Predicting Chronic Hepatitis B Disease Progression and Outcomes in the Gambia using Machine Learning Algorithms

Dorcas Asare [1,4] Ismaël Koné [2,3] Bubacarr Bah [2] Gibril Ndow [2] Maud Lemoine [2] Rohey Bangura [2] Lamin Bojang[2]

[1]Stellenbosch University [2]MRC Unit The Gambia at LSHTM [3]Université Virtuelle de Côte d'Ivoire(UVCI) [4]African Institute for Mathematical Sciences

Abstract

Hepatitis B virus (HBV) infection remains an important global health problem with high mortality and morbidity. There is a problem in the early detection of HB patients at high risk and its contributing factors. The study is aimed to predict the mortality outcome of hepatitis B patients in The Gambia, and also, to identify the significant associated factors. Six (6) machine learning models was used to predict the mortality of HB patients namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), and K Nearest Neighbor (K-NN). The f1-score for LR, DT, RF, NB, SVM, and K-NN models were $0.53, 0.96, 0.97, 0.53, 0.53,$ and 0.75 respectively with RF showing the best performance for predicting mortality outcome. The model can be deployed in clinical routine to further test its effectiveness and adjust it to better detect high-risk patients.

Introduction

- HBV disease which is inflammation of the liver puts people at high risk of death in its progression. According to WHO, Hepatitis B affects approximately 257 **million** people (3.5% of the world's population) globally of which 60 **million** are from the Africa region [3].

