A Comparison of Data Mining Algorithms for Liver Disease Prediction on Imbalanced Data

Ain Najwa Arbain[1], B. Yushalinie Pillay Balakrishnan[2]
[1, 2] School of Computing,
[1, 2] Asia Pacific University of Technology & Innovation,
Kuala Lumpur, Malaysia
[1]ainnajwa630@gmail.com, [2]yushalinie97@hotmail.com

Received: 12-Dec-18; Accepted: 30-Jan-19; Published: 09-Feb-19
Corresponding author email: [1]ainnajwa630@gmail.com

***Abstract*— Liver is one of the most important organs in the human body but due to unhealthy lifestyle and excessive alcohol intake, liver disease has been increasing at an alarming rate globally hence it calls for an immediate attention to predict the disease before it is too late. However, medical data is often associated to be imbalanced and complex. Hence, the aim of this project is to investigate the data mining algorithm to predict liver disease on imbalanced data through random sampling. Results are compared and analysed based on accuracy and ROC index. K-Nearest Neighbour (k-NN) outperforms the other algorithms such as Logistic Regression, AutoNeural and Random Forest with the accuracy of 99.794%. As a conclusion, the model proposed in this research is performing better than past researchers conducted on Andhra Pradesh liver disease dataset.**

***Keywords*— Liver disease Prediction; Imbalanced data; Data Mining; Classification; SAS Enterprise Miner.**

1. Introduction

Liver is considered to be one of the important organs in the human body with core functions such as producing enzymes, processing waste products, and removing worn out tissues or cells [1]. However, liver is often exposed to multiple diseases that can eventually lead to liver damage and even worse, failure. According to World Gastroenterology Organisation and World Health Organization, 35 million people die due to chronic diseases and liver failure is one of the concerned diseases mentioned) [2][3]. It is further supported that more than 50 million adults will be affected with chronic liver disease and it calls for an immediate attention for actions in a conference held in Paris that discussed the alarming trends of liver disease globally.

According to American Liver Foundation, there are specific stages in order to diagnose a patient with liver disease in which it starts with an inflammation in the liver that could eventually makes the liver to become bigger than its normal size [4]. Next, a patient will be experiencing fibrosis in which the inflamed liver will begin to scar. If it is left untreated, the liver will be at a severe scarred stage known as Cirrhosis. Final stage, is liver cancer or failure. Genetics, unhealthy diet, immune system and viruses are some of the factors that could contribute to liver disease [5]. Severe scarring or Cirrhosis would reduce the ability for liver to self-heal as it is no longer possible to regain its main functionalities [6]. Patients that are diagnosed with the disease will need to seek immediate consultation with medical experts to prevent it from getting worse hence, in order to aid medical experts in providing faster and more accurate diagnosis for patients, data mining algorithms could be used to predict the occurrence of liver disease in patients at the early stages.

Data mining is a useful method to seek for patterns, hidden trends from voluminous data and it has been widely used in predicting different diseases in the medical industry. The is due to the increasing medical data being collected. This serves as a motivation for different researchers to utilize the data to enhance services for public and predict disease before it is too late. It is supported by Standford Medical Report that medical data is increasing at a staggering rate each year, it is estimated to increase to 2, 314 exabytes by 2020 [7]. However, the problem with medical data is it is imbalanced. Therefore, the aim of this project is to investigate the data mining algorithm to predict liver disease on imbalanced data through random sampling. The comparison of four different data mining algorithms such as Logistic Regression, Random Forest, k-Nearest Neighbour (kNN), and Artificial Neural Network were applied on the balanced data.

2. Related Works

According to Cleveland Clinic, there are various types of liver disease disorder that will eventually resulted in liver being severely scarred and unable to regenerate healthy cells at a normal rate in which it is known as Cirrhosis [5]. Liver disease can be diagnosed by analysing the level of enzymes in the blood with an early diagnosis of liver problems which may increase patient's survival rate [8]. However, disease is often diagnosed when it is too late, the amount of virus that has been spread across the liver might be at an alarming rate hence, it is important to have an early prediction. This has been a recurring issue mentioned across different researches that [9] [10]. As a result, the amount possibility to reduce the risks of getting liver failure might be low. Hence, various past researchers are focusing on providing a better solution to the issue through the usage of data mining algorithms.

Data mining within medical industry is not new, it has been widely applied by various researchers in developing classification and predictive models for early disease

predictions such as heart disease, diabetes, and cancers. There are also numerous researchers focusing on the prediction of liver disease and the aim of conducting the analysis varies from one research to another.
The liver patients are not easily discovered in an early stage therefore, automatic classification algorithms are needed for liver diseases and classification techniques are very general in various automatic medical diagnoses tools [11]. Classification techniques such as Naïve Bayes, Back propagation, k-Nearest Neighbour (k-NN), Support Vector Machines (SVM) and C4.5 Algorithm were used. From this study, k-NN algorithm has given a better accuracy with all feature set combination.

The existing studies have proposed a comparative model among different classification algorithms in order to select the best algorithm in predicting liver disease [12] [13]. It is also supported that by comparing each algorithm's performance will enable researchers to focus on the classification accuracy in producing a reliable and efficient prediction or system. The findings were evaluated using difference evaluating criteria such as sensitivity, precision, recall, sensitivity and specificity [13]

Moreover, feature selection has been chosen as one of the most beneficial methods by different researchers in enhancing the performance of model produced [14] [15] [16]. Feature selection is known as a process of choosing a set of useful attributes from the available attributes within a data set. This is because some attributes might not be very relevant to the analysis of the research hence, instead of using all the attributes, only selected attributes with highest worth will be chosen for modelling. Improved performance in terms of accuracy is one of the advantages of using feature selection.

