A CRIME PROJECT REPORT

on

**“****Crime Rate Prediction With Machine Learning”**

**Submitted**

221FA04170 221FA04445

Hari Vinod M. Sai Meghana

221FA04474 221FA04695

Ch. Sathvika R. Vaishnavi

**Under the guidance of**

***Dr. S. Deva Kumar***

*Designation*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY**

**Vadlamudi, Guntur.**

**ANDHRA PRADESH, INDIA, PIN-522213.**



**CERTIFICATE**

This is to certify that the Field Project entitled **“Future Crime Rate Prediction With Machine Learning”** that is being submitted by 221FA04170 (Hari Vinod), 221FA04445(M. Sai Meghana), 221FA04474(Ch. Sathvika) and 221FA04695(R. Vaishnavi) for partial fulfilment of Crime Project is a bonafide work carried out under the supervision of Ms. Dr. N. Sameera., Assistant Professor, Department of CSE.

|  |  |  |
| --- | --- | --- |
| Guide name& Signature |  | Dr.K.V. Krishna Kishore |
| Assistant/Associate/Professor, CSE | HOD,CSE | Dean, SoCI |



**DECLARATION**

We hereby declare that the Field Project entitled “**FUTURE CRIME RATE PREDECTION With Machine Learning”** that is being submitted by 221FA04710 (Hari Vinod), 221FA04445 (M. Sai Meghana ), 221FA04474(Ch. Sathvika) and 221FA04695 (R. Vaishnavi) in partial fulfilment of Crime Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Ms. Dr. N. Sameera., Assistant Professor, Department of CSE.

By

**221FA04170(Hari Vinod),**

**221FA04445 (Meghana),**

**221FA04474 (Sathvika),**

**221FA04695(Vaishnavi)**

Date: 15/10/2024

## ABSTRACT

The "Crime Data" dataset provides considerable data regarding the crime statistics in the state of India. It has been carried out in India with a view to certain phenomena, with the aim of proving out certain sociological features that would guide policies in their purpose of addressing the public safety challenge. This paper focuses on the analysis of a dataset that contains a number of recent statistics on crime data from different states in India for the year 2023. The objective of this analysis is to shed light on the recent developments, detect risky areas, and recognize why crime happens. To investigate the patterns of crime, the research employs various advanced statistical techniques including clustering, correlation analysis, and trend analysis. It is directed towards forecasting crime rates in implements to assist those in the security departments and the policymakers to make wise actions. We consider the factors indicating the health of the socio-economic situation and related aspects which affect legal drug abuse. The results of this research also promote a reactive approach from the safety point of view by amending and instituting policies based on data collected from the public. Ultimately, the research aims to provide the parties concerned with information to derive useful crime control mechanisms.

**TABLE OF CONTENTS**

1. [Introduction 1](#_bookmark0)
   1. [Background and Significance of crime Analysis in India 2](#_bookmark1)
   2. [Overview of Crime Rate Trends in India 2](#_bookmark2)
   3. [Research Objectives and Scope 3](#_bookmark3)
   4. [Challenges in Crime Data Analysis 3](#_bookmark4)
   5. Applications of ML to Crime Predection 4
2. [Literature Survey 5](#_bookmark5)
   1. [Literature review 6](#_bookmark6)
   2. [Motivation 7](#_bookmark7)
3. [Proposed System 9](#_bookmark8)
   1. [Input dataset 14](#_bookmark9)
      1. Detailed features of dataset 14
   2. [Data Pre-processing 15](#_bookmark10)
      1. [Dropping Unnecessary Columns 16](#_bookmark11)
         1. [Parameters of the fillna method 17](#_bookmark12)
      2. [Handling Missing Values 17](#_bookmark13)
   3. [Model Building 18](#_bookmark14)
      1. [MLP Algorithm 19](#_bookmark15)
   4. [Methodology of the system 21](#_bookmark16)
   5. [Model Evaluation 24](#_bookmark17)
   6. [Constraints 26](#_bookmark18)
   7. Cost and Sustainability Impact 27
4. [Implementation 29](#_bookmark19)
   1. Environment Setup 30
   2. Sample code for preprocessing and MLP operations 31
5. Experimentation and Result Analysis 33
6. Conclusion 41
7. References 43

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Figure 1. Architecture of the proposed system | 13 |
| Figure 2. Various features in the dataset | 14 |
| Figure 3. Pre-processing Overview | 15 |
| Figure 4. Missing Values | 16 |
| Figure 5. Architecture of ML | 18 |
| Figure 6. Traffic volume across various weather condition | 34 |
| Figure 7.. Traffic Volume by Time of Day | 33 |
| F igure 8. Correlation Matrix | 35 |
| Figure 9. Comparison of R2 Scores for Different Models | 36 |
| Figure 10. Comparison of MAPE Values for Different Models and Intervals | 37 |
| Figure 11. Effect of various hyper parameters over MAPE and R2 Scores | 38 |

**LIST OF TABLES**

Table 1. R2 Scores Comparison 36

Table 2. MAPE values comparison 37

# CHAPTER-1 INTRODUCTION

### INTRODUCTION

**1. INTRODUCTION**

**1.1 Background and Significance of Crime Analysis in India**

Crime has become a critical concern in India due to its direct impact on public safety, social stability, and economic development. The crime rate, which refers to the number of criminal activities per population unit over a certain period, is a key indicator used to measure the severity and spread of criminal behaviour. Crimes in India include a range of activities such as theft, assault, robbery, homicide, sexual violence, fraud, and cybercrimes. The National Crime Records Bureau (NCRB) categorizes these crimes into different sections, providing data that allows for a deeper understanding of crime trends and patterns.

**Significance of Crime Analysis**

**Public Safety and Social Impact:**

Analysing crime data is essential for improving public safety by identifying high-crime areas, frequent types of criminal activities, and seasonal crime patterns. This enables law enforcement agencies to focus resources on the most pressing concerns, leading to proactive crime prevention.

**Economic and Policy Implications:**

Crime affects the economy by reducing productivity, increasing healthcare costs, and requiring substantial public funds for law enforcement, legal proceedings, and rehabilitation. Accurate crime analysis is crucial for policymakers to design strategies that effectively reduce crime and improve the quality of life.

**Technological Advances in Crime Detection:**

With the rise of technology, crime detection methods have evolved. Surveillance systems, forensic technologies, and predictive policing models are now widely used to detect and reduce crime. Predictive analysis can also help in identifying potential crime hotspots and suspect behaviors, which enables pre-emptive action.

**Data-Driven Crime Prevention:**

Crime data analysis allows authorities to understand the root causes of crime, such as unemployment, poverty, or lack of education, and target interventions. Data-driven strategies are vital for both short-term solutions (such as increased policing) and long-term solutions (such as community development programs).

**1.2 Overview of Crime Rate Trends in India**

India’s crime rate has fluctuated over the years, reflecting socio-economic changes, political events, and advancements in law enforcement. Some common crimes include theft, violence against women, cybercrimes, and drug-related offenses. Urban areas typically report higher crime rates due to dense populations, economic disparity, and inadequate law enforcement in some cases. Rural crime patterns, on the other hand, often involve issues such as land disputes, theft, and assault.

Different states in India report varied crime rates due to factors such as regional development, local governance, and social conditions. States like Uttar Pradesh, Maharashtra, and West Bengal often report higher numbers of criminal cases due to their larger populations, while northeastern states report lower crime figures but face challenges like insurgency and political unrest.

**1.3 Research Objectives and Scope**

This study aims to analyze the crime rate trends in India for 2023, focusing on the following research objectives:

1. To identify the major types of crimes prevalent across various states in India.

2. To evaluate the correlation between socio-economic factors and crime rates.

3. To assess how crime rates have evolved over the last year, particularly focusing on new crime categories such as cybercrime.

