WOD7001 Principal of Data Science

1/2022/2023

Group 2 - Dr Maizatul Akmar Binti Ismail

HEPATITIS C PREDICTION MODEL FOR EUROPEAN REGION

Group Assignment 2

Name	Student No
Nazmus Sakib	S2135114/1
Jia Wei Eng	S2172813/1
Pui Yee Sum	S2177713/1
Shi Sun Lai	S2175592/1
Wen Hui See	S2176564/1

Table of Content

Contents

Project Background

This project is to implement a hepatitis C predictive model using machine learning which can be referred by physicians in the hospitals and health clinics to identify the prognosis of hepatitis C patients in the European region according to the attributes given. Across the European region, the adult general population with unknown HCV infection status was targeted.

And, with the attributes that signifies health features and status, this could assist physicians in suggesting early interventions to prevent the occurrence of chronic stage for suspected hepatitis C patients. The understanding of association between attributes and target patients would further speed up the diagnosis process from the medical professionals, providing treatment without delay with effective customised treatment plans for each patient.

Project Objectives

- To examine the main features that contribute to the prediction of hepatitis C virus.
- To implement a hepatitis C virus prediction model in the European region.
- To evaluate several machine learning techniques in predicting hepatitis C virus.

Data Modelling

Diving into the modelling section, 1 parametric (Naive Bayes (NB)) and 2 non-parametric (K-Nearest Neighbors (KNN), Decision Tree (DT)) classifiers are selected for the HCV detection.

The classification model was trained on the HCV dataset using the target feature, Category with the 8 highly associated features, AL, ALP, ALT, AST, BIL, CHE, CHOL and GGT that were selected after conducting the correlation matrix test. The dataset was divided into two parts by using the random splitting method whereby 70% of the dataset (430 samples) are utilized to train the model and the remaining 30% of the dataset (185 samples) are used for validation purposes (testing set).

In order to assess the performance of the models, the outcome regarding the actual and predicted results obtained by the classification models can be depicted using a confusion matrix. In addition, accuracy, precision, recall, F1 and AUC scores are measured and

compared among each proposed machine learning algorithm for performance evaluation. The results of the modelling will be discussed in the data interpretation section.

Codes: https://colab.research.google.com/drive/1t7yoOC_-esT10PIzTEeeesho3YcjJn6x

Data Interpretation

As mentioned in the data exploration section using OSEMN Framework, the correlation matrix is plotted to identify the relationship between the attributes and the target features. Spearman's rank coefficient is applied since all the continuous attributes are not normal, and the target attribute of this research, Category, is an ordinal attribute. Multicollinearity exists when the relationship attribute is higher than 0.8 (P. Vatcheva & Lee, 2016). The multicollinearity attribute will affect the accuracy of the modelling; therefore, it should be excluded from the modelling. The correlation for each attribute is plotted. Detailed information on the correlation and p-value matrices are attached in Appendix I and Appendix II, respectively.

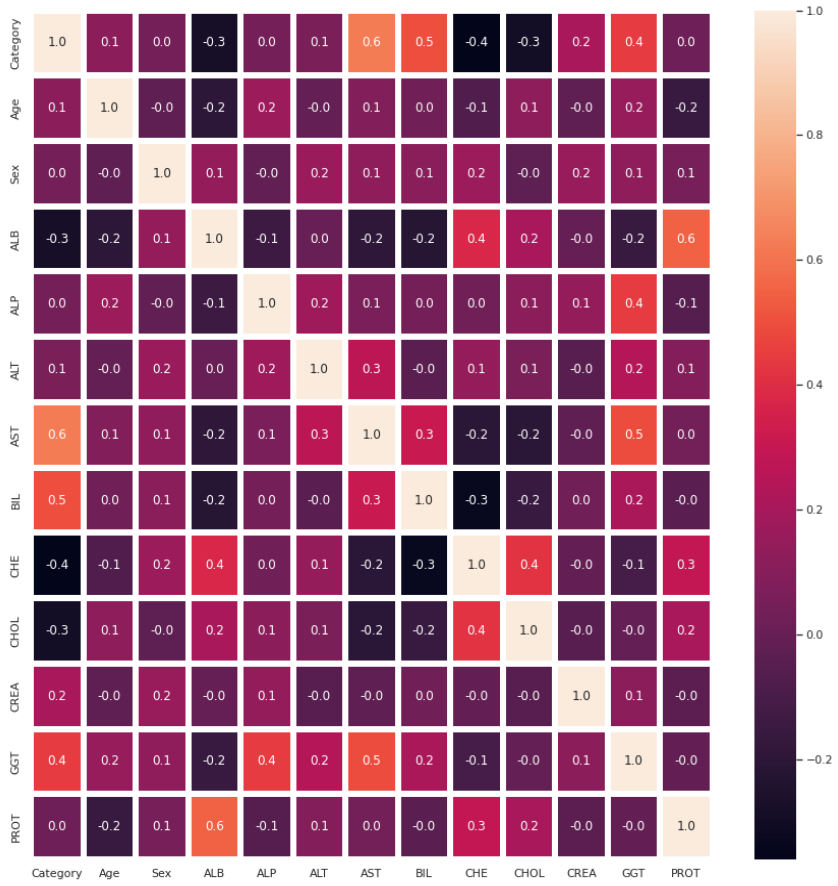

Figure 1: Correlation Matrix

Based on the correlation matrix in Figure 1, there is no correlation higher than 0.8 or lower than -0.8. Therefore, there is no multicollinearity in this dataset.

In addition, Age (correlation = 0.06, p-value=0.13), Sex (correlation = 0.07, p-value=0.09) and PROT (correlation=0.07, p-value=0.06) are very weakly correlated towards the Category. The p-value also showed more than the significant value (0.01) which indicates the spearman rank correlation coefficient is not significant (Chen, 2002), which indicates no correlation between these three attributes and the Category attributes.

