Group 8

Liver Disease Prediction System

A report submitted as per the requirements of ST 3082 / DS 3003

Tharindu Fernando – 15522
Hiruni Kudagama – 15680
Kaveesha Vidushinie - 15572

Abstract

Currently liver disease pose a significant threat on the health of humans worldwide. Liver issues sometimes go unnoticed, and they are a silent killer with a huge impact. This proves the necessity of an accurate and efficient system able to predict the occurrence of liver disease. This report executes a comprehensive advanced analysis of liver disease data as means of addressing this global health issue. The findings of this project will enable early detection of disease ensuring timely medical intervention.

The data considered for analysis were taken from the online platform Kaggle and contains various attributes of patients with and without liver disease. These were mainly clinical variables. Individual variable analysis along with multivariate analysis performed to acquire a thorough understanding of the data. This project utilizes statistical methods and machine learning techniques to identify influential characteristics which are then used for liver disease prediction.

Multiple models are fitted to the data to address our objective of predicting liver disease and the effectiveness of these models are explored by comparing different evaluation metrics. The results show that the model built with the random forest algorithm on the original data with 100% accuracy is the best model for liver disease prediction.

Table of Contents

Abstract ii
Table of Contents ii
1. Introduction 4
2. Description of the Question 4
3. Description of the dataset 4
Data preprocessing 5
4. Important results of descriptive analysis 5
5. Important results of advanced analysis 7
5.1 Selected models and justification of choice 7
5.2 Evaluation of Classification Models on Original Training Set. 8
5.3 Handling Class Imbalance and Evaluation of Models on Data Set after applying different sampling techniques 9
5.4 Handling Multicollinearity with PCA 11
5.5 Clustering Approach 11
5.6 Selection of best model 14
5.7 Selection of optimal model 16
6. Output of study 17
7. Discussion and conclusions 18
8. Appendix 18

List of Figures

Figure 1 Pie Chart Of Dicease Distribution 6
Figure 2 Distibution Og Gender By Dicease 6
Figure 3 Explained Variance Ratio Vs Pc's 11
Figure 4 Scree Plot 11
Figure 5 Silhoutte Plot - Kprototype Clusters 12
Figure 6 Wcss Vs Clusters - Kprototype 12
Figure 7 Wcss Vs Clusters - Kmeans 12
Figure 8 Silhoutte Plot - Kmeans Clusters 12
Figure 9 Wcss Vs Clusters - Kmedioids 13
Figure 10 Figure 8 Silhoutte Plot - Kmedioids Clusters 13
Figure 11 Important Feature Selection Under 1st Model 15
Figure 12 Important Feature Selection Under 2nd Model 15
Figure 13 Important Feature Selection Under 3rd Model 16

List of Tables

Table 1 Description Of Variables 5
Table 2 Comparison Between Means Of The Clinical Variables With Healthy Livers And Unhealthy Livers 6
Table 3 Performance Metrics Obtained From Original Data Set 8
Table 4 Performance Metrics Obtained From Smote Data Set 9
Table 5 Performance Metrics Obtained From Downsampled Data Set 10
Table 6 Class Distribution Withing Clusters - Kprototype 12
Table 7 Class Distribution Withing Clusters - Kmeans 13
Table 8 Class Distribution Withing Clusters - Kmedioids 13
Table 9 Summary Of Best Model 14
Table 10 Selection Of Optimal Model After Feature Selection 16

1. Introduction

A significant global health issue, the prevalence of liver disease is on the rise. Existing diagnostic methods are limited, and they require specific equipment and trained individuals. The golden standard of diagnosing liver disease currently is liver biopsy which is an invasive procedure. Thus the need of an accurate and interpretable model for liver disease prediction is evident. Early detection of liver disease enables timely intervention helping medical professionals treat patients efficiently. This project engages in a descriptive analysis and advanced analysis of a diverse dataset of individuals with and without liver disease. We utilized prior literature on the research topic to gain a good understanding of the domain and data and conducted analysis to fit multiple models and identify the best model to predict the occurrence of liver disease. This is achieved using the Python programming language and the Jupyter Notebook platform. A data product capable of efficiently predicting liver disease is the expected output of this project. The findings of this project will be of importance to doctors and patients alike as well as other interested parties such as pharmaceutical companies.

2. Description of the Question

This report aims to explore the capability of various machine learning algorithms to predict occurrence of liver disease in a diverse set of patients. It targets identifying best model for the same purpose. Medical professionals can use the identified best model to navigate treatment plans, and researchers and pharmaceutical companies with a focus on preventive medicine will also benefit from the findings of this research. However, our target audience for this project are medical professionals as we aim to enable early detection of disease ensuring timely intervention. This project does not address "how" to treat liver disease but rather "when" to treat it.

The main questions that this project aims to answer are:

- Which variables have a higher impact on accurately predicting liver disease and to what extent?
- What machine learning models can predict liver disease in patients and which is the best model for the same?

