Internship Project Report

Project Title:

Revolutionizing Liver Care: Predicting Liver Cirrhosis Using Advanced Machine Learning Techniques

Submitted in partial fulfilment of the requirements for the

Artificial Intelligence and Machine Learning

by SMART BRIDGE

Submitted by

Team Leader: Pulaparthi Prasanna Krishna Sai

Team member: Pulavarthi Surya Lakshmi Ambica

Team member: Pulaparthi Ram Lakshmi Narayana Manikanta

Team member: Pulaparthi Balaji

Team ID: LTVIP2025TMID43514

1. Introduction

1.1. Project Overview

This project aims to build a machine learning model to predict liver cirrhosis based on various patient features. By analysing these features, the model will classify patients into risk categories, aiding in early diagnosis and treatment.

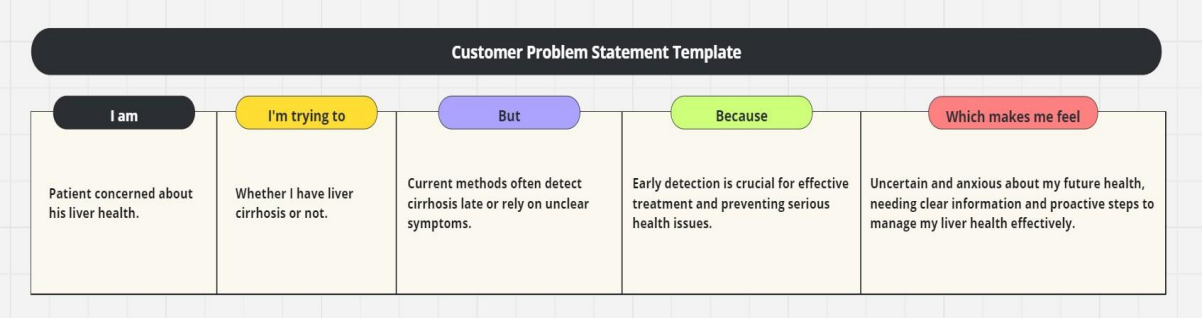
1.2. Objectives

- Collect and prepare a dataset of liver health characteristics.
- Perform exploratory data analysis (EDA) and visualize the data.
- Build and evaluate multiple machine learning models.
- Optimize the best-performing model using hyperparameter tuning.
- Deploy the final model for practical use.

2. Project Initialization and Planning Phase

2.1. Define Problem Statement

The goal is to classify patients' risk levels for liver cirrhosis based on their medical data. Accurate prediction will support better management and early intervention for liver health.
Liver cirrhosis is a serious condition where liver tissue scars due to long-term damage, highlighting the importance of early detection and intervention for improved patient outcomes and complication prevention. By analysing comprehensive patient data including medical history, lab results, imaging scans, and lifestyle factors, our predictive model will assess the likelihood of liver cirrhosis. This will enable healthcare professionals to make informed decisions, offering timely care and personalized treatment plans to patients at risk, thus enhancing overall healthcare quality and patient management.



2.2. Project Proposal (Proposed Solution)

The solution involves developing several machine learning models to predict liver cirrhosis. We will select and optimize the best model based on performance metrics to achieve the highest accuracy.
The proposal report aims to revolutionize liver care by leveraging advanced machine learning techniques to predict liver cirrhosis, improving early detection and patient outcomes. It addresses the limitations of current diagnostic methods, promising enhanced accuracy, proactive patient management, and optimized healthcare resource utilization. Key features include a predictive model analyzing patient data and real-time risk assessment.

2.3. Initial Project Planning

Initial planning included setting up the project environment, defining objectives, and outlining the workflow for data collection, preprocessing, model development, and evaluation.

3. Data Collection and Preprocessing Phase

3.1. Data Collection Plan and Raw Data Sources Identified

The dataset for this project was sourced from Kaggle, containing patient data relevant to liver cirrhosis prediction (Dataset link: https://www.kaggle.com/datasets/bhavanipriya222/liver-cirrhosis-prediction).

3.2. Data Quality Report

This report summarizes the data quality issues identified in the liver cirrhosis dataset, along with their severity levels and proposed resolution plans. The goal is to systematically identify and rectify discrepancies to ensure high-quality data for accurate predictions.