ALB (correlation = -0.16), ALP (correlation = correlation=-0.15), ALT (correlation = -0,12), CHE (correlation=-0.18), CHOL (correlation = -0.27), CREA (correlation = -0.12) are weakly negative correlated towards category attribute whereas AST (correlation = 0.50), BIL (correlation = 0.39), GGT (correlation = 0.41) have moderate positive correlation towards category attribute.

Therefore, objective 1 has been achieved. We conclude that Age, Sex, CREA and PROT attributes are not the main features that contribute to the result of the prediction of HCV. Therefore, Age, Sex, CREA and PROT are excluded from the dataset for modelling purposes. Concurrently, the Category with the 8 highly associated features, AL, ALP, ALT, AST, BIL, CHE, CHOL and GGT were selected for the modelling process instead to have higher accuracy in predicting HCV.

In order to achieve the second and third objective of this project, this section compares the performance of the 3 machine learning algorithms (KNN, DT, NB). As indicated in Table 1, train and test scores of each model are calculated to determine if the model is overfitting, underfitting or well-fitting.

Model	Test Score	Train Score
K-Nearest Neighbors	0.9243	0.9442
Decision Tree	0.9514	0.9488
Naive Bayes	0.8973	0.9140

Table 1. Train and test scores of each model

The train and test scores of KNN, DT and NB are fairly comparable. As a result, there are no signs of overfitting or underfitting, indicating that the 3 models are well-fitted.

To examine the classification results, accuracy, recall, precision, f1-measure performance metrics are calculated. The results obtained from the models are summarised in Table 2.

Model	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors	0.9243	0.9061	0.9243	0.9087
Decision Tree	**0.9514**	**0.9538**	**0.9514**	**0.9419**
Naive Bayes	0.8973	0.9174	0.8973	0.9065

Table 2. Performance of HCV detection by machine learning

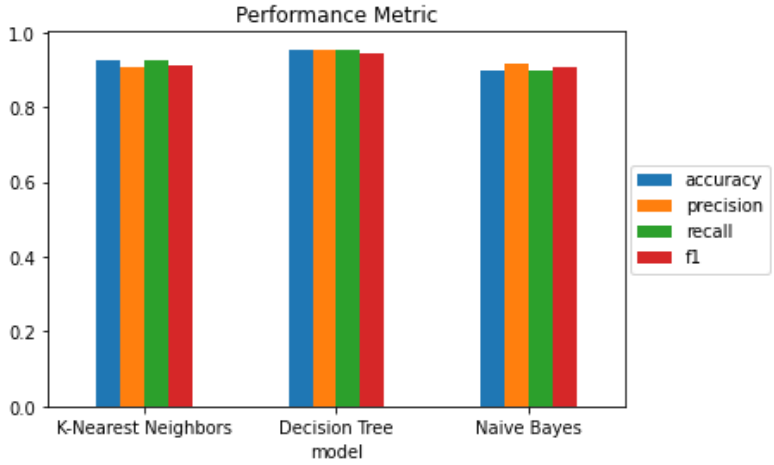

Figure 2: Accuracy, Precision, Recall and F1-measure metric of HCV detection performance.

To visualise the results for better comparison and understanding, figure 2 is displayed from the results in Table 2. The findings demonstrate that all the 3 proposed classification models have high accuracy predicting HCV diseases, with DT outperforming the other 2 models due to its highest accuracy (95%), precision (95%), recall (95%) and f1 measure (94%) values.

The weighted and macro Area Under Curve (AUC) for each model were also shown in Table 3. This project will take the AUC score highly into account for model evaluation as past research supported that Area Under Curve (AUC) scores are considered as being more resistant to data imbalance in comparison to Precision, Recall, and F1-measure (Fawcett, 2006) which is more relevant to the dataset used in this project.

model	Weighted AUC	Macro AUC
K-Nearest Neighbors	0.8155	0.7884
Decision Tree	0.7888	0.7161
Naive Bayes	**0.9457**	**0.9332**

Table 3. Weighted and Macro AUC of each model

As shown from Table 3, the results denoted that NB is the best performance model as it has the highest value for both the Weighted AUC (0.9457) and Macro AUC (0.9332). The ROC curves for the NB model are also shown in Figure 3.

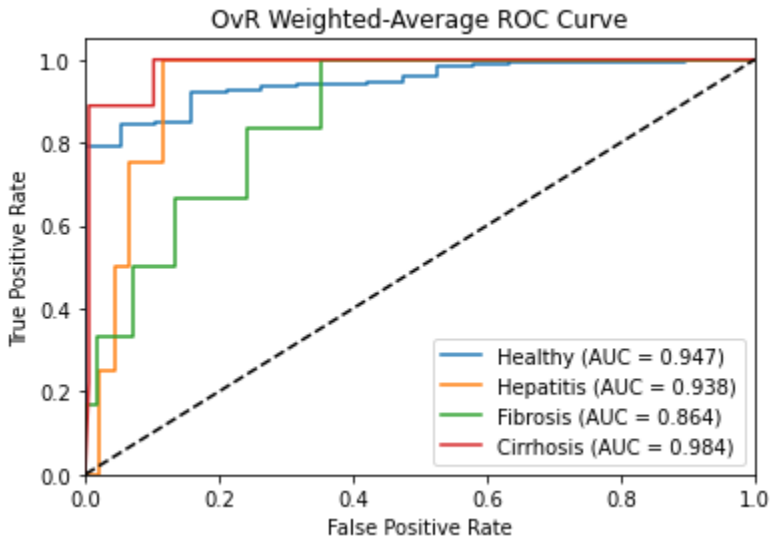

Figure 3: ROC curves of Naive Bayes model

In this project, DT generated the best results in terms of Accuracy, Recall, Precision, and F1-measure. However, NB has the highest Weighted and Macro AUC scores among the 3 models. Keeping in mind that the AUC scores were given priority in the selection of the best performance model since it is more robust to data imbalance, while at the same time the Accuracy, Recall, Precision, and F1-measure of all three models showing good performance which are greater than 0.8 can still be produced. As a result, NB is selected as the best model in predicting HCV since it has the highest value of AUC scores with good results (>0.8) in terms of Accuracy, Recall, Precision, and F1-measure.