This has been proven through the analysis of results produced, the classification accuracy of Random Forest algorithm increases when applying feature selection, 71.8696% as compared to not using feature selection, 70.3259% [15]. While, the classification accuracy achieved was 100% using Random Forest algorithm with feature selection and findings analysed analyses different numbers of feature sets while applying the algorithm over 2 different datasets which are BUPA dataset and Andhra Pradesh dataset [14] [16]. The accuracy achieved was better in Andhra Pradesh because there are more relevant attributes can be chosen rather than in BUPA dataset which only has small number of attributes that are not very relevant to the analysis.

Sampling method before applying decision tree algorithm is taken as the priority in research [9] [17] [18]. This approach is applied due to high amount of error rate and there are insufficient records within the collected data set. To overcome the issue, oversampling and under-sampling methods are discussed in the analysis. Oversampling is duplicating minority class several times and combined with majority class while under-sampling is when the major class is reduced or dropped to a smaller size. However, under-sampling could be a least beneficial method to take because there will be a huge number of information that could be lost from the data set and the information could be very beneficial and important to produce an accurate model. Oversampling in minority class is chosen to overcome the issue of splitting nodes in minority class and it is done by increasing the number of records in minority class by negative [17] [18].

On study proposed C4.5 algorithm to generate tree which are interpretable by the healthcare practitioners [18]. This algorithm is used to for continuous and discrete values of datasets. Decision tree was built from a set of training data set. It is discovering the interest rule or relations which in turn lead to improve the understanding of the process. Affording that C4.5 gives the better result compare to the other algorithms and the future work would be improved C4.5 could be derived with various parameters. To make the decision the highest normalized gain attribute is chosen. From the dataset that they use for this algorithm, over fitting happens

Three different supervised machine algorithms derived from the WEKA data mining tool which include Naïve Bayes, KStar, and FT Tree are used [19]. WEKA contains tools for data pre-processing, regression, association rules, classification, clustering and visualization from a dataset. Conferring the values of the dataset the accuracy is analysed and calculated efficiently. From this experiment, FT Tree algorithm shows the better performance algorithm. To calculate he accuracy for other algorithms it takes some time compare to FT Tree algorithm. This proves that FT Tree algorithm use a key role in improving the classification accuracy of the dataset. Refer the table at below for the further information.

The findings are evaluated through different performance criteria such as accuracy, precision, sensitivity and time to build the model. Accuracy, precision, recall, time of execution and sensitivity are some of the example of important criteria to be evaluated from results produced. On study concluded K* algorithm outperforms the other models because it has the highest accuracy, precision and sensitivity [20]. While, Support Vector Machine (SVM) outperforms Naïve Bayes in terms of accuracy and time execution [21]. Grading algorithm outperforms the other meta learning algorithms in terms of accuracy and time execution [22]. This is because good result is produced when it has high accuracy, high sensitivity with low error rate and low time execution.

Most of the researchers conclude a single model as the best model however, in IntelliHealth, it is a framework that combines 7 different algorithms to predict diseases such as liver disease and diabetes [23]. It was developed to reduce performance bottleneck. The hybrid methods have proven to produce highest accuracy and sensitivity as compared to a single model. Although the purpose of each research differs from one another, most of the researcher concludes that increasing trend of liver diseases due to unhealthy diet act as the main motivation and concern to be solved [13] [14] [15] [22].

The Table 1 shows a summarized version of algorithms applied and the accuracy achieved in past findings [10] [12] [13] [14] [15] [17] [18] [19] [20] [22] [24].

Table 1: Past findings on liver disease prediction

Author	Year	Algorithm	Accuracy	Data set
Reena	2010	FT Tree	97.10%	UCI repository
Ramana, Surendra and Ventateswarlu	2011	k- NN	97.47%	Andhra Pradesh and UCLA of Irvine
Ramana, Surendra and Ventateswarlu	2012	ANOVA and MANOVA		Andhra Pradesh and BUPA
Ramana, Surendra and Babu	2012	Modified rotation forest proposed with: -For UCI dataset: Multi-layer perception & random subset feature selection -For India dataset: K-NN & correlation based feature selection	74.78% 73%	UCI and India liver dataset
Jin, Kim and Kim	2014	Logistic regression	91.3%	Andhra Pradesh
Gulia, Vohra and Rani	2014	Random forest	71.8936%	Andhra Pradesh
Sindhuja and Priyadarsini	2016	CART decision tree	70.10%	UCLA of Irvine
Baitharu and Pani	2016	Multi-layer Perceptron	71.59%	Not mentioned
Sindhuja, Jemina and Priyadarsini	2016	C4.5	Not mentioned	Not mentioned
Ghosh and Waheed	2017	KStar (K*)	98.5%	Andhra Pradesh
Sharmila, Daruman and Venkatesan	2017	Fuzzy Neural Network	91%	Andhra Pradesh
Pasha and Fatima	2017	Grading algorithm	71.36%	Andhra Pradesh
Nahar and Ara	2018	Decision Stump	70.67%	Andhra Pradesh

3. Methodology

Knowledge discovery in database (KDD) is a repeating multi-stage process for extracting valuable, non-common information from large databases. KDD is to finding knowledge in high-level data and useful patterns in a data. We need KDD for terabytes of data and impractical to mined for interesting patterns to wide range of organization. From using this methodology, we can extract knowledge from data to research in design statistic, pattern recognition, database management, machine learning, artificial intelligence, web discovery solutions, and to deliver advanced business intelligence. The steps of KDD process shown in Figure 1.

Knowledge Discovery Process (KDD) Diagram.

