Constructing a Machine learning model to determine the patient survival rate in Cirrhosis of Liver

University of Pittsburgh- School of Health and Rehabilitation Sciences

2244 HI 2454 SEC1010 Data Science and Health Informatics

Final Project

Guided By : Leming Zhou

BY:

Sree Gayatri Anusha Mylavarapu

Introduction

Cirrhosis of the liver is a progressive, irreversible condition that results in severe liver damage due to the replacement of healthy cells by scar tissue. This scar tissue is what causes the loss of function of the liver. Usually liver diseases like Cirrhosis, ascitis, hepatitis B,C , non-alcoholic fatty liver etc represent a major public burden worldwide affecting millions of individuals and having significant health costs associated. It is important to know how many patients survive with liver injury so that treatment planning and overall patient care can be improved. Currently, the only treatment for liver injury is liver transplantation.

Despite advances in the medical and health science, accurately predicting patient survival rates are difficult and based on the current methods of testing like clinical tests, laboratory tests and psychometrics there may be lack of accuracy which is required for personalized predictions and with the use of machine learning and using the available datasets we can make predictions.

Motivation -

The motivation of this project is to use the machine learning techniques and accurately determine the patient survival rates for cirrhosis of liver. Machine learning can help us to use a wide range of data like patient data, demographic data , clinical data etc.

This project stems for the accurate prediction of the survival rates to improve the prognosis and overall care of the patients with this liver injury.

1. Treatment plan - Predicting the survival rate allows the healthcare professionals to curate tailored treatments and reduce the risk factors.
2. Find resources - If the patient survival rate is known then the hospitals can find the necessary resources required for the patient in time without any interventions.

3. Decision making - Clinicians can determine if the patient would need an immediate transplant or can be cured by any other methods and improve patient outcome.

Objective -

The main objective is to provide the clinicians and healthcare professionals about the insights of the factors that affect the survival prognosis of liver injury and enhance the decision making but getting to know the prognosis of the liver injury.

Some major objectives that can be seen here are -

- Develop a model to determine the patient survival rate.
- To do this we need to collect and prepare the data
- Select a target variable and select the features we need and which are relevant and develop a machine learning model.
- Evaluate the models performance by finding the accuracy , precision, recall and F1 score to enhance the predictive performance.
- By achieving these objectives the project aims to provide better patient outcome and quality of life.
- The goal of the model is to provide clinicians with translational insight into factors that affect survival prognosis for liver injury. The goal is to enhance clinical decision making and patient care, as well as to contribute to a better understanding of the prognosis of liver injury.

Description-

Prediction of patient survival rates in cirrhosis is essential as it can be a guiding hand to the clinicians and help in treatment decisions , allocate resources and improve the patient outcomes.

Although machine learning can provide us with a lot of information and predictions it is important to know how to accurately predict the chances of survival for individuals with cirrhosis.

Currently the methods generally used are clinical examinations , lab testings and other evaluations. These methods are basic but may lack precision and may not fully capture all the factors.

This is where machine learning comes in hand to address these challenges and these can access large scale data sets with extreme diversity yet providing outcomes. Utilizing these models can provide all the needed information to clinicians and hospitals and also to the patients.

Overview of Data -

The data source I am using for this project is taken from kaggle datasets and it has 418 rows and 20 columns and this data is taken from Mayo clinic from 1974 to 1985 which is a 10 year span and in this 312 people took part in the trial and the other 112 patients did not but participated in providing information. A brief description of the dataset indicates that it includes important variables such as patient personality, number of days of observation, patient condition, type of medication, age, sex, presence of ascites, liver internal weight gain, spiders, edema, as well as various labs bilirubin levels, fatty acids , albumin and other measurements. The data set is structured to capture a comprehensive view of the patient's health status and medical history. Each entry in the dataset represents a unique patient with cirrhosis, and the target variable may be the patient's survival status or length of stay Dataset abundance and source the inclusion of

survival makes it ideally suited for training a predictive model to assess patient survival rates in cirrhosis.