Data Product

Data Visualisation using Power BI

In the context of data product, Power BI is used to visualise the dataset. The visualisation can be accessed via the following web link: https://app.powerbi.com/links/kJaIRZR_Lx?ctid=facd9fe9-b6aa-43e8-8f56-72d8888ac026&pbi_source=linkShare&bookmarkGuid=bc76f807-1422-4f04-ad67-94b095e78cba

Figure 3, 4 & 5 illustrates the number of observations by Sex and Category. The compositions by Category are 87.8% Healthy, 3.9% Hepatitis, 3.4% Fibrosis, and 4.9% Cirrhosis. We can further drill down the Category composition by selecting Sex from the slicer. The Category compositions by male are 85.9% Healthy, 5.3% Hepatitis, 3.5% Fibrosis, and 5.3% Cirrhosis. The Category compositions by females are 90.7% Healthy, 1.7% Hepatitis, 3.4% Fibrosis, and 4.2% Cirrhosis.

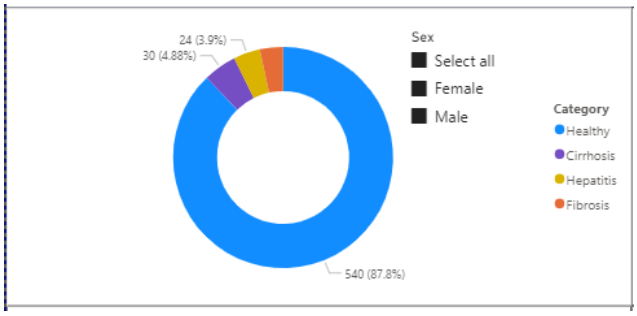

Figure 4: Total Count by Category

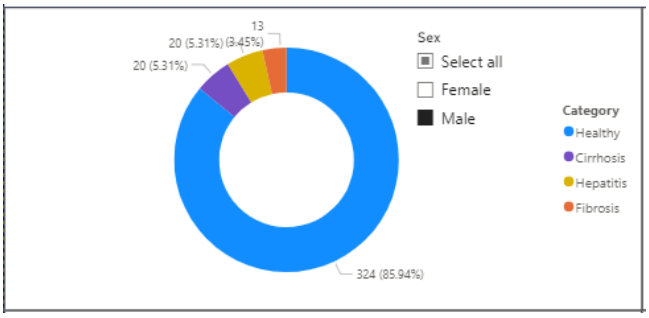

Figure 5: Male Count by Category

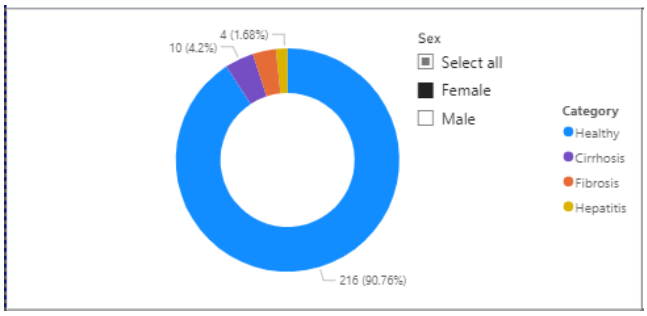

Figure 6: Female Count by Category

Based on Figure 6, Age 46 has the highest count. It can be further drilled down by Category by clicking on the category of donut chart. For example, both Figure 7 and Figure 8 shows the same count highlighted for Age 46 and Age 48 highlighted, ie. 27 count.