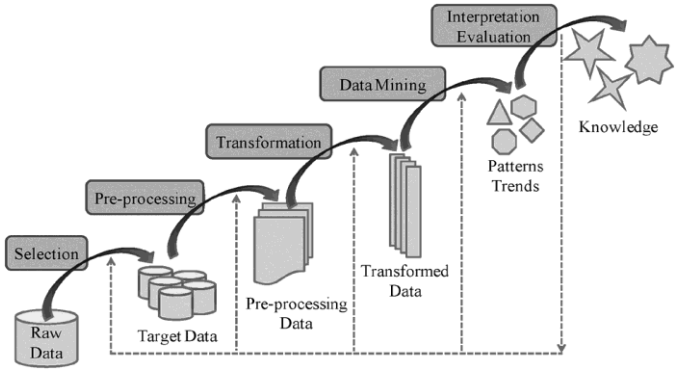

Figure 1: Stages in KDD

The chosen methodology for this project is Knowledge Discovery from Databases (KDD). KDD can be described as the overall process of discovering knowledge from data and highlights the "high-level" application of particular data mining methods. Based on the figure above, KDD consists of **6 stages or phases** which are:

- Stage 1: Data Selection
- Stage 2: Data Pre-processing
- Stage 3: Data Transformation
- Stage 4: Data Mining
- Stage 5: Interpretation and Evaluation
- Stage 6: Knowledge Discovery

These stages will be further explained in the section below. The aim of KDD methodology is to find useful and hidden patterns, and new understanding of the data by applying data mining techniques and few required processes such as cleansing and transforming data in order to increase the quality of data.

- Stage 1: Data Selection

Data may come from different sources such as databases and data warehouses in which data resides within the source could be having low quality data that is resulted from increasing number of missing values, duplicated values and inconsistent data. This will eventually affect the analysis in the later stages

hence this issue must be resolved before further analysis is conducted. The domain must be specified beforehand to further understand the targeted field of research. This can ensure the researcher to focus on the target data, identifying the problem within the data set and resolve the issues. The output is a specific targeted data selected from various sources that is relevant for the analysis.

Indian liver disease patients data set is selected as the main data set. The data set is collected from a region in India known as Andhra Pradesh. It contains a total of 11 attributes or parameters including a target data and 584 patient records classified into liver patient and non-liver patient. Refer Section 4 for the dataset description.

- Stage 2: Data Pre-processing

Data pre-processing is a very important stage in KDD methodology in which it involves handling missing values, noisy data and inconsistent data. According to Press (2016) [38], data scientists spend 60% of the project timeline to clean and organize the data. This is because the result of analysis will be affected if the quality of data is low. Hence, in this stage, pre-processing can help in increasing the quality of the data to ensure that the result will be meaningful and useful when discovering new knowledge. There are several ways to replace missing values within the data set such as manually replacing each missing value, replace with the attribute mean or simply delete the attribute however the last option is not advisable. The output of this stage is a higher quality data that are cleaned from problems mentioned above. There are only several issues that need to be addressed such as:

- missing values
- possible outliers detected after visualizing the data.

Furthermore, outliers are detected from histogram visualization in which further outlier analysis will be conducted to figure out the reason behind such irregular readings recorded within the data set whether it is real values or errors resulted from data entry.

- Stage 3: Data Transformation

Cleansing the data from issues stated in the previous stage was not enough hence data must be transformed into forms that will be suitable for data mining to perform efficiently. Reduction on data can also be the preferred method in transforming and it can be further classified into dimensionality reduction and numerosity reduction (Han, Kamber, and Pei, 2012) [25]. For example, normalization of data could be used as a method to ensure huge range of data can be converted into smaller range (0 to 1). As a result, the data will be easier to be interpreted and understand by user. The transformed data is now ready for data mining.

- Stage 4: Data Mining

In this stage, data will be divided into training set and validation set to a certain user-specified ratio in order to train and validate the accuracy of the model. Machine learning algorithms can be applied as a process to seek for patterns. Machine learning can be classified as supervised and unsupervised learning.

Supervised learning is used when there is a prior knowledge about the data, such as provided in the dataset a target attribute hence it can be applied to make prediction. Examples of supervised learning is classification and regression. While unsupervised learning is used to describe the data when there is no prior knowledge given. In this process, data containing similar characteristics with be clustered together. The distances between centroid in each cluster will be calculated iteratively until there is no changes. For example, K-means clustering. Algorithms that can be applied to seek for patterns are Naïve Bayes, Regression, Artificial Neural Network.

Logistic Regression

Logistic regression is a multivariable analysis in which it uses multiple variables to predict a single outcome (Park and Hyeoun-Ae, 2013) [26]. It is used to describe the data and its relationship between variables within the data. Hence, there will be multiple independent variables and only one dependent variable. There are 2 types of logistic regression model which are binary logistic regression and multinomial logistic regression (Park and Hyeoun-Ae, 2013) [26]. When the outcome or target data is in binary or dichotomous format, it is known as binary logistic regression. It is closely related to the analysis of this project; the outcome is either "yes" or "no". While when the outcome consists of more than two categories, it is known as multinomial logistic regression.

Logistic regression is closely related to odds. Odds is the ratio of probability of event occurring over probability of event not occurring and it can be seen from the equation below:

Odds (event) = $\frac{pi}{1-pi}$

According to Park and Hyeoun-Ae (2013), the impact of the independent variables is often described in terms of odds [26]. In addition, the probability of an event occurring depends on the independent variables. P is known as the probability of an instance belonging to Class 0 (Jin, Kim and Kim, 2014) [13]. It will be classified into Class 0 when p > 0.5 and classified into Class 1 when the result of p < 0.5. The expression below shows the equation of logistic regression:

$$logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

The above equation is derived from the simple logistic regression equation. This is because the above equation can cater multiple variables until $B_k X_k$.