Key Variables in this dataset are -

- Demographic data
- Laboratory data
- Medication history data
- Status of survival which is my target variable : This variable has three entities which is D- Death (If the patient survived or not), C - Censored or CL - Censored due to liver injury.
- Each entry in this dataset shows a patient with cirrhosis exhibiting different factors that influenced their liver injury.
- By analyzing this data we can identify the main factors and features and improve prognosis and accuracy.

Methods of Data Analysis -

As mentioned in the proposal the 8 week plan of rationale of the project goes as follows

Week 1 - Data Exploration and Preprocessing

- During the first week of exploratory data analysis (EDA). I have taken time to understand the structure of the data, understanding the data length and shape , identifying the missing values and gaining insights about the features and their distribution.

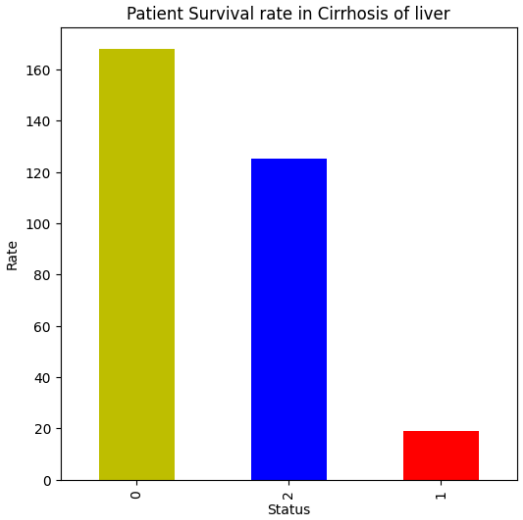
- Handling the missing values included dropping rows and columns which were not necessary.
- This dataset consists of 418 rows and 20 columns including ID, number of days observed , Status ,age ,sex, ascites etc..
- The data.info function provides a summary of the dataset and the data type. This helps in understanding the structure of the dataset like identifying missing values or knowing about the distribution of numerical features.
- The data.describe function gives the count , mean ,min, max values and provides an understanding of the range of values and also helps in detecting outliers.
- Visualizing the data and understanding the relationship between features and target variable
- Here my target variable is Status (D,C or CL)

Week 2 - Data Preprocessing and feature selection and engineering

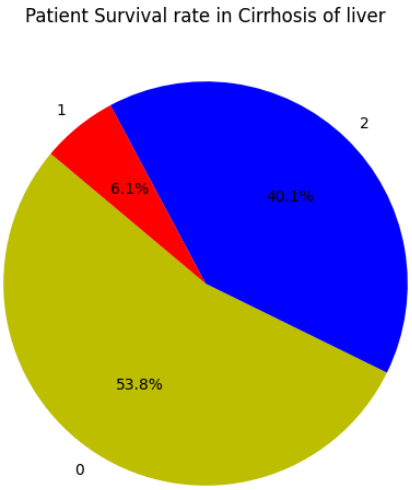
- I have encoded categorical variable using label encoding
- In this step I have used the data.dropna feature even though null valued rows account for 25% of the dataset. Dropping makes more sense than filling as multiple columns are NULL for those rows.
- Performed normalization to ensure feature selection as this feature selection provides better variables for patient survival analysis
- This preprocessing step is done to clean and prepare the data to run further machine learning models. By handling missing values and identifying extra or duplicate values and removing them can create a reliable predictive model.

Week 3 - Data Visualisation and model selection

- I have created a bar chart and a pie chart for data visualization and converted the Status into numerical values for easier prediction so 0 is Censored , 1 is Censored due to liver injury and 2 is death



As we can observe that Censored is the highest and Censored due to liver injury is the lowest



Death is 40.1% , Censored is 53.8% and Censored due to liver injury is 6.1%

- Model selection is generally done by splitting the data to training and testing sets. I have used Gaussian Naive Bayes in this project.
- Evaluation metrics and models performance is established by using metrics like accuracy, precision , recall and F1 score.

