SEYBOLD Report
ISSN: 1533 - 9211
https://seyboldreport.net/

ITERATIVE IMPUTATION PREPROCESSING TECHNIQUES FOR HANDLING MISSING DATA IN LIVER DISEASE PREDICTION

K.Sindhya
Research scholar ,Department of Computer Science, Rathnavel Subramaniam College of Arts and Science, Sulur, Coimbatore, Tamil Nadu, India

M.Suganya
Associate Professor and Head, Department of Information Technology, Rathnavel Subramaniam College of Arts and Science, Sulur, Coimbatore, Tamil Nadu, India

S.Santhana Megala
Assistant Professor School of Computer Studies, Rathnavel Subramaniam College of Arts and Science, Sulur, Coimbatore, Tamil Nadu, India

Abstract: - The liver serves as the body's primary organ for detoxifying toxins, which makes it crucial for ensuring survival. The liver may be harmed if it contracts a virus, is the target of the body's immunological response, or is exposed to toxins. The burden of liver illness in the nation is substantial given that India alone accounted for 18.3% of the two million deaths brought on by liver disease worldwide in 2015. Early discovery and the proper therapy at the appropriate time are the only ways to solve the problem. Clinical results are more dependent on data than models. It can be particularly challenging to identify the appropriate target (response variable) and features for classification problems in medical diagnostics. Another common problem in real-world data science applications is missing values in a data set. For handling the missing values two hybrid strategy were proposed, MSMOTE and MFSMOTE employing imputation algorithm along with SMOTE. As evaluation measures, confusion matrix, precision, recall, and f1-score were used. With an accuracy of 87.80% in predicting disease and 11.38% in predicting no disease, Extra Tree - MFSMOTE does well overall.
Keywords: Imputation Techniques, Missing Values, Preprocessing SMOTE, Liver Disease.

1. Introduction

The second-largest internal organ in the human body is the liver. It is essential to the human body's ability to produce protein, coagulate blood, and process cholesterol, glucose, and iron. Liver is important to maintain survival because it also functions to remove poisons from the body. Many bodily activities cannot be carried out effectively when the liver is not functioning, which results in serious harm to the body. If the liver becomes infected with a virus, is targeted by its own immune system, or is exposed to chemicals, it may suffer damage. Hepatotrophic viruses such the hepatitis B virus (HBV), hepatitis C virus, and hepatitis delta virus can lead to chronic liver damage, which can be fatal.

Hepatic disease is another name for liver disease. Hepatic illnesses cause symptoms like nausea, vomiting, weariness, back pain, fatigue, and weight loss. They also cause stomach pain and

swelling. Jaundice (a yellowing of the skin and eyes), fluid in an atypical cavity, pale stools, and particularly enlarged spleen and gallbladder are seen in certain people. Imaging and liver function tests can assess for liver damage and aid in the diagnosis of acute or chronic liver disorders [8]. Acute liver disease is defined as having a history of the condition for no longer than six months.

India is quickly realizing that liver ailments are a public health priority. Due to the fact that India alone was responsible for 18.3% of the two million deaths caused by liver disease worldwide in 2015, the burden of liver disease in the country is enormous. Since 1980, chronic liver diseases (CLDs), which include cirrhosis and its consequences, have contributed more to mortality in India than in China, the other Asian nation with a sizable population, where it has remained steady or even decreased (Fig.1.1).

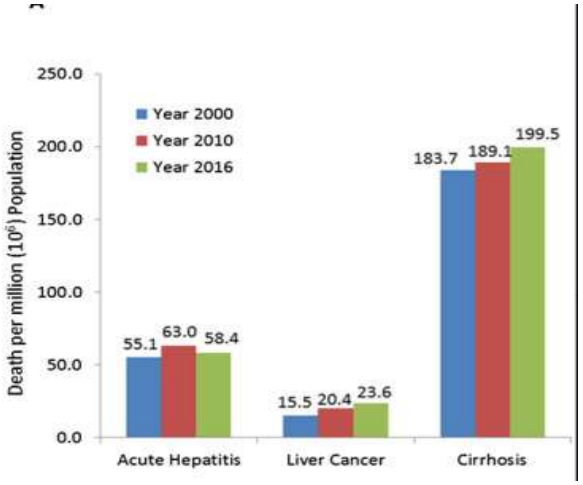


Fig: 1.1 Death Rate due to Liver Disease

Important demographic drivers of this trend include the country's growing population and rising life expectancy. Encouragement-based health system response techniques [5] have been adopted in India in line with this. Since 2018, India has had a federally sponsored national viral hepatitis control programme in place, which consists of both early detection (linkage to care, screening populations at risk, drug distribution, and CLD surveillance) and preventive (vaccination, blood safety) efforts. Additionally, a national NAFLD control programme was just begun in 2021 with initiatives that incorporate the management of liver disease more generally into a programme for the management of other non-communicable illnesses [4].

Epidemiology of liver disease is changing in India. Improved preventative measures awareness, more effective connection to care for early-stage liver disease, and better screening techniques all potentially be valuable interventions.

