**Healthcare Analytics for Disease Prediction**

**INTERNSHIP PROJECT REPORT**

OF PROJECT

**BACHELOR OF TECHNOLOGY**

**Department of Computer Science and Engineering**

## SUBMITTED BY

## Ms. Akanksha Arjun More

**March 2024**

**Contents**

1. [**Introduction 4**](#_30j0zll)
2. [**Objectives 6**](#_1fob9te)
3. [**Scope of Project 6**](#_3znysh7)
4. [**Problem Statement 6**](#_2et92p0)
5. [**Literature Survey 7**](#_tyjcwt)
6. [**Methodology 10**](#_3dy6vkm)
7. [**Facilities required for proposed work 17**](#_1t3h5sf)
8. [**Schedule of the project (Gantt chart) 18**](#_4d34og8)
9. [**Implementation and Output 19**](#_2s8eyo1)
10. [**References 24**](#_17dp8vu)

# Introduction

Healthcare analytics has emerged as a pivotal tool in revolutionizing healthcare systems worldwide. By harnessing the power of data analysis, healthcare providers can glean valuable insights to enhance patient care, optimize resource allocation, and drive informed decision-making. In this era of digital transformation, the integration of advanced analytics techniques with healthcare data holds immense promise in improving clinical outcomes and operational efficiency.

Liver cirrhosis stands as a significant healthcare challenge, representing the advanced stage of scarring (fibrosis) of the liver caused by various liver diseases and conditions. Chronic liver diseases such as hepatitis and excessive alcohol consumption are primary contributors to the development of liver cirrhosis. The progression of liver cirrhosis is marked by distinct stages, ranging from early fibrosis to advanced cirrhosis, each carrying different prognostic implications.

Given the complex and multifactorial nature of liver cirrhosis, accurately predicting its progression is of paramount importance for clinicians. Early identification of the disease stage enables timely interventions, personalized treatment strategies, and improved patient outcomes. However, the prediction of liver cirrhosis progression poses several challenges, including the integration of diverse clinical data, the presence of missing values, and the complexity of disease trajectories.

Against this backdrop, this project endeavors to leverage healthcare analytics techniques to develop a robust predictive model for liver cirrhosis progression. By harnessing machine learning algorithms and advanced data analytics methodologies, we aim to unravel hidden patterns within healthcare data to predict the likelihood of liver cirrhosis progression accurately. Through this endeavor, we aspire to contribute to the early detection, proactive management, and improved prognosis of liver cirrhosis patients, thereby advancing the field of predictive medicine and enhancing healthcare delivery.

# Objectives

1. Data Collection and Preprocessing: Gather and clean healthcare data related to liver cirrhosis.
2. Exploratory Data Analysis (EDA): Analyze the dataset to understand its structure and characteristics.
3. Disease Prediction Model Development: Build machine learning models to predict liver cirrhosis stages.
4. Model Evaluation: Assess the performance of the predictive models.
5. Insights for Early Intervention and Prevention: Provide actionable insights for proactive management of liver cirrhosis.

# Scope of Project

1. Analyze healthcare data to predict the likelihood of a disease.
2. Utilize machine learning algorithms for classification.
3. Provide insights for early intervention and prevention.

# Problem Statement

To implement a healthcare analytics solution for disease prediction, focusing on leveraging machine learning algorithms to analyze patient data and accurately predict the progression of liver cirrhosis.

# 5. Literature Survey

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr. No** | **Title of paper** | **Name of Authors** | **Year of Publication** | **Summary Points** |
| **1** | Big Data Analytics for Prediction Modelling in Healthcare Databases | Ritu Chauhan and Eiad Yafi | 2021 | This paper checks and makes healthcare better for patients. It studies how the length of telomeres connects to the risk of getting cancer. It explores using fluorescence to find specific nodes in cancer patients. Also, it talks about using big data in healthcare for things like predicting and analyzing data. |
| **2** | Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease | P. Sujatha  and K. Mahalakshmi | 2020 | The paper looked at previous studies on using machine learning to predict heart disease. It found that many studies have used different algorithms like decision tree, naïve bayes, SVM, KNN, logistic regression, and random forest. These studies have shown that random forest is often the most accurate for predicting heart disease. The paper also noted that with more data, better models can be built, and future work could involve using different machine learning techniques with larger datasets for more accurate predictions of heart disease. |
| **3** | Predictive Analytics Model Based on Multiclass Classification for Asthma Severity by Using Random Forest Algorithm | Wasif Akbar, Sehrish Saleem, Wei-Ping Wu, Arslan Javed, Muhammad Faheem and Muhammad Asim Saleem | 2020 | The paper explores big data applications in healthcare, including predictive modeling, real-time data management, and machine learning on streaming health data using technologies like Hadoop and Spark. |
| **4** | Big Data Analytics in Healthcare | Guorong Chen and Mohaiminul Islam | 2019 | The paper discusses the use of big data analytics in the healthcare industry. It highlights how researchers are using large amounts of data to find new solutions for diseases, leading to advancements in healthcare technology and better patient care. The article also addresses the challenges of adopting big data analytics in healthcare, such as data cleaning, storage, security, and stewardship. |
| **5** | Prediction of Diabetes Using Machine Learning Algorithms in Healthcare | Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid and Munam Ali Shah | 2018 | The paper discusses various research papers and conferences related to the use of predictive analytics and machine learning in the diagnosis and prediction of diseases such as heart disease, diabetes, chronic kidney disease, breast cancer, and asthma. It also mentions the use of data mining techniques and neural networks for medical diagnosis. |