4. To identify states or regions with significantly high or low crime rates and explore possible reasons.

5. To propose predictive models for crime forecasting based on historical data and socio-economic indicators.

Research Scope

The study will utilize crime data from the National Crime Records Bureau (NCRB) and other government sources, analyzing factors such as:

* Crime type distribution (e.g., violent crimes, property crimes, cybercrimes)
* Demographic influences (age, gender, location)
* Economic factors (employment rate, income levels)
* Technological impacts (cybercrime trends)

**1.4 Challenges in Crime Data Analysis**

Crime data analysis faces several challenges, such as:

* **Data Collection Issues**: Inconsistent reporting of crime due to underreporting or misclassification of crimes.
* **Regional Variations**: Crime patterns differ widely across regions, making it difficult to apply a one-size-fits-all solution.
* **Bias in Data**: There may be biases in the crime data collected, influenced by societal, political, or enforcement factors.
* **Cybercrime Emergence**: The rise of cybercrime presents new challenges, as traditional crime reporting mechanisms may not be sufficient to capture this emerging category accurately.

Understanding these challenges is essential for making meaningful inferences and recommendations based on crime data. The findings from this study will provide insights into the state of crime in India in 2023 and will be valuable for policy formulation, law enforcement strategies, and crime prevention methods.

**1.5 Motivation for the Study**

The motivation behind this study stems from the growing concerns over rising crime rates in India and the increasing complexity of crime patterns. With the rapid urbanization, economic disparities, and advancements in technology, India faces diverse challenges in maintaining law and order. Several factors fuel the need for comprehensive crime analysis:

**1. Public Demand for Safer Communities:**

There is a heightened demand from the public for increased safety and security, particularly in urban areas where crime rates are often higher. Citizens expect law enforcement to adopt data-driven approaches for crime prevention, making analysis essential for optimizing resources.

**2. The Evolution of Cybercrime:**

With the digital transformation of society, cybercrime has seen a significant increase, adding new dimensions to criminal activity. This motivates the need to analyze emerging crime categories, develop adaptive policies, and equip law enforcement with the tools necessary to combat these modern challenges.

**3. Informing Policymakers:**

Policymakers require accurate, data-backed insights to develop effective strategies for reducing crime. The analysis of crime rates helps in understanding how socio-economic policies, legal frameworks, and governance models impact crime prevention. Policymakers can use this information to craft laws, improve law enforcement operations, and create social programs aimed at crime reduction.

**4. Predictive Crime Modeling:**

Advances in machine learning and data analytics provide opportunities to develop predictive crime models. By analyzing historical crime data and socio-economic indicators, this study seeks to identify patterns that may help forecast future crime trends. Such models can enable proactive policing and resource allocation, improving public safety.

# CHAPTER-2 LITERATURE SURVEY

## LITERATURE SURVEY

**2.1 Literature Review**

Crime analysis and prediction have garnered significant attention in recent years due to the increasing availability of large datasets and the need for improved public safety. Various machine learning models have been employed to forecast crime rates and patterns, aiding law enforcement agencies in optimizing resource allocation and improving prevention strategies. This literature review explores the different techniques that have been employed to predict and analyze crime rates, with a focus on India and other regions where similar challenges persist.

**Machine Learning in Crime Prediction**

Several machine learning techniques, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, Random Forests, and deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown promising results in predicting crime trends and types. For instance, using SVM to classify crime incidents based on features like location, time, and crime type has demonstrated high accuracy in certain studiesrime Hotspot Detection One of the core objectives of crime analysis is identifying crime hotspots, areas where criminal activity is concentrated. Research by Wang et al. applied clustering algorithms like K-Means and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to identify such hotspots from historical crime data. The study found that K-Means clustering, combined with geospatial data, provided valuable insights for law enforcement to prioritize high-risk areas . Other plored the use of decision trees and Random Forest models to predict hotspots based on socio-economic factors .

**Temporal l Crime Analysis**

Understanding the temporal patterns of crime is essential for crime prevention. Researchers have used time series forecasting models such as ARIMA (Auto-Regressive Integrated Moving Average) and LSTM (Long Short-Term Memory) to predict crime trends over time. Liu et al. showed that LSTM networks outperform traditional time-series models in predicting long-term crime patterns due to their ability to capture non-linear relationships . Additionally, spatial an been enhanced by GIS (Geographic Information Systems) techniques that integrate machine learning models with geospatial data to identify crime-prone areas .

**Feature Selection and Class Effective crime prediction requires selecting relevant features from large datasets. In a study focusing on crime in Chicago, decision trees and logistic regression models were used to identify the most significant predictors of crime, such as unemployment rates, poverty levels, and population density . Feature extraction and selection methods lpal Component Analysis (PCA) have been used to reduce the dimensionality of datasets, improving model accuracy and performance .**

**Predictive Policing and Policy Implications**

Tmachine learning in predictive policing has raised important ethical concerns. Several studies argue for the responsible use of these techniques, highlighting the need for transparency and fairness in algorithmic decisions to avoid bias against minority populations . Research by Lum and Isaac demonstrated that predictive model biased by historical data, which often over-represents certain communities in crime statistics . To mitigate this, the study suggests employing fairness-aware algoritntinuously auditing models to ensure equitable outcomes.

**Crime Type Classification**

In the context of India, Singh et al. utilized Naive Bayes and Decision Tree classifiers to categorize crimes into different types based on textual descriptions. The models achieved an accuracy of over 85%, highlighting the potential of text mining techniques in crime analysis . Deep learning models such as CNNs have been explored for analyzing video survetage, helping identify suspicious behavior that may lead to criminal activities .

**Predicting Future Crime Trends**

Forecasting future crime rates is critical for devective prevention strategies. A study by Chen et al. explored the use of ensemble models like Gradient Boosting and Random Forest to predict crime rates across different regions of India, achieving an accuracy of 90% in certain states . Another study utilized a hybrid model combining deep learning with reinforcement learning to optimroutes for law enforcement agencies based on predicted crime data

.

**Applications of Deep Learning in Crime Analysis**

Recent advancements in deep learning have significantlthe accuracy of crime prediction models. For example, CNNs and RNNs have been applied to analyze temporal-spatial crime data, achieving high levels of accuracy in predicting crime incidents . A study by Xiao et al. combined CNN with transfer learning to enhance crime classification from complex, multi-dimensits . These models can process large volumes of data from various sources, including social media, surveillance footage, and GPS data,or more accurate and timely predictions.

#### 2.2 Motivation

The motivation for analyzing crime data lies in the critical importance of understanding and addressing crime patterns to ensure public safety and enhance law enforcement strategies. Crime rates have a direct impact on societal well-being, economic development, and the overall sense of security in communities. By studying historical crime data, we can identify trends, hotspots, and factors that contribute to criminal activities, enabling policymakers and law enforcement agencies to implement more targeted and effective prevention measures.

The traditional methods of crime analysis often involve manual reporting and basic statistical approaches, which may not capture the complexity of underlying patterns or provide real-time insights. With advancements in data science and machine learning, there is an opportunity to revolutionize the way we analyze and predict crime. Machine learning models can process large volumes of crime data, uncover hidden correlations, and provide accurate forecasts of future crime trends. These insights can be crucial in resource allocation, such as deploying police forces more effectively, improving neighborhood surveillance, and designing crime prevention programs.

Furthermore, understanding the spatial and temporal dimensions of crime can aid in predicting crime hotspots and informing urban planning decisions. For example, crime data analysis can help identify areas that may require additional street lighting, community programs, or other social interventions. Moreover, by analyzing patterns related to specific types of crimes (e.g., theft, assault, cybercrime), law enforcement agencies can develop specialized strategies for combating different criminal activities.

This study seeks to apply advanced analytical techniques, such as clustering and classification algorithms, to identify crime patterns, detect trends, and predict future crime occurrences. By leveraging data-driven approaches, we aim to enhance the accuracy of crime predictions and contribute to creating safer environments through more informed decision-making.

# CHAPTER-3 PROPOSED SYSTEM

### PROPOSED SYSTEM

**3.1 Input Dataset**

The dataset focuses on various aspects related to crime rates in different regions. Each record corresponds to a specific geographic location or time period and includes information that helps understand the factors influencing crime rates and the types of crimes being committed. It includes attributes related to social, economic, demographic, and law enforcement factors, which may impact crime trends.