3. Description of the dataset

The dataset considered contains information of 30689 individuals with and without liver disease. Each person in the dataset has 11 attributes which are known to be significant contributors to liver disease. Majority if these variables are clinical variables. A brief description of them are as follows:

	Variable	Description
1	Age of the patient	Age of the patient
2	Gender of the patient	Gender of the patient
3	Total Bilirubin	Total amount of direct and indirect bilirubin (in mg/dl)
4	Direct Bilirubin	Amount of direct bilirubin (in mg/dl)
5	Alkphos Alkaline Phosphotase	Amount of Alkaline Phosphotase (which is an enzyme produced by the liver) present (in iu/l)
6	Sgpt Alamine Aminotransferase	Amount of Alamine Aminotransferase (SGPT) present (in iu/l)
7	Sgot Aspartate Aminotransferase	Amount of Aspartate Aminotransferase (SGOT) present (in iu/l)
8	Total Proteins	Amount of total proteins present (in mg/dl)
9	ALB Albumin	Amount of Albumin (which is a protein produced by the liver) present (in mg/dl)
10	A/G Ratio Albumin and Globulin Ratio	The Albumin and Globulin Ratio
11	Result	Whether the individual has liver disease or not 1: Liver patient ; 2: Not a liver patient

TABLE 1 DESCRIPTION OF VARIABLES

Data preprocessing

Prior to analyzing the data, it needed to be preprocessed to derive meaningful and clear interpretations. This was achieved by performing the following:

- The response variable, "Result" was renamed as "Liver_disease" and recoded to 1's and 0's (1 signifying has liver disease and 0 marking does not have liver disease) for easier usage.
- The columns were renamed to remove blank spaces in the column names for easier handling.
- The count of missing values were checked. Total missing values accounted for only 9% of the total data. Therefore, missing values were dropped as there was no significant information loss when done so.
- The data was split into train and test sets and the following were performed on training data.
- Duplicate was investigated
- Gender variable is converted into binary variable (1:male 2: female)

4. Important results of descriptive analysis

- Descriptive analysis of the response variable, "price" helped understand it better. Figure 5.1 reveals its unbalanced Ness, which indicates that higher number of individuals in the dataset do not have liver disease. A few different techniques, namely SMOTE, upsampling and downsampling were used to balance the data.

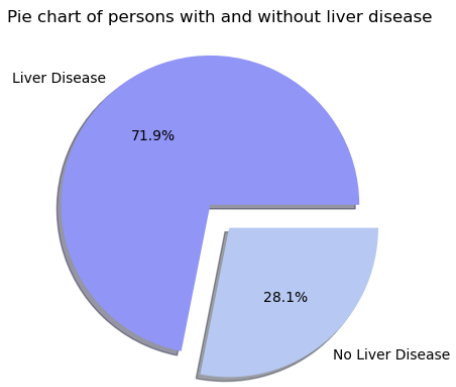

FIGURE 2 PIE CHART OF DICEASE DISTRIBUTION

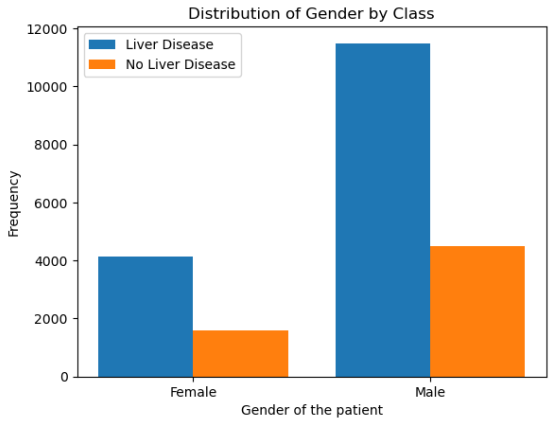

FIGURE 1 DISTIBUTION OG GENDER BY DICEASE

- Majority of patients with and without liver disease are male, which aligns with the findings of previous researchers. According to BMC public health, interestingly the death rate among males was estimated to be 1.51 times higher than that of females globally.
- It was noted that total bilirubin increases as direct bilirubin increase, which is to be expected as the direct bilirubin amount is included in the total bilirubin amount. There is a positive relationship between these two variables despite the gender of the individual and the health of their liver.
- The distribution of the 'Age' variable is approximately normal as opposed to the distributions of the most other quantitative variables, which are all right skewed. This suggests that majority of the individuals included in the dataset have smaller values for most of the clinical variables.
- However, 'Albumin' and 'Total proteins' also have relatively normal distributions, whether a patient has liver disease or not.
- Comparison between means of the clinical variables of the individuals with healthy livers and unhealthy livers is seen in table 4.1.