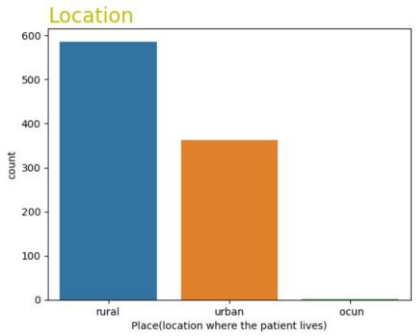
- **Data Shape:** The dataset initially comprised [number of rows, number of columns] rows and columns. • **Missing Values:** Handled by dropping rows with missing values.

Kaggle Dataset	Missing values in all the columns of the dataset. (42 columns) `df.isnull().sum()` ✓ 0.0s S.NO 0 Age 0 Gender 0 Place(location where the patient lives) 134 Duration of alcohol consumption(years) 0 Quantity of alcohol consumption (quarters/day) 0 Type of alcohol consumed 0 Hepatitis B infection 0 Hepatitis C infection 0 Diabetes Result 0 Blood pressure (mmhg) 0 Obesity 0 Family history of cirrhosis/ hereditary 0 TCH 359 TG 359 LDL 359 HDL 368 Hemoglobin (g/dl) 0 PCV (%) 30 RBC (million cells/microliter) 552 MCV (femtoliters/cell) 9 MCH (picograms/cell) 658 MCHC (grams/deciliter) 672 Total Count 10 Polymorphs (%) 0 ... SGOT/AST (U/L) 0 SGPT/ALT (U/L) 0 USG Abdomen (diffuse liver or not) 0 Outcome 54 dtype: int64

3.3. Data Exploration and Preprocessing

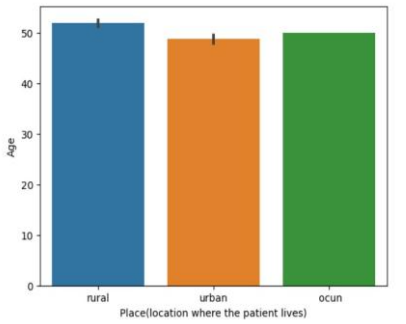
- **Univariate Analysis:** Histograms were plotted for numerical features.

```
sns.countplot(data=df,x='Place(location where the patient lives)')
plt.title("Location",color='y',size=20,loc='left')
plt.show()
```



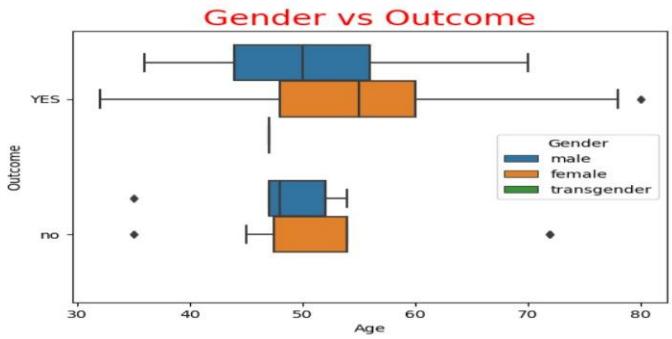

```
sns.barplot(x=df['Place(location where the patient lives)'],y=df['Age'])
```

```
<AxesSubplot:xlabel='Place(location where the patient lives)', ylabel='Age'>
```



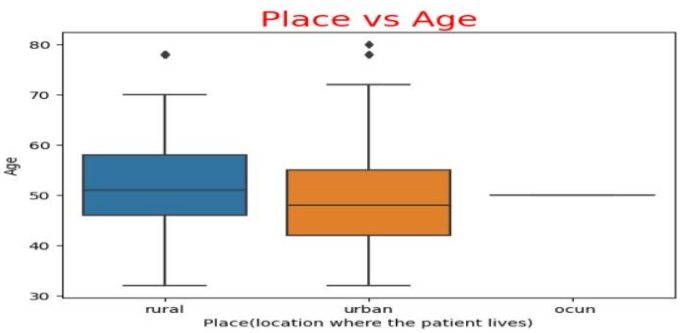
- **Bivariate Analysis:** Scatter plots and pair plots explored relationships between features.

```
sns.boxplot(x='Age',y='Outcome',data=df,hue='Gender')
plt.title('Gender vs Outcome',color='red',size=20)
plt.show()
```