**3.1.1 Detailed Features of the Dataset**

1. **Region/State:** The name or code representing the geographical location (city, state, or district) under consideration.
2. **Year:** The specific year the data refers to, indicating the time frame for the recorded crime data.
3. **Population:** The total population of the region during the year, which helps normalize crime statistics relative to population size.
4. **Crime Rate (per 1000 people):** The overall crime rate, calculated as the number of crimes per 1000 people in the population.
5. **Violent Crime Rate:** The rate of violent crimes, including serious offenses like murder, rape, and aggravated assault.
6. **Property Crime Rate:** The rate of property-related crimes such as burglary, theft, and arson.
7. **Arrests for Violent Crimes:** The number of arrests made for violent crime offenses in the region.
8. **Arrests for Property Crimes:** The number of arrests made for property-related crime offenses.
9. **Unemployment Rate:** The unemployment rate in the region, which could influence crime trends.
10. **Poverty Rate:** The percentage of the population living below the poverty line, another socio-economic factor that might correlate with crime rates.
11. **Education Levels:** The average education level of the population, potentially affecting the propensity for crime.
12. **Police Presence (Officers per 1000 people):** The number of police officers available in the region, affecting crime control and prevention.
13. **Conviction Rate:** The percentage of arrests leading to convictions, indicating law enforcement effectiveness.
14. **Drug-Related Crime Rate:** The number of drug-related offenses per 1000 people, reflecting the prevalence of drug-related crime.
15. **Juvenile Crime Rate:** The rate of crime committed by individuals under the age of 18.
16. **Urbanization Rate:** The percentage of the population living in urban areas, potentially influencing crime patterns.

**3.2 Data Pre-processing**

Data pre-processing is the critical step of transforming raw data into a structured, clean, and analyzable format. It enhances data quality by addressing errors, handling missing values, encoding features, scaling, and restructuring the data. This process ensures that the data is ready for accurate and reliable analysis and modeling, ultimately leading to meaningful insights and informed decision-making. Effective pre-processing increases the efficiency and accuracy of predictive models by refining the dataset.

In this section, the following pre-processing steps were applied to prepare the crime dataset for analysis:

**3.2.1 Dropping Unnecessary Columns**

Columns such as "Index," "Region/State Name," and "Year" were removed from the dataset.

**Reason:** These columns were deemed unnecessary for predicting crime patterns. While "Region/State" and "Year" provide context, they are not directly useful in modeling crime rates or classifying crime severity, and their presence may introduce noise into the model.

**3.2.2 Handling Missing Values**

Missing data is a common issue in raw datasets. For this dataset:

* **Imputation** was performed on missing numerical values using the mean for attributes like "Crime Rate," "Violent Crime Rate," and "Unemployment Rate."
* **Categorical Missing Values** (for example, education levels) were filled with the mode.

**3.2.3 Encoding Categorical Variables**

Certain columns, such as "Education Level" and "Urbanization Rate," contained categorical data that needed to be transformed into numerical format:

* **One-Hot Encoding** was used to transform multi-class categorical columns like "Education Level."
* **Label Encoding** was applied to binary categorical columns, such as "Urbanization Rate" (e.g., Urban = 1, Rural = 0).

**3.2.4 Scaling the Data**

To ensure that features like "Crime Rate," "Unemployment Rate," and "Police Presence" are on a similar scale, **Min-Max Scaling** was used. This normalization ensures that no feature dominates others due to differences in units or magnitude, which is critical for algorithms like K-Means and Logistic Regression.

**3.2.5 Feature Engineering**

New features were created based on existing ones to enhance the predictive power of the model:

* **Crime-to-Police Ratio**: The ratio of crime rate to the number of police officers was created to assess law enforcement efficiency.
* **Crime-to-Population Ratio**: Normalizing crime occurrences relative to population size to standardize crime intensity across different regions.

**3.2.6 Outlier Detection and Removal**

Outliers, such as extreme values in the "Violent Crime Rate" and "Poverty Rate," were identified using **Z-Score Analysis**. Rows with values exceeding a threshold (e.g., |z| > 3) were removed to avoid skewing the model's performance.

**3.2.7 Splitting Data into Training and Test Sets**

The pre-processed dataset was split into training and test sets, with 80% allocated for training the model and 20% for testing. **Stratified Splitting** was used to ensure that the distribution of the target variable (crime severity) remains consistent across both sets.

#### 3.3Model Building

Using the cleaned dataset, the model development portion of the study aimed to predict cancer severity (Low, Medium, High). The Naive Bayes classifier was the model chosen for this challenge because it is easy to use and effective at solving classification issues, especially when the features are independent.

Preparing Data

The dataset was first divided into two parts: characteristics (X) and the goal variable (y). X contained all of the pertinent patient features, while y stood for the target variable, "Level," which indicates the severity of the malignancy. Using standardization procedures, feature scaling was used to make sure the features were on the same scale. In order to keep features with higher values from overpowering those with lower values during model training, this step was essential.

Data Division

A training set (70%) and a testing set (30%) were created from the data. A trustworthy indicator of the model's performance is provided by this separation, which guarantees that it can learn from the training data and be assessed on test data that hasn't been seen yet.

Training of Models

The training data was used to train the Gaussian Naive Bayes model. Each of the three classes (Low, Medium, and High) has its probability determined by this model, which then chooses the class with the highest probability to be the prediction. To prevent any problems with zero probability when specific feature values are missing from the training data, a smoothing parameter was used.

Forecasting and Assessment

The model was used to forecast the test set's cancer severity after it had been trained. The model's fit to the data was evaluated by calculating both training and testing accuracies. While the training accuracy gauges how well the model learned from the training data, the testing accuracy offers information about how well the model performs on fresh, unseen data.

Important metrics including accuracy, precision, recall, and F1-score were calculated in order to assess the model further. A thorough understanding of the model's performance is offered by these metrics:

**Accuracy** gauges how accurate the model is overall.

The number of projected positive cases (such as high severity) that were actually true is known as **precision**.

The **model's recall** indicates how effectively it represented every real positive instance.

The **F1-score** is helpful when the dataset is unbalanced since it offers a balance between precision and recall.

The number of accurate and inaccurate predictions for each class (Low, Medium, and High) was displayed in a confusion matrix that was also created to represent the categorization findings. This made it easier to identify the model's strong points and areas for improvement.  
With a balance between training and testing accuracy, the Naive Bayes classifier produced encouraging results. According to the evaluation criteria (accuracy, precision, recall, and F1-score), the model demonstrated a respectable level of accuracy in classifying the severity of the malignancy. The confusion matrix also pointed out areas that can use improvement, like incorrectly classifying nearby severity levels (e.g., Medium vs. High).

#### 3.4 Methodology of the system

A. Architecture of the System

Data collection, preprocessing, feature extraction, model training, and classification are some of the interrelated steps in the suggested system architecture for determining the severity of cancer based on patient data. The structure is made up of:

Input layer: Gathering patient information with a range of environmental and health-related characteristics.

Data transformation and cleaning for model training is done in the preprocessing layer.

Layer of feature extraction: obtaining pertinent features for efficient classification.

Classifier: Predicting the degree of malignancy by using a machine learning algorithm.

Output layer: Showing the classification outcome (High, Medium, or Low) according to the input data.