The only way to resolve the issue is by early detection and appropriate treatment at right time. But because it involves skilled medical professionals and takes a long time, diagnosing liver problems is always a difficult task. Machine learning is crucial in the diagnosis and treatment of diseases to help healthcare professionals. It can be utilized to draw out useful data from medical datasets and create a model to pinpoint the patients. Machine learning and data mining techniques have been used in numerous studies to identify people with liver problems.

1.1.Motivation

The clinical outcomes rely more on data than they do models. Finding the right target (response variable) and attributes for classification issues in medical diagnostics is particularly difficult [6]. Missing values in a data set are another frequent issue in data science applications in real-world settings. In medical research, missing data are a persistent issue that can be caused by a variety of factors, including participant dropouts or mistakes made by lab staff. Missing data reduce statistical power and increase the risk of bias in medical research [3]. This research aims at reduce the wrong prediction through missing data.

1.2.Contribution

- To handle the missing values at the pre-processing stage an two Imputation techniques introduced MSMOTE and MFSMOTE.
- Oversample the minority class using the synthetic minority oversampling technique (SMOTE) to control over fitting
- Analyze various ML methods to judge which one performs better at diagnosing liver disease.

1.3.Organization of paper

The study aims at reducing the missing value issues which is a greater risk of predictions. This paper is organized in following way as well: Section 2 discusses the related works, Section 3 deals with methodology of research proposed and the Section 4 describes about the result and analysis. Finally, Section 5 ends up with conclusion.

2. Related Works

In statistical analysis, missing data are frequently encountered, and imputation techniques based on random forests (RF) are increasingly common for addressing missing data, particularly in biomedical research. When there are non-normally distributed variables, interactions, or non-linearities and the data are MAR, (Hong & Lynn, 2020) [2] compares the imputation accuracy of missForest and CALIBERrfimpute. A case study based on clinical information for patients with hepatocellular carcinoma and a series of simulation experiments were used to conduct the assessment (HCC). Although RF-based imputation can have strong predictive accuracy, the findings of the simulated experiments and the case study demonstrate that it can also result in significantly biassed inference when the imputed variables are utilised in future regression studies. With outcome-dependent MAR and highly skewed data, Missforest can perform poorly since its predicted value does not go beyond the range of the imputed variable's observed values. Although CALIBERrfimpute can somewhat mitigate this issue by taking samples from the conditional distribution of the RF projected value, regression results still suffer from significant bias and inadequate confidence interval coverage. As a result, RF-based imputation, particularly missForest, should not be utilised as a magic bullet to replace missing data.

In the four discrete real-world datasets, (Cihan & Ozger, 2019) [1] proposes a novel approach for missing value imputation based on the Artificial Bee Colony algorithm. Bayesian Optimization is incorporated into the Artificial Bee Colony algorithm at the proposed Artificial Bee Colony Imputation (ABCimp) approach. The suggested method's performance is contrasted with that of the other six well-known approaches: Mean, Median, k Nearest

Neighbor (k-NN), Multivariate Equation by Chained Equation (MICE), Singular Value Decomposition (SVD), and MissForest (MF). The Naive Bayes algorithm is employed as the classifier, and the classification error and root mean square error are used as assessment criteria for the performance of the imputation methods. The empirical findings demonstrate that, at varied missing rates ranging from 3% to 15%, state-of-the-art ABCimp outperforms the other most widely used imputation techniques.

(Sovilj et al., 2016) [12], regression estimation with missing data is a topic of discussion. First, original data with missing values are subjected to a combination of Gaussians. Second, a multiple imputation approach is used to perform a large number of imputations. A unique approach based on the Gaussian Mixture Model and Extreme Learning Machine is created to generate accurate estimations for the regression function (approximation). The data distribution is modeled using a Gaussian Mixture Model that is modified to handle missing values, and an Extreme Learning Machine is utilized to create a multiple imputation technique for the final estimation. The final estimation is improved above the mean imputation that was conducted just once to complete the data by using multiple imputation and an ensemble technique over numerous Extreme Learning Machines. Compared to basic methods, the proposed methodology requires more time to run, but the overall improvement in accuracy justifies this trade-off.

(Rahman & Islam, 2013) [7], provide two novel methods for the imputation of missing variables, both categorical and numerical. The methods make use of decision trees and forests to locate horizontal data sets segments where records have stronger attribute and similarity correlations. The missing data are then imputed using the correlations and similarity measures. An innovative method is used to combine several segments in order to improve the quality of the imputation. We empirically compare our strategies with a few already existing ones using nine publicly accessible data sets and four widely used evaluation criteria. The experimental results show that our strategies, which are based on statistical analyses like confidence intervals, are clearly superior.

The process of calculating values for missing data elements is known as data imputation(Silva-Ramírez et al., 2015) [10] focuses on the creation of automated data imputation models for monotone patterns of missing values that are based on artificial neural networks. It suggests two imputation methods: a single method that uses a multilayer perceptron trained under distinct learning rules, and a multiple method that uses a multilayer perceptron combined with k-nearest neighbors. A perturbation experiment using monotone missing data patterns generated at random was conducted on 18 actual and virtual databases. On these data sets, an empirical test was conducted using both methodologies (single and multiple imputations), as well as three traditional single imputation techniques: mean/mode imputation, regression, and hot-deck. Therefore, five imputation techniques were used in the tests. The findings, taking into account several performance metrics, showed that, in comparison to conventional tools, both ideas improve the automation level and data quality while providing an acceptable degree of performance.