Chemicals	Diseased	Healthy
A/G Ratio Albumin and Globulin Ratio	0.9112900953966323	1.022018994596365
ALB Albumin	3.047058070298995	3.320386441788112
Alkphos Alkaline Phosphotase	318.034573276138	221.15621418044867
Direct Bilirubin	1.9949932774185288	0.39235303749795325
Sgot Aspartate Aminotransferase	140.71771560279146	41.18781725888325
Sgpt Alamine Aminotransferase	100.76663038606824	34.07057475028655
Total Bilirubin	4.326102823484218	1.1365645980022925
Total Protiens	6.456751392534733	6.523268380546913

TABLE 2 COMPARISON BETWEEN MEANS OF THE CLINICAL VARIABLES WITH HEALTHY LIVERS AND UNHEALTHY LIVERS

- Based on results acquired upon performing chi squared test for association we discovered that all variables except for ‘Gender’ are significantly associated with the response variable ‘Liver disease’. Further, there is a strong relationship between ‘Total proteins and ‘Albumin’, and ‘Direct Bilirubin’ and ‘Total Bilirubin’, indicating multicollinearity in the data.

5. Important results of advanced analysis

The data set was split into training (80% of original data) and test (20%) tests in the pre-processing stage and a few select models were fitted on the training data and tested on the test data.

5.1 Selected models and justification of choice

1. **Logistic Regression**: A simple, interpretable technique that provides probabilities of class membership, makes it useful for understanding the likelihood of disease presence based on various features.
2. **Support Vector Machine (SVM)**: SVMs are able to find optimal hyperplanes to separate classes, potentially capturing intricate decision boundaries between liver disease and non-liver disease patients.
3. **Decision Tree**: Decision trees are intuitive and capture non-linear relationships between features and the target variable, making them suitable for medical diagnosis tasks like predicting liver disease. They can handle both numerical and categorical data.
4. **Random Forest**: Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and generalization and is robust to overfitting.
5. **XGBoost**: A powerful gradient boosting algorithm, it is known for its performance and scalability. Effectively handling imbalanced datasets and it captures complex interactions between features, making it well-suited for predicting liver disease where class distribution may be highly skewed.
6. **KNN**: KNN is simple and easy to understand, thereby well suited for initial exploration of the data and potentially capturing localized patterns relevant to liver disease diagnosis.
7. **Adaboost**: Adaboost is an ensemble learning method that combines multiple weak classifiers to create a strong classifier. Adapting to the complexity of the dataset, it potentially improves predictive performance when data is highly imbalanced.

8. **Gradient boost**: This method builds a series of decision trees sequentially, with each tree correcting errors made by previous ones.
9. **MLP (Multi-Layer Perceptron):** MLP is an artificial neural network capable of learning complex relationships between input features and the target variable. It's apt for predicting liver disease where there are non-linear relationships and interactions between multiple variables.

5.2 Evaluation of Classification Models on Original Training Set.

Upon completion of preprocessing, we fitted the above models to the data. The table below presents the performance metrics obtained from these models:

Models Fitted for the original training Set							
Model	Accuracy	f1-score	recall	Confusion Matix			
Logistic Regression	0.7243	0.558	0.564			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	308	1248
					Actual Positive	278	3700
Support Vector Machine	0.7188	0.418	0.5			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	0	1556
					Actual Positive	0	3978
Decision Tree	0.9996	1	1			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1555	1
					Actual Positive	1	3977
Random Forest	0.9998	1	1			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1555	1
					Actual Positive	0	3978
Xgboost	0.9995	0.9999	0.9999			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1554	2
					Actual Positive	1	3977
KNN	0.9736	0.968	0.972				
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1508	48
					Actual Positive	98	3880
Adboost	0.8536	0.806	0.788				
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	994	562
					Actual Positive	248	3730
Gradient boost	0.9995	0.9999	0.9999				
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	15554	2
					Actual Positive	1	1977
MLP (multi-layer perceptron	0.9561	0.946	0.945				
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1431	125
					Actual Positive	118	3860

TABLE 3 PERFORMANCE METRICS OBTAINED FROM ORIGINAL DATA SET

The **Random Forest** model emerged as the top performer with evaluation metrics, accuracy: 99.98%, F1-score: 100% and recall: 100%.

5.3 Handling Class Imbalance and Evaluation of Models on Data Set after applying different sampling techniques

We observed a substantial class imbalance within our dataset, with approximately 71.9% prevalence of liver disease within the studied population. To address this imbalance, we applied several techniques which were SMOTE (Synthetic Minority Over-sampling Technique), upsampling and downsampling. The same set of algorithms were then used to fit various models and these were evaluated using the previously-mentioned metrics.

5.3.1 Smote dataset

Models Fitted for the SMOTE training Set							
Model	Accuracy	f1-score	recall	Confusion Matix			
Logistic Regression	0.6409	0.629	0.697			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1283	273
					Actual Positive	1714	2664
Support Vector Machine	0.6243	0.62	0.713			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1423	113
					Actual Positive	1946	2032
Decision Tree	0.9989	0.999	0.999			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1554	2
					Actual Positive	4	3974
Random Forest	0.9996	1	0.999			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1554	2
					Actual Positive	0	3978
Xgboost	0.9998	1	1			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1555	1
					Actual Positive	0	3978
KNN	0.9817	0.978	0.985			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1544	12
					Actual Positive	89	3889
Adboost	0.8211	0.804	0.858			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1464	92
					Actual Positive	898	3080
Gradient boost	0.9995	0.999	0.999			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1554	2
					Actual Positive	1	3977
MLP (multi-layer perceptror	0.9827	0.979	0.988			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1554	2
					Actual Positive	94	3884

TABLE 4 PERFORMANCE METRICS OBTAINED FROM SMOTE DATA SET

It was identified that the **best model fitted on the SMOTE dataset is XGBoost.**

5.3.2 Upsampled dataset

Next, we evaluated the performance of several models on the upsampled dataset. However, none of the models showed adequate performance compared to the other methods.