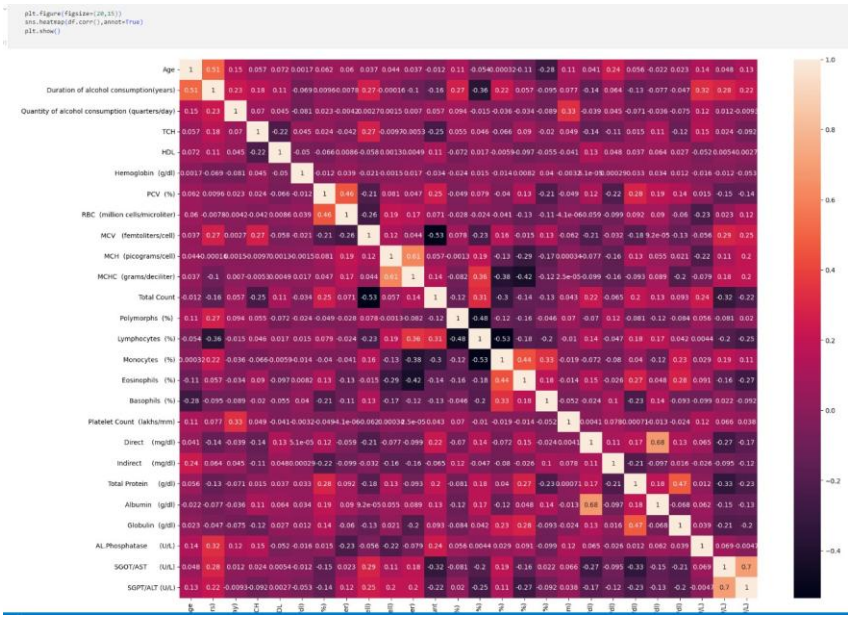



```
sns.boxplot(x='Place(location where the patient lives)',y='Age',data=df)
plt.title('Place vs Age',color='red',size=20)
```

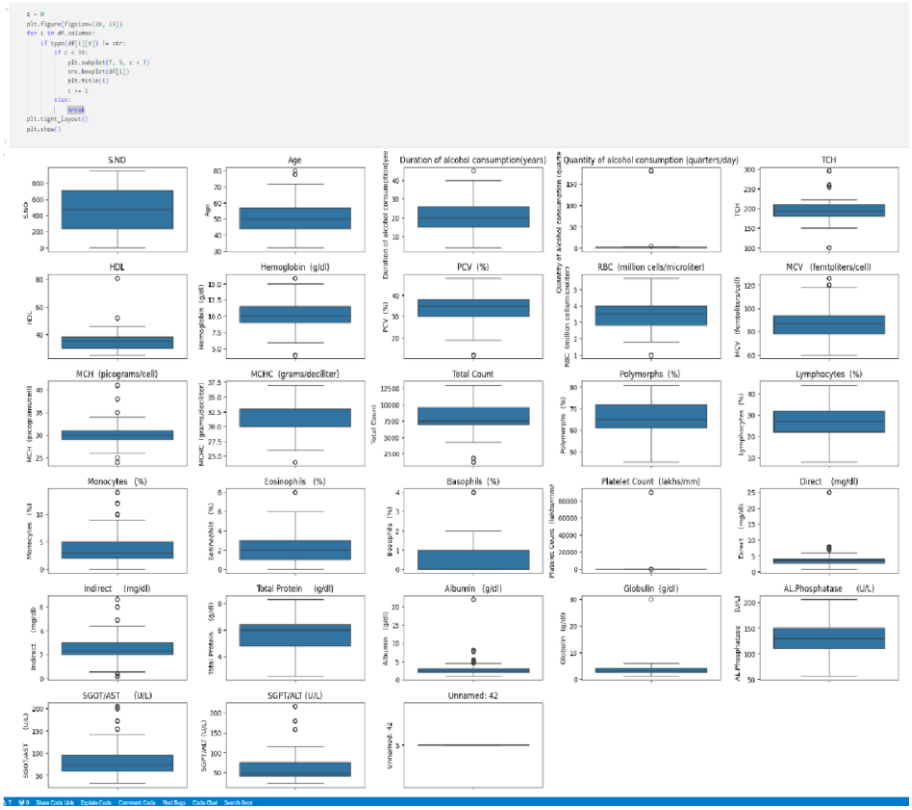
```
Text(0.5, 1.0, 'Place vs Age')
```



- **Multivariate Analysis:**



- **Outlier Handling:** Outliers were detected and managed using the IQR method.