3. Research Methodology

The proposed model is designed in such a way presented in fig. 3.1.

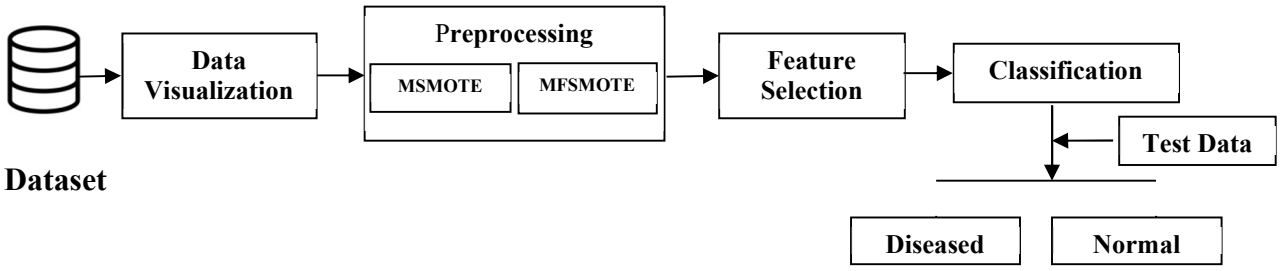


Fig: 3.1 Proposed Systems

3.1. Data Visualization

3.1.1. Dataset

The UCI ML Repository provided the multivariate dataset. The data collection includes demographic information, such as age, as well as laboratory results for blood donors and Hepatitis C patients. Category (blood donors vs. Hepatitis C and its progression, including "simply" Hepatitis C, Fibrosis, and Cirrhosis) is the desired attribute for classification. The dataset has 14 columns and a total of 615 entries (as shown in fig. 3.2).

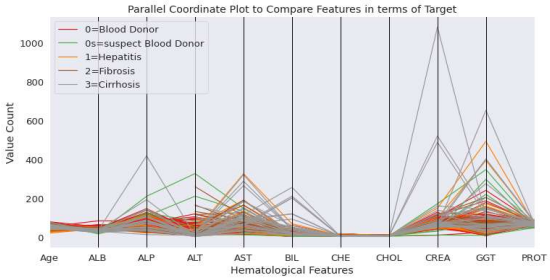
	Unnamed: 0	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
0	1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
1	2	0=Blood Donor	32	m	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
2	3	0=Blood Donor	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.20	86.0	33.2	79.3
3	4	0=Blood Donor	32	m	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
4	5	0=Blood Donor	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7

Fig: 3.2 Dataset

3.1.2. Feature Distribution

To comprehend the distribution of features both individually and in terms of goal features, visualize the dataset. For the comparison of numerical features with the target feature, a parallel coordinate plot was used. We can observe from the Parallel Coordinate plot that an excessively high value (for example, >200) denotes the presence of liver illness. Additionally, it demonstrates that when the values are close to or above 200, the likelihood of the disease stage being cirrhosis is higher than that of other stages (as shown in fig. 3.3). Sankey Plot was used to display the mean distribution for many Categories. From the fig 3.4, it means for the severable variables "cirrhosis," "fibrosis," "hepatitis," and "suspect disease" are significantly higher or lower than the means for the severable variables "no disease."

ISSN: 1533 - 9211



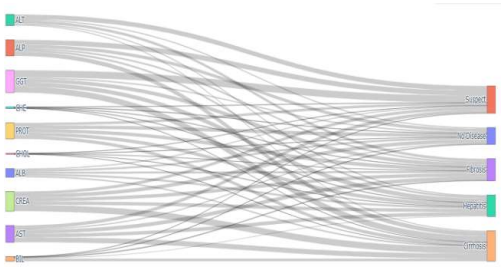
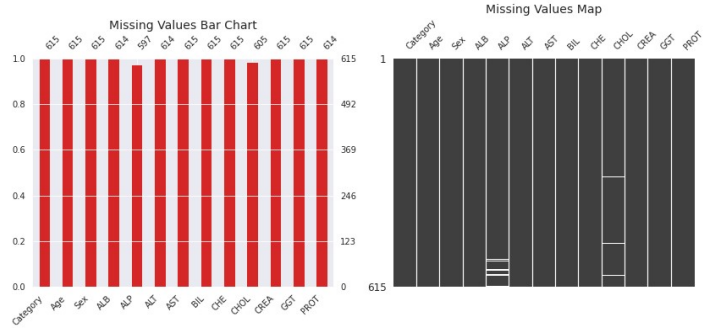


Fig: 3.3 Comparison of features in terms of Target **Fig: 3.4 Mean distributions of different categories**

3.2.Preprocessing

At the preprocessing stage, all the fundamental preprocessing operations were carried out, including the elimination of the unnamed field and the testing for duplicate rows and unique values. From the observation it is clear that the dataset having Missing values and Outlier. Fig 3.5 visualized the missing values and fig 3.6 shows the outlier of the dataset. Several values in the ALP, ALT, AST, BIL, CREA, and GGT can clearly be characterized as extreme values, which suggest that there may be outliers. Extreme values in features typically lead to the diagnosis of hepatitis.