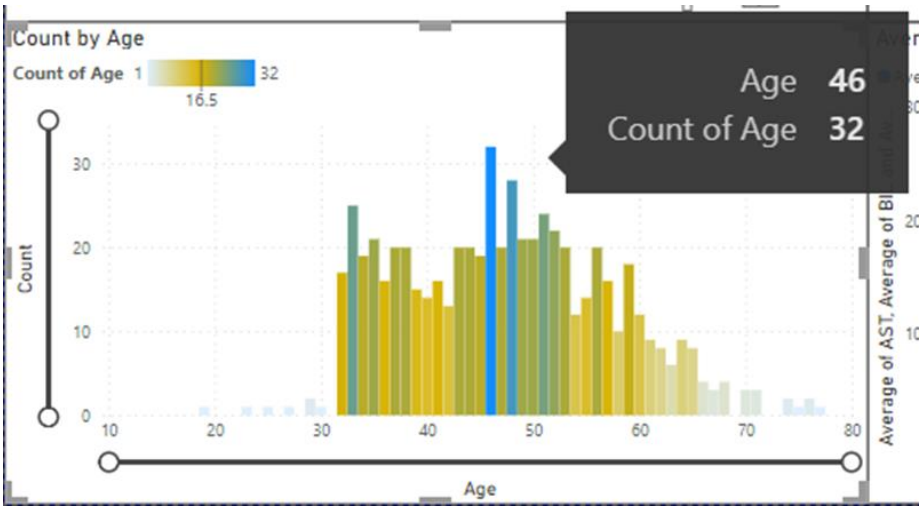

Figure 7: Total Count by Age

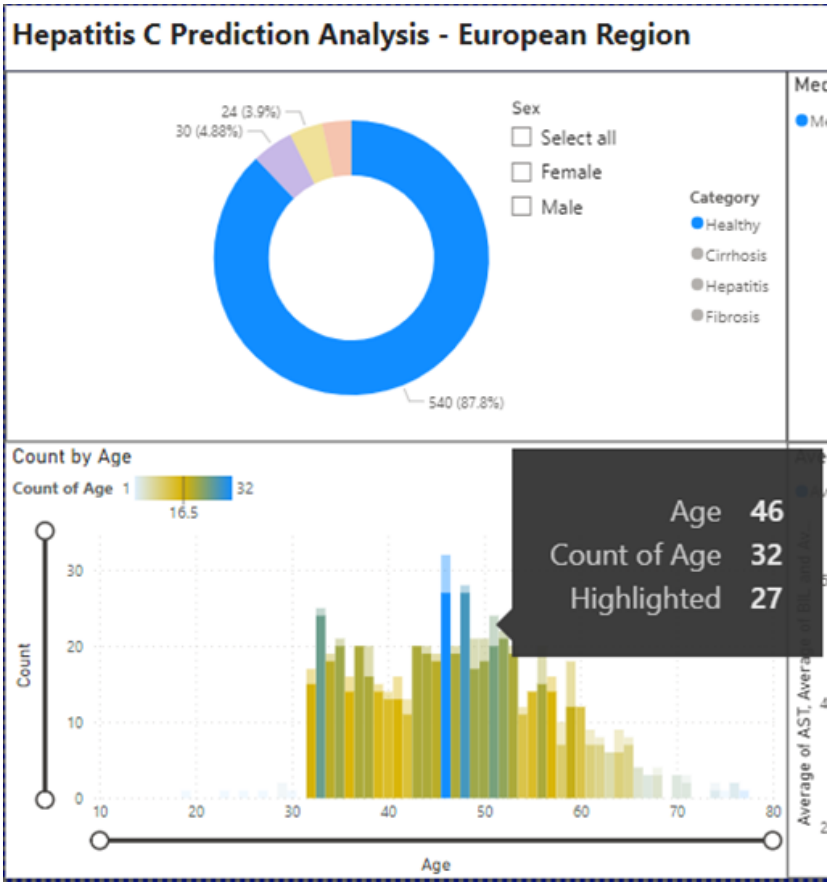

Figure 8: Count by Age 46 & Healthy

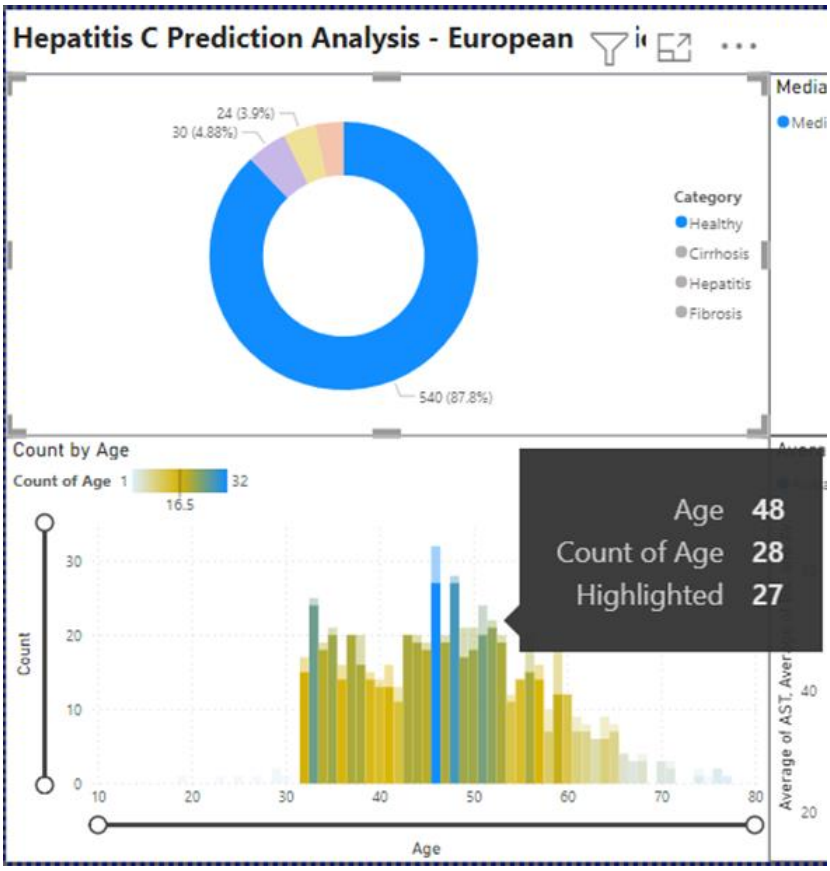

Figure 9: Count by Age 48 & Healthy

Figure 9 illustrates the Median of attributes that are moderately positively correlated towards attribute Category. Figure 10 shows that Median of AST for Cirrhosis is 92.90. By filtering the attribute Sex, Median of AST for Cirrhosis for both male and female are 98.65 and 71.00 respectively. Male classified as Cirrhosis has Median of AST higher than the overall sample.