# Methodology

The methodology of this project:

1. **Data Collection and Preprocessing**: This step involves gathering relevant healthcare data related to liver cirrhosis and preprocessing it to handle missing values, outliers, and inconsistencies. This may include acquiring data from medical records, clinical databases, or other sources, and then cleaning and transforming it to make it suitable for analysis.
2. **Exploratory Data Analysis (EDA)**: EDA is conducted to gain insights into the structure and characteristics of the dataset. This involves examining variable distributions, identifying patterns, and understanding feature relationships. EDA helps in understanding the data better and aids in feature selection and engineering.
3. **Disease Prediction Model Development**: Machine learning models such as logistic regression, decision trees, random forests, or support vector machines are employed to develop predictive models for liver cirrhosis stage classification. These models utilize the pre-processed data to learn patterns and relationships between features and the target variable (i.e., liver cirrhosis stage).
4. **Model Evaluation**: The performance of the predictive models is assessed using appropriate metrics such as accuracy, precision, recall, and F1-score. This step involves splitting the data into training and testing sets, training the models on the training set, and evaluating their performance on the testing set. Model selection may involve comparing the performance of different algorithms and tuning hyperparameters to optimize performance.
5. **Documentation**: The entire process, including data collection, preprocessing steps, model development, evaluation results, and insights obtained, is documented. Documentation ensures transparency, reproducibility, and clarity in understanding the methodology and results of the project.

Overall, the methodology follows a structured approach from data collection and preprocessing to model development, evaluation, and documentation, aiming to develop a robust predictive model for liver cirrhosis progression using healthcare analytics techniques.

# Facilities required for proposed work

* Software Requirements:

1. Data Analysis Tools: Python (with pandas, NumPy, scikit-learn, TensorFlow)
2. Integrated Development Environment (IDE): Jupyter Notebook, PyCharm, Anaconda Navigator
3. Database Management System (DBMS): MySQL, PostgreSQL, SQLite
4. Statistical Software: R, SPSS
5. Version Control System: Git
6. Documentation Tools: Microsoft Word, LaTeX, Markdown editors

* Hardware Requirements:

1. Memory (RAM): Minimum 8GB
2. Storage: At least 50GB free disk space
3. Processing Power: Multi-core processor (e.g., Intel Core i5 or higher)
4. Graphics Processing Unit (GPU): Optional for deep learning
5. Internet Connection: Stable and high-speed
6. Monitor: High-resolution display

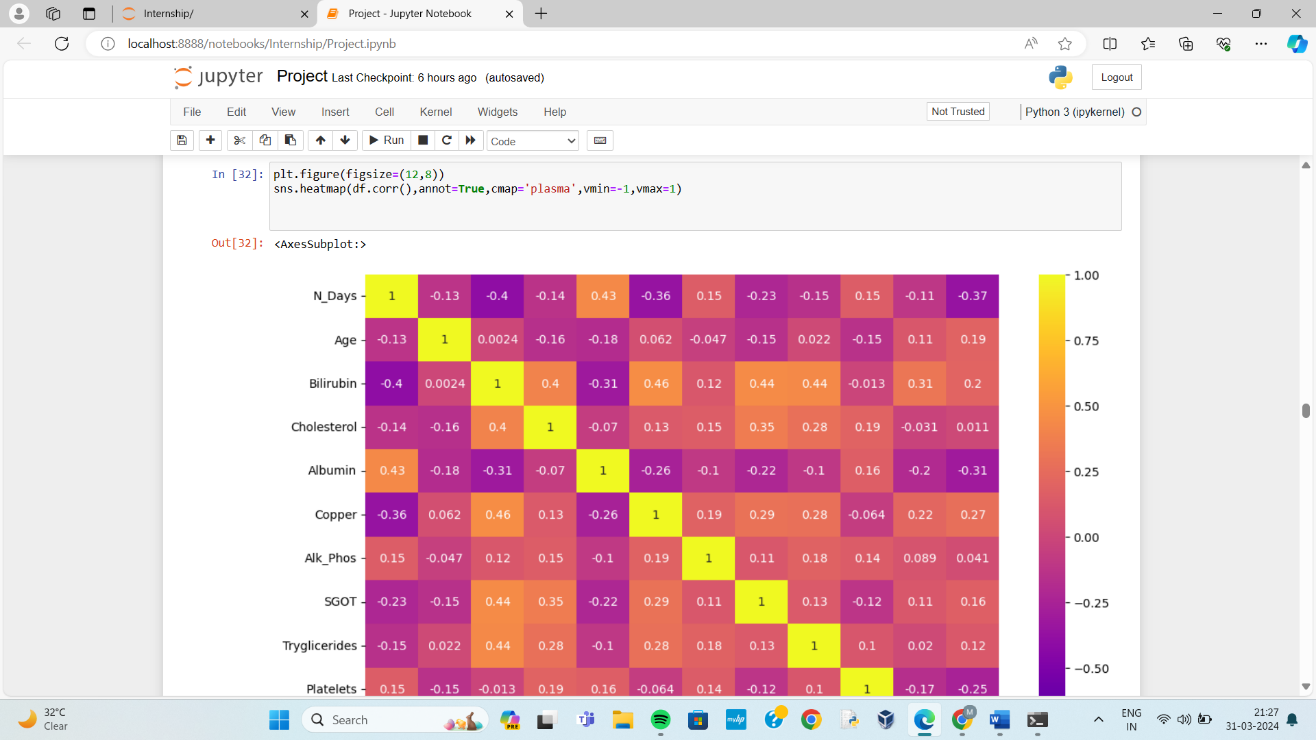
* Additional Considerations:
  1. Data Security: Compliance with HIPAA, GDPR
  2. Backup and Recovery: Implement backup solutions
  3. Collaboration Tools: Slack, Microsoft Teams, Zoom
  4. Model Deployment Environment: Cloud platforms (AWS, Azure, GCP)
  5. Ethical Considerations: Patient confidentiality, fairness, transparency