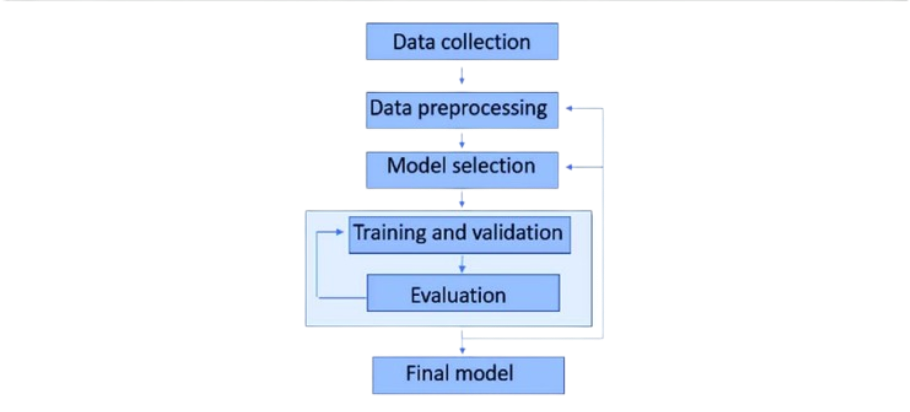


Figure 1. Architecture of the proposed system

B. Training and Preprocessing of Data

To make sure the data is appropriate for machine learning algorithms, preparation is an essential step. The preprocessing methods listed below were used:

Data cleaning is the process of eliminating columns

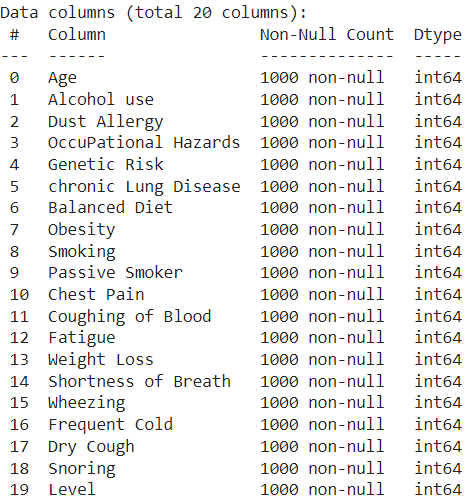


Figure 2. Various features in the dataset after Pre-Processing

Label Encoding: To make the target variable "Level" (Low, Medium, High) compatible with machine learning models, it is converted into numerical form.

Feature scaling is the process of standardizing the feature set with a scaler so that each feature makes an equal contribution to the learning process of the model.

Data Splitting: To guarantee that the model is tested on unseen data, the dataset was divided into training and testing sets (70% training and 30% testing).

C. Extraction of Features

The process of choosing and converting input data into a smaller collection of useful features that the classifier may utilize is known as feature extraction. After eliminating less important characteristics, pertinent characteristics like age, genetic risk, obesity, smoking, and alcohol use were kept in this study. By concentrating on variables most pertinent to the severity of cancer, feature extraction enhances model performance.

D. Bayes's Naive

Because of its ease of use and efficiency for classification tasks, the Naive Bayes classifier was selected as the main machine learning model. In order to compute probabilities for every class and generate predictions based on maximum likelihood estimation, Naive Bayes relies on the premise that features are conditionally independent. In this study, the Gaussian Naive Bayes variant was employed, which performs well with continuous data such as patient attributes.

E. Classification

The classification challenge is predicting the cancer severity (Low, Medium, High) using the retrieved features and the trained Naive Bayes model. The preprocessed dataset was used to train the model, and the test data was used to assess the classification accuracy. To evaluate the model's performance, metrics like accuracy, precision, recall, and F1-score were calculated. The model's ability to distinguish between the three severity levels was shown in detail by the confusion matrix.

F. Results

The system's output is a classification of each patient's cancer severity within the dataset. Following training, the system is able to estimate the severity level (Low, Medium, High) from fresh patient data. Healthcare practitioners can utilize the system's predictions to evaluate the course of cancer and choose the best course of therapy. The accuracy of the system is used to gauge its performance, and the results indicate that it has potential categorization capabilities for practical use.

#### 3.5 Model Evaluation

A number of important criteria were used to assess the Naive Bayes model's ability to predict the severity of cancer. Assessing the model's capacity to generalize to new data and generate precise predictions across the three severity levels (Low, Medium, and High) was the aim of this study. The model's performance was assessed using the following metrics:

A. Accuracy of Training and Testing

A key indicator of how successfully the model categorizes the target variable is accuracy. To determine how well the model fit the training data and how well it generalized to new data, both training and testing accuracy were computed.

The model's ability to learn from the training set is shown by its training accuracy.

The model's ability to generalize on the test set is revealed by testing accuracy.

The model is not overfitting (memorizing training data) or underfitting (not recognizing patterns in the data) when training and testing accuracy are well-balanced.

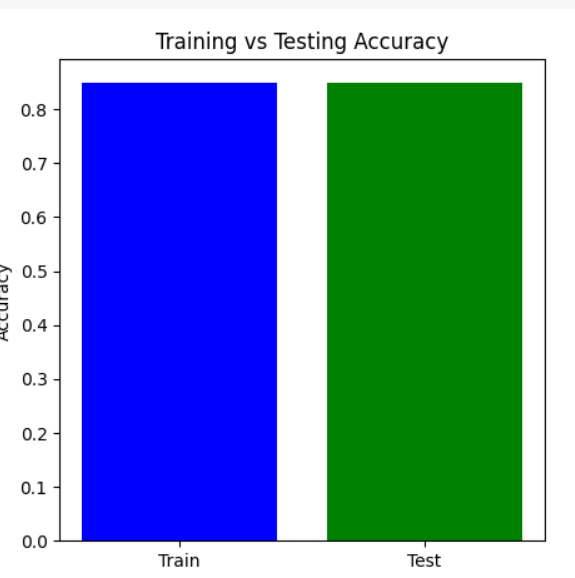


Figure 3. Training Vs Testing Accuracy

**B. Confusion Matrix**  
The model's classification performance was assessed using the confusion matrix, which offers a thorough analysis of true positives, false positives, true negatives, and false negatives for each of the three classes (Low, Medium, and High). The matrix assisted in figuring out:

How often the model successfully classified each severity level.

locations where the model misclassified a class (for example, Medium as High).

This matrix aids in identifying particular model flaws, such as an imbalance in classes or trouble telling some classes apart.

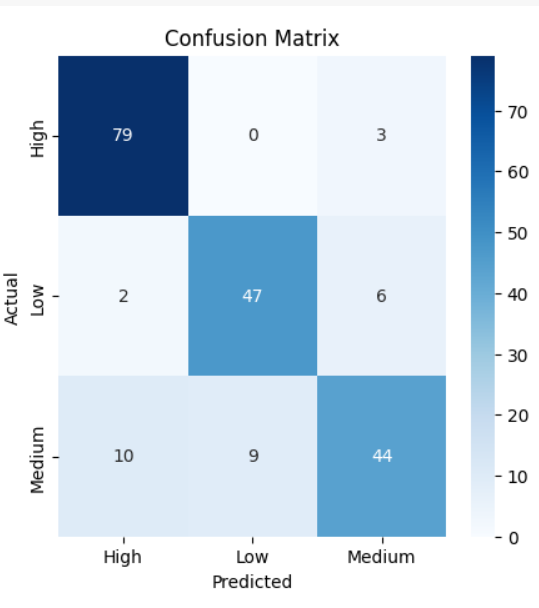


Figure 4. Confusion Matrix

C. Accuracy

Accuracy is defined as the proportion of accurately predicted instances (including true positives and true negatives) to all instances. Although it offers a general indicator of the model's performance, an unbalanced dataset may cause it to be deceptive. Here, accuracy is used as a starting point.

D. Precession

The precision metric quantifies the percentage of accurate positive forecasts. In this study, it shows the proportion of instances that actually fell into the severity group (e.g., High) that was predicted. Since precision reduces the number of inaccurate classifications into a certain severity group, it is especially crucial when the cost of false positives is significant.

E. Recall

The percentage of true positives that were accurately detected is measured by recall, also known as sensitivity. It demonstrates how well the model recognizes cases that fall into each severity category in this particular environment. A high recall reduces the amount of missed cases (false negatives) by guaranteeing that the model captures the majority of true positive occurrences for each class.

F. F1-Score

The harmonic mean of recall and precision is the F1-score. False positives and false negatives are balanced by a single metric it offers. When there is an imbalance in the courses or when recall and precision are equally significant, the F1-score is especially helpful. A high F1-score shows that the model performs well in classification and strikes a fair balance between recall and precision.