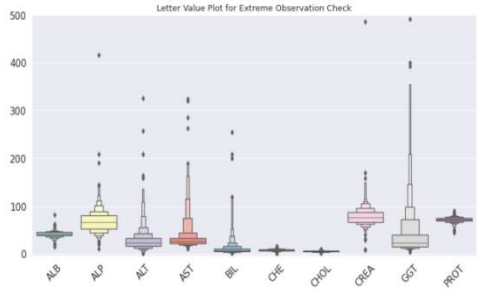


Fig: 3.5 Missing Values **Fig: 3.6 Outliers**

Imputation is used to handle missing values (Iterative Imputer-MICE and MissForest) [9]. By examining information from other columns and attempting to estimate the best prediction for each missing value, the Multivariate Imputation By Chained Equations approach, or MICE, makes it simple to impute missing values in a dataset. Another machine learning-based data imputation algorithm that uses the Random Forest algorithm is called MissForest. Missing values are imputed using "missForest," especially when mixed-type data is involved. It can be used to infer continuous and/or categorical data, including intricate relationships and nonlinear ones. Fig. 3.7 shows that neither the MICE nor the MissForest Imputed datasets show a significant change in distribution following imputation. After imputation, the extreme value observations are identical. Similar numbers of extreme observations have been made. These unusual observations may be the result of human or mechanical error.

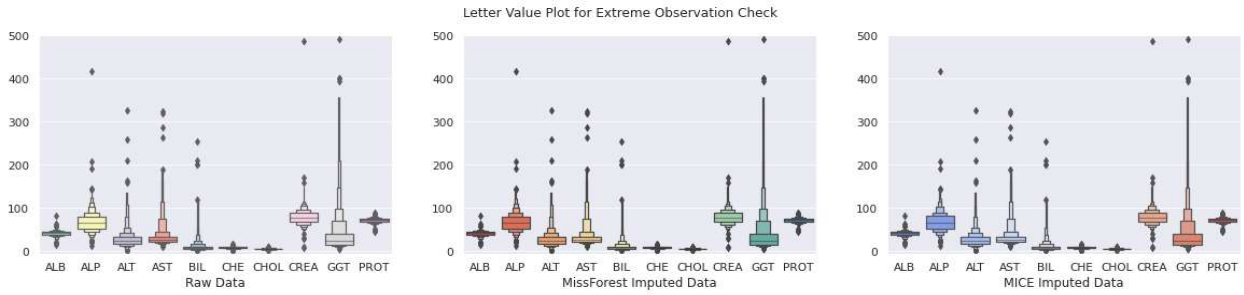


Fig: 3.7. Observation after Imputation

There is no normal distribution in the sample data and the model overfits. To handle the overfitting SMOTE is used. A statistical method for evenly expanding the number of cases in dataset is Synthetic Minority Oversampling Technique (SMOTE) [11]. The component creates new instances from minority situations that you specify as input that already exist.

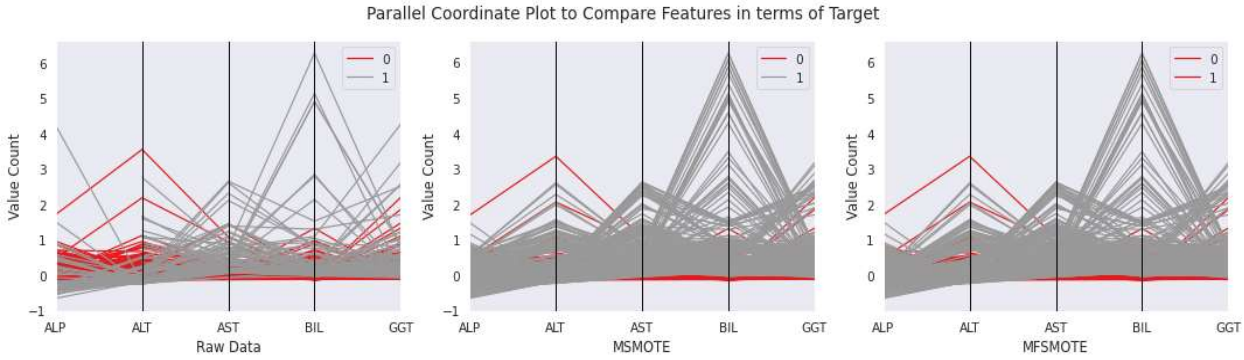


Fig. 3.8 Observation after SMOTE

After SMOTE, the datasets differed; some features were given varying degrees of priority in datasets as shown in fig. 3.8. Disease categories are more resilient than realize. The mean for "Disease" in characteristics is much higher/lower than the mean for "No Disease," as can be seen. After SMOTE, the dataset is now balanced across features, as is evident.

3.3.Feature Selection

Features selected from the imputed dataset based on the importance. Fig. 3.9 shows the features importance from the imputed dataset of MICE and missForest.