Week 5 and 6 - Model Training and Evaluation

- I have explored different machine learning models like random forest, decision tree and SVM - support vector machines.
- I have trained each model using training and testing sets and compared the performance of each model using these sets and also done evaluation metrics.
- Implementation of techniques to the trained models to understand the most important features in predicting the survival rates
- The most important features of my dataset are ‘ID', ‘N_days’,’Age’,’Bilirubin’, ‘Albumin’, ‘Platelets’, ‘Prothrombin’.

Week 7 - Model Validation and Comparison

- I have performed cross- validation techniques such as k- fold to obtain a reliable models performance.
- I have compared the model's performance with other models to find the best model for prognosis of survival rates.

Week 8 - Documentation

- Documentation of the entire process including data preprocessing, feature selection, and discussion of the entire process.
- The 8 week approach covers essential steps like data analysis , model development and evaluation.

Results -

I have obtained different results for different models

The overall accuracy of the model is 0.78571

Results obtained in Naive Bayes: Accuracy - 0.651, Precision - 0.562, Recall - 0.552 , F1 score - 0.552

SVM : Accuracy-0.683, Precision - 0.457, Recall - 0.521, F1 score - 0.486

Decision tree classifier - Accuracy - 0.556 , precision - 0.459, Recall - 0.452, F1 score 0.452

Cross validation scores - [0.666 0.6984 0.72580 0.6935 0.677419]

Mean cross- validation score : 0.692370711

The mean cross-validation score of 69.24% suggests that the model has a moderate level of predictive performance for this particular dataset and problem.

Alternative approach - If the first approach fails, an alternative is to explore more sophisticated machine learning algorithms, such as gradient growth , to capture complex relationships in the data. If domain experts work together it can lead to a more targeted

approach. Ensemble methods can be used by combining forecasts from multiple models to deal with potential errors. Finally, analyzing other relevant datasets to enhance the existing information may provide a comprehensive approach to more robust predictive modeling.

Distribution -

Accuracy:

- The overall accuracy across all models is reported as 0.78571 or 78.57%, which indicates a reasonably good predictive performance.
- Among the individual models, the Naive Bayes model achieved the highest accuracy of 0.651 or 65.1%.
- The SVM model had an accuracy of 0.683 or 68.3%.
- The Decision Tree Classifier model had the lowest accuracy of 0.556 or 55.6%.

Precision:

- The Naive Bayes model had the highest precision of 0.562, indicating a moderate ability to avoid false positives.
- The SVM model had a lower precision of 0.457, suggesting a higher rate of false positives.
- The Decision Tree Classifier model had a precision of 0.459, similar to the SVM model.

Recall:

- The SVM model had the highest recall of 0.521, meaning it correctly identified a higher proportion of true positive cases (survivors).

- The Naive Bayes model had a recall of 0.552, slightly higher than the Decision Tree Classifier model's recall of 0.452.

F1-score:

- The Naive Bayes model achieved the highest F1-score of 0.552, indicating a balanced performance between precision and recall.
- The SVM model had a lower F1-score of 0.486, reflecting a trade-off between precision and recall.
- The Decision Tree Classifier model had the lowest F1-score of 0.452, also indicating a lower balance between precision and recall.

Cross-Validation Scores:

- The cross-validation scores ranged from 0.666 to 0.72580, with a mean cross-validation score of 0.692370711 or approximately 69.24%.
- These cross-validation scores suggest that the models have a moderate level of predictive performance on unseen data and a reasonable ability to generalize.

Overall, the distribution of evaluation metrics across the different models highlights their varying strengths and weaknesses. The Naive Bayes model demonstrated a balanced performance, while the SVM model prioritized recall (identifying true positives), and the Decision Tree Classifier maintained a balance between precision and recall, albeit with lower overall scores.