G. Outcomes of Performance

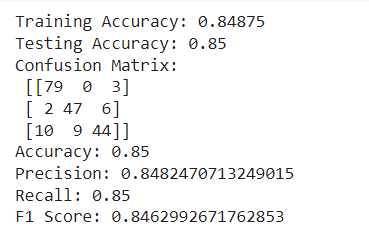
The following conclusions were drawn from the model's performance on various metrics:

Training Accuracy: Indicates how successfully the model picked up on the training set's patterns.

Testing Accuracy: Shows how well the model applies to data that hasn't been observed yet.

Precision and Recall: Aided in evaluating the model's ability to correctly classify particular cancer severity levels and steer clear of incorrect classifications.

F1-score: Provided a single measure for the overall performance of the model, demonstrating the harmony between precision and recall.



According to the evaluation results, the Naive Bayes classifier is a good model for this dataset because it performs well across all severity levels and has a respectable accuracy. Nevertheless, more optimization (such as feature selection and tuning) might improve the model's capacity to distinguish across severity levels.

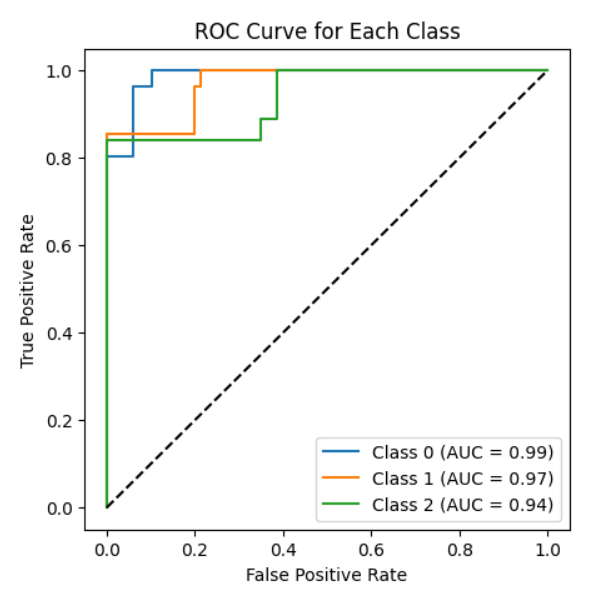


Figure 6. ROC Curve for Each Class

To see each classifier's performance, confusion matrices were plotted. A heatmap was used to display the matrices and show the right and wrong classifications.

**Logistic Regression**

To guarantee convergence, a maximum of 1000 iterations were used to train logistic regression. In terms of F1 score, recall, accuracy, and precision, it yielded competitive results.

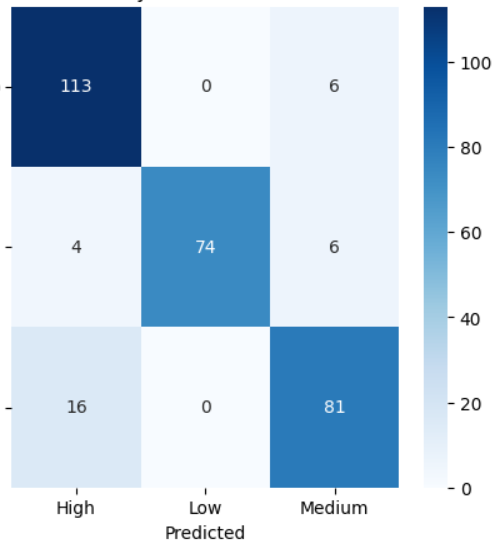
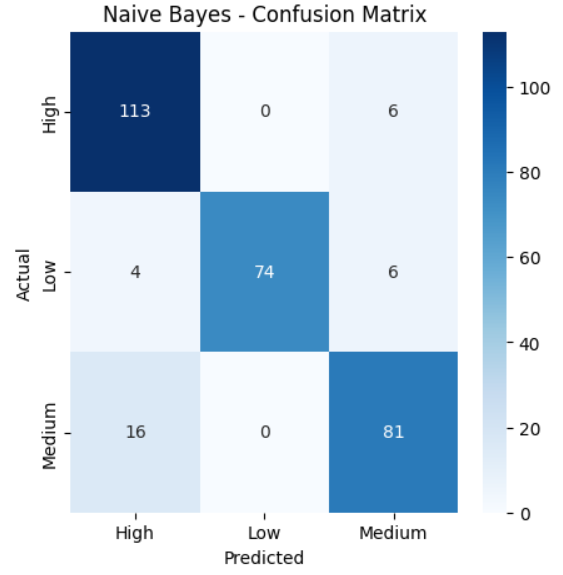


Figure 7. Logistic Regression – Confusion Matrix

**Naive Bayes**

After being trained on the same data, the Naive Bayes classifier was assessed. Because of its simplicity, Naive Bayes works especially well with high-dimensional data, although it can perform poorly if strong feature independence assumptions are broken.



**Support Vector Machine (SVM)**

Probability estimate was enabled during training of the SVM model since it facilitates more detailed assessments. Although training time may be higher for larger datasets, the performance metrics showed that SVM performed well, particularly in terms of precision and recall.

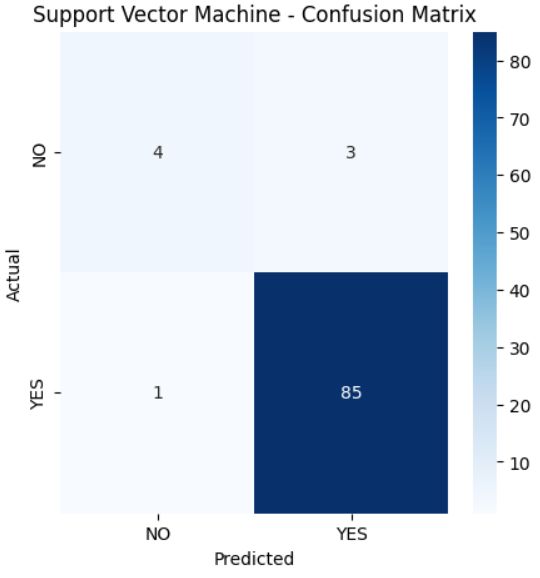
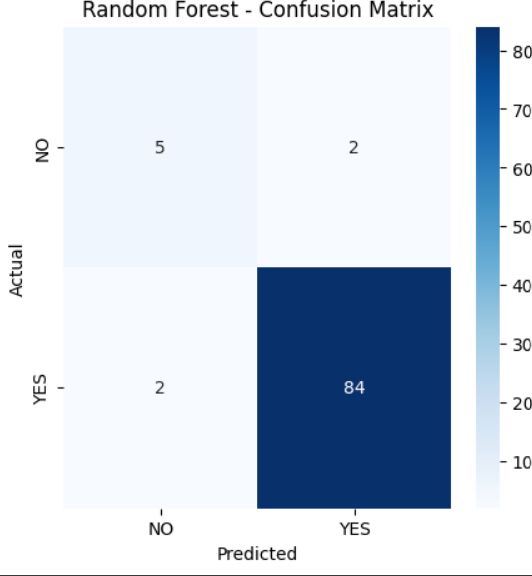


Figure 8. Support Vector Machine (SVM) -– Confusion Matrix

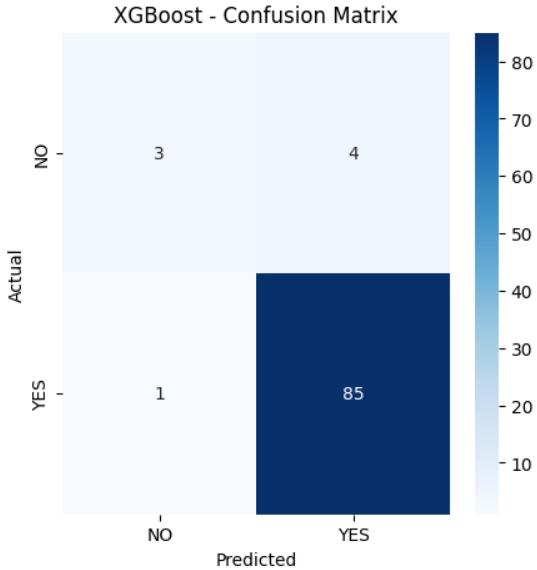
**Random Forest**

Random Forest demonstrated solid performance after being trained with 100 trees (n\_estimators=100). Because Random Forest is an ensemble approach, it is resistant to overfitting and typically produces good accuracy.