5.3.3 Downsampled dataset

Models Fitted for the Doensampled training Set							
Model	Accuracy	f1-score	recall	Confusion Matix			
Logistic Regression	0.6449	0.635	0.71			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1335	221
					Actual Positive	1744	2234
Support Vector Machine	0.6122	0.608	0.703			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1419	137
					Actual Positive	2009	1969
Decision Tree	0.9989	0.999	0.999			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1555	1
					Actual Positive	3	3973
Random Forest	0.9998	1	1			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1555	1
					Actual Positive	0	3978
Xgboost	0.9996	1	0.999			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1555	1
					Actual Positive	1	3977
KNN	0.9275	0.916	0.948			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1547	9
					Actual Positive	392	3586
Adboost	0.8339	0.816	0.863			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1444	112
					Actual Positive	807	3171
Gradient boost	0.9993	0.999	0.999			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1554	2
					Actual Positive	2	3976
MLP (multi-layer perceptror	0.9389	0.927	0.946			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1495	61
					Actual Positive	277	3701

TABLE 5 PERFORMANCE METRICS OBTAINED FROM DOWNSAMPLED DATA SET

These results indicate that the **best model fitted on the downsampled dataset is Random Forest**.

5.4 Handling Multicollinearity with PCA

As multicollinearity was observed in our dataset, Principal Component Analysis (PCA) was applied as a remedy.

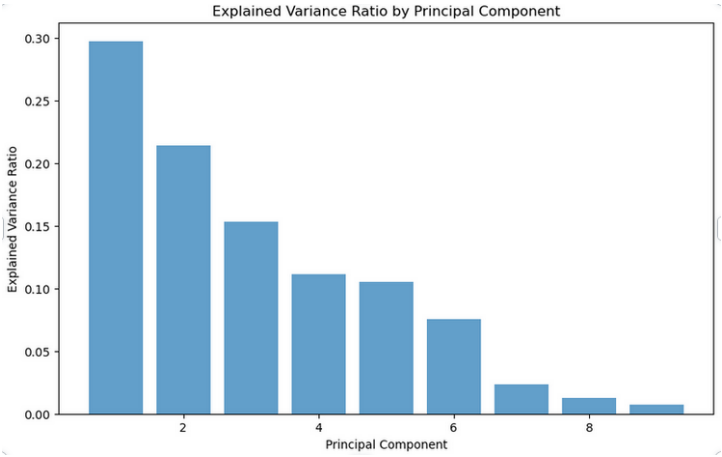

FIGURE 4 EXPLAINED VARIANCE RATIO VS PC'S

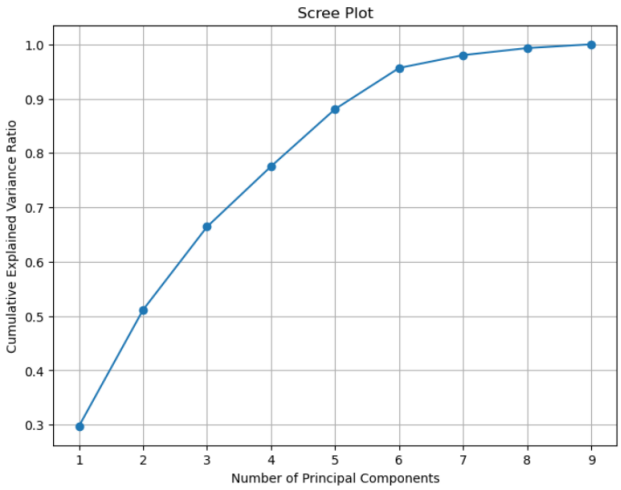

FIGURE 3 SCREE PLOT

The Explained Variance Ratio is the proportion of variance explained by the ith principal component with respect to the total variance. An elbow point at 5 PCs was observed, where the explained variance ratio begins to level off or the rate of decrease slows down significantly. Further an elbow point was observed in the scree plot. Based on the analysis of the Explained Variance Ratio and the Scree plot, the optimal number of principal components to retain was determined as 5. This choice strikes a balance between retaining sufficient variance in the data while reducing dimensionality.

Model comparison after applying PCA

After applying PCA, the number of variables reduced to 6, containing 5 principal components and 1 categorical variable (Gender). We proceeded to fit nine models on the original, SMOTE, upsampled, and downsampled datasets after applying PCA. However, the models obtained after applying PCA did not achieve the same level of accuracy as those without PCA.