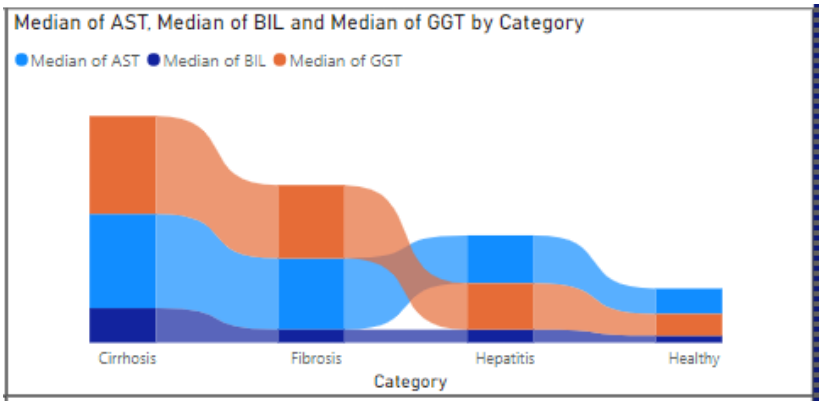

Figure 10: Median of AST, BIL and GGT by Category

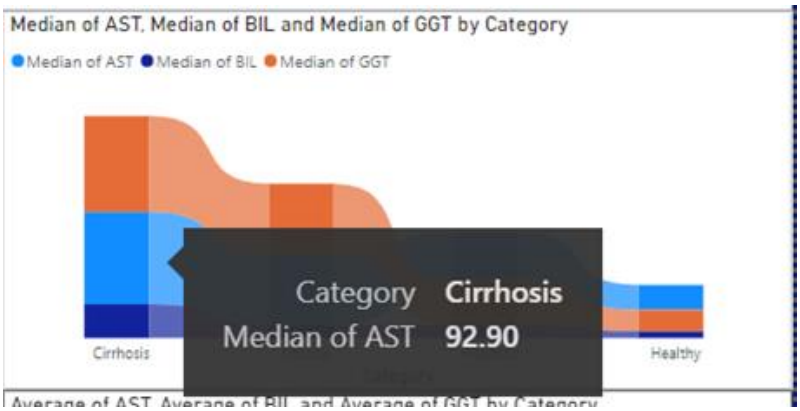

Figure 11: Median of AST for Cirrhosis

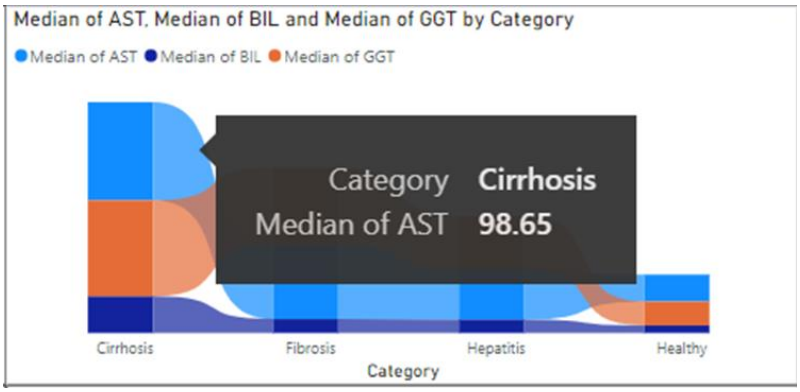

Figure 12: Median of AST for Male classified as Cirrhosis

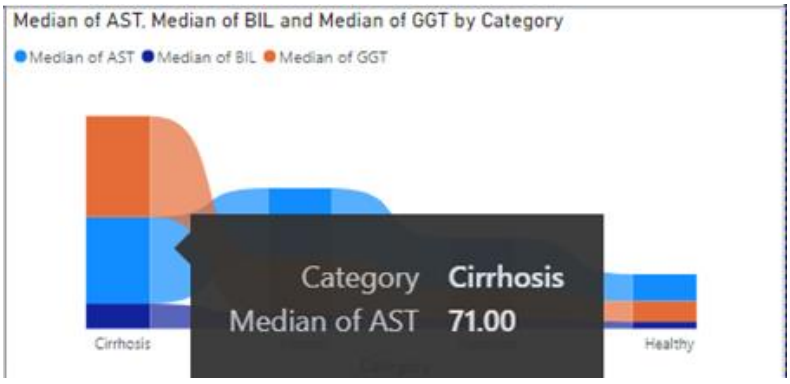

Figure 13: Median of AST for Female classified as Cirrhosis

Figure 13, 14 and 15 illustrate the Mean and Median of the main features (attributes) selected for Hepatitis C prediction model. The Mean and Median of AST, GGLT and BIL for unhealthy groups are significantly higher than healthy group. Therefore, we can conclude that AST, GGT and BIL are important features that contribute to the accuracy of the prediction model.