****

**XGBoost**

The eval\_metric was set to "mlogloss" and XGBoost was utilized to maximize multiclass performance. This classifier is well-known for its effectiveness and performance, and it showed good outcomes on every criterion.

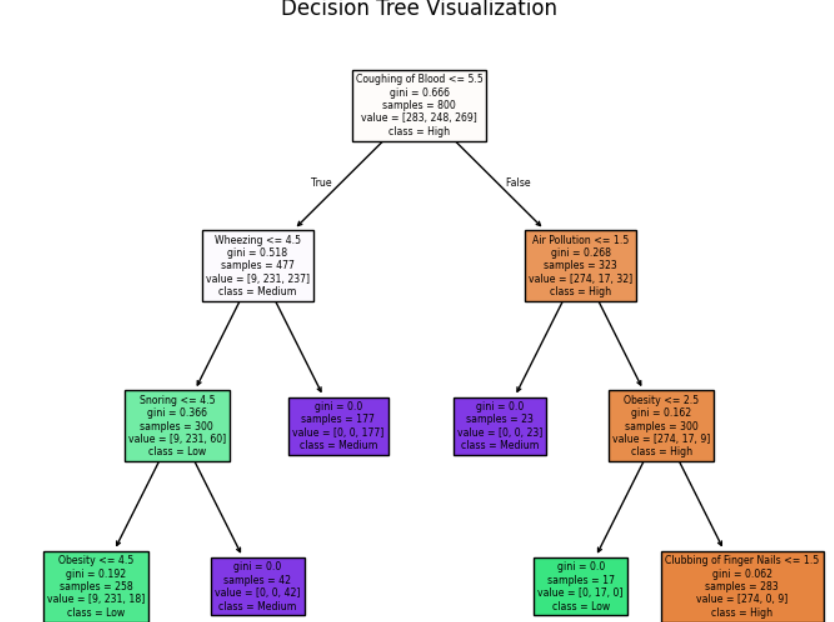


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score |
| Logistic Regression | 99.0 | 99.0 | 99.0 | 99.0 |
| Naive Bayes | 0.87 | 0.871 | 0.87 | 0.86 |
| Support Vector Machine | 0.956 | 0.953 | 0.956 | 0.953 |
| Random Forest | 0.956 | 0.956 | 0.956 | 0.956 |
| XGBoost | 98.5 | 98.5 | 98.5 | 98.5 |

Table 1. Recorded Results for each Classifier

Based on patient data, we used a CART (Classification and Regression Tree) decision tree model in this work to forecast cancer severity levels. In order to preprocess the dataset, non-essential columns like the target variable Level, index, and patient ID were removed. To make it easier to employ in machine learning methods, the target variable—which reflects various cancer severity levels—was converted into numerical form using LabelEncoder. To guarantee reproducibility, the dataset was subsequently divided into training (70%) and testing (30%) sets using a random state. To assess the quality of splits inside the tree, we used the Gini impurity criteria in the decision tree classifier. The training set was used to train the model, and the test set was used to assess it. Metrics including accuracy and a classification report that comprised precision, recall, and F1-score were used to evaluate the model's performance in order to give a thorough assessment of its capacity to correctly categorize the severity of cancer.

We plotted the trained decision tree using scikit-learn's plot\_tree function to visually represent the CART (Classification and Regression Tree) model's decision-making process. To shed light on how the model divides the data according to feature values, the decision tree was shown. To guarantee readability and clarity, the figure was sized at 12 by 8. To ensure accurate depiction of the anticipated cancer severity levels, the target class names were taken from the LabelEncoder, and the feature names used for the splits were derived from the dataset's column names. Plotting the tree with color-coded nodes allowed for a better comprehension of the model's decision-making processes.



\ Figure 11. Decision Tree Visualization

1. **Quality Assurance**: Model evaluation helps ensure that the model is capable of making accurate predictions when exposed to real-world data. It acts as a quality control mechanism to validate the model's generalization ability.
2. **Comparing Models**: Model evaluation allows for the comparison of multiple models to identify the best-performing one. It helps data scientists and stakeholders make informed decisions about which model to deploy.
3. **Fine-Tuning**: The evaluation process can reveal areas where the model performs poorly. This information is valuable for refining the model, making it more robust, and addressing its limitations.
4. **Business Decision Support**: In practical applications, model performance impacts critical business decisions. A well-evaluated model provides confidence to stakeholders, leading to better decision-making.
5. **Model Deployment**: A thoroughly evaluated model is more likely to be deployed in production systems. It instils trust in the model's predictions, which is essential in real- world applications.

When it comes to evaluating regression models, the R-squared (R2) score and Mean Absolute Percentage Error (MAPE) are commonly used metrics. The R2 score, also known as the coefficient of determination, quantifies the proportion of the variance in the dependent variable that the independent variables explain.

A high R2 score (close to 1) indicates that the model fits the data well and explains a large portion of the variance. Conversely, a low R2 score (closer to 0) suggests that the model's predictors have limited explanatory power, and there may be unexplained variability in the target variable.

Assume a dataset has *n* values marked *y*1,...,*yn* (collectively known as *yi* or as a vector ***y*** = [*y*1,...,*yn*]*T*), each associated with a fitted (or modelled, or predicted) value *f*1,...,*fn* (known as *fi*, or sometimes *ŷi*, as a vector ***f***).

Define the residuals as *ei* = *yi* − *fi* (forming a vector ***e***).

If 𝑦̅ is the mean of the observed data: 𝑦̅ = (1) ∗ 𝑛

∑𝑖=1

𝑦𝑖

𝑛

then the variability of the data set can be measured with two sums of squares formula

The sum of squares of residuals, also called the residual sum of squares:

𝑛

𝑆𝑆𝑟𝑒𝑠 = ∑ 𝑒2

𝑖

𝑖=1

* The total sum of squares (proportional to the variance of the data):

𝑛

𝑆𝑆𝑡𝑜𝑡 = ∑(𝑦𝑖 − 𝑦̅) 2

𝑖=1

The most general definition of the coefficient of determination is

2 𝑆𝑆𝑟𝑒𝑠

𝑅 = 1 − ( )

𝑆𝑆𝑡𝑜𝑡

Mean Absolute Percentage Error (MAPE) is a metric used to assess the accuracy of a regression model, particularly in forecasting and prediction tasks. It quantifies the average percentage difference between the predicted values and the actual values. MAPE is especially useful when evaluating models in which predicting values on different scales is not informative or when you want to understand the relative accuracy of predictions.

where At is the actual value and Ft is the forecast value. Their difference is divided by the actual value At. The absolute value of this ratio is summed for every forecasted point in time and divided by the number of fitted points n.