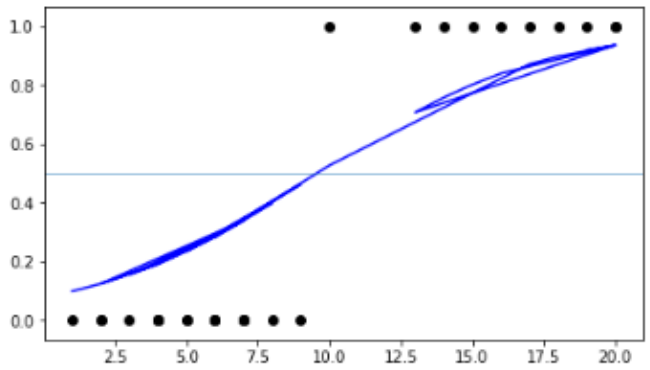

Figure 2: Sigmoid graph [35]

The Figure 2 shows the sigmoid function of the logistic regression in order to gain the S-shaped as shown.

Random Forest

Random forest creates a group of methods that consist in building an ensemble or so called, forest of decision trees from a randomized variant of tree induction algorithm [13] [27]. It is utilizing bagging, a resampling method to create a random sample set of data to build a decision tree. It is considered one of the most powerful algorithms as it is highly capable in performing both classification and regression [28]. Based on this approach, multiple decision trees will be produced, and the best decision tree will be evaluated and chosen based on majority vote.

The forest of decision trees will be developed through bagging in which instances will be put into a "bag" and based on this, different instances can be selected to form a tree. However, overfitting could be one of the main problems in Random Forest.

Auto Neural Network

Neural Network is also called as an Artificial Neural Network (ANN) which is an information processing paradigm that is inspired by the way biological nervous systems, like the brain and process information [29]. ANN was inspired from a real neural system like a human brain. Neural Network have a network structure consisting of artificial neurons (nodes) and neuronal connections (weights). Neural network needs several training-like inputs that we are going to supply and processing element that can be output. It is better to standardize the data with different range value of the data.

In 1943, the first artificial neuron was produced by a neurophysiologist Warren McCulloch and a logician Walter Pits [29]. The technology that use by them when available at that time did not allow to do too much. Neural Network is used to extract patterns and detect trends that are too multipart to be noticed by either humans or various computer techniques where ability to derive complicated meaning and rough data.

There are six type of neural network which are feed forward, radial basis function, kohonen self-organizing, convolution, modular and recurrent neural network. Feed Forward neural networks show the information is propagated from the inputs to the outputs. It moves in one direction only where it gets and move forward. Computations of No non-linear functions from n input variables by compositions of Nc algebraic functions and no role for time where there is no cycle between inputs and outputs. Feed forward contains of three type which are single-layer perceptron (SLP), multi-layer perceptron (MLP) and backpropagation**.**

Single-layer perceptron is the simplest type of neural network where they use to classify linear separable like 1 and 0. The sums of all the weighted inputs and if the sum is above the threshold, the output equals to 1. The initial weights are assigned randomly. It is similar to logistic regression.

Multi-Layer Perceptron for prediction of the class label of tuples is learn from the set of weights. It consists of an input layer, an output layer and more than one hidden layer. MLP composes of multiple layer of nodes in a directed graph with every connected layer to the next one except for the input node, and every node is a neuron with a non-linear stimulation function. MLR can differentiate data where the data is not linearly separable. It is a universal approximator and only operates on the numerical data type.

Backpropagation performs on a multi-layer feed forward neural network. The final output is been pass back to the input layer. It can also have more than one hidden layer. It takes a random weight of various layers corresponding the output and detect complex non-linear function. The number of hidden units can be change according to your preference but must be less than the number of attributes in the dataset. The weight is updated with each iteration where they move in the direction of the minimum of loss function.

In SAS Enterprise Miner, we can find AutoNeural node which it performs automatic configuration of neural network Multilayer perceptron model and it conducts limited searches for a better network configuration [30]. The AutoNeural searches over several network configurations to find one that best describes the relationship in a dataset and then trains that network [31]. We can specify the architecture at the property of the AutoNeural node suck like single layer, block layer, funnel layer and cascade. Single Layer, the hidden nodes are added in parallel. Block Layers, the hidden nodes are added as additional layers with the number of neurons. Funnel Layers, to form a funnel pattern the hidden nodes are added to the existing layer and to a new layer. Cascade, to train cascade network models only. We can change the property of train action to either train, increment and search for better performance.

K-Nearest Neighbour (K-NN)

K-Nearest Neighbour algorithm is the supervised learning algorithm is used to statistical estimation and pattern recognition. K-NN is a lazy learner because it differs from the classifiers previously. The entire memory will store in and at a time it will take out and move it to based class. The presented training data is simply stored when a new query instance is encountered, the related instances is retrieved from memory and used to classify the new query instance. Hamming distance is used for categorical data while Euclidean distance is used for continuous data. All the training samples is calculated using any distance measure which are Euclidean and Hamming [34].

The target is classified by a majority of its neighbours. K value is the number of neighbours which always have to be in positive integer and cannot be more than the number of attributes from the datasets. In SAS Enterprise Miner, K-NN is also known as Memory-Based Reasoning (MBR) to categorize and predict observation. MBR node uses K-nearest neighbour algorithm where neighbours are determined by shortest Euclidean distance and to combine the results of the neighbours, democracy method.

- Stage 5: Interpretation and Evaluation

In this step, data pattern is evaluated. Mined pattern is interpreted to essential knowledge. Since new knowledge may even be in conflict with knowledge that before the process began, interpretation may request that we analyse possible strife with previously disclose knowledge to make them understandable such as summarization, transformation, removing redundant pattern and visualization. Model evaluation is used to evaluate how to predict the accuracy of the datasets, confusion matrix and ROC chart is a graphical plot that illustrates the diagnostic ability of a binary classifier system used to measure the quality of the classification models. Accuracy is the proportion of the total number of predictions that were shown correctly. In ROC chart there are two performance is measure be done which is sensitivity and specificity. Sensitivity is the proportion of the real positive classes which are correctly identified. Specificity is the proportion of the real negative classes which are correctly identified. The classification model is compared where a confusion matrix is to show the number of correct and incorrect predictions was made

- Stage 6: Knowledge Discovery

After discovering new knowledge, in this the knowledge will be presented to user as the new enhancement in the domain knowledge.