5.5 Clustering Approach

Cluster analysis, a statistical technique used to group similar objects or observations based on their characteristics was implemented utilizing various methods namely, “kmeans” algorithm, the “kprototype” algorithm (an extension of k-means that can handle both numerical and categorical data when kmeans handles only numerical) and “kmedoids” Clustering.

5.5.1 kprototype clustering

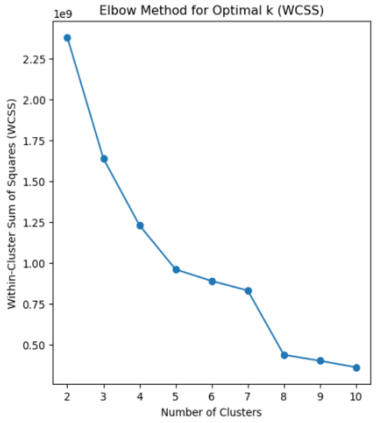

FIGURE 6 WCSS VS CLUSTERS - KPROTOTYPE

FIGURE 5 SILHOUTTE PLOT - KPROTOTYPE CLUSTERS

- The elbow point analysis and silhouette plots suggest the presence of 4 clusters in the dataset.

Kprototypes			
Silhoutte Scores	0.7		
cluster	with Liver Disease	without Liver Disease	Total Count
0	0.6805	0.31948	18890
1	0.9472	0.0527	1366
2	1	0	1211
3	1	0	259

TABLE 6 CLASS DISTRIBUTION WITHING CLUSTERS - KPROTOTYPE

- However, within these clusters, the distribution of the two classes (Liver Disease and Non-liver Disease) is highly disproportionate, indicating the inappropriateness of fitting models within clusters due to potential bias or unreliable results.

5.5.2 KMeans Clustering

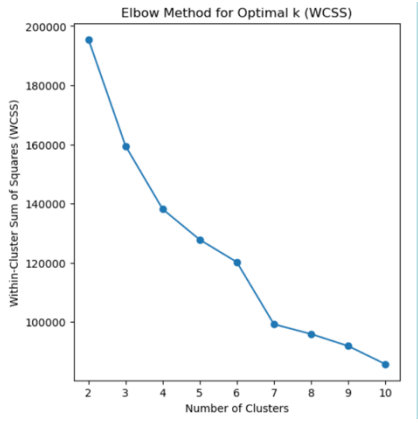

FIGURE 7 WCSS VS CLUSTERS - KMEANS

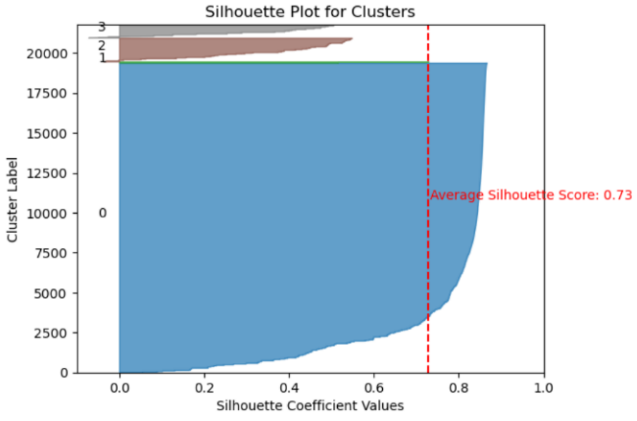

FIGURE 8 SILHOUTTE PLOT - KMEANS CLUSTERS

- Similar to KPrototype, the elbow point and silhouette analysis suggest 4 clusters.

Kmeans			
Silhoutte Scores	0.73		
cluster	with Liver Disease	without Liver Disease	Total Count
0	0.6805	0.31948	18890
1	0.9472	0.0527	1366
2	1	0	1211
3	1	0	259

TABLE 7 CLASS DISTRIBUTION WITHING CLUSTERS - KMEANS

- However, the disproportionate distribution of classes within clusters raises concerns regarding the reliability of fitting models within clusters.