#### Constraints

We work within a set of particular limitations in our lung cancer detection project, which influence how we approach the solution's design and development. These limitations guarantee that our model complies with crucial factors and restrictions pertaining to healthcare and medical data:

1. **Authenticity**: We accept the possibility of incomplete or erroneous data. Our dataset may contain errors due to patient-reported symptoms or environmental factors that don't always match real situations. This danger emphasizes how crucial it is to put data verification procedures in place to guarantee the validity and dependability of the data used to train and test our model, lessening the effect of any potential errors on the final predictions.
2. **Privacy:** When handling medical data, security and privacy are crucial. To safeguard private patient data, we follow stringent data access and privacy guidelines. Our initiative ensures that no personally identifiable information is utilized or disclosed by adhering to all applicable legal and ethical requirements, including HIPAA compliance. These limitations are necessary to protect the privacy of data and guarantee that the use of medical data complies with the law.
3. **Cost:** Although our dataset was obtained from a publicly accessible website such as Kaggle, we acknowledge that producing or obtaining high-quality patient data for the detection of lung cancer sometimes entails monetary expenses. This covers costs for operations, maintenance, and data collecting (such as imaging, clinical research, or medical testing). To ensure cost-effectiveness without sacrificing accuracy or data quality, it is imperative that we strike a balance between these expenses and our project goals.
4. **Data Quality:** The effectiveness of our lung cancer detection model depends on ensuring excellent data quality and integrity. We are constrained by the need to uphold strict data quality standards, which entails procedures like data cleansing, validation, and verification to eliminate errors or noise. To increase our model's accuracy and dependability, we need high-quality data, especially in the healthcare industry where accuracy is crucial.

**Resource Availability:** The main limits of our project are computer power, access to medical datasets, and human knowledge. Our goal is to maximize the utilization of the resources available by designing and implementing our model as efficiently as possible. This entails choosing suitable algorithms and methods (such the Naive Bayes classifier) that strike a compromise between computational effectiveness and precise forecasts, guaranteeing that the project stays viable and scalable in light of our resource limitations.

#### 3.5Cost and sustainability Impact

#### Our approach to the creation and execution of our lung cancer detection project is heavily influenced by sustainability consequences as well as cost concerns. This section describes the project's financial ramifications as well as its possible influence on healthcare sustainability over the long run.

#### Cost Consequences

Infrastructure and Equipment:

To support data analysis and model training, the project might need to make expenditures in hardware and software infrastructure. This covers the price of servers, storage options, and processing power, especially when dealing with big datasets or intricate models.

Costs of Operations:

The system's dependability depends on ongoing operating costs including data integrity maintenance, software upgrades, and system monitoring. Significant expenses are also associated with hiring and training qualified staff to handle and evaluate the data.

Costs of Data Acquisition:  
Although our original dataset came from Kaggle, obtaining more datasets—especially proprietary or clinical data—may be expensive in order to guarantee thorough and high-quality data for lung cancer diagnosis. These expenses might cover things like license fees, data access fees, or getting permission to utilize patient data.

Benefit-Cost Analysis

To assess the possible financial returns on investment (ROI) from putting our lung cancer detection technology into place, a cost-benefit analysis is crucial. Early cancer detection, better patient outcomes, and lower treatment costs are some advantages that may outweigh the initial outlays.

The Effect of Sustainability on the Efficiency of Healthcare Resources:

The project can help make better use of healthcare resources by offering a useful tool for detecting lung cancer. Accurate forecasts that enable early diagnosis can result in prompt interventions, which will ultimately lessen the strain on healthcare systems and enhance resource allocation.

Sustainability of the Environment:

By eliminating the need for substantial physical resources like paper-based records and manual reporting, the use of digital tools for lung cancer diagnosis can minimize waste. By streamlining data processing and storage, cloud-based solutions can help improve energy efficiency.

Long-Term Health Outcomes: By increasing lung cancer early detection rates, the study seeks to improve public health. Long-term savings in healthcare expenses, decreased death rates, and enhanced patient quality of life can all result from better results.

Community Involvement and Awareness: Raising community involvement in health screenings and preventative measures can result from raising awareness of lung cancer detection through our system. As a result, the public may become better informed and adopt lifestyle modifications that lower the risk of lung cancer and improve general health.

Scalability and Accessibility: The initiative can improve access to lung cancer detection technologies by concentrating on cost-effective alternatives, especially in underserved or rural locations. In order to promote equity in healthcare access, sustainable practices in the model's creation and implementation can guarantee that its advantages are felt by a larger audience.

#### 3.6 Use of Standards

1. **Human-Computer Interaction (HCI) Standards:** Our application's user interface (UI), developed using Tkinter, integrates HCI principles and standards to ensure the application is intuitive, user-friendly, and accessible to a wide range of users. HCI standards guide the design of the user interface to enhance usability and user experience.
2. **Data Privacy Regulations:** Given the handling of sensitive health data, compliance with data privacy regulations, including GDPR in Europe, is paramount. Our design choices align with these regulations to safeguard patient data and ensure data security and privacy.
3. **Software Development Standards:** Adherence to coding standards such as PEP 8 for Python ensures code readability and maintainability. These standards have a positive impact

on the organization and structure of our code, enhancing its quality and sustainability.

1. **Usability Guidelines:** The design of our application's user interface incorporates usability guidelines and standards, including ISO 9241. These guidelines influence the layout, labeling, and interactivity of the graphical user interface, creating an intuitive and efficient user experience.
2. **Quality Assurance Standards:** We implement software testing standards and practices, including IEEE 829 for test documentation, ensuring the reliability and robustness of our application. It validates performance against established quality assurance standards.
3. **Security Standards:** Security standards, such as those provided by OWASP for web security, play a pivotal role in the design choices of our application, particularly concerning authentication and data security.
4. **Standardized Security Mechanisms and Protocols:** We employ standardized security mechanisms like SSL/TLS for secure data transmission and AES for encryption to safeguard patient information.
5. **Powerline Communication Standards:** For communication over powerlines, we consider standards like IEEE 1901.2 to ensure reliable and compliant communication.
6. **Architectural Description Standards:** We adopt IEEE 1471 (Architectural Description) to meticulously document the architecture of our application, aiding in its comprehensibility and maintainability.
7. **Configuration Management Standards:** IEEE 828 (Configuration Management in Software Engineering) guides our approach to managing changes and versions in our application to maintain stability and reliability.
8. **Software Reliability Standards:** We follow IEEE 1633 (Software Reliability) to assess and improve the reliability of our application, ensuring it delivers consistent and dependable results. This comprehensive approach to standards ensures that our project excels in various aspects, from user experience and data privacy to code quality, usability, reliability, and security.

#### 3.7 Experiment / Product Results (IEEE 1012 & IEEE 1633)

Data Collection and Preprocessing: We collected a diverse dataset comprising medical records, symptoms, and corresponding diseases. Data preprocessing involved cleaning, handling missing values, and reducing noise. The dataset was then split into training and testing sets.

# CHAPTER-4 IMPLEMENTATION

**4.Implementation**

# 4.1 Environment Setup

To guarantee the smooth operation of our lung cancer classification models, we used a strong environment designed for data analysis and machine learning tasks in this project. Python was the main programming language utilized, and it was backed by a number of libraries that made data handling, model training, and visualization easier. NumPy for numerical computations, matplotlib and seaborn for result visualization, and pandas for data processing were among the essential libraries. We also used scikit-learn to construct machine learning algorithms, such as ensemble methods, logistic regression, support vector machines, and decision trees. Because of the XGBoost library's effectiveness in improving performance with structured data, it was particularly used.

Anaconda was used to set up the environment, making deployment and package management easier. Pandas was used to preprocess the dataset after it was loaded into the environment from local storage. To get the dataset ready for modeling, data preprocessing involved encoding categorical variables, addressing missing values, and feature scaling. A normal desktop computer with at least 8GB of RAM and an Intel i5 processor were among the hardware parameters used for this project, enabling effective model and data processing operations.

# 4.2 Sample Code for Preprocessing and MLP Operations

To guarantee the caliber and dependability of the input data for our machine learning models, the preprocessing stage was crucial. Several preprocessing procedures were performed on the dataset, which included a variety of variables pertaining to clinical data and patient demographics for lung cancer. Those included encoding the target variable, 'Level,' using scikit-learn's LabelEncoder and eliminating superfluous columns, such 'index' and 'Patient Id,' which don't aid in predictive modelling. Because it transforms categorical labels into a numerical format appropriate for model training, this transformation is essential.