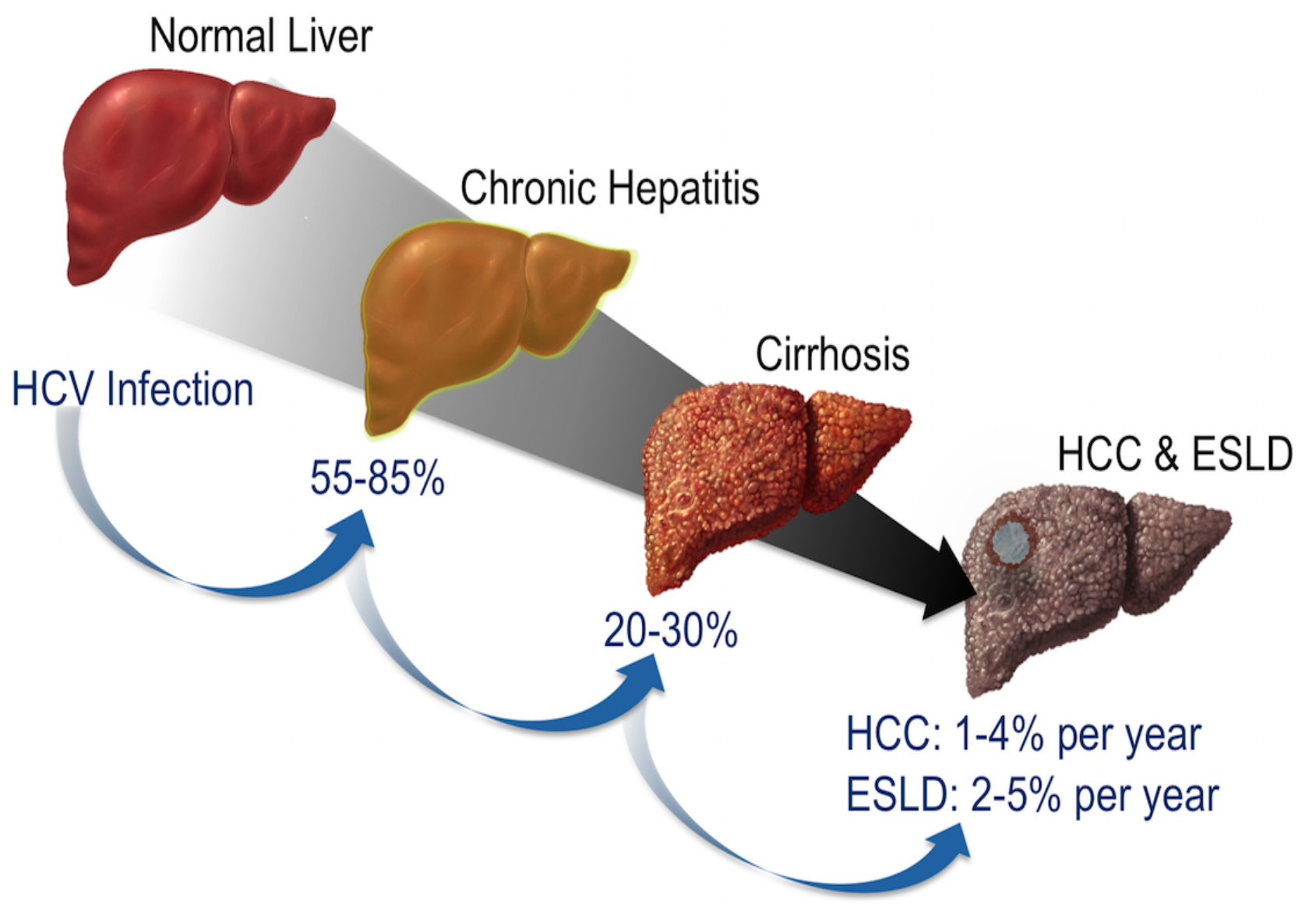

Figure 1. HBV disease progression

- HBV causes approximately $820,000$ deaths every year [3].
- The project is part of the PROLIFICA (Prevention of Liver Fibrosis and Cancer in Africa) program in The Gambia which seeks to alleviate the burden of HBV disease. The longitudinal study of HBV patients was set up in 2011 with approximately $2,000$ adults.
- The objective of the study is to use different machine learning models to predict mortality and find the most important associated factors based on the PROLIFICA data.

PROLIFICA Dataset

- A follow-up study was done in 2018-present of which $325(21\%)$ patients died out of 1545

Figure 2. 1 out of 5 patients die from HBV disease in the Gambia

- The dataset has over 100 features that contain information on patients' demographic, clinical, and family history of HBV.

