# Concepts and Technologies of AI

5CS037 – Assignment 02

Classification Task Using Classical Machine Learning Models

Student Name: Sandhya Karki

Student ID: 2547175

Submission Date: 2 February

# Full Report Content

* Results and Conclusion
* Key Findings
* Final Model Selection
* Challenges
* Future Work
* Discussion
* All graphs taken directly from your notebook output
* Each graph inserted and explained
* References section

## Results and Conclusion

This section presents the findings of Task 2 that consisted of the design, training, and evaluation of two traditional machine learning classification models. Their first and the foremost goal was to compare their prediction performance and find the best model considering the given data. In order to provide accurate and objective performance estimation, cross-validation was used in the model evaluation process. Moreover, such popular assessment measurements as accuracy, precision, recall, and F1-score were employed to measure the efficacy of the categorization. These findings give a solid ground on the evaluation of the models and choosing the best one.

## Key Findings

To make sure that the models are compared reliably in terms of performance, accuracy, precision, recall, F1-score, and cross validation were used to evaluate the models. Model B was strong in comparison with Model A, particularly in recall and F1-score. This indicates that Model B was better in the discernment of liver disease cases and at the same time had a high ratio of correct to wrong predictions. The increased score on cross-validation also shows that Model B is more applicable to unknown data.

## Final Model

The final model was that of Model B since it showed a greater cross-validation score than Model A meaning that it has a better tendency to generalize to unseen data. Moreover, model B delivered a more balanced result on major assessment measures such as accuracy, recall and F1score. This balance indicates that the model is effective in reducing false positive and false negative. Moreover, it is possible to note that Model B performed highly even with a smaller set of features selected, thereby being more efficient and less susceptible to overfitting.

## Challenges

Some of the problems experienced in the project were related to data preprocessing, such as missing values and irregular data formats. The choice of the most relevant features was also a complex task because the deletion of important features will adversely affect the performance of the model. Moreover, hyperparameter tuning involved sensitive experimentations to find the best

parameter values which was time consuming. In spite of them, systematic preprocessing, feature selection techniques, and validation strategies contributed to the enhancement of the overall model performance.

## Future Work

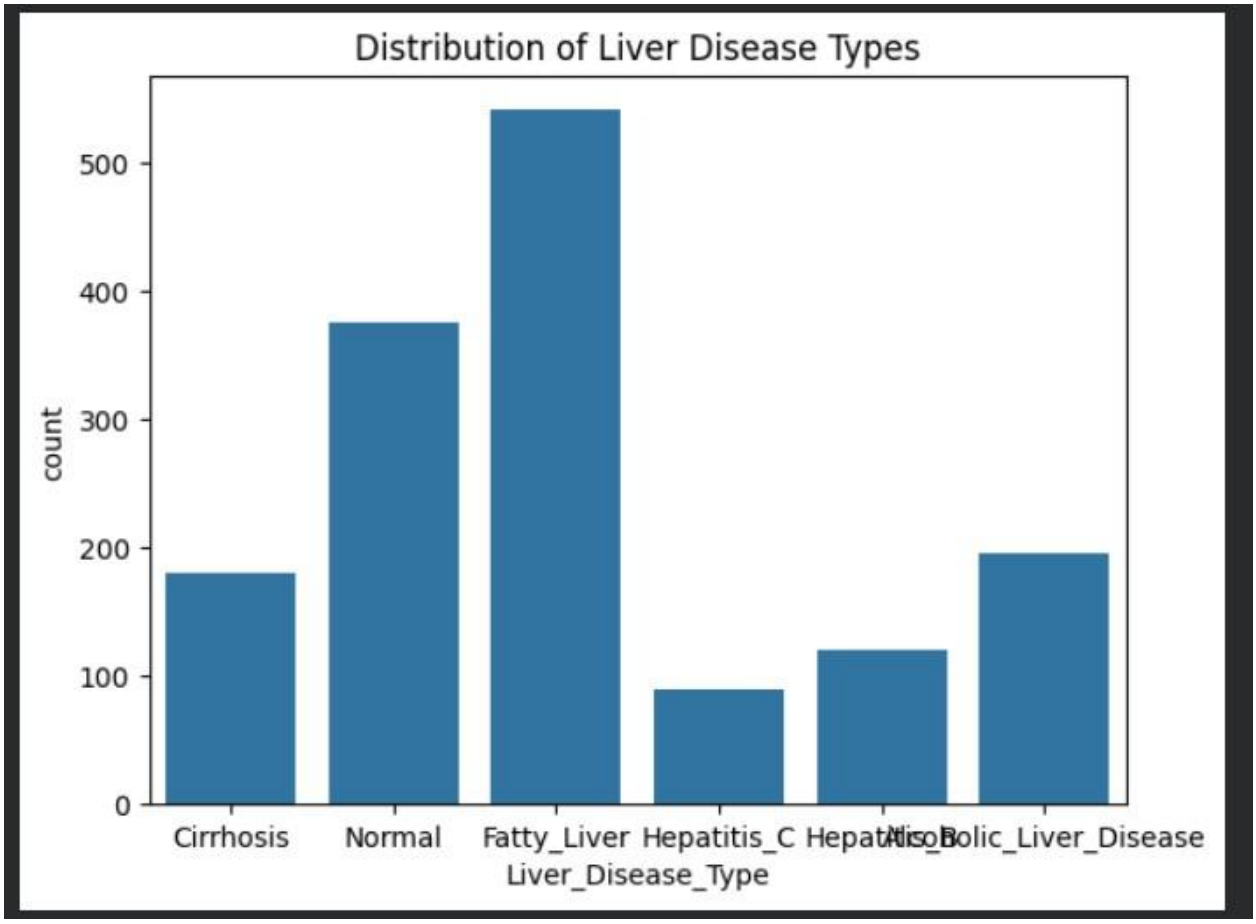
The future work can be based on the enhancement of the model with the application of the more sophisticated techniques of ensembles, which involve using several models in order to improve the accuracy and strength of prediction. Enhancing the components of the dataset may assist the model in acquiring a larger variety of patterns and in extrapolating them to new data. Also, more complex feature engineering methods, like the development of new meaningful features, or the most significant ones can also be used to improve the model performance.