	Features	Importance% of MICE_df	Importance% of MissForest_df	Selected_by_RFE
0	Age	3.739862	3.804882	False
1	Sex	0.248680	0.282146	False
2	ALB	2.328031	2.239931	False
3	ALP	8.440734	8.711602	True
4	ALT	12.770949	12.887889	True
5	AST	33.391208	32.943828	True
6	BIL	7.932497	7.641219	True
7	CHE	8.213410	8.102146	False
8	CHOL	4.075872	4.354860	False
9	CREA	3.434964	3.210800	False
10	GGT	10.574358	10.883434	True
11	PROT	4.849435	4.937262	False

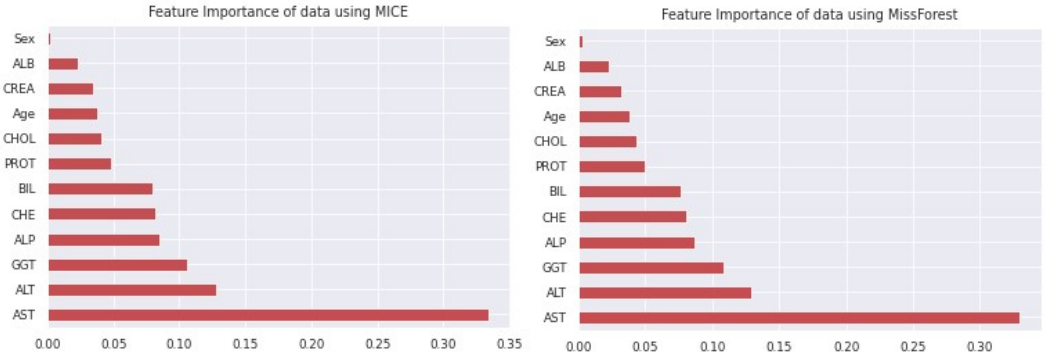


Fig. 3.9 Feature Importance

3.4.Classification

Model is trained for classification with 3 base classifiers such Random Forest, Extra Tree and Multilayer Perceptron. Training the imputed and SMOTE data on the classifier and validated the best performing classifier.

3.4.1. Random Forest

The broad category of ensemble-based learning techniques includes random forest classifiers. They are easy to set up, operate quickly, and have had great success across a wide range of industries [13]. The main idea behind the random forest method entails building a lot of "simple" decision trees during training and using a majority vote (mode) across them for classification. This voting method corrects for the unfavourable tendency of decision trees to overfit training data, among other things. Random forests apply the general bagging strategy to each individual tree in the ensemble during the training phase. Bagging continually chooses a random sample from the training set with replacement and then fits trees to these samples. No pruning is done as the trees grow. A free parameter that is easily learned automatically utilising the so-called out-of-bag error is the ensemble's number of trees.

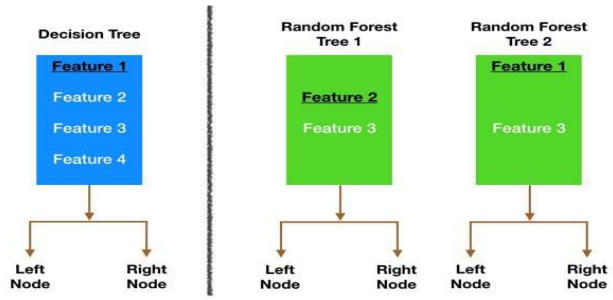


Fig: 3.10. Node Splitting in Random Forest

When splitting a node in a typical decision tree, all potential features and choose the one that results in the greatest gap between the observations in the left node and those in the right node. In contrast, only a random subset of features is available to each tree in a random forest (as shown in fig.3.10). In the end, this leads to less correlation between trees and increased diversity by forcing even more variety across the model's trees.

3.4.2. Extra Tree

The Extra Tree Regression (ETR) approach was proposed by Geurts et al. and was built from the Random Forest (RF) model. The Extra Tree Regression (ETR) algorithm creates a collection of unpruned judgements or regression trees (as shown in fig. 3.11) in accordance with the traditional top-down method [15]. The Train Using AutoML tool employs the ensemble supervised machine learning technique known as extra trees, sometimes known as excessively randomised trees, which makes use of decision trees. Extremely Randomized Trees Classifier, also known as Extra Trees Classifier, is a form of ensemble learning technique that combines the findings of various de-correlated decision trees gathered in a "forest" to produce its classification outcome. The only way it differs conceptually from a Random Forest Classifier is in how the decision trees in the forest are built.

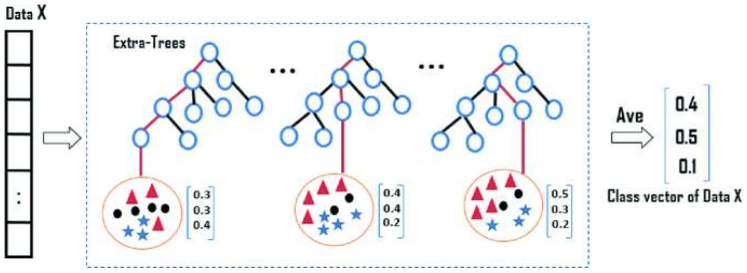


Fig: 3.11. Extra Tree

3.4.3. Multilayer Perceptron

MLP is the abbreviation for multi-layer perception. It is made up of dense, completely connected layers that may change any input dimension into the desired dimension. A neural network with numerous layers is referred to as a multi-layer perception. In order to build a neural network, we combine neurons so that some of their outputs are also their inputs. As seen in multi-layer perceptron fig. 3.12 includes three inputs, which results in three input nodes, and three nodes for the hidden layer. There are two output nodes since the output layer produces two outputs. In the diagram above, the nodes in the input layer forward their output to each of the three nodes in the hidden layer, and in a similar manner, the hidden layer processes the data before sending it to the output layer. The nodes in the input layer receive input and forward it for further processing.