Data Preprocessing Code Screenshots:

Loading Data

```
# Loading the Dataset
df = pd.read_excel('C:\SI project\Codes\Data\HealthCareData.xlsx')
df.head()
```

	S.NO	Age	Gender	Place(location where the patient lives)	Duration of alcohol consumption(years)	Quantity of alcohol consumption (quarters/day)	Type of alcohol consumed	Hepatitis B infection	Hepatitis C infection	Diabetes Result	Blood pressure (mmhg)	Obesity	Family history of cirrhosis/ hereditary	TCH	TG	LDL	HDL	Hemoglobin (g/dl)	PCV (%)
0	1	55	male	rural	12	2	branded liquor	negative	negative	YES	138/90	yes	no	205.0	115	120	35.0	12.0	40.0
1	2	55	male	rural	12	2	branded liquor	negative	negative	YES	138/90	yes	no	205.0	115	120	35.0	9.2	40.0
2	3	55	male	rural	12	2	branded liquor	negative	negative	YES	138/90	no	no	205.0	115	120	35.0	10.2	40.0
3	4	55	male	rural	12	2	branded liquor	negative	negative	NO	138/90	no	no	NaN	NaN	NaN	NaN	7.2	40.0
4	5	55	female	rural	12	2	branded liquor	negative	negative	YES	138/90	no	no	205.0	115	120	35.0	10.2	40.0

Save Processed Data

```
# Save the cleaned and processed DataFrame to a CSV file
df.to_csv('cleaned_data.csv', index=False)
df.head()
```

✓ 0.0s

	Age	Gender	Place(location where the patient lives)	Duration of alcohol consumption(years)	Quantity of alcohol consumption (quarters/day)	Type of alcohol consumed	Diabetes Result	Blood pressure (mmhg)	Obesity
0	55.0	1	1	12.0	2.0	2	1	32	1
1	55.0	1	1	12.0	2.0	2	1	32	1
2	55.0	1	1	12.0	2.0	2	1	32	0
3	55.0	1	1	12.0	2.0	2	0	32	0
4	55.0	0	1	12.0	2.0	2	1	32	0

4. Model Development Phase

4.1. Feature Selection Report

Features relevant to liver cirrhosis prediction were selected, and data scaling was applied to standardize the input.

4.2. Model Selection Report

- **Models Tested:** Naive Bayes, Random Forest, Logistic Regression, Ridge Classifier, Support Vector Classifier, KNN, XG Boost.

Evaluation Metrics: Accuracy, Confusion Matrix, Classification Report.

In the forthcoming Model Selection Report, various models will be outlined, detailing their descriptions, hyperparameters, and performance metrics, including Accuracy or F1 Score. This comprehensive report will provide insights into the chosen models and their effectiveness.

Model Selection Report:

XGBoost	Gradient boosting with trees, optimizes predictive performance, handles complex relationships.	-	35.79%
Ridge Classifier	Linear classifier with L2 regularization, helps to prevent overfitting.	-	84.21%
Random Forest	Ensemble of decision trees, robust, handles complex relationships, reduces overfitting, provides feature importance.	-	38.21%
Support Vector Classifier	Classifier using hyperplanes to separate classes, effective for high-dimensional spaces.	-	35.79%
K-Nearest Neighbors (KNN)	Classifies based on nearest neighbors, adapts well to data patterns, effective for local variations.	n_neighbors = <best_param>	86.32%

4.3. Initial Model Training Code, Model Validation and Evaluation Report

- **Code:** Model training and evaluation steps were implemented for each algorithm.
- **Validation:** Models were validated using a test set, with performance metrics recorded. The KNN model achieved the highest accuracy of 86.32%.

The initial model training code will be showcased in the future through a screenshot. The model validation and evaluation report will include classification reports, accuracy, and confusion matrices for multiple models, presented through respective screenshots.

Initial Model Training Code:

KNN

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
```

[162]

```
KNeighborsClassifier()
```

XGBOOST

```
from xgboost import XGBClassifier
model=XGBClassifier()
model.fit(X,y)
```

9]