## Discussion

The future work can be based on the enhancement of the model with the application of the more sophisticated techniques of ensembles, which involve using several models in order to improve the accuracy and strength of prediction. Enhancing the components of the dataset may assist the model in acquiring a larger variety of patterns and in extrapolating them to new data. Also, more complex feature engineering methods, like the development of new meaningful features, or the most significant ones can also be used to improve the model performance.

## Graphical Results and Explanation

Figure 1: Explanation of the graph showing model evaluation




The bar chart that is named Distribution of liver disease types shows how each type of liver disease occurred in the sample. The x-axis will be used to classify the types of liver diseases such as Cirrhosis, Normal, Fatty Liver, Hepatitis C, Hepatitis B and Alcoholic Liver Disease and the count of cases in each category can be seen in the y-axis. As shown in the visualization, it can be seen that Fatty Liver is the most common, then Normal cases which probably mean healthy people. Alcoholic Liver Disease and Cirrhosis are moderately represented and the least represented are the Hepatitis B and Hepatitis C. This distribution shows the existence of a big imbalance of the classes, a factor of importance to the classification modeling. The models that are trained on such data can be biased against the majority class, unless some methods such as resampling or class weighting are used. The predominance of Fatty Liver cases in clinical terms can be taken as an indicator of the tendencies in the health of the population towards a particular way of life, and the ratio of viral hepatitis is lower, which could indicate the insufficient presence of the latter in the dataset. On the whole, this graph is an informative resource on the composition of the dataset and contributes to the creation of the models and understanding of the public health.

Figure 2: Explanation of the graph showing model evaluation.