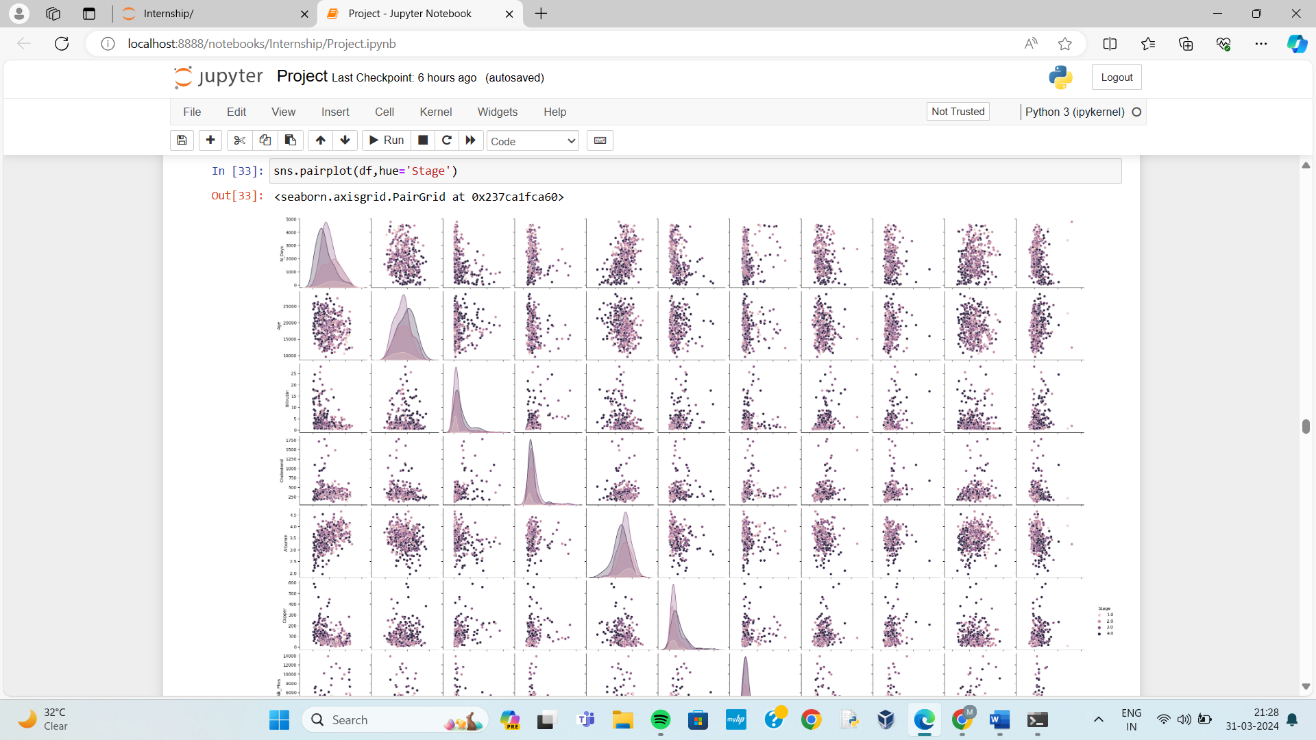
# Schedule of the project (Gantt chart)