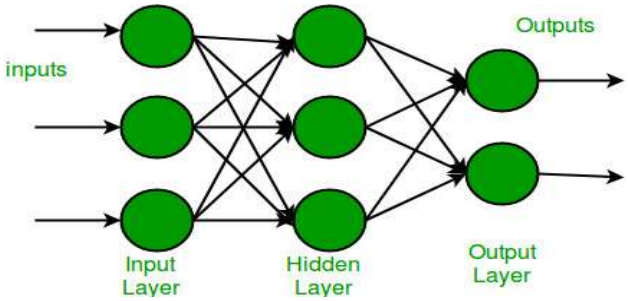


Fig: 3.12. Multilayer Perceptron

4. Result and Analysis

To analysis the performance of the model confusion matrix is used. Precision, Recall and F1 Score are used as the evaluation metrics of the models.

4.1. Confusion Matrix

Confusion matrix [14] is used to gauge how well categorization models work. It can be applied to multinomial and binary classification. Positive and negative values for the target variable are shown in the matrix as true positive (TP), true negative (TN), false positive (FP), and false negative (FN). In TP, the model predicted a positive outcome, and the actual outcome was positive, in TN, the model projected a negative outcome, and in FP, the model expected a positive outcome, but the actual outcome was negative. FP is another name for a type 1 mistake. FN's model projected a negative outcome, but the outcome was positive. FN is another name for type 2 error.

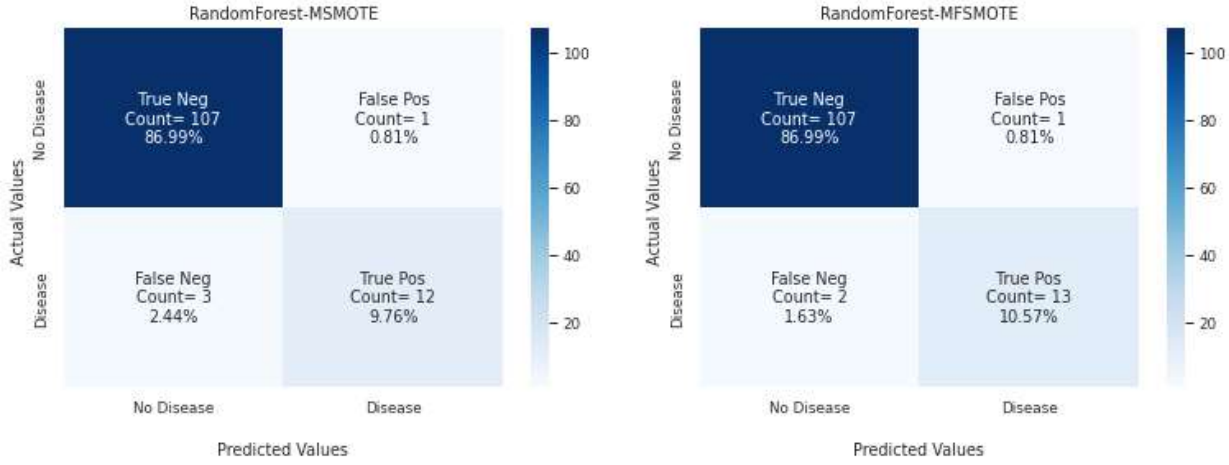


Fig: 4.1 Confusion Matrixes for Random Forest

From fig. 4.1 it is clear that Random Forest- MFSMOTE performs well and gives a better accuracy of prediction.