4. **Analysis and Discussion**

4.1 **Experimentation**

SAS Enterprise Miner is chosen as the software to be used for data exploration, pre-processing, development of models using different data mining techniques and the results are compared in order to select the best model throughout the research. It streamlines the process of data mining in order to produce a highly accurate descriptive and predictive models based on the analysis of the huge amount of data from different enterprise (SAS, 2018) [32]. The steps in data mining process are known as SEMMA:

i. Sample.
ii. Explore.
iii. Modify.
iv. Model.
v. Asses.

4.2 **Data Selection**

Medical field has been producing and storing a huge amount of data in repositories hence with the availability of obtaining medical data to perform data mining processes to seek patterns became the motivation for this analysis. In this stage, data is selected to ensure only data that is relevant to the research will be collected from the available sources. Data set on liver patients is obtained from a region in India called Andhra Pradesh [14]. The Table 2 shows the attributes' description.

Table 2: Data description

Attribute name	Data Type
Alamine_Aminotransferase	Interval
Total_Bilirubin	Interval
Direct_Bilirubin	Interval
Alkaline_phosphotase	Interval
Asparate_aminotransferase	Interval
Age	Interval
Albumin	Interval
Albumin_and_Globulin_Ratio	Interval
Total_Proteins	Interval
Gender	Nominal
Dataset	Nominal

Original data set contains mixture of data type; initially attribute, Gender is in Nominal data type indicated using Female or Male. Hence, the attribute is converted into Interval data type to ensure that the data mining algorithms can be performed efficiently.

4.3 **Data Pre-processing**

Data pre-processing is a very essential stage in KDD methodology as the stage focuses on preparing the data with the aim of improving the quality of data. The elements of data quality involves accuracy, completeness and consistency [25].

The missing values (Albumin_and Globulin_Ratio) are replaced using the attribute mean. In order to perform this task, impute node is selected and it is connected to the dataset file. The impute node will replace the missing values based on the user-specified method.

Based on the property diagram, Default input method is changed to Mean. This will enable the missing values within the attribute to be replaced by its mean value, this approach is used because missing values are identified as an interval data type. As a result, all missing values are imputed using mean value of 0.947064. Next, it is important to check every attribute mean and standard deviation.

4.4 **Data Transformation**

In order to perform data transformation, there are several tasks that will need to be done such as identifying each attribute range, kurtosis and skewness value and this information are provided in StatExplore.

4.4.1 Skewness and Kurtosis

Skewness and kurtosis are very helpful in understanding the distribution of data within a data set. The normal range of skewness and kurtosis value is from (-3, 0, 3), this indicates that the data is normally distributed. It is important to address highly skewed data because when data is not normally distributed, it will affect the performance of data mining algorithm used in later stages and certain algorithm such as regression will be affected if kurtosis value is high than its normal range. The Table 3 shows attributes' skewness and kurtosis value and range.

Table 3: Table for transformation

Variable	Minimum	Median	Maximum	Skewness	Kurtosis
Alamine_Aminotransferase	10	35	2000	6.549192	50.57945
Alkaline_Phosphotase	63	208	2110	3.765106	17.75283
Aspartate_Aminotransferase	10	42	4929	10.54618	150.9199
Total_Bilirubin	0.4	1	75	4.907474	37.16379
Direct_Bilirubin	0.1	0.3	19.7	3.212403	11.35253

The distribution of the identified attributes is highly skewed than the normal distribution range, hence using transformation node, these values will be transformed based on a certain criterion in the property, refer Table 4:

Table 4: Transformation criteria

Criteria 1:	Criteria 2:
• Kurtosis > Skewness • Skewness in positive range • Transformation option: Log10, Log and Reciprocal	• Kurtosis > Skewness • Skewness in negative range • Transformation option: Square
Purpose: To transform data that is highly skewed into a less skewed distribution [36]. This is done by minimizing the stretch data into smaller range.	Purpose: It will stretch the small range of data into less skewed distribution.

Transformation Criteria 1 is chosen because it matches the above criteria. Hence, the property is changed into Log10 for the each of the identified attribute.

After transforming the data, the newly transformed values are identified to ensure that each attribute is within the acceptable range of normal distribution. As a result, the transformed attributes are now within normal distribution range, (-3, 0, 3). There are also changes found in the variable range, mean and standard deviation.

4.5 **Sampling method**

Sampling method is used to address imbalanced data through methods such as oversampling, under-sampling or random sampling. This is because imbalanced data will affect the model performance as the result might be overfitted. The overfitting occurs when the model produced learns the data and noise in training set to the extend that it is negatively impact the performance of model on the validation set [33].

In this research, random sampling is chosen as the approach before model development from sample node. The property size type chosen is 100% percentage and criterion selected is equal. This will enable the data to be randomly sampled into equal amount of data in target class. Based on the observation, it is highly imbalanced, having to 416 count of target data 1 (Liver patient) and 167 count for target data 2 (Non-liver patient).

After using sampling method, the new sample is now having equal amount of data for both target data, 1= Liver patient and 2 = Non-liver patient. Hence, the newly balanced data is now ready to be modelled.