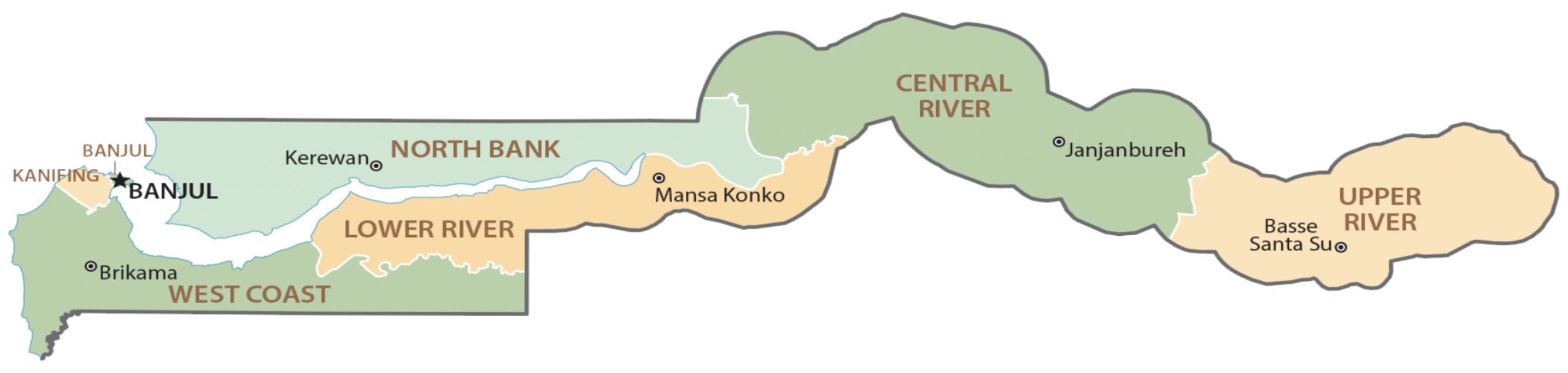

Figure 3. Map of The Gambia

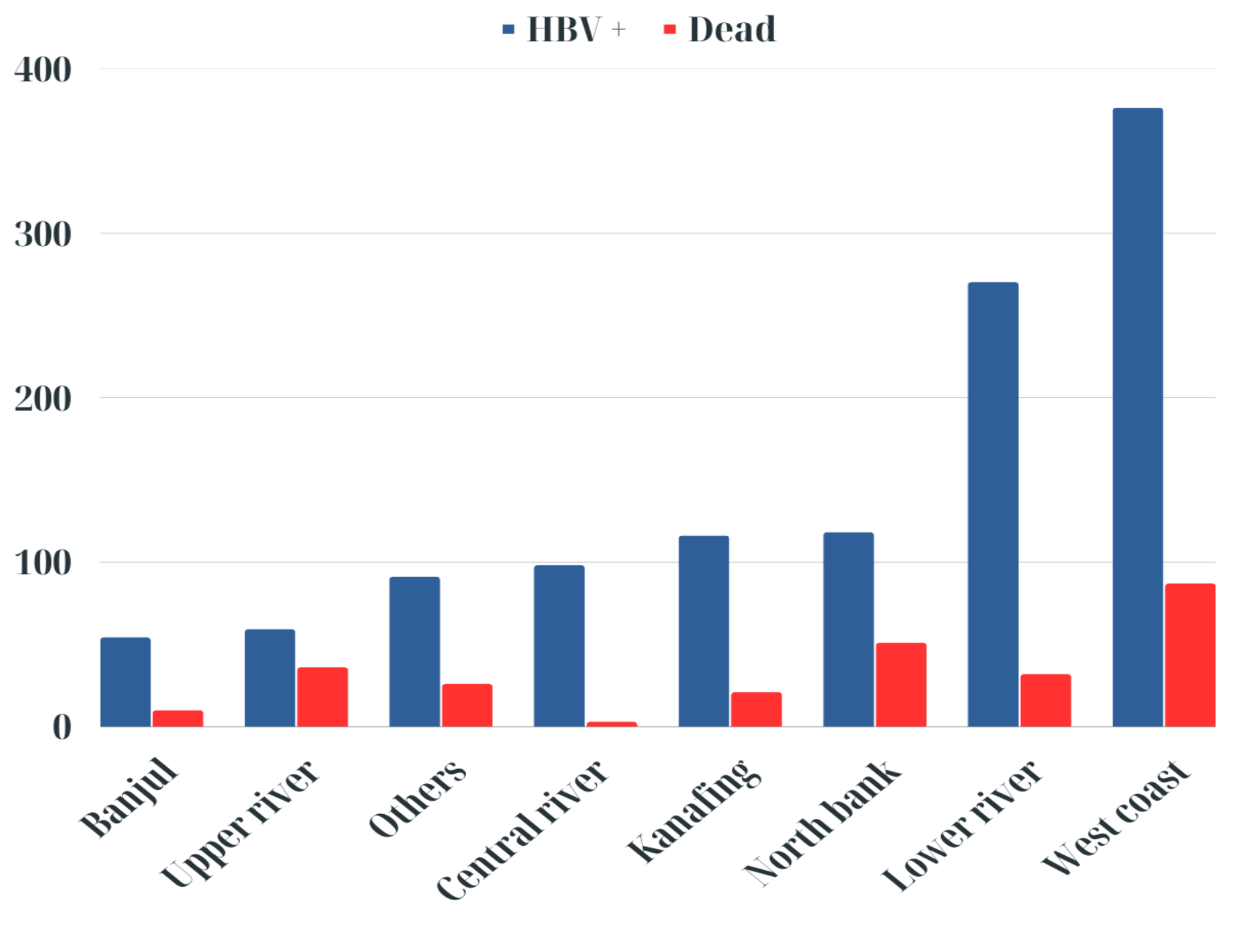

Figure 4. Number of HBV positives and deaths in The Gambia regions

Methods

After pre-processing of the data, 40 features remained and 681 patients were obtained for further analysis. 571 patients were alive whiles $110(16\%)$ were dead. Feature selection techniques like correlation and mutual information were used to filter out important features.