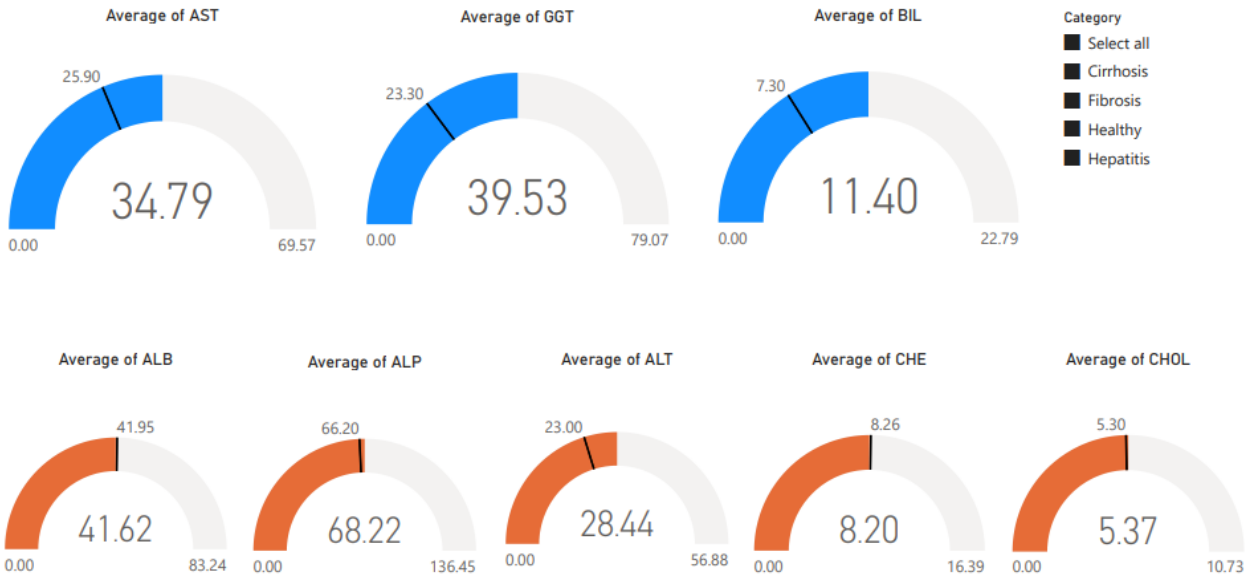

Figure 14: Median and mean of attributes that have correlation towards all category

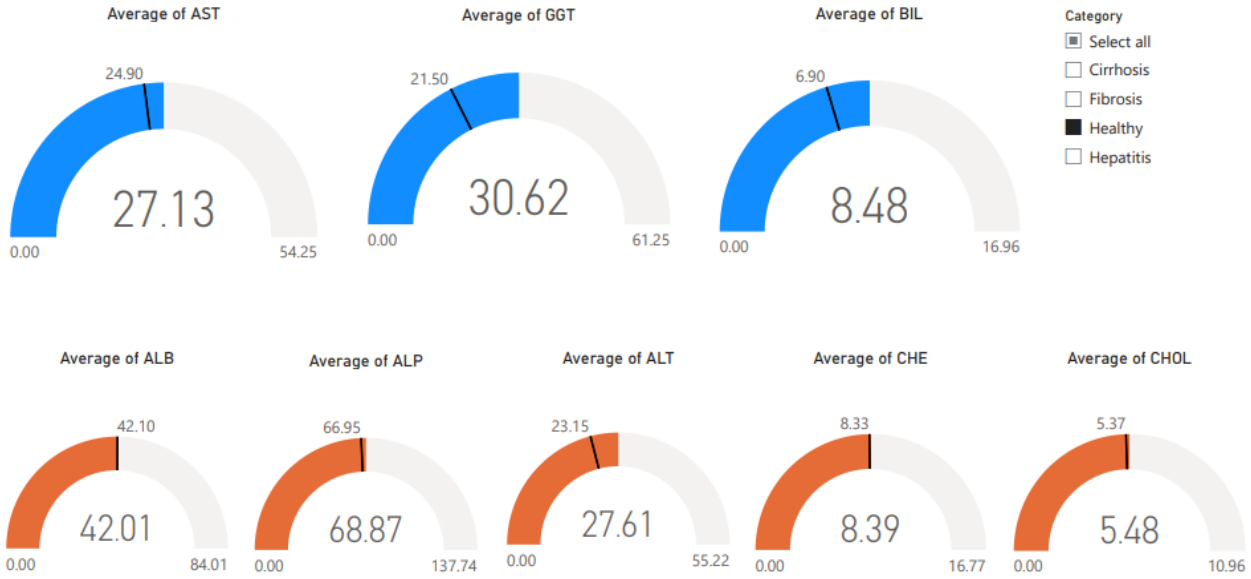

Figure 15: Median and mean of attributes that have correlation towards healthy group

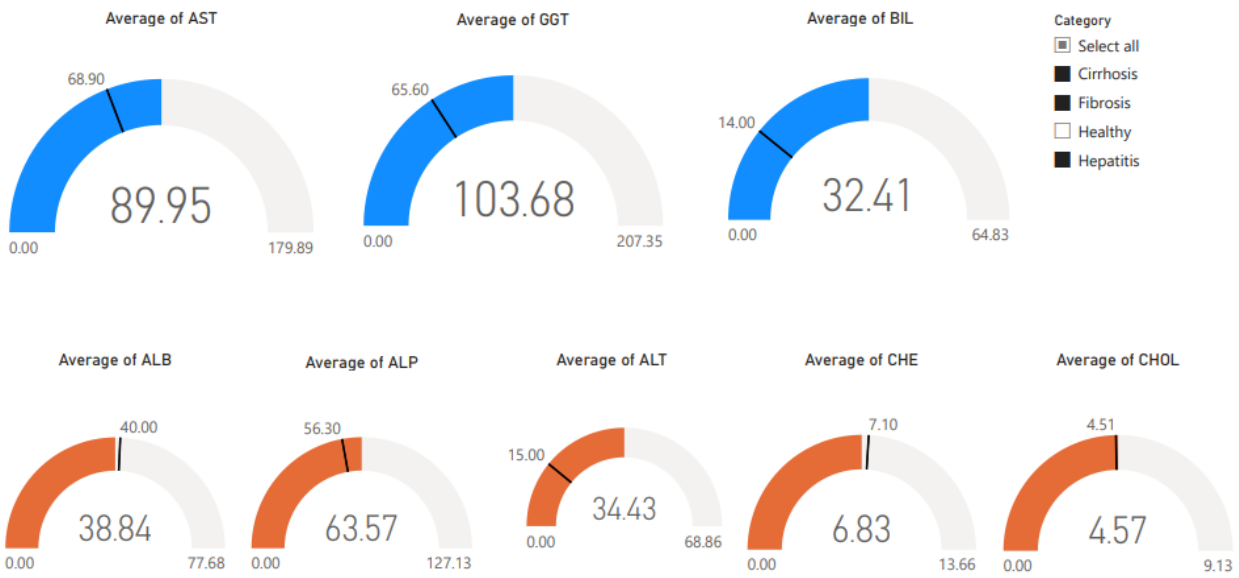

Figure 16: Median and mean of attributes that have correlation towards unhealthy group