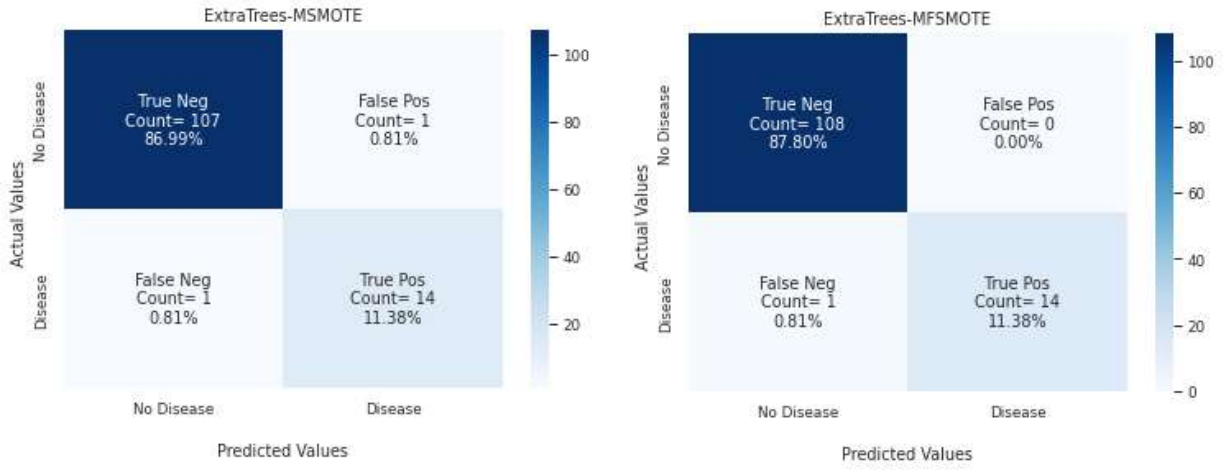


Fig: 4.2 Confusion Matrixes for Extra Tree

From fig. 4.2 it is clear that Extra Tree- MFSMOTE performs well and gives a better accuracy of prediction.

ISSN: 1533 - 9211

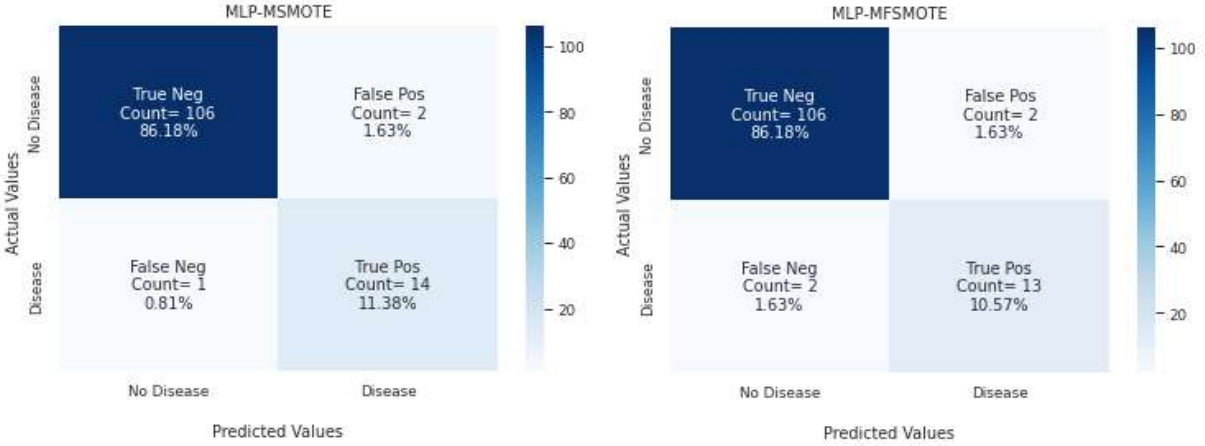


Fig: 4.3 Confusion Matrixes for Multilayer Perceptron

From fig. 4.3 it is clear that MultiLayer Perceptron- MSMOTE performs well and gives a better accuracy of prediction. Overall the Extra-Tree - MFSMOTE performs better with an accuracy of 87.80% and 11.38% for No Disease and Disease respectively.

4.2.Evaluation Metrics

Based on the evaluation metrics such as precision, recall and f1-score Extra Tree – MFSMOTE outperforms than the other classifiers. Which also have a precision of 0.99, recall of 1 and f1-score of 0.995.