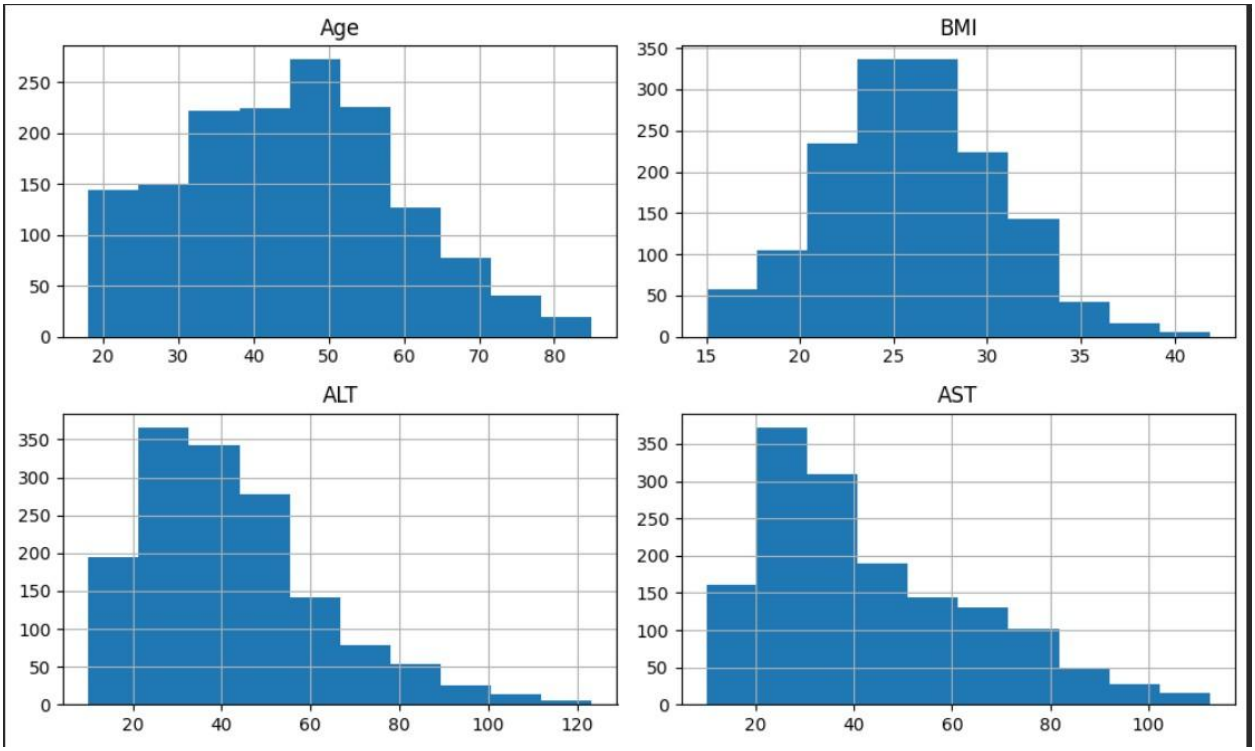
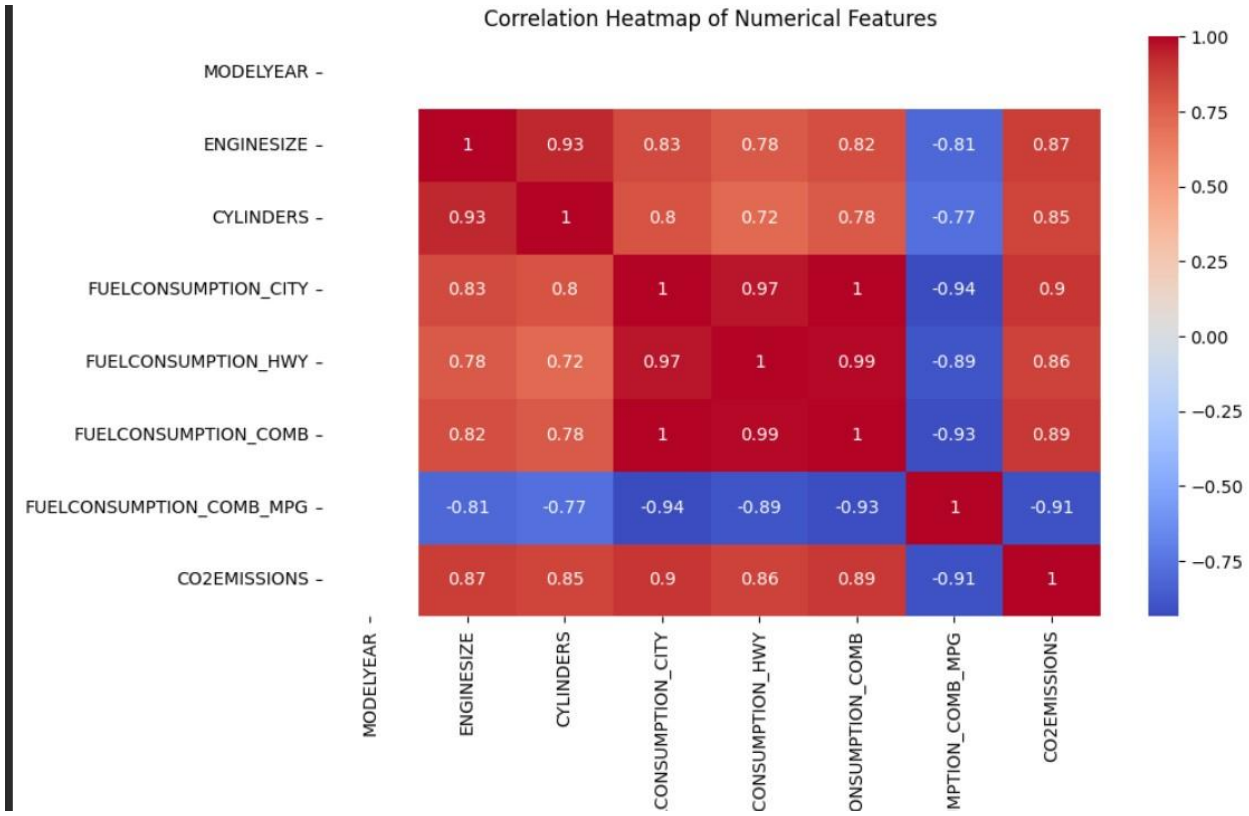



