), *Why is data cleaning important?*, https://xaltius.tech/why-is-data-cleaning-important/, (accessed: 29 of June 2020)
6. Tara Boyle(2019), *Dealing with Imbalanced Data.* http://tiny.cc/DealingImbalancedData, (accessed 29 of June 2020)
7. Rapidminer GmbH (2000), *Performance Binominal Classification.* http://tiny.cc/RapidMinerBinomial, (accessed 28 of June 2020)
8. Faraggi, D. and Reiser, B. *"Estimation of the area under the ROC curve."* Statistics in Medicine, vol. 21, 2002, pp. 3093-3106, doi:10.1002/sim.1228
9. Grzybowski M, Younger JG. *"Statistical methodology: III. Receiver operating characteristic (ROC) curves."* Academic Emergency Medicine, vol. 4,8, 1997, pp. 818-826, doi:10.1111/j.1553-2712.1997.tb03793.x
10. Fan, Jerome, et al. *"Understanding Receiver Operating Characteristic (ROC) Curves."* Canadian Journal of Emergency Medicine, vol. 8, no. 1, 2006, pp. 19–20, doi:10.1017/S1481803500013336.
11. Mayo Clinic Staff (2017), *Acute Liver Failure.* http://tiny.cc/MayoClinicALFDiabetes, (accessed 28 of June 2020)
12. Hashem B. El–Serag, James E. Everhart, *"Diabetes increases the risk of acute hepatic failure"*, Gastroenterology, vol. 122, Issue 7, 2002, pp. 1822-1828, ISSN 0016-5085, doi:10.1053/gast.2002.33650.