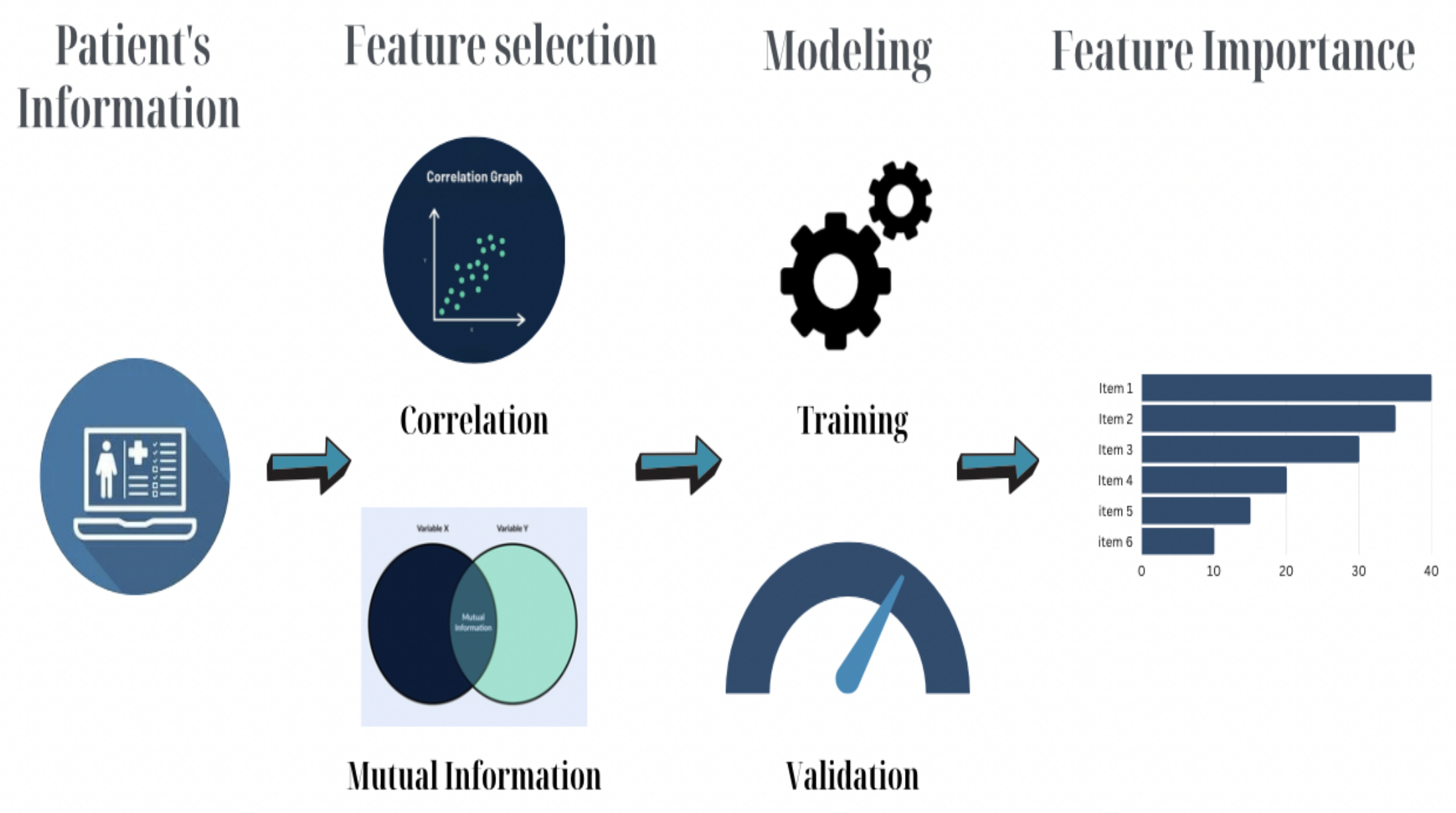

Figure 5. Diagram summarizing the whole process

- Correlation measures the strength of association between features [1] whiles mutual information explains the amount of information obtained from the target variable[4].
- The data was split into training and validation for modeling.