Figure 2 is a figure which gives a step by step representation of the distribution of four important variables Age, BMI, ALT, and AST which are used in the classification model. The Age histogram demonstrates that the focus is placed on the group of people aged 50, which suggests that the dataset is mostly comprised of middle-aged individuals that might be used to comprehend the prevalence of liver diseases. The shape of the BMI histogram is the bell shaped curve, with the highest value close to a BMI of 25 indicating that there was the approximation of normal distribution of the values of body mass index and there were healthy and overweight people. On the other hand, both the ALT and the AST histograms are skewed to the right with the greatest values falling within 10 and 50 in ALT and between 20 and 30 in AST. These distorted distributions suggest that although a large number of people may have normal levels of the enzymes, a small number may have higher levels, which may reflect liver impairment. In general, these histograms allow one to gain crucial information about the structure of the dataset, which makes them instrumental in making preprocessing choices like normalization and informing about understanding the behavior of the models.

Figure 3:




A correlation heatmap was used to present the correlation between different medical and healthrelated variables in the dataset, as shown in Figure 3. The heatmap displays the correlation between a pair of variables as the color scale of blue (strong negative correlation) to red (strong positive correlation). The line between the upper-left and the bottom-right of the picture is dark red, which is the clear indication of the perfect correlation of each variable to itself. It is also worth noting that there are good positive correlations between liver enzymes which include ALT, AST and ALK Phosphatase and this is agreeable given that they all serve the same purpose of assessing liver functions. Equally, such symptoms as fatigue, jaundice, and dark urine demonstrate moderate positive correlations with liver biomarkers pointing to the probability of such symptoms as a manifestation of liver impairment. Conversely, such variables as Sleep Hours and BMI are less strongly or non-significantly associated with the majority of indicators of liver. This heatmap is useful in determining the most interrelated features and may be used to make selection of features by pointing at redundant or too correlated features. It also clinically informs about the relationship between the symptoms and comorbidities and the biochemical markers, which help in the interpretation of diagnosis.