4.6 **Data Mining**

In this stage, data is modelled using 4 data mining algorithms, Logistic Regression, Random Forest, AutoNeural and k-Nearest Neighbour (k-NN). Secondly, data is partitioned into training set and validation set using data partition node into 80:20. Hence, 80% of the data will be allocated for training and 20% to validate the accuracy of classification. More data will

be allocated for training than validation set, it is to ensure the model produced is not overfitting the training set [32].

4.6.1 Logistic Regression

Backward selection of Logistic Regression is chosen in predicting liver disease. This is because backward selection will take into account all variables within the Andhra Pradesh liver dataset as the training data begins. On each iteration, less important attribute will be removed from the training set until the stopping criteria is met. The backward selection model and default optimization is chosen for this analysis.

The analysis is further done on the model produced based on effect plot (Figure 3), output and *p* value.

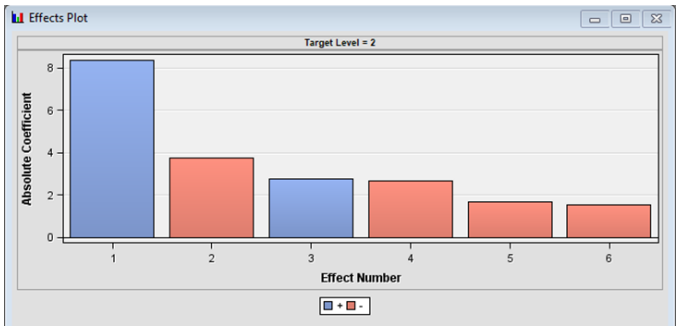

Figure 3: Effect plot

The effect plot indicates the slope of the line and few other variables that are chosen in backward selection model. The first bar shows the y-intercept of the line, having absolute coefficient of 8.37697.

The final output of Logistic Regression in predicting liver diseases, a total of 5 attributes has been selected from backward selection method, which are Albumin, Albumin globulin ratio, Alamine Aminotransferase, Total Bilirubin and Total Proteins. These attributes have been selected because it has the highest significance and worth for the analysis. This can be evaluated by the p value (< 0.05). will have the highest significance in the model. The *p* value < 0.05 express that there is a strong evidence against null hypothesis [41]. Based on the outcome produced, all the attributes match the p value criteria.

As a result, the misclassification rate for training and validation are identified to ensure that the model is not overfitting. When validation misclassification rate is bigger than training misclassification rate (Validation > Training) the model is overfitted. The misclassification rate for training is 0.319549 (99.680451%) whereas validation misclassification rate is 0.235294 (99.764706%). However, the analysis also covers Stepwise selection model and default model.

The model in stepwise selection method stops at step 2 having only 2 variables to predict liver disease, Bilirubin and Alamine Aminotransferase and the accuracy is slightly lower than backward selection method. The 2 variables are insufficient in predicting the occurrence of liver disease because there are various high significance attributes identified in backward selection. Hence, it will have an impact in accurately predicting liver disease only based on the 2 selected attributes in stepwise.

4.6.2 Random Forest

Random Forest is an ensemble that creates an ensemble of multiple decision trees at random and it is known as one of the high-performance nodes in SAS Enterprise Miner. The algorithm offers a direct relationship between the outcome and the number of trees; the more the number of trees, the more accurate the outcome produced [34]. Firstly, Random Forest is known as HP Forest node in SAS. Hence, the property will be changed accordingly such as number of trees that can be generated, maximum depth of tree and variable importance method.

The maximum number of trees is set to 100 and loss reduction are selected as the variable importance method. Hence, 100 trees will be generated and variables that are not important to the analysis will be removed using loss reduction method. In loss reduction method, variables worth more than 0 will be input while the others will be removed. The initial result shows it is overfitting after reaching 30 trees, hence, the misclassification rate shows no changes until the 100th tree.

The result produced overfitting after the 30th tree. In order to treat this, the number of trees to 30 and results are analysed. Even after reducing the number of trees to 30, the result is overfitting.

Random Forest is not suitable for this analysis because there is insufficient data in the data set. This may be caused by the inability to develop other sampling method in SAS such as oversampling method. This is because oversampling will enable more data to be created, without losing any valuable information from the dataset. However, due to this limitation of only using random sampling or equal balancing, the model starts to overfit at a very early stage, hence, the usage of Random Forest is not the best method to use when it comes to insufficient data

4.6.3 Auto Neural

The reason for to use auto neural node is because it searches over several network configuration and it calculates the require more performance than the neural network node. The auto neural train property in SAS Enterprise Miner, where it can be change for better model evaluation. At the final training property is to allow the model to converge which indicate whether the final model should be trained again. Using the default values of the properties, it shows the accuracy of (0.221) 99.779% of the validation and (0.316) 99.69%. The researchers have also tried changing the train property of architecture and train action.

4.6.4 k-Nearest Neighbour (k-NN)

k-NN algorithm can be also called as Memory-Based Reasoning (MBR) node where it assigns to observe new data using the average of the nearest neighbour. It classified based on the majority of each k-nearest neighbour. Before using the MBR node, the data is partitioned in the ratio of 80:20 prior modelling. This node works in multiple dimensions and requires the inputs to have equal variance and no correlation.

The MBR node property, where some changes have been made for better accuracy of the model. In the method property it contains two type of method which is RD-tree and Scan to determines what data representation is used to store the training set observations and then retrieve the nearest neighbour. RD-tree uses a tree structure to organize data observation in various dimension space whereas scan method chooses the nearest neighbour of the smallest linear distance function of all the possible k neighbour. The number of neighbour can be change to the value of less than the number of attributes of the datasets. This dataset contains of 11 attributes. No overfit is found and the lesser the number of neighbour, higher the valid rate get but not much difference with the default value. The method of RD-Tree and number of neighbour = 5 shows the better accuracy of 99.79% valid rate. RD-tree works better than the scan method.