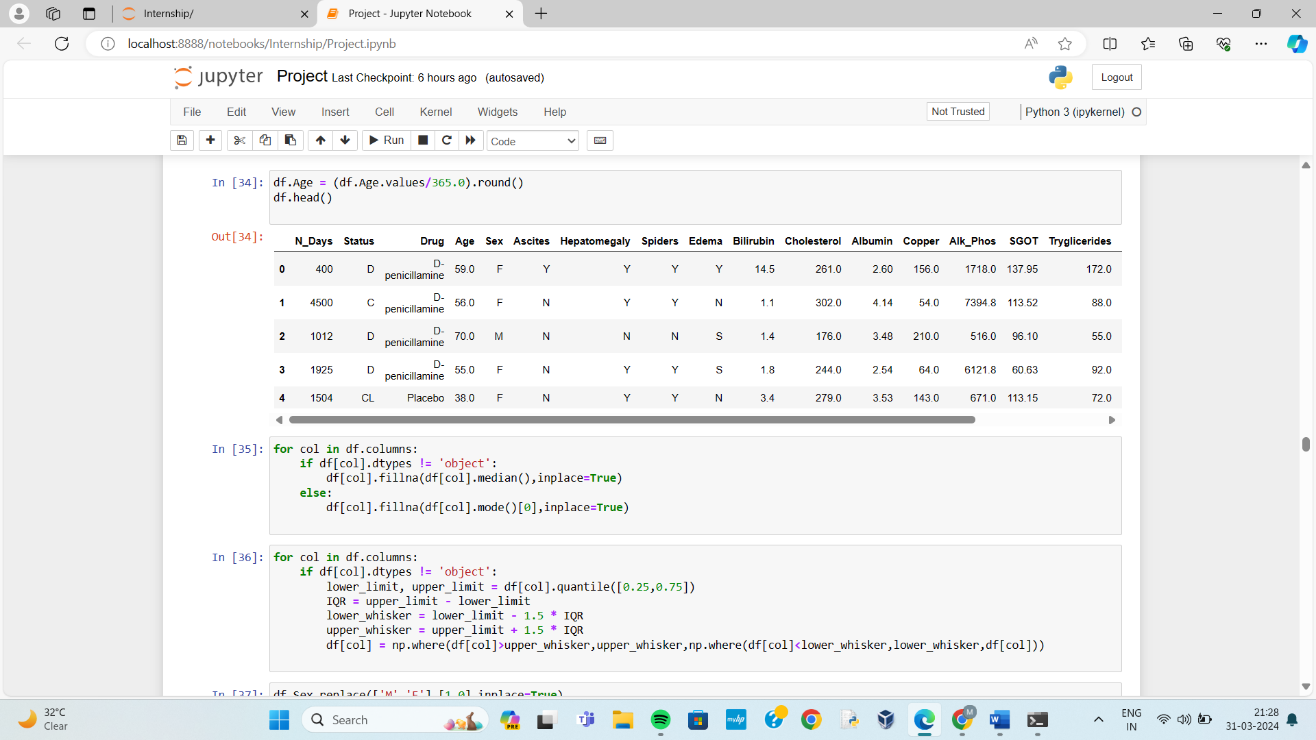
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Months/Activities** | **Dec** | **Jan** | **Feb** | **Mar** |
| **PHASE 1: USER EXPERIENCE STRATEGY**  **RESEARCH** |  |  |  |  |
| Kick-off meeting (half-day working session) |  |  |  |  |
| Audit of current architecture, metrics + research |  |  |  |  |
| **PHASE 2: INFORMATION ARCHITECTURE** |  |  |  |  |
| Features+ functionality set + strategy development |  |  |  |  |
| Data Understanding |  |  |  |  |
| Data preparation |  |  |  |  |
| Detailed wireframes |  |  |  |  |
| **PHASE 3:INTERFACE DESIGN PROTOTYPE** |  |  |  |  |
| Initial design development |  |  |  |  |
| Refined design |  |  |  |  |
| Data Modelling and Evaluation |  |  |  |  |
| **PHASE 4: DEVELOPMENT OF MODEL** |  |  |  |  |
| Coding |  |  |  |  |
| Frontend Development |  |  |  |  |
| Backend Development |  |  |  |  |
| **PHASE 5: USER VALIDATION TESTING+REFINEMENT** |  |  |  |  |
| Definition of recruiting criteria + screener |  |  |  |  |
| Recruiting occurs |  |  |  |  |
| Test planning Facilitation/research occurs |  |  |  |  |
| Analysis occurs |  |  |  |  |
| Presentation of recommendations |  |  |  |  |
| Refinement of deliverables |  |  |  |  |
| **PHASE 6: TEMPLATES, STANDARDS + DEVELOPMENT** |  |  |  |  |
| Templates + standards documentation |  |  |  |  |
| Technical  development |  |  |  |  |
| Testing+ launch |  |  |  |  |