5.5.3 KMedioids Clustering

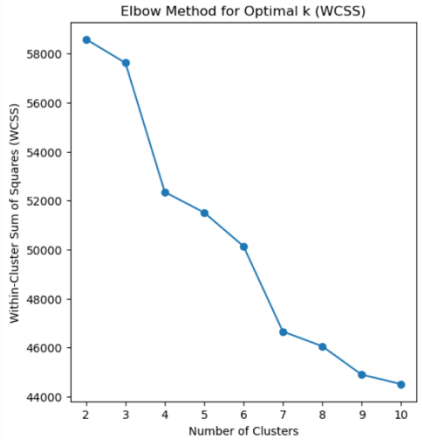

FIGURE 9 WCSS VS CLUSTERS - KMEDIOIDS

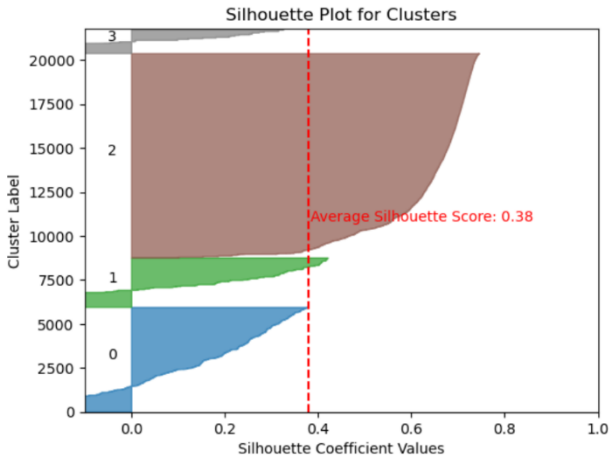

FIGURE 10 FIGURE 8 SILHOUTTE PLOT - KMEDIOIDS CLUSTERS

The above figures indicate the presence of 4 clusters based on the elbow point and silhouette scores.

Kmedoids			
Silhoutte Scores	0.37		
cluster	with Liver Disease	without Liver Disease	Total Count
0	0.82919046	0.17080954	5954
1	0.893214286	0.106785714	2800
2	0.587267402	0.412732598	11608
3	1	0	1364

TABLE 8 CLASS DISTRIBUTION WITHING CLUSTERS - KMEDIOIDS

- While there is some level of separation between clusters, the disproportionate distribution of classes within clusters suggests caution in fitting models within clusters.

5.5.4 Summary on usage of clustering

- All three types of clustering methods used suggest the presence of 4 clusters in our dataset.
- Within the clusters the two classes, i.e., Liver Disease and Non liver Disease are highly disproportionate suggesting inappropriateness of fitting models within the clusters as it may leads to biased or unreliable results
- Since our best model achieved 100% accuracy on the test set and 99.99% accuracy on the train set, there's no need to delve further into clustering with the aim of enhancing the model's predictive accuracy.

5.6 Selection of best model

We consider the model with highest Accuracy,F1 Score and Recall value as our best Model. We have obtained 3 different models with Same Accuracy,F1Score and Recall values. Further, the difference between the training and testing Accuracy is investigated to check whether the model is overfitted or not.

	With All Variables		
	Evaluation Matrices	Training	Test
Model 1 Random Forest on ORIGINAL Training Set	Accuracr	0.9999	0.9998
	F1 Score	1.0000	1.0000
	Recall	1.0000	1.0000
Model 2 XGBoost on SMOTE Training Set	Accuracr	0.9999	0.9998
	F1 Score	1.0000	1.0000
	Recall	1.0000	1.0000
Model 2 Random Forest on DOWNSAMPLE Training Set	Accuracr	0.9999	0.9998
	F1 Score	1.0000	1.0000
	Recall	1.0000	1.0000

TABLE 9 SUMMARY OF BEST MODEL

5.6.1 Important feature selection under 1st model

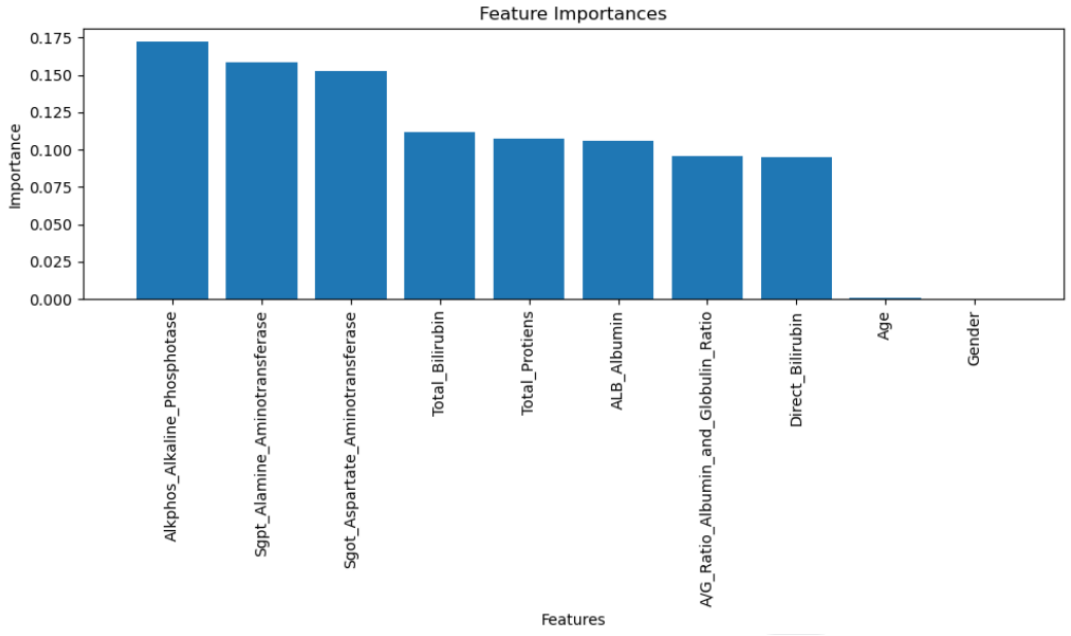

FIGURE 11 IMPORTANT FEATURE SELECTION UNDER 1ST MODEL

According to this plot Gender and Age has least significance for predicting Liver Disease. **Therefore we have fitted the model again to only the top 8 important variables.**