5. Results and discussion

The performance of each classification algorithm is evaluated on the Andhra Pradesh liver disease dataset. The dataset contains 583 liver disorder patient records with 11 attributes. For the purpose of this study, SAS Enterprise Miner, an open source machine learning software is used. 4 types of classification algorithms that were used to predict liver disease in this research are Random Forest, *k*-NN, AutoNeural and Logistic Regression. In order to evaluate the performance of each model, accuracy and ROC chart and index are used.

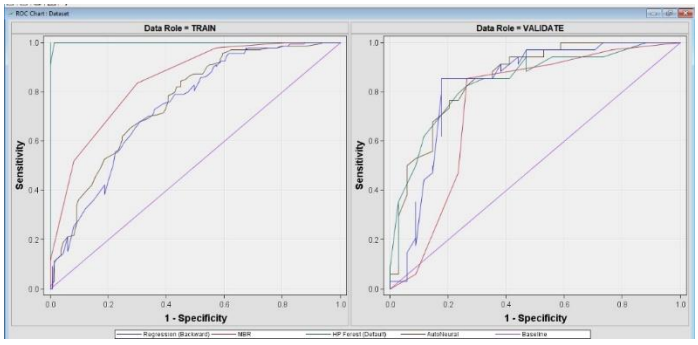

Figure 4: ROC chart

Figure 4 shows the ROC chart for each model. In order to evaluate the highest sensitivity level of the classification algorithm, the line graph should be higher than the baseline in the graph. The green graph line from the train graph is Random Forest, it shows highest value of overall classification algorithm and the second highest is *k*-NN which is the red graph line.

Table 4: Mode performance

Classification Algorithms	Valid Rate (Accuracy)	Train Rate	Train - ROC
Random Forest	99.779%	99.996%	1
K-NN	99.794%	99.766%	0.845
AutoNeural	99.779%	99.684%	0.759
Logistic Regression (Backward)	99.764%	99.680%	0.741

The Table 4 shows, Random Forest gives the overall best classification result than others including in ROC chart. Although, Random Forest give the best accuracy, the model is overfitting because the value of the misclassification rate (0.003759) for train is lesser than misclassification rate (0.220588) for valid. Overfitting in the model starts at the 30th tree in which it produces similar accuracy until the 100th tree. Therefore, Random forest is not the suitable algorithm to be use in this research. As compared to other algorithm, it is proven that K-NN gives a better accuracy of prediction with the accuracy of 99.794% and ROC train chart of 0.845.

6. Conclusions

In conclusion, predicting liver disease at the early stages is a crucial step in increasing the survival rate of liver disease patients before it becomes severe. The liver disease dataset is obtained from Andhra Pradesh, a region in India containing. It contains 583 records of patient either classified into liver patient and non-liver patient. Class imbalance is one of the issues found in the data set apart from missing values and high range. Hence, this research focuses on applying data mining algorithm on imbalanced dataset to predict the occurrence of liver disease in patients. Random sampling is done before applying data mining algorithms to ensure that the target class is balanced. This is important as it will affect the performance of each model.

K-Nearest Neighbour (k-NN) has the highest accuracy rate and ROC index as compared to other algorithm such as Logistic Regression, Random Forest and AutoNeural. The accuracy rate is 99.794%, the second highest algorithm is AutoNeural with 99.779% accuracy and lastly, Logistic Regression with backward selection method with an accuracy of 99.764%. Due to insufficient data in the data set, Random Forest algorithm is overfitted and is not the most suitable algorithm in this research despite having a perfect result from the ROC chart. Future research could be done on utilizing oversampling method to the dataset to address this issue.

References

[1] "Liver Cancer | CDC", Cdc.gov, 2018. [Online]. Available: https://www.cdc.gov/cancer/liver/index.htm. [Accessed: 25- July- 2018].

[2] "Global Burden Of Liver Disease: A True Burden on Health Sciences and Economies!! | World Gastroenterology Organisation", Worldgastroenterology.org. 2018. [Online]. Available: http://www.worldgastroenterology.org/publications/e-wgn/e-wgn-expert-point-of-view-articles-collection/global-burden-of-liver-disease-a-true-burden-on-health-sciences-and-economies. [Accessed: 25-July-2018].

[3] World Health Organization (2018), "Age-standardized death rates of liver cirrhosis". [Online]. Available: http://www.who.int/gho/alcohol/harms_consequences/deaths_liver_cirrhosis/en/. [Accessed: 26- July- 2018].

[4] American Liver Foundation, (2018), "Liver Disease Statistics". [Online]. Available: https://liverfoundation.org/liver-disease-statistics/ [Accessed on: 25-July-2018].

[5] Cleveland Clinic, (2018). "Understanding Liver Disease". [Online]. Available: https://my.clevelandclinic.org/ccf/media/files/Digestive_Disease/DDC_Liver_Brochure.pdf [Accessed on: 20-July-2018]

[6] Piedmont Healthcare, (2018), "How quickly liver can repair itself". [Online]. Available: https://www.piedmont.org/living-better/how-quickly-the-liver-can-repair-itself [Accessed on: 28-July-2018].

[7] Standford Medicine Report, (2017), "Harnessing the power of data in health". [Online]. Available: https://med.stanford.edu/content/dam/sm/sm-news/documents/StanfordMedicineHealthTrendsWhitePaper2017.pdf [Accessed on: 1-August-2018].

[8] Babu, B. V. R. and P. M. P., "Liver Classification Using Modified Rotation Forest", International Journal of Engineering Research and Development, vol. 1, pp. 17-24, June 2012.

[9] Sontakke, S., Lohokare, J. and Dani, R., (2017), "Diagnosis of Liver Diseases using Machine Learning", International Conference on Emerging Trends and Innovation in ICT (ICEI), p.129 – 133.