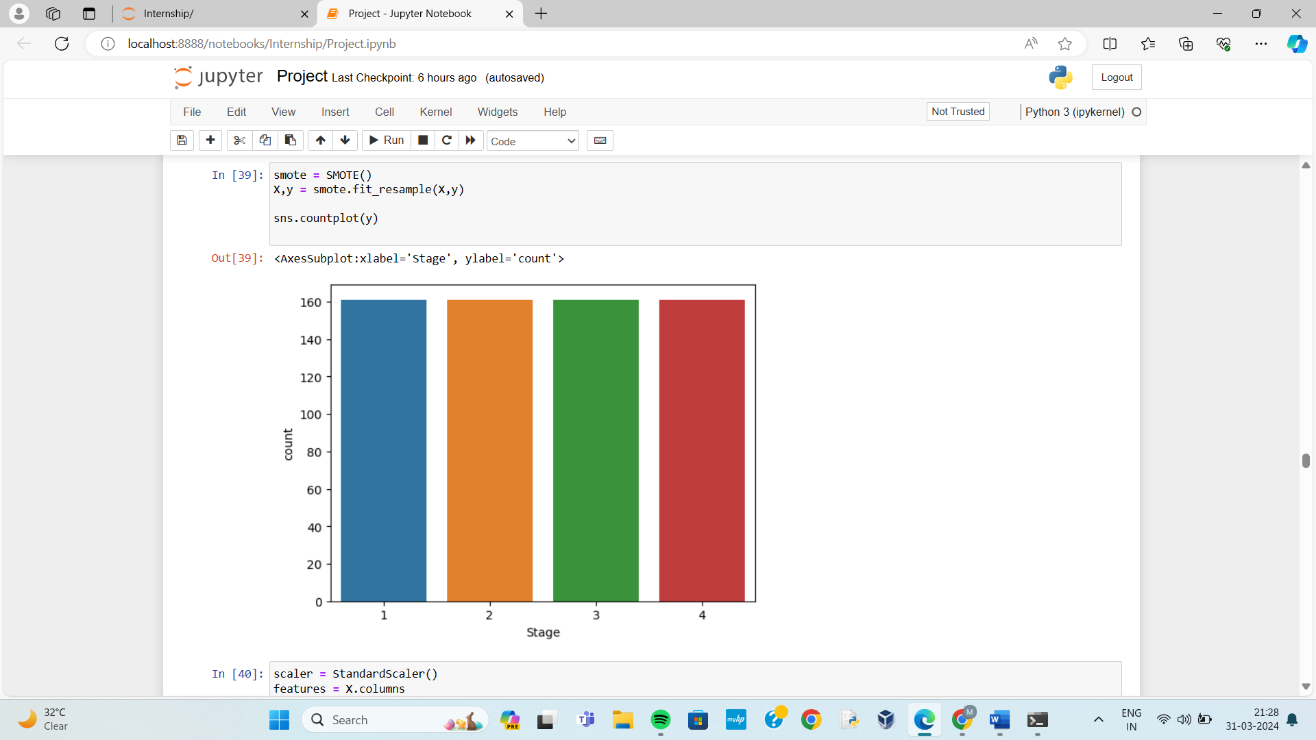
Figure 6: Gantt Chart

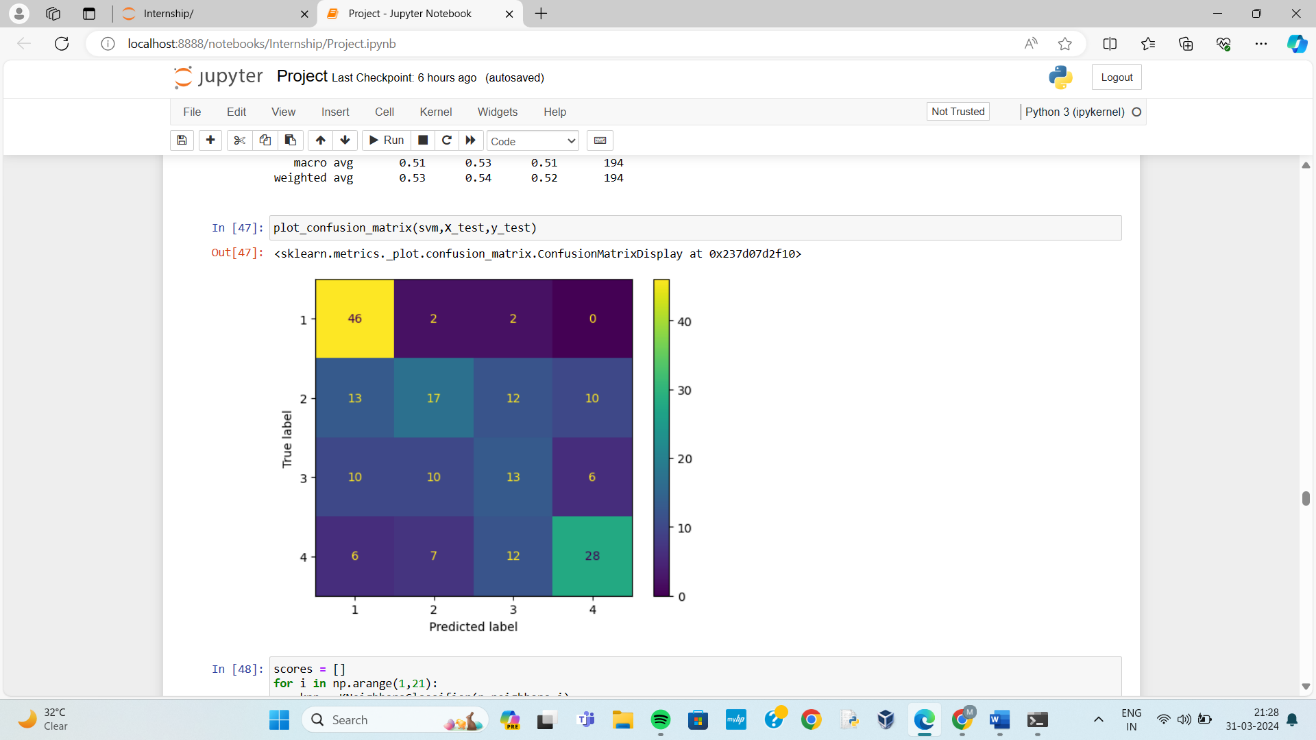
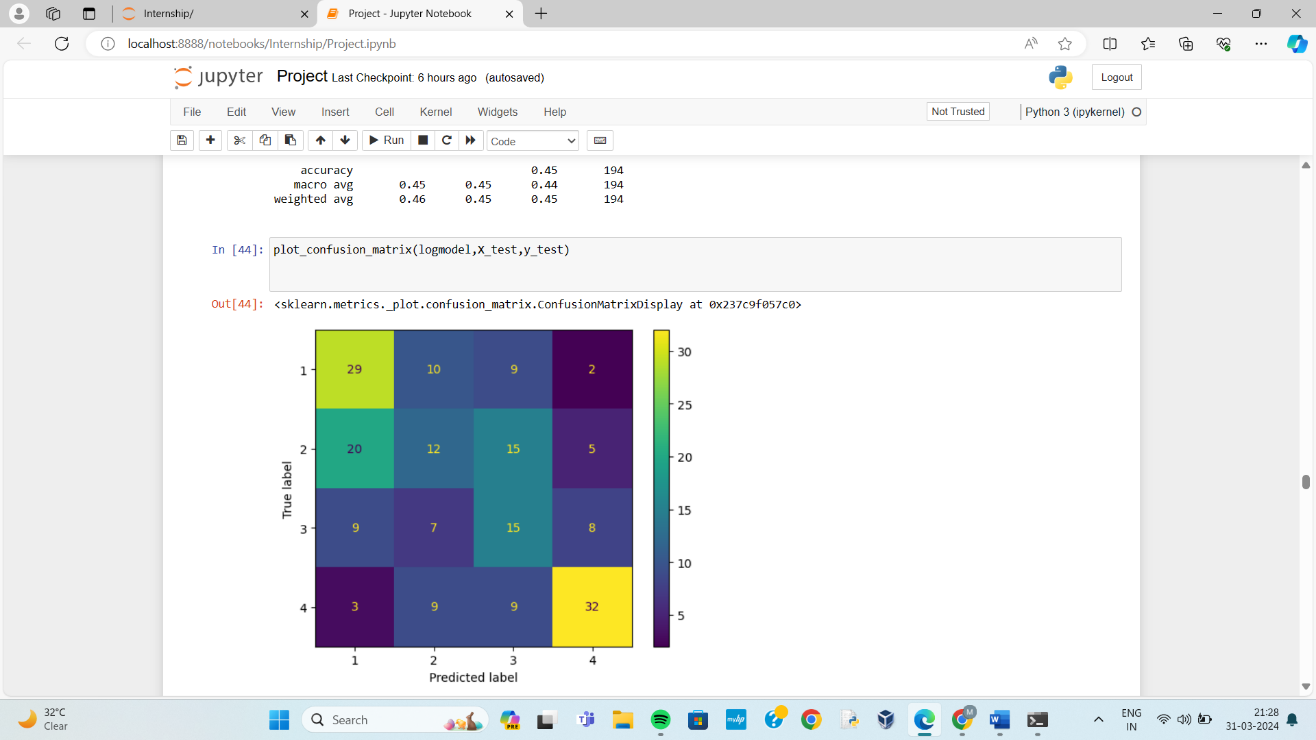
# Implementation and Output