5.6.2 Important feature selection under 2nd model

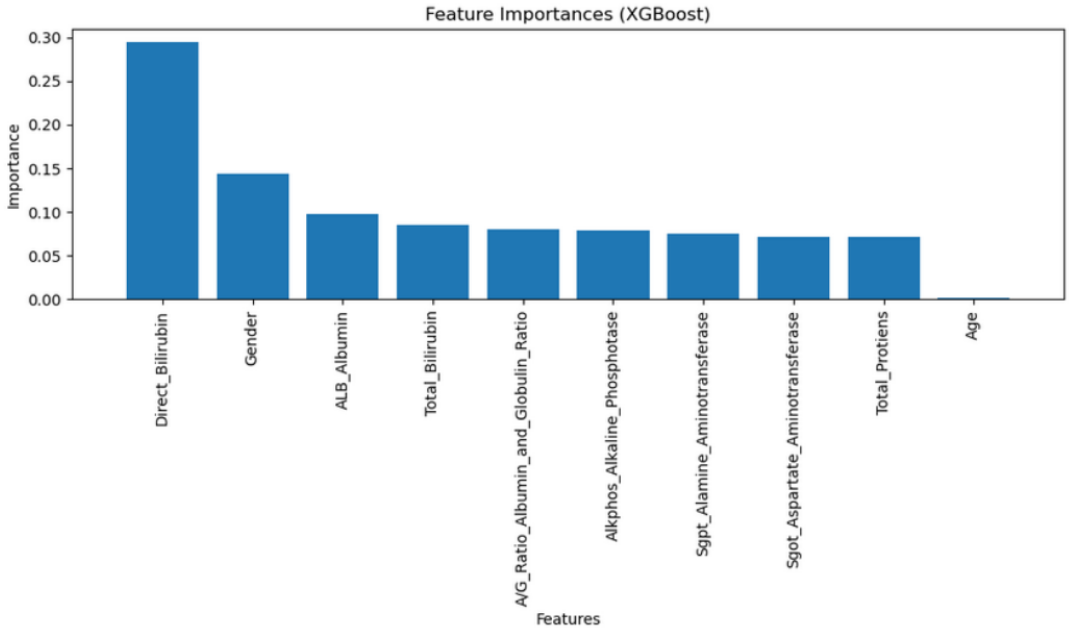

FIGURE 12 IMPORTANT FEATURE SELECTION UNDER 2ND MODEL

According to this plot Age has least significance for predicting Liver Disease, and so **we have fitted the model again to only the top 9 important variables.**

5.6.3 Important feature selection under 3rd model

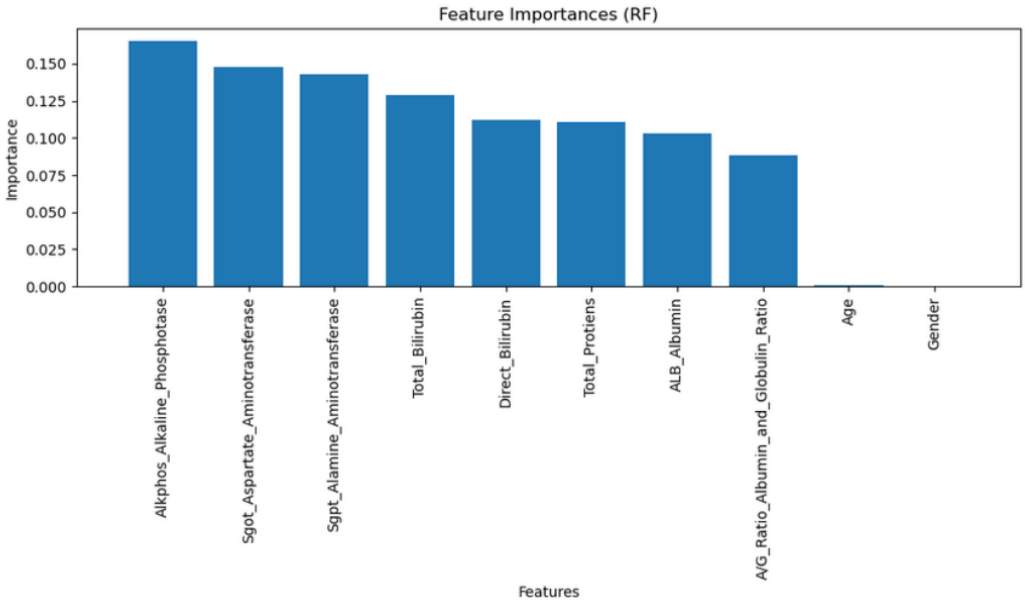

FIGURE 13 IMPORTANT FEATURE SELECTION UNDER 3RD MODEL

According to this plot Gender and Age has least significance for predicting Liver Disease, **therefore we have fitted the model again to only the top 8 important variables.**