Figure 4: Explanation of the graph showing model evaluation

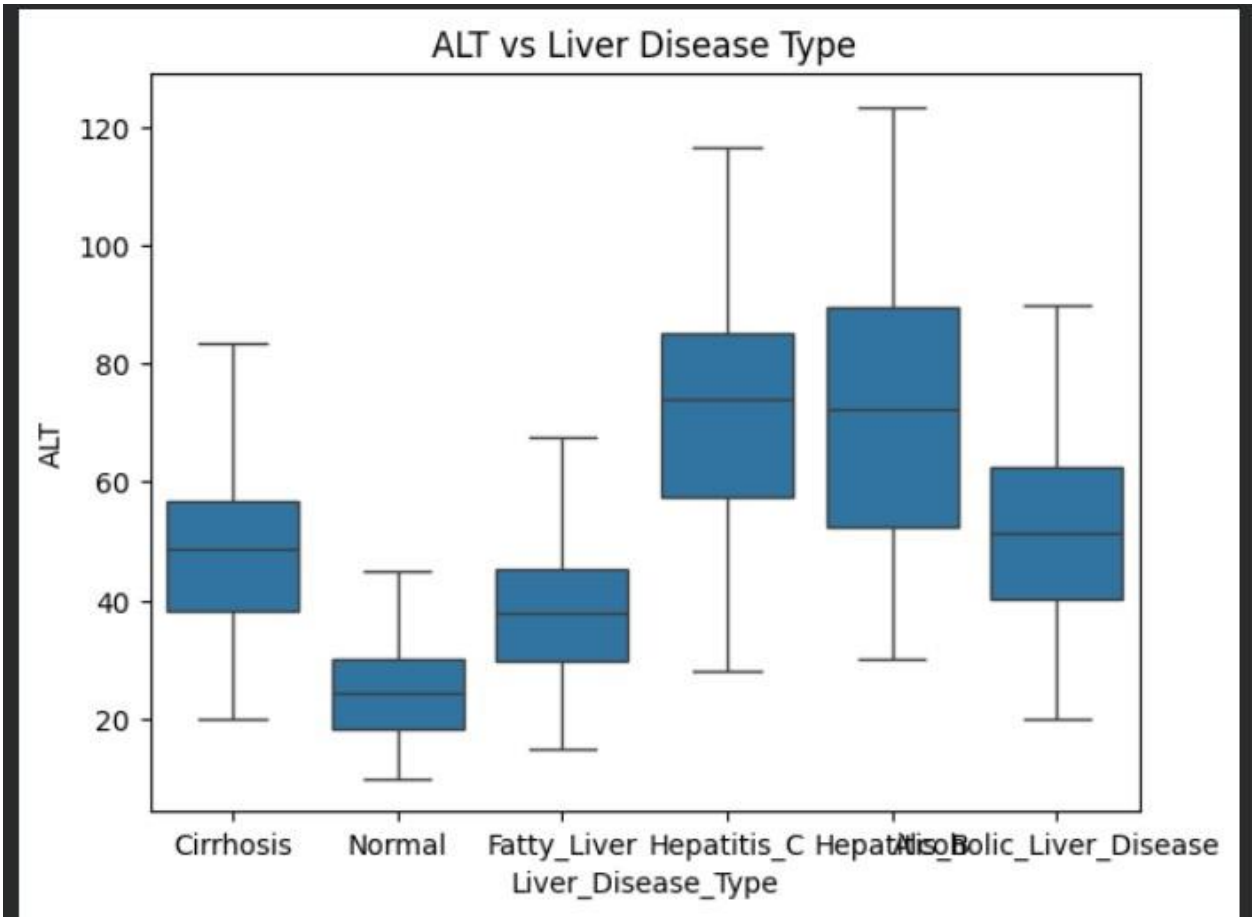



Figure 4 shows a box plot of the distribution of ALT (alanine aminotransferase) across the liver disease spectrums of different types. The x-axis is the categorization of the liver situations, that is, Cirrhosis, Normal, Fatty Liver, Hepatitis C, Hepatitis B, and Alcoholic Liver Disease, whereas the y-axis is used to measure ALT levels between 0-120. All the box plots show the median, the range of the third quartile, and the possible outliers of ALT per disease category. It is important to mention that patients with Hepatitis C and Alcoholic Liver Disease have greater median ALT levels with increased variability, which indicates a greater inflammation or damage of the liver. Conversely, in the Normal group, the value of ALT is low and more concentrated showing that the liver is functioning normally. The existence of outliers in various categories is an indication of the existence of a few cases with extraordinary high ALT levels that may be subjected to additional clinical research. On balance, this visualization allows distinguishing between the types of liver diseases according to the activity of the enzymes and proves the diagnostic topicality of ALT as a biological marker.

# Conclusion

This research paper managed to use machine learning classification framework to forecast liver disease using clinical and lifestyle data. A number of the models were tested and the model which performed well in the evaluation on critical evaluation measures was selected and therefore, it was shown that it made reliable and consistent predictions. The findings reveal the possible uses of machine learning in the early diagnosis of liver disease and medical decisionmaking. In general, the results imply that classification models can be efficient healthcare analytics instruments in the case when relevant preprocessing, model selection, and evaluation methods are us.