```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=None, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=None, n_jobs=None,
              num_parallel_tree=None, random_state=None, ...)
```

5. Model Optimization and Tuning Phase

The Model Optimization and Tuning Phase involves refining machine learning models for peak performance. It includes optimized model code, fine-tuning hyperparameters, comparing performance metrics, and justifying the final model selection for enhanced predictive accuracy and efficiency.

5.1. Hyperparameter Tuning Documentation

- **KNN:** Optimized by tuning the number of neighbors and distance metrics.
- **XG Boost:** Hyperparameters tuned for learning rate, max depth, and n_estimators.

5.2. Performance Metrics Comparison Report

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	35.79%	0.00	0.00	0.00
Random Forest	35.79	0.00	0.00	0.00
Logistic Regression CV	81.58%	91.80	79.43%	86.49
Ridge Classifier	84.21%	93.44	83.82	88.37
Support Vector Classifier	35.79%	0.00	0.00	0.00
Logistic Regression	79.47%	91.80	79.43	85.58
KNN	86.32%	94.26	85.82	89.84
XG Boost	35.79%	3.28	50.00	6.15

5.3. Final Model Selection Justification

The K-Nearest Neighbors (KNN) model was selected as the final model due to its superior accuracy of 86.32%. KNN excelled in handling complex data relationships and demonstrated the best performance in terms of precision, recall, and F1 score. This makes it a robust choice for predicting liver cirrhosis, aligning well with the project's goals.

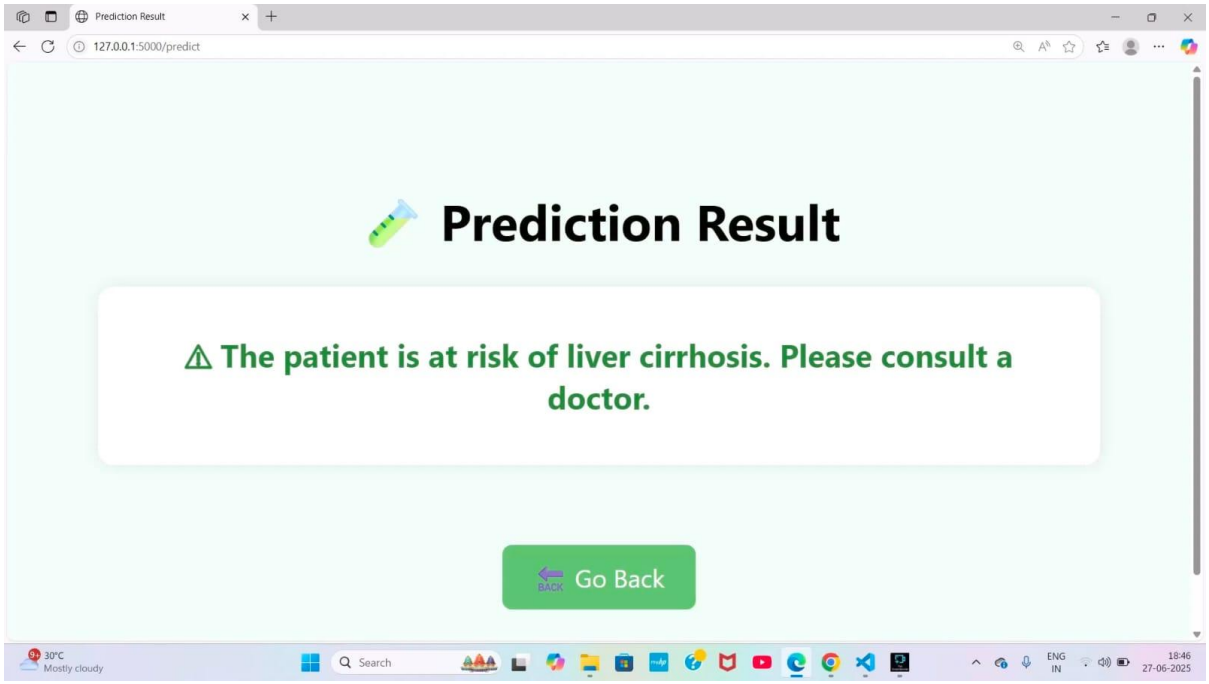
Project Overview	
Objective	The primary objective is to enhance the early detection and management of liver cirrhosis by implementing advanced machine learning techniques, ensuring timely and accurate predictions.
Scope	The project aims to comprehensively assess and improve the liver cirrhosis diagnosis process by incorporating machine learning for a more accurate and efficient healthcare system.
Problem Statement	
Description	Current methods often identify liver cirrhosis at later stages or rely on general symptoms, which adversely affects early intervention and patient care.