5.7 Selection of optimal model

	With All Variables			Number of Important variables	Only consisdering selected Variables		
	Evaluation Matrices	Training	Test		Evaluation Matrices	Training	Test
Model 1 Random Forest on ORIGINAL Training Set	Accuracr	0.9999	0.9998	8	**Accuracr**	**0.9999**	**1.0000**
	F1 Score	1.0000	1.0000		**F1 Score**	**1.0000**	**1.0000**
	Recall	1.0000	1.0000		**Recall**	**1.0000**	**1.0000**
Model 2 XGBoost on SMOTE Training Set	Accuracr	0.9999	0.9998	9	Accuracr	0.9999	0.9996
	F1 Score	1.0000	1.0000		F1 Score	1.0000	1.0000
	Recall	1.0000	1.0000		Recall	1.0000	1.0000
Model 2 Random Forest on DOWNSAMPLE Training Set	Accuracr	0.9999	0.9998	8	Accuracr	0.9999	0.9998
	F1 Score	1.0000	1.0000		F1 Score	1.0000	1.0000
	Recall	1.0000	1.0000		Recall	1.0000	1.0000

TABLE 10 SELECTION OF OPTIMAL MODEL AFTER FEATURE SELECTION

Hence, we have obtained **the best model is as random forest model fitted on original data set after selecting top 8 important variables.**

6. Output of study

The main output of our study is the data product which is a website that interested parties can use to predict the occurrence of liver disease in patients.

Description:

(01) Airflow Services:

airflow-trigger: Initiates the data pipeline execution.

airflow-worker: Executes tasks defined in the Directed Acyclic Graph (DAG).

airflow-scheduler: Schedules the DAG runs.

airflow-webserver: Provides a user interface for monitoring and managing the pipeline.

(02)API Service:

api-service: A Flask REST API deployed on port 5050, facilitating model predictions through an '/predict' endpoint. This service dynamically downloads the latest model from an S3 bucket for prediction.

* Directed Acyclic Graph (DAG) Tasks:

1)Import Data (import_data_task): Fetches data from a training S3 bucket.

2) Validate Schema (load_schema_train): Verifies the format of loaded CSV files against a predefined schema.

3) Create Feature Store (creating_feature_store): Conducts feature engineering on validated data, splits it into training and testing sets (x_train, x_test, y_train, y_test), and stores them in a feature store bucket.

4) Download Data (download_data): Retrieves data from the feature store bucket.

5) Model Training (model_training): Utilizes MLflow to train three different models, tracking their parameters and storing them in metadata store (Amazon RDS) and artifacts in S3. MLflow is configured with Amazon RDS for metadata storage, S3 for artifact storage, and an EC2 instance for remote tracking.

6) Model Evaluation (model_evaluation): Evaluates models by scraping MLflow, selecting the model with the lowest accuracy, and storing it in a model serving S3 bucket.

7) Model Monitoring (model_monitoring): Utilizes Evidently services to monitor model drift.

Workflow and Automation:

The Airflow DAG is scheduled to run daily, enabling continuous integration of new training data added to the S3 bucket.

Upon triggering, the DAG updates the feature store and retrains the model with the expanded dataset.

MLflow is leveraged to analyze and select the best-performing model based on accuracy, serving it in the model serving S3 bucket.

Evidently generates daily reports, triggered by the DAG, facilitating model drift detection and monitoring.

Deployment:

The entire system is deployed using Docker Compose, ensuring portability and scalability.

AWS EC2 hosts the Docker containers, providing a reliable and scalable infrastructure.

This comprehensive system ensures efficient data processing, model training, deployment, and monitoring, facilitating informed decision-making and predictive analytics in real-time applications

Link to access data product:

https://drive.google.com/file/d/1B14DGztgxwTj3OOw7Xy08t_ysZ0mNJN9/view

7. Discussion and conclusions

Initial examination of the data revealed class imbalance in the response variable "liver disease". We addressed this issue by implemeting SMOTE, upsampling and downsampling. The advanced analysis was conducted on the original datset as well as the balanced datasets. Various models were constructed, with the optimal model being selected by comparing accuracy, f1-score and recall among other evaluation metrics.

The **Random Forest** model emerged as the top performer based on the following metrics:

Accuracy (99.98%): This high accuracy indicates that the model correctly predicts the outcome for almost all instances in the dataset.

- F1-score (100%): The F1-score reflects a perfect balance between precision and recall for both classes. This implies that the model effectively identifies positive cases while minimizing false positives and false negatives.
- Recall (100%): With a recall of 100%, the model captures all positive cases without missing any, demonstrating its robustness in identifying instances of interest.

8. Appendix

Link to access codes:

https://github.com/tharindu-frd/Group-8-final-project.git