It's important to note that the choice of the most suitable model may depend on the specific requirements and priorities of the application. For example, in a medical setting, a higher recall (identifying more true survivors) might be prioritized over precision to avoid missing potential

survivors. Conversely, in other scenarios, a higher precision (minimizing false positives) might be more crucial.

The cross-validation scores provide further validation of the models' moderate predictive capabilities and generalization potential, indicating their potential for practical applications with appropriate fine-tuning and optimization.

Limitations -

1. Data Quality Issues: Incomplete or inaccurate data can affect the model's reliability.
2. Bias and Generalization: Biased datasets may limit the model's applicability to diverse patient populations.
3. Complexity of Cirrhosis: Predicting cirrhosis outcomes is challenging due to its multifaceted nature.
4. Limited Predictive Power: Unforeseen factors may limit the model's ability to accurately forecast outcomes.
5. Ethical Considerations: Privacy concerns and ethical implications must be addressed when using patient data.
6. Interpretability and Transparency: Complex models may lack transparency, hindering understanding and trust.
7. Dynamic Data Nature: The model needs regular updates to account for evolving patient conditions.

Conclusion-

Cirrhosis of the liver is a condition which has limited treatment options. This study aimed to develop a machine learning model to predict the survival rate. I have trained multiple models including Naive Bayes, SVM and Decision Tree Classifier and evaluate metrics like accuracy, precision, recall and F1 score. Of all, Naive Bayes had a balanced performance with an accuracy of 0.651.

In conclusion, while the developed model represents a significant step forward in predicting survival rates in cirrhosis patients, ongoing refinement and validation are necessary to enhance its predictive power and clinical utility further. By addressing limitations, incorporating feedback from clinical experts, and embracing advances in machine learning techniques, we can continue to improve the model's efficiency and effectiveness in improving patient outcomes and advancing liver disease management.

Dataset I used -

https://www.kaggle.com/datasets/joebeachcapital/cirrhosis-patient-survival-prediction

REFERENCES

- *Cirrhosis patient Survival prediction*. (2023, October 17). Kaggle. https://www.kaggle.com/datasets/joebeachcapital/cirrhosis-patient-survival-prediction
- Department of Health & Human Services. (n.d.). *Cirrhosis of the liver*. Better Health Channel. https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/cirrhosis-of-the-liver
- Huang, D. Q., Terrault, N. A., Tacke, F., Gluud, L. L., Arrese, M., Bugianesi, E., & Loomba, R. (2023). Global epidemiology of cirrhosis — aetiology, trends and predictions. *Nature Reviews Gastroenterology & Hepatology*, *20*(6), 388–398. https://doi.org/10.1038/s41575-023-00759-2
- Jocelyndumlao. (2023, December 8). *Predicting cirrhosis outcomes*. Kaggle. https://www.kaggle.com/code/jocelyndumlao/predicting-cirrhosis-outcomes#--Import-Modules
- Schuppan, D., & Afdhal, N. H. (2008). Liver cirrhosis. *The Lancet*, *371*(9615), 838–851. https://doi.org/10.1016/s0140-6736(08)60383-9
- Sharma, B., & John, S. (2022, October 31). *Hepatic cirrhosis*. StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK482419/
- DATA FILE - https://www.kaggle.com/datasets/fedesoriano/cirrhosis-prediction-dataset
- Prediction of long-term survival among patients with cirrhosis using time-varying models David Goldberg and Yalda Zarnegarnia https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10241498/

- Predicting mortality among patients with liver cirrhosis in electronic health records with machine learning Aixia Guo 1, Nikhilesh R Mazumder 2 3, Daniela P Ladner 3 4, Randi E Foraker 1 5 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8407576/#:~:text=Deep%20learning%20models%20(DNN)%20can,by%20the%20MELD%2DNa%20score