[10] Nahar, N. and Ara, F., (Feb 2018), "Liver Disease Prediction by using Different Decision Tree Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. 8, pp.1 – 9.

[11] Bendi Venkata Ramana, P. M. P. B. and P. N. B. V. , (Feb 2011), "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis", International Journal of Database Management Systems, vol. 3, pp. 101-114.

[12] Sharmila, S. L., Dharuman, C. and Venkatesan, P., (2017), "Disease Classification Using Machine Learning Algorithms - A Comparative Study", International Journal of Pure and Applied Mathematics. 114(6), pp.1–10.

[13] Jin, H., Kim, S. and Kim, J., (April 2014), "Decision Factors on Effective Liver Patient Data Prediction", International Journal of Bio-Science and Bio-Technology, vol. 6, pp.167 – 178

[14] Ramana, B. V., Babu, M. S. P. and Venkateswarlu, N. B., (Feb 2011),"A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis", International Journal of Database Management Systems (IJDMS), vol. 3, pp.101 – 114.

[15] Gulia, A., Vohra, R. and Rani, P., (April 2014), "Liver Patient Classification Using Intelligent Techniques", International Journal of Computer Science and Information Technologies, vol. 5, pp. 5110 – 5515.

[16] Pathan, A., Mhaske, D., Jadhav, S., Bhondave, R. and Rajeswari, K., (Feb 2018), "Comparative Study of Different Classification Algorithms on ILPD Dataset to Predict Liver Disorder", International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol 6, pp.388 – 394.

[17] Sindhuja, D. and Priyadarsini, (2016), "Liver disease analysis and accuracy prediction using machine learning techniques", International Science Press. 9(26). pp. 379-384.

[18] Priyadarsini, D. S. and R. J., "A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder", International Journal of Computer Science and Mobile Computing, pp. 483-488, 2016.

[19] Reena, P. and G., "Analysis of liver disorder using data mining algorithm", Global Journal of Computer Science and Technology, 2010.

[20] Ghosh, S. R. and Waheed, S., (2017), "Analysis of classification algorithms for liver disease diagnosis", Journal of Science, Technology and Environment Informatics, vol. 5, pp.360 – 370.

[21] Vijayarani, S. and Dhayanand, S., (2016), "Liver Disease Prediction using SVM and Naïve Bayes Algorithms", International Journal of Science, Engineering and Technology Research (IJSETR), vol. 4, pp.816 – 820.

[22] Pasha, M. and Fatima, M., (2017), "Comparative Analysis of Meta Learning Algorithms for Liver Disease Detection", Journal of Software, vol. 12, pp.923 – 933.

[23] Bashir, S., Qamar, U. and Khan, F. H., (2016), "IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework", Journal of Biomedical Informatics, pp. 185 – 200.

[24] Baitharu, T. P. and Pani, S. K., (2016), "Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset", International Conference on Computational Modeling and Security (CMS 2016), pp. 862 – 870.

[25] Han, J., Kamber, M. and Pei, J., (2012), "Data Mining Concepts and Techniques".

[26] Park and Hyeoun-Ae, "An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain", J Korean Acad Nurs, vol.43, No. 2, pp. 154-164, April 2013.

[27] Louppe, G., (2014), "Understanding Random Forest", University of Liege.

[28] Augmented Startups, (2017), "Random Forest – Fun and Easy machine learning".

[29] Siganos, C. S. and D., (2018), "Neural Network". [Online] Available: https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#Introduction%20to%20neural%20networks

[30] Tadeusz Lasota, M. M. and B. T., "Comparative Analysis of Neural Network Models for Premises Valuation Using SAS Enterprise Miner", pp.337-348,2009.

[31] Support SAS, (2017), "SAS The Power To Know". [Online] Available: http://support.sas.com/documentation/cdl/en/emgsj/61207/HTML/default/viewer.htm#n1docbb4hkr3nwn1ukqzqmme48yn.htm [Accessed 9-August-2018].

[32] SAS Official Website, (2018), "SAS". [Online]. Available: https://www.sas.com/en_us/home.html [Accessed on: 15-July-2018].

[33] Brownlee, J., (2016), "Overfitting and Underfitting with Machine Learning Algorithms". [Online]. Available: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/ [Accessed on: 9-August-2018].

[34] Synced ,(2017), "How Random Forest Algorithm Works in Machine Learning". [Online]. Available at: https://syncedreview.com/2017/10/24/how-random-forest-algorithm-works-in-machine-learning/ [Accessed on: 7 August 2018] .

[35] Agrawal A., (2017), "Logistic Regression Simplified".[Online]. Available: https://medium.com/data-science-group-iitr/logistic-regression-simplified-9b4efe801389 [Accessed on: 5 August 2018].

[36] Lane, D. M., (2018), "Log Transformations". [Online]. Available: http://onlinestatbook.com/2/transformations/log.html [Accessed on: 7-July-2018].

[37] Medline Plus,(2017), "Liver Disease". [Online]. Available: https://medlineplus.gov/liverdiseases.html [Accessed on: 18-July-2018].

[38] Press, G.,(2016), "Cleaning Big Data: Most of the time-consuming, least enjoyable data science task, survey says". [Online]. Available: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#499ec55e6f63 [Accessed on: 7-July-2018].

[39] UCSF Health, (2018), "Total protein". [Online]. Available: https://www.ucsfhealth.org/tests/003483.html [Accessed on: 16-July-2018].

[40] WebMD, (2018), "Liver Disease medical descriptions". [Online]. Available: https://www.webmd.com/ [Accessed on: 15-July-2018]

[41] Rumsey, J. D., (2018), “What P-value tells you about statistical data”. Available: https://www.dummies.com/education/math/statistics/what-a-p-value-tells-you-about-statistical-data/