User Interface for Naive Bayes Prediction Model

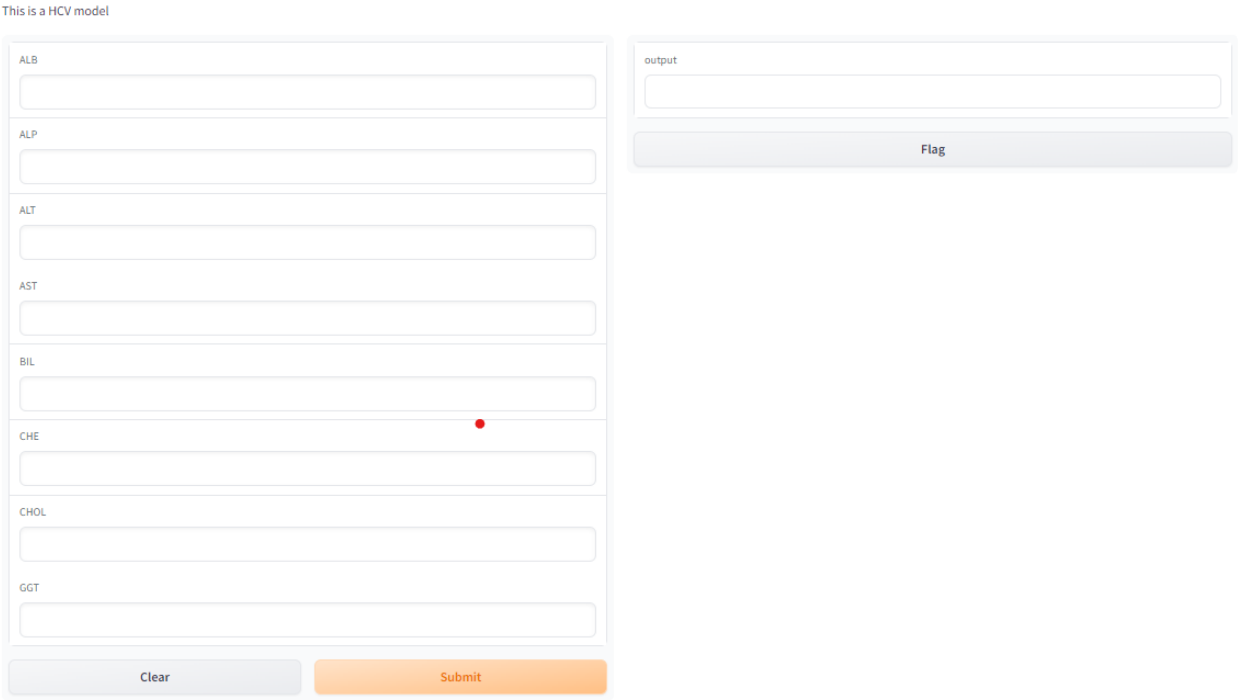

Figure 17: User Interface for Naive Bayes Prediction Model

https://huggingface.co/spaces/Sun33/HCV_model

Figure 16 illustrates the user interface for the Naive Bayes prediction model. The user is required to insert the values of ALB, ALP, ALT, AST, BIL, CHE, CHOL and GGT. After inserting the related values, the user can click the “Submit” button. The prediction result will then be displayed in the output cell. There will be 4 types of outcome which are Healthy, Hepatitis, Fibrosis and Cirrhosis. The user may consult with the relevant medical institution for relevant treatment if the result predicted is either Hepatitis, Fibrosis or Cirrhosis.

Insights and Conclusion

This research concluded that NB is the best model for HCV diagnosis in the European region compared to DT and KNN due to its highest AUC scores and good performance in terms of accuracy, precision, recall and F1 score.

Firstly, to achieve the objective of this research, main features had been selected and attributes that are less significant being removed for the modelling process to improve the HCV classification accuracy. The research would then implement the suggested classification model (KNN, NB, DT) and was trained on the HCV dataset using the target feature, Category with the 8 highly associated features, AL, ALP, ALT, AST, BIL, CHE, CHOL and GGT. To evaluate the performance and select the best for HCV prediction in real life practice, the accuracy, precision, recall, F1 and AUC scores were then measured and compared among each proposed machine learning algorithm.

According to the spearman rank correlation coefficient, this study suggests that there is no relationship between the variables PROT, age, and gender and the target characteristic, Category. In addition, based on the results of the Kruskal Wallis test, the levels of creatinine in the healthy and sick groups are nearly equal (CREA). As a result, this study drew the conclusion that age, gender, PROT, and CREA are not the major factors that affected how accurately the disease would have been predicted in the European Region. In other words, individuals in the European Region have a similar likelihood of being diagnosed with HCV. By establishing and understanding the patterns between factors associated with HCV infection, it sheds light on the literature of the medical field. Doctors can perform medical screening by focusing on the 8 associated features to enhance diagnosis capabilities and hence lower screening-related medical costs. The study outcomes can also serve as a tool to train nurses or medical students in the

diagnosis procedure of hepatitis C. With better and more accurate prediction and classification, patients and the infected regions can be discovered efficiently at the early stages. This can ultimately lead to a significant reduction in transmission and prevalence of Hepatitis C virus and hence reduce the HCV infection rate in the European region.

Diving into the modelling section, results generated from the 3 classifiers (KNN, DT, NB) are generally promising as these models achieved above 80% of accuracy scores with AUC score above 70% as well. However, to select the best model to predict HCV infection in the European region, NB still stands out due to its highest AUC scores compared to KNN and DT. As the data set used in this study is highly imbalanced, thus AUC scores were given more weight in the selection of the best performance model as it tends to be more robust. At the same time, NB would still obtain acceptable accuracy, precision, recall, F1 score with highest AUC score compared to KNN and DT.