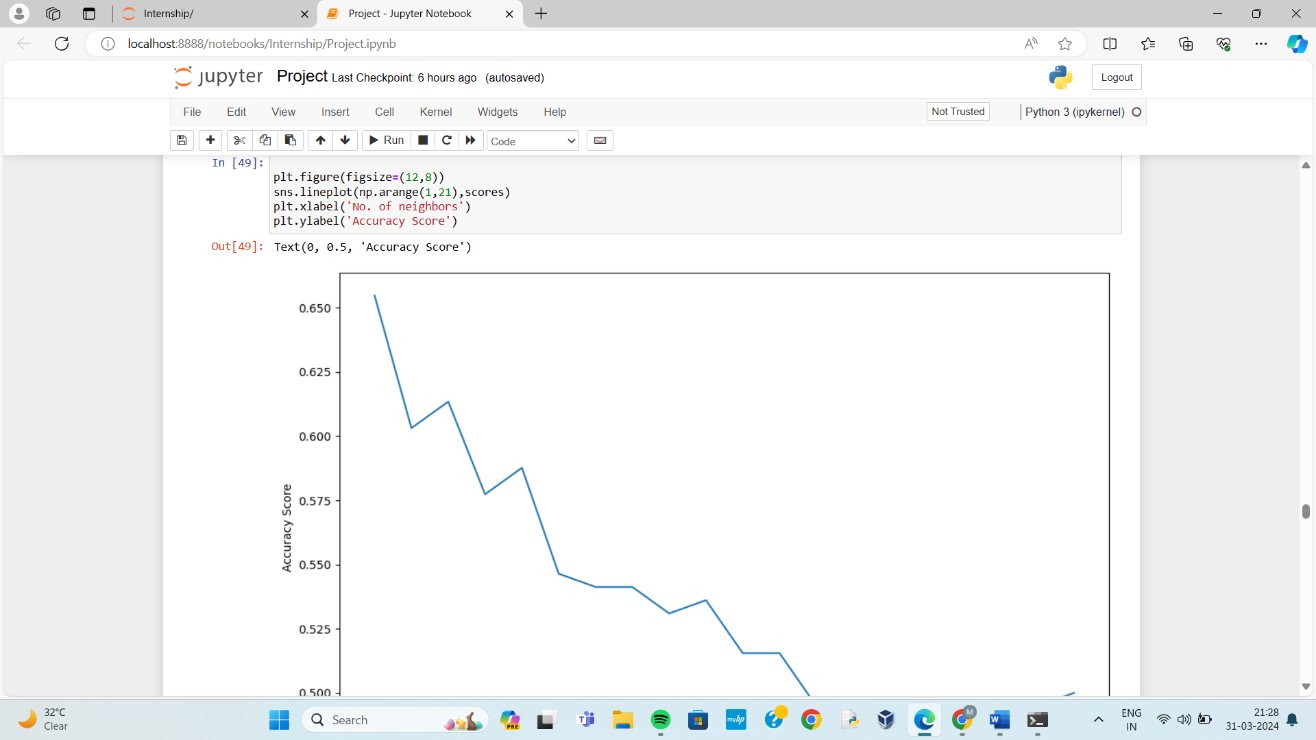


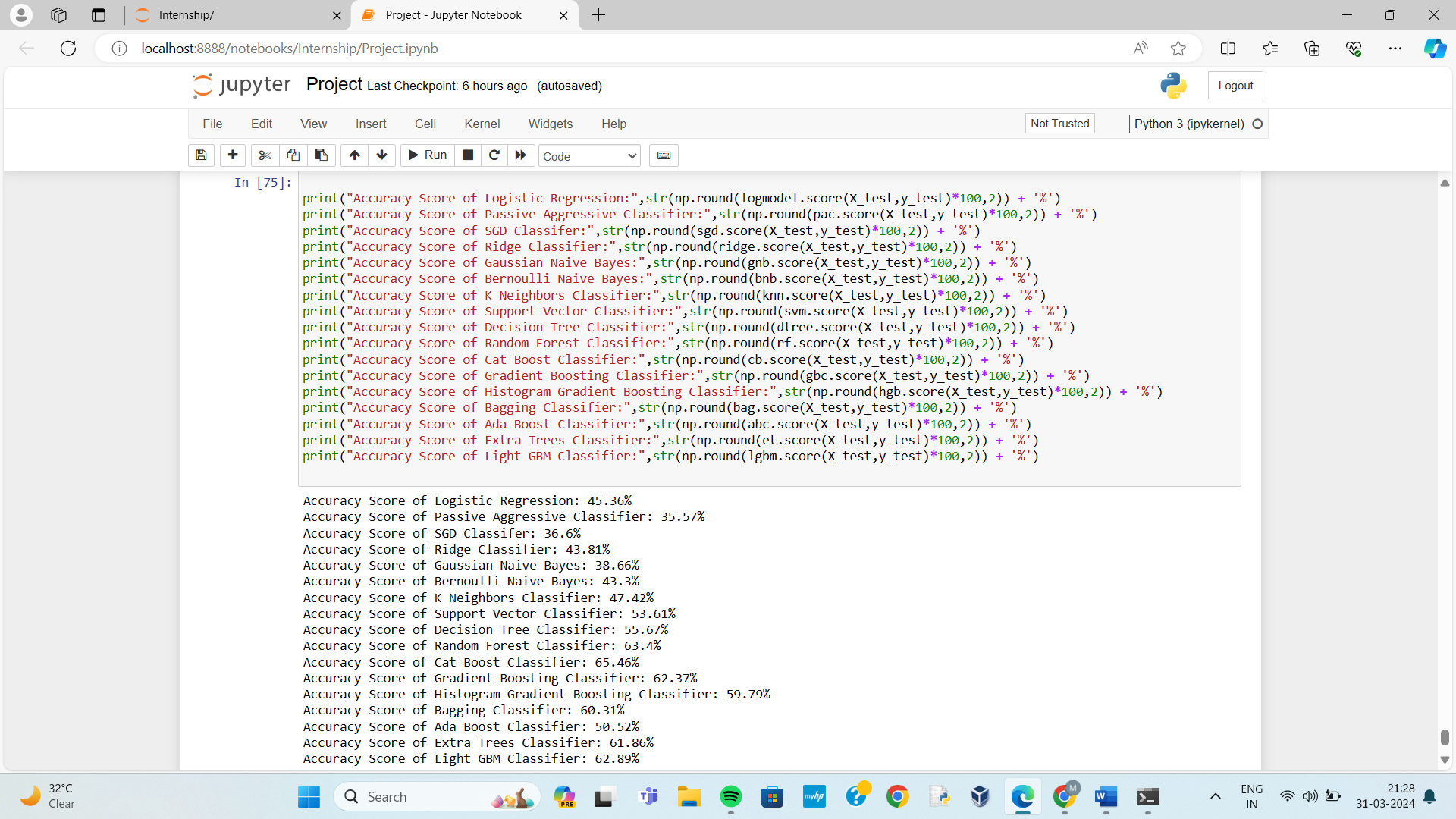