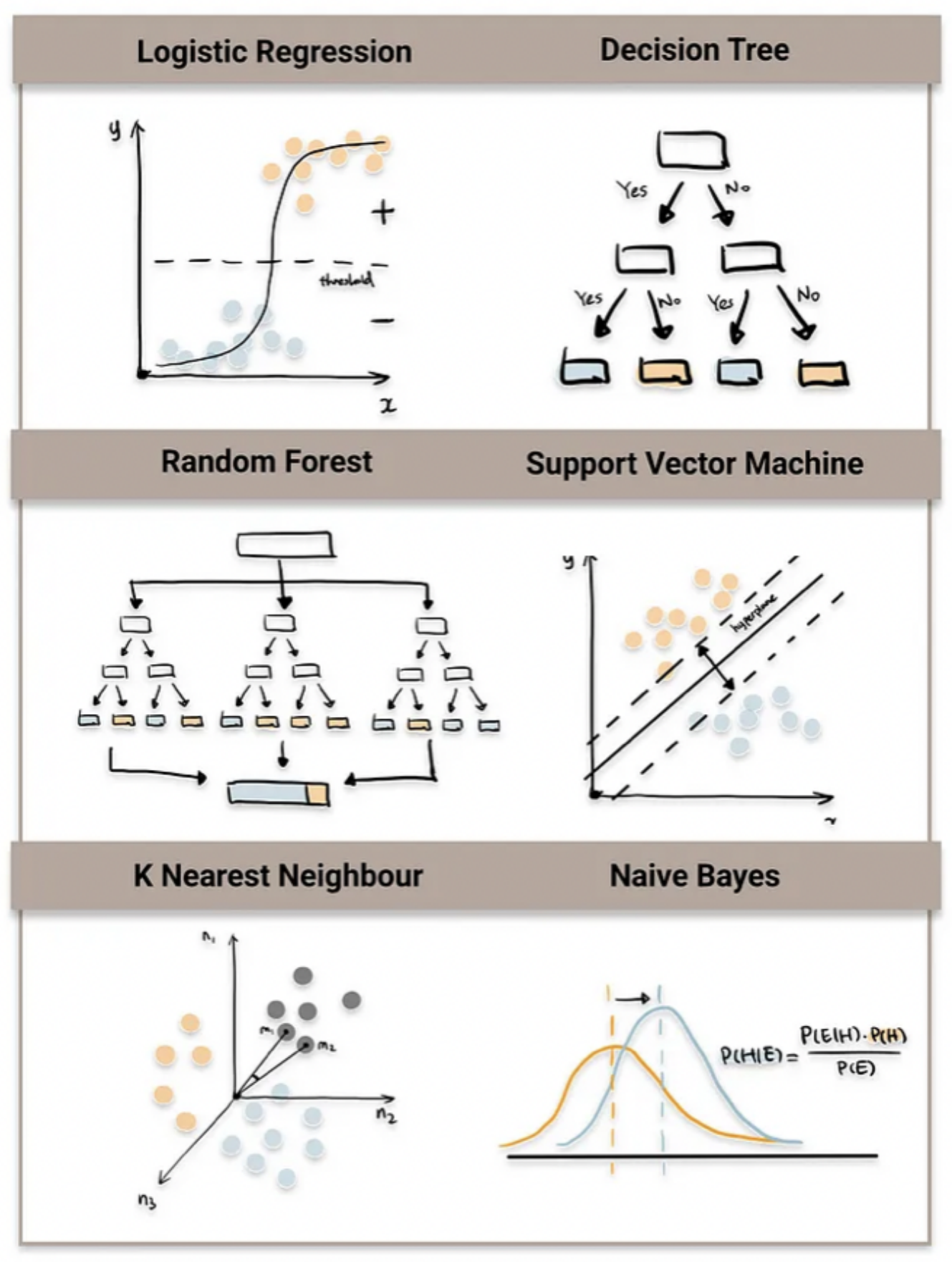

Figure 6. Machine learning classification models [2]

- Logistic regression is used to estimate probabilities by using a logistic function.
- A decision tree models decisions to predict the value of the target variable by using a flowchart-like tree structure.
- Random forest consists of multiple decision trees to select output by averaging all of the tree outputs.
- Support vector machine classifies data points into the target classes with the help of a hyperplane.
- The k-Nearest Neighbors algorithm uses proximity to make classifications about the group a data point belongs to.
- Naive Bayes is a probabilistic classifiers which uses Bayes theorem to assign class labels to instances.

Results

Table 1. Summary of the predictive performance of each model

Model	Accuracy	Precision	Recall/Sensitivity	F-score
Logistic regression	0.66	0.44	0.66	0.53
Decision tree	0.96	0.96	0.96	0.96
Random forest	0.97	0.98	0.97	0.97
SVM	0.66	0.44	0.66	0.53
Naive Bayes	0.66	0.44	0.66	0.53
K-nearest neighbour	0.77	0.76	0.77	0.75

Conclusion

The best-performing model was the Random Forest with an f1-score of 0.97. The 5 most contributing factors were fibroscan results, patients with a normal ultrasound scan, albumin levels, patients with HCC, and space-occupying lesions. We plan to study HB patients' evolution and driving factors as well.

Future work

Aside from predicting the mortality outcome in hepatitis B patients and finding significant factors related to the outcome, we also want to predict HBV disease progression (cirrhosis, hepatocellular carcinoma (HCC), and end-stage liver disease (ESLD)). We want to identify the significant factors associated with the progression of CHB disease.

References

[1] Agustin Garcia Asuero, Ana Sayago, and AG González. The correlation coefficient: An overview. *Critical reviews in analytical chemistry*, 36(1):41–59, 2006.

[2] Destin Gong. Top 6 machine learning algorithms for classification. *Accessed: Dec*, 4:2022, 2022.

[3] World Health Organization et al. *Global hepatitis report 2017*. World Health Organization, 2017.

[4] Brian C Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357, 2014.