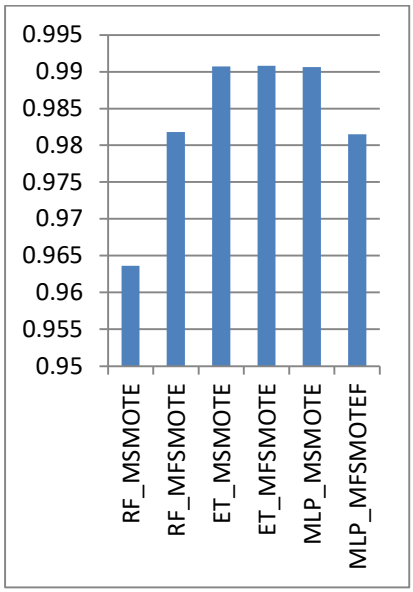


Fig: 4.4 Precision

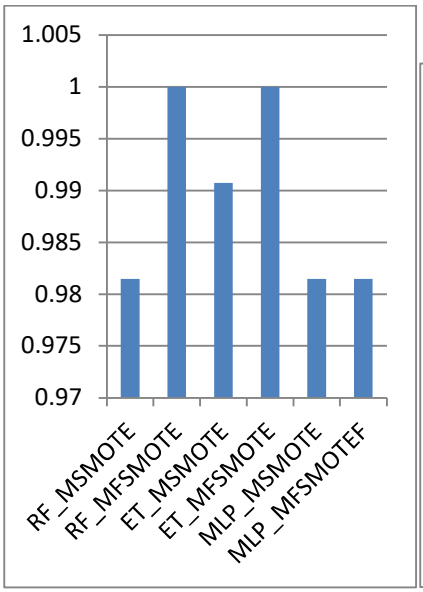


Fig: 4.5 Recall

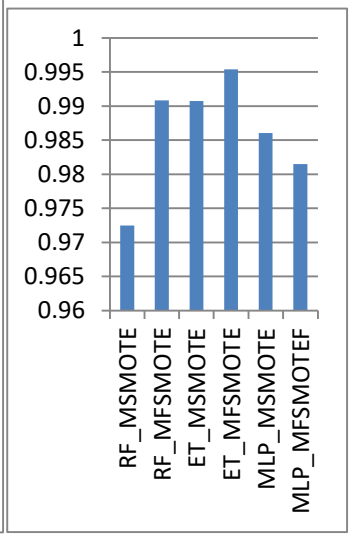


Fig: 4.5 F1-Score

5. Conclusion

The most frequent cause of death and one of the most hazardous diseases, liver disease has been noted to be. The goal here is to forecast early detection by using classifiers and handling the missing values. In order to solve the issues with missing value imputation, a new approach is created. Model suggested MSMOTE and MFSMOTE using the imputation algorithm, and

classifiers to create a robust method to estimate the missing data. Utilizing the Synthetic Minority Oversampling Technique, over fitting of the data was addressed (SMOTE). Imputed data trained with base classifiers such as Random Forest, Extra Tree and Multilayer perceptron. Overall the Extra Tree with MFSMOTE outperforms well both in disease and no disease prediction. It holds a c\score of 87.80% of accuracy in predicting as Disease, 11.38% as No Disease, which also have a precision of 0.99, recall of 1 and f1-score of 0.995.

Reference

1. Cihan, P. and Ozger, Z.B. (2019) "A new heuristic approach for treating missing value: ABCimp," Elektronika ir Elektrotechnika, 25(6), pp. 48–54. Available at: https://doi.org/10.5755/j01.eie.25.6.24826.
2. Hong, S. and Lynn, H.S. (2020) "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction," BMC Medical Research Methodology, 20(1). Available at: https://doi.org/10.1186/s12874-020-01080-1.
3. Keerthana, P.S.M. et al. (2020) "A prediction model of detecting liver diseases in patients using logistic regression of machine learning," SSRN Electronic Journal [Preprint]. Available at: https://doi.org/10.2139/ssrn.3562951.
4. Mohammed Majeed Hameed, Mohamed Khalid AlOmar, Faidhalrahman Khaleel, Nadhir Al-Ansari, "An Extra Tree Regression Model for Discharge Coefficient Prediction: Novel, Practical Applications in the Hydraulic Sector and Future Research Directions", Mathematical Problems in Engineering, vol. 2021, Article ID 7001710, 19 pages, 2021. https://doi.org/10.1155/2021/7001710
5. Mondal, D., Das, K. and Chowdhury, A. (2022) "Epidemiology of Liver Diseases in India," Clinical Liver Disease, 19(3), pp. 114–117. Available at: https://doi.org/10.1002/cld.1177.
6. Mostafa, F. et al. (2021) "Statistical machine learning approaches to liver disease prediction," Livers, 1(4), pp. 294–312. Available at: https://doi.org/10.3390/livers1040023.
7. Rahman, M.G. and Islam, M.Z. (2013) "Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques," Knowledge-Based Systems, 53, pp. 51–65. Available at: https://doi.org/10.1016/j.knosys.2013.08.023.
8. Razali, N. *et al.* (2020) "A data mining approach to prediction of liver diseases," *Journal of Physics: Conference Series*, 1529(3), p. 032002. Available at: https://doi.org/10.1088/1742-6596/1529/3/032002.
9. Richman, M.B., Trafalis, T.B. and Adrianto, I. (2009) "Missing data imputation through machine learning algorithms," Artificial Intelligence Methods in the Environmental Sciences, pp. 153–169. Available at: https://doi.org/10.1007/978-1-4020-9119-3_7.
10. Silva-Ramírez, E.-L., Pino-Mejías, R. and López-Coello, M. (2015) "Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron

and K-nearest neighbours for monotone patterns," Applied Soft Computing, 29, pp. 65–74. Available at: https://doi.org/10.1016/j.asoc.2014.09.052.
11. Singh, J., Bagga, S. and Kaur, R. (2020) "Software-based prediction of liver disease with feature selection and classification techniques," *Procedia Computer Science*, 167, pp. 1970–1980. Available at: https://doi.org/10.1016/j.procs.2020.03.226.
12. Sovilj, D. et al. (2016) "Extreme learning machine for missing data using multiple imputations," Neurocomputing, 174, pp. 220–231. Available at: https://doi.org/10.1016/j.neucom.2015.03.108.
13. T. Pratheebhda, Mrs. V. Indhumathi, Dr. S. Santhana Megala, "An Empirical Study On Data Mining Techniques And Its Applications", 2021.
14. V. Indumathi, S.Santhana Megala, R.Padmapriya, Classify Bully Text With Improved Classification Model Using Grid Search With Hyperparameter Tuning, Advances and Applications in Mathematical Sciences, Vol. 21, Is. 9, pp. 4973-4980, 2022.
15. V. Indumathi and S. Santhanamegala, Enhanced Classification Model on CyberBullying detection with Multi-Label, 2021.