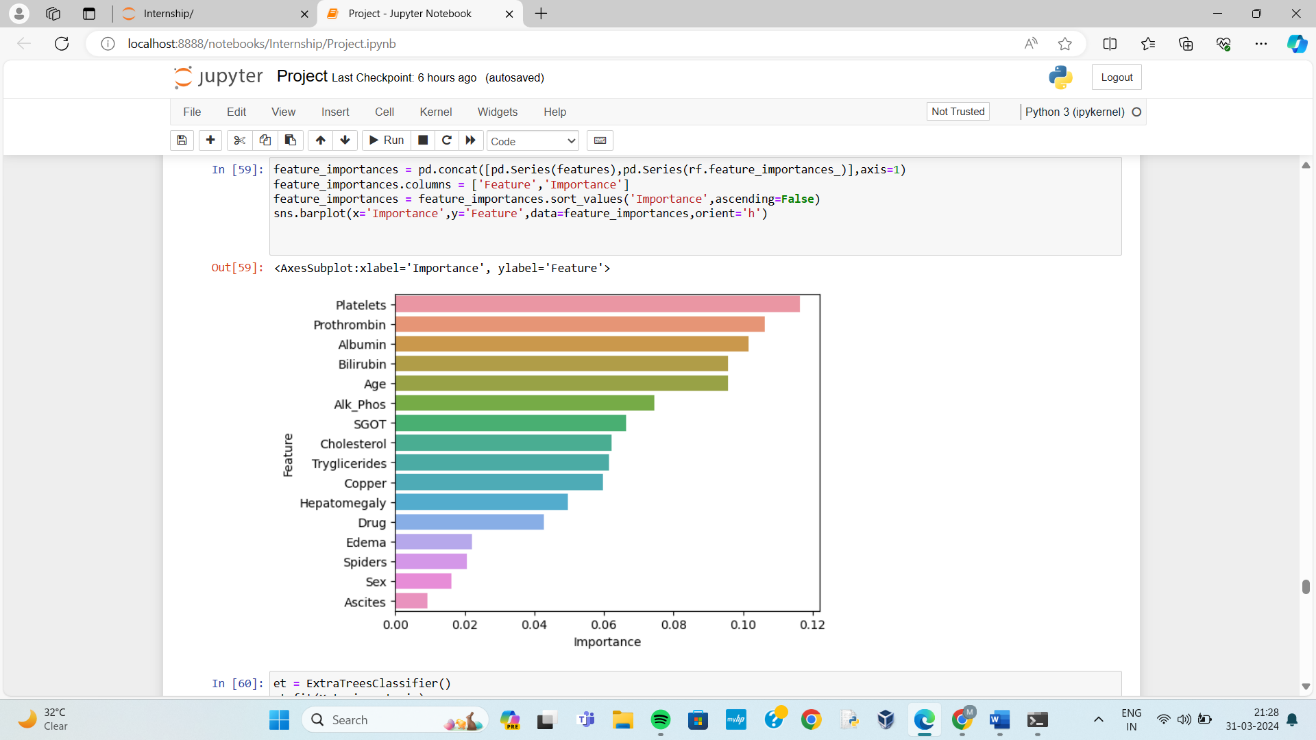