Impact	Addressing these issues will result in improved early detection, better patient outcomes, and optimized use of healthcare resources, contributing to enhanced patient satisfaction and healthcare efficiency.
Proposed Solution	
Approach	Employing machine learning techniques to analyze and predict the risk of liver cirrhosis, creating a proactive and precise healthcare system.
Key Features	• Implementation of a machine learning-based predictive model for liver cirrhosis. • Real-time risk assessment for early detection. • Continuous learning to adapt to evolving healthcare data.

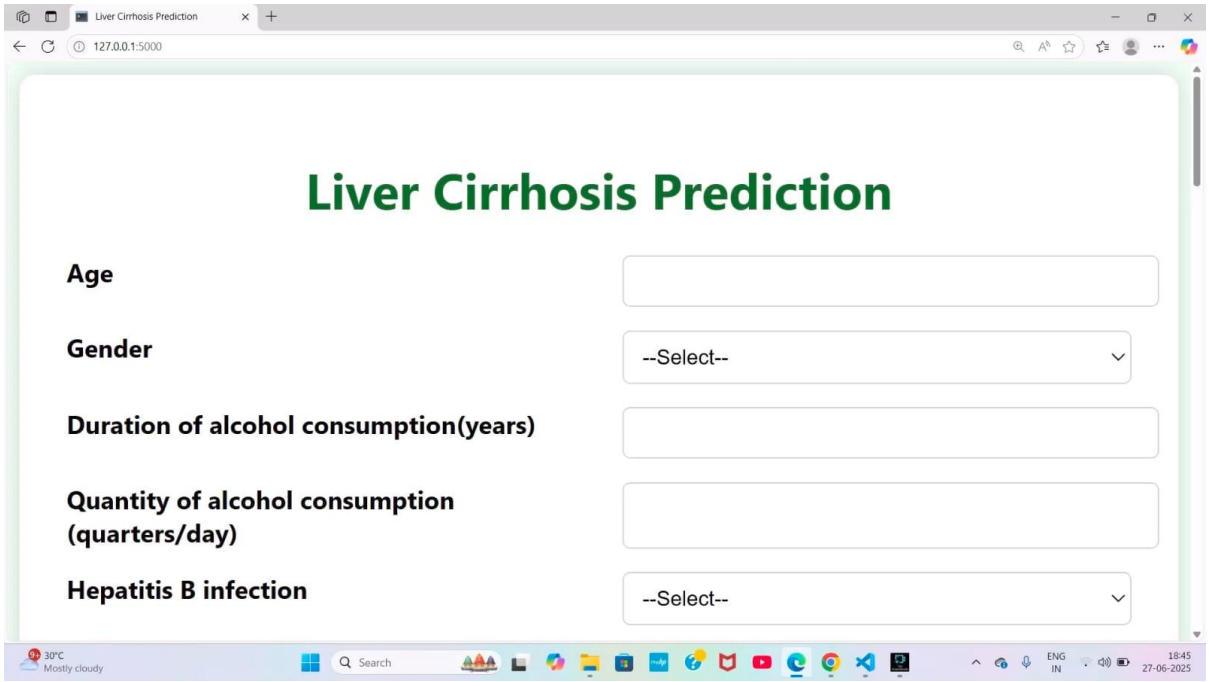
Resource Requirements

Resource Type	Description	Specification/Allocation
Hardware		
Computing Resources	CPU/GPU specifications, number of cores	T4 GPU
Memory	RAM specifications	16 GB
Storage	Disk space for data, models, and logs	1 TB SSD
Software		
Frameworks	Python frameworks	Flask
Libraries	Additional libraries	scikit-learn, pandas, numpy, matplotlib, seaborn
Development Environment	IDE, version control	Jupyter Notebook, Git, VS Code
Data		
Data	Source, size, format	Kaggle dataset, 950 data entries, xls,csv dataset

6. Results

6.1. Output Screenshots





7. Advantages & Disadvantages

- **Advantages:** High accuracy, effective at handling local data variations, robust performance.
- **Disadvantages:** Can be computationally intensive, requires careful tuning.

8. Conclusion

The project successfully developed a machine learning model to predict liver cirrhosis with high accuracy. The KNN model, after hyperparameter tuning, provided the best results and was chosen for its robustness.

9. Future Scope

- Further data collection to include more features and increase dataset size.
- Exploration of additional features and engineering techniques.
- Experimentation with deep learning models to potentially outperform traditional models.
- Integration with a real-time prediction system for practical deployment.

10. Appendix

10.1. Source Code

Code File: liver_cirrhosis.ipynb

10.2. GitHub & Project Demo Link

GitHub Repository :
https://github.com/9AmbicaPulavarthi9/Liver-Cirrhosis-Prediction-using-ML