With that, the implementation of the machine learning model - NB in the European region offers some preliminary proof that these patients might be detected earlier in their journey, which would result in anticipated improvements in outcomes. Despite the fact that many of these pre-diagnosis interactions are linked to known risk factors, it is unclear that using a common or usual risk screening programme will be sufficient given the low specificity of the risk factors. NB as the best machine learning model in this research may recognise complicated, frequently subtle correlations as opposed to usual screening approach, potentially leading to a much more effective method of detecting undiagnosed patients.

With these findings, authorities can encourage the public in the European region to get regular health screenings constructed by the 8 highly associated features when examined for hepatitis C. In addition, authorities can also provide more support towards public health agencies to ensure that they develop and improve sufficient capability and competencies, which includes their laboratories. Those who have been identified as lacking capabilities or in need of improvement shall provide opportunities for training to assure they can fulfil their responsibilities towards the hepatitis C patients. With constant monitoring from the authorities, implementation of the action plan can only be sustained, the goal for preventing hepatitis C or to be detected earlier can also be established.

Reference

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Appendix

Appendix I: Spearman Correlation Matrix

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
Category	1.0000	0.0612	0.0682	-0.1604	-0.1475	-0.1207	0.5020	0.3862	-0.1750	-0.2729	-0.1194	0.4091	0.0743
Age	0.0612	1.0000	-0.0375	-0.1603	0.1761	-0.0462	0.0869	-0.0049	-0.0528	0.1707	-0.0482	0.1004	-0.1007
Sex	0.0682	-0.0375	1.0000	0.2125	0.0266	0.3132	0.3117	0.2584	0.1970	-0.0339	0.5026	0.3120	0.0872
ALB	-0.1604	-0.1603	0.2125	1.0000	-0.0547	0.1825	0.0264	0.1023	0.3270	0.1398	0.2429	0.0321	0.5120
ALP	-0.1475	0.1761	0.0266	-0.0547	1.0000	0.2098	0.0600	-0.0730	0.1249	0.1430	0.0496	0.1550	0.0246
ALT	-0.1207	-0.0462	0.3132	0.1825	0.2098	1.0000	0.4973	0.1405	0.3287	0.1669	0.2973	0.4185	0.2164
AST	0.5020	0.0869	0.3117	0.0264	0.0600	0.4973	1.0000	0.3560	0.1018	-0.0597	0.1631	0.5054	0.1706
BIL	0.3862	-0.0049	0.2584	0.1023	-0.0730	0.1405	0.3560	1.0000	-0.0679	-0.1157	0.1897	0.2473	0.1436
CHE	-0.1750	-0.0528	0.1970	0.3270	0.1249	0.3287	0.1018	-0.0679	1.0000	0.3971	0.2098	0.1559	0.2928
CHOL	-0.2729	0.1707	-0.0339	0.1398	0.1430	0.1669	-0.0597	-0.1157	0.3971	1.0000	0.0676	0.0644	0.1558

CREA	-0.1194	-0.0482	0.5026	0.2429	0.0496	0.2973	0.1631	0.1897	0.2098	0.0676	1.0000	0.1470	0.1320
GGT	0.4091	0.1004	0.3120	0.0321	0.1550	0.4185	0.5054	0.2473	0.1559	0.0644	0.1470	1.0000	0.1835
PROT	0.0743	-0.1007	0.0872	0.5120	0.0246	0.2164	0.1706	0.1436	0.2928	0.1558	0.1320	0.1835	1.0000

Appendix II: Spearman’s p-value Correlation Matrix

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
Category	0.0000	0.1296	0.0910	0.00010	0.0002	0.0027	0.0000	0.0000	0.0000	0.0000	0.0030	0.0000	0.0655
Age	0.1296	0.0000	0.3529	0.00010	0.0000	0.2523	0.0313	0.9043	0.1913	0.0000	0.2329	0.0128	0.0124
Sex	0.0910	0.3529	0.0000	0.00000	0.5109	0.0000	0.0000	0.0000	0.0000	0.4011	0.0000	0.0000	0.0307
ALB	0.0001	0.0001	0.0000	0.00000	0.1756	0.0000	0.5142	0.0112	0.0000	0.0005	0.0000	0.4270	0.0000
ALP	0.0002	0.0000	0.5109	0.17560	0.0000	0.0000	0.1372	0.0704	0.0019	0.0004	0.2196	0.0001	0.5426
ALT	0.0027	0.2523	0.0000	0.00000	0.0000	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
AST	0.0000	0.0313	0.0000	0.51420	0.1372	0.0000	0.0000	0.0000	0.0116	0.1395	0.0000	0.0000	0.0000
BIL	0.0000	0.9043	0.0000	0.01120	0.0704	0.0005	0.0000	0.0000	0.0926	0.0041	0.0000	0.0000	0.0004
CHE	0.0000	0.1913	0.0000	0.00000	0.0019	0.0000	0.0116	0.0926	0.0000	0.0000	0.0000	0.0001	0.0000
CHOL	0.0000	0.0000	0.4011	0.00050	0.0004	0.0000	0.1395	0.0041	0.0000	0.0000	0.0938	0.1104	0.0001
CREA	0.0030	0.2329	0.0000	0.00000	0.2196	0.0000	0.0000	0.0000	0.0000	0.0938	0.0000	0.0002	0.0010
GGT	0.0000	0.0128	0.0000	0.42700	0.0001	0.0000	0.0000	0.0000	0.0001	0.1104	0.0002	0.0000	0.0000
PROT	0.0655	0.0124	0.0307	0.00000	0.5426	0.0000	0.0000	0.0004	0.0000	0.0001	0.0010	0.0000	0.0000