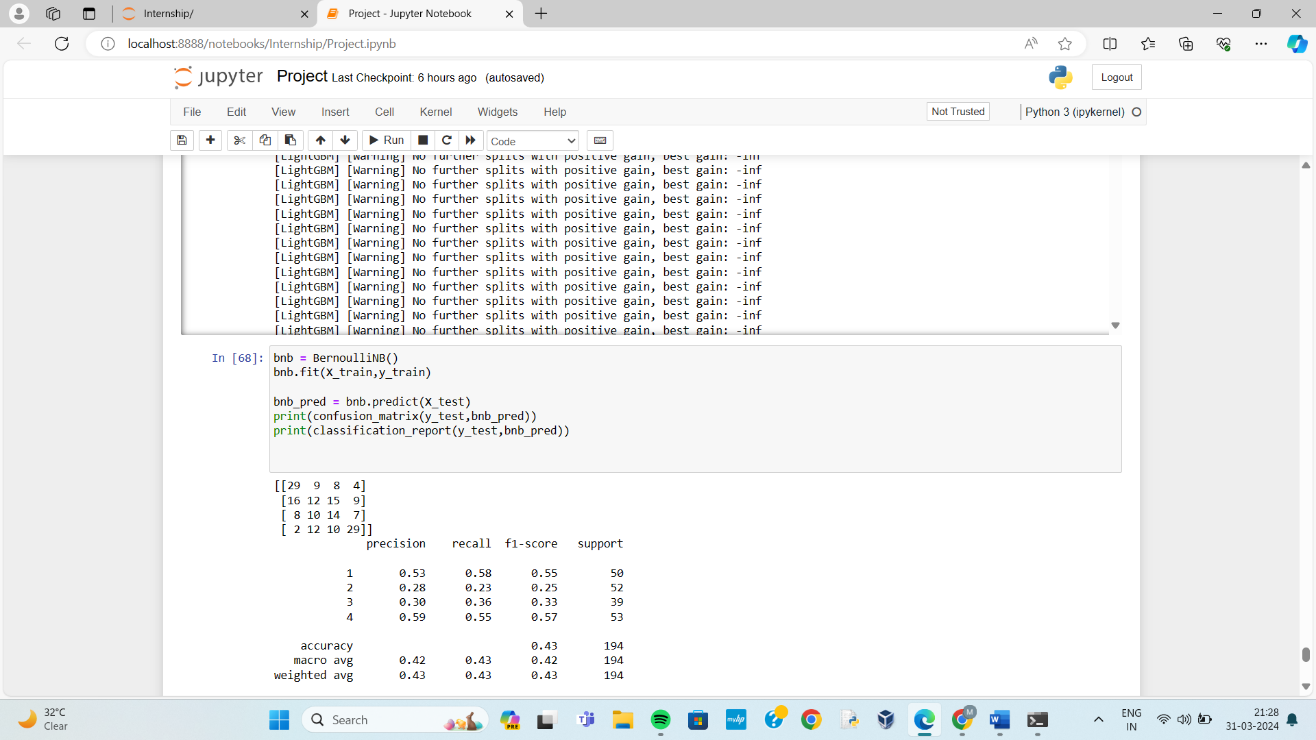












# References

# Big Data Analytics for Prediction Modelling in Healthcare Databases: - Ritu Chauhan and Eiad Yafi (2021)

# Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease: - P. Sujatha and k. Mahalaxmi (2020)

1. Predictive Analytics Model Based on Multiclass Classification for Asthma Severity by Using Random Forest Algorithm: - Wasif Akbar, Sehrish Saleem, Wei-Ping Wu, Arslan Javed, Muhammad Faheem and Muhammad Asim Saleem. (2020)
2. Big Data Analytics in Healthcare: - Guorong Chen and Mohaiminul Islam (2019)
3. Prediction of Diabetes Using Machine Learning Algorithms in Healthcare: - Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid and Munam Ali Shah (2018)

# 