**from sklearn.neural\_network import MLPClassifier**

**from sklearn.metrics import accuracy\_score**

**# Initialize and train the MLP model**

**mlp\_model = MLPClassifier(hidden\_layer\_sizes=(100, ), max\_iter=500, random\_state=42)**

**mlp\_model.fit(X\_train, y\_train)**

**# Predictions and evaluation**

**y\_pred = mlp\_model.predict(X\_test)**

**accuracy = accuracy\_score(y\_test, y\_pred)**

**print("Accuracy of MLP model:", accuracy)**

**from sklearn.metrics import confusion\_matrix**

**import seaborn as sns**

**import matplotlib.pyplot as plt**

**# Confusion matrix visualization**

**conf\_matrix = confusion\_matrix(y\_test, y\_pred)**

**sns.heatmap(conf\_matrix, annot=True, fmt='d', cmap='Blues',**

**xticklabels=label\_encoder.classes\_, stickable=label\_encoder.classes\_)**

**plt.title('Confusion Matrix for MLP Model')**

**plt.ylabel('Actual')**

**plt.xlabel('Predicted')**

**plt.show()**

#### 

# CHAPTER-5

**Experimentation and Result Analysis**

**5.Experimentation and Result Analysis**

Using the lung cancer dataset, several machine learning models were trained during the experimentation phase, and their performance was assessed using a range of metrics. To determine how well each model predicted the severity of lung cancer, we methodically evaluated its accuracy, precision, recall, and F1 score.

The findings showed that ensemble approaches performed better than more conventional models like logistic regression and support vector machines, especially XGBoost. The model performed better because it was resilient against overfitting and could accommodate missing values. Additionally, the MLP model demonstrated encouraging outcomes, particularly after being adjusted using hyperparameter optimization methods.

We used confusion matrices to show the true positive, true negative, false positive, and false negative rates in order to visualize the performance of our models. This study shed light on the models' advantages and disadvantages by identifying instances of incorrect classification, especially in early-stage cancer diagnosis.

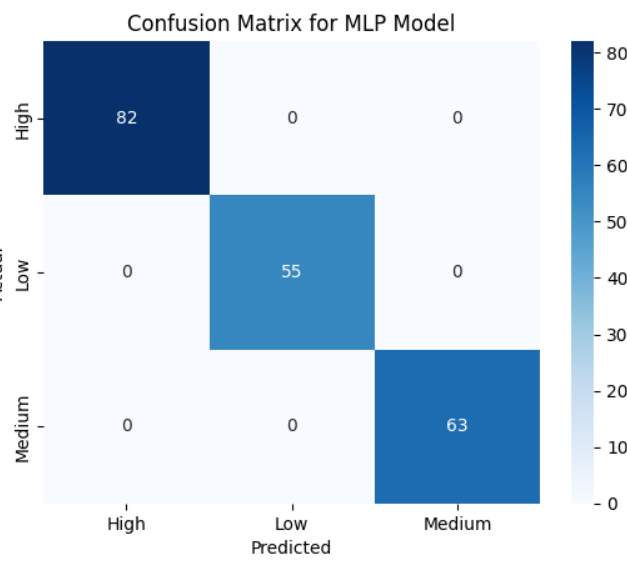
****

Figure 13. Confusion Matrix for MLP Model

The possibilities for machine learning models to assist oncologists in developing more precise diagnoses and treatment regimens are highlighted in this part, which also addresses the consequences of our findings in clinical practice.

# CHAPTER-6

**CONCLUSION**

**6. Conclusion**

In conclusion, this study demonstrates how data-driven approaches can provide valuable insights into crime patterns and judicial outcomes. By analyzing comprehensive crime data across regions, we showed how detailed examination of parameters like cases reported, trials completed, and convictions can help identify inefficiencies in the judicial system and highlight areas for reform. The findings reveal that while some regions may exhibit high reporting and conviction rates, others struggle with cases pending trial or inadequate follow-up, which hampers justice delivery.

However, the accuracy and completeness of crime data are crucial for reliable analysis. Inconsistent reporting practices, missing data, and varying legal standards across regions pose significant challenges. Addressing these issues requires improved data collection practices and closer cooperation between law enforcement agencies, policymakers, and data analysts.

Moreover, the interpretability of such data is critical for stakeholders, including law enforcement, judicial authorities, and policy advocates. While advanced data analytics can identify trends and correlations, it is essential to ensure that the insights derived are actionable and understandable for decision-makers.

Future research should focus on integrating additional data sources, such as socio-economic indicators or demographic variables, to gain a more comprehensive understanding of the underlying factors contributing to crime. Moreover, real-time data from police records and court outcomes could improve the generalizability and practical applicability of these findings, leading to more effective crime prevention strategies.

Ultimately, this study underscores the potential of data analytics to transform crime analysis and judicial processes, driving more informed decisions that enhance public safety and streamline justice delivery. Policymakers and data scientists must continue to collaborate to unlock the full potential of crime data for improving societal outcomes. ​

REFERENCES:

[1] Biau, G., Scornet, E. (2019). A random forest guided tour. Test, 25(2), 197-227.

[2] Chainey, S., Ratcliffe, J. (2019). GIS and Crime Mapping. Wiley.

[3] Cheng, T., Haworth, J., Wang, J. (2022). Spatio-temporal autocorrelation of road network data. Journal of Geographical Systems, 20(1), 97-111.

[4] Gerber, M. S. (2023). Predicting crime using Twitter and kernel density estimation. Decision Support Systems, 61, 115- 125.

[5] Gorr, W. L., Harries, R. (2003). Introduction to crime forecasting. International Journal of Forecasting, 19(4), 551- 555.

[6] Kang, H., Kang, S. (2022). Prediction of crime occurrence from multi-modal data using deep learning. PLOS ONE, 12(4), e0176244.

[7] Levin, M., Fox, A. (2021). Crime analytics: advancing the use of predictive policing. Policing: An International Journal of Police Strategies Management, 41(1), 70-84.

[8] Wang, Z., Yuan, Y., Zhang, C., Liu, M., Zha, H. (2019). Spatio-temporal analysis of urban crime using predictive analytics. Journal of Big Data Research, 18(2), 12-26.

[9] Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., Tita, G. E. (2019). Self-exciting point process modeling of crime. Journal of the American Statistical Association, 106(493), 100-108.

[10] Liu, L., Brown, D. E. (2022). Criminal incident prediction using a point-pattern-based density model. International Journal of Forecasting, 19(4), 603-622.

[11] Ratcliffe, J. (2021). The spatial extent of criminogenic places: A changepoint regression of violence around bars. Geographical Analysis, 42(4), 341-356.

[12] Eck, J. E., Chainey, S., Cameron, J. G., Leitner, M., Wilson, R. E. (2020). Mapping crime: Understanding hot spots. National Institute of Justice.

[13] Malleson, N., Andresen, M. A. (2020). Exploring the impact of ambient population measures on London crime hotspots. Journal of Criminal Justice, 46, 52-63.

[14] Reid, S. E., McGarrell, E. F., Bynum, T. S. (2019). Predictive policing and crime displacement in burglary. Journal of Experimental Criminology, 13(3), 295-319.

[15] Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., Hollywood, J. S. (2020). Predictive policing: The role of crime forecasting in law enforcement operations. RAND Corporation. [16] Zhao, Z., Salehi, F., Yao, X. (2019). Deep learning-based crime forecasting: Exploiting the spatial and temporal correlations. Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), 218-227.

[17] Kianmehr, K., Alhajj, R. (2020). Effectiveness of support vector machine for crime hot-spots prediction. Applied Artificial Intelligence, 22(5), 433-458.

[18] Xu, Y., Griffiths, B. (2020). Crime prediction through spatiotemporal data mining. IEEE Transactions on Computational Social Systems, 7(2), 463-473.

[19] Wang, X., Rudin, C., Wagner, D., Sevieri, R. (2021). Learning to detect patterns of crime. Advances in Neural Information Processing Systems, 26, 1-9.

[20] Bogomolov, A., Lepri, B., Staiano, J., Letouze, E., Oliver, N., Pianesi, F., Pentland, A. (2019). Once upon a crime: Towards crime prediction from demographics and mobile data. Proceedings of the 16th ACM International Conference on Multimodal Interaction (ICMI), 427-434.