CENTER FOR DEVELOPMENT OF ADVANCED COMPUTING
ACTS Chennai

PROJECT REPORT

ON

MEDICINE OVERDOSE PREDICTION USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for

the Diploma in Big Data Analytics (DBDA)

Submitted by:

ADITYA MISHRA (Roll No.: 250260825001)

VIDYUTH RAMACHANDRAN (Roll No.: 250260825014)

Batch: DBDA-Feb2025

Under the Guidance of:

Dr. Sumithra Radhakrishnan

(Mentor, C-DAC ACTS Chennai)

Submitted to:

Centre for Development of Advanced Computing (C-DAC)

Advanced Computing Training School (ACTS), Chennai

Date of Submission: 08/08/2025

ABSTRACT

The opioid epidemic represents a critical public health challenge worldwide, with devastating consequences for individuals and communities. To address this crisis, there is a growing need for predictive models that can identify individuals at risk of opioid use and misuse. This research aims to develop a data-driven approach to forecast opioid-related outcomes by leveraging advanced machine learning techniques on comprehensive datasets.

The study utilizes a diverse set of features, including demographic information, medical history, prescription records, and social determinants of health, to build a robust predictive model. The dataset encompasses a large and representative sample of individuals, ensuring reliability during both training and validation phases. To address ethical concerns, privacy-preserving techniques are employed to manage sensitive information securely.

The proposed model not only focuses on identifying individuals at risk of opioid abuse but also differentiates between therapeutic use and potential misuse—an essential distinction for enabling personalized healthcare strategies and targeted interventions. Model performance is rigorously evaluated using key metrics such as sensitivity, specificity, and the area under the ROC curve.

Additionally, the research emphasizes model interpretability by identifying key contributing factors to the predictions. This transparency builds trust among healthcare professionals, policymakers, and the public. The findings of this study aim to inform early intervention programs, improve prescription practices, and support the creation of evidence-based policies to combat the opioid epidemic.

In conclusion, this work contributes meaningfully to ongoing efforts to mitigate opioid misuse by presenting a scalable, interpretable, and ethical predictive solution for proactive public health strategies. According to recent CDC data, drug overdose deaths remain a leading cause of mortality, reinforcing the need for real-time, data-driven tools in public health ([CDC, 2025](https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm)

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 About the Project

Drug overdose has become a critical public health crisis and is now the leading cause of death for individuals under the age of 50 globally. A significant challenge for city officials and public health organizations is the lack of adequate and timely data, which hinders their ability to understand and address the full scale of the opioid crisis. This project focuses on developing a predictive model that leverages machine learning to forecast drug overdose events. The model will analyze various factors to estimate the level of drug consumption, identify the types of drugs being used, and pinpoint the geographic areas most affected by overdoses.

According to provisional statistics published by the Centers for Disease Control and Prevention (CDC), over **80,000** people died from drug overdoses in the U.S. in 2024, with opioids being a major contributor. These numbers highlight the need for advanced tools capable of predicting and preventing such crises." (Source: https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm)

1.2 Project Objective

The primary objective of this project is to investigate, develop, and analyze several machine learning models for predicting drug use and potential overdoses. To achieve this, the project will integrate and analyze diverse data obtained from multiple sources, including sewage-based drug epidemiology, healthcare records, social media data mining, and law enforcement data. The resulting analysis aims to provide actionable insights that can help policymakers formulate more effective strategies and programs to combat fatal opioid overdoses and support affected communities.

1.3 Problem Statement

The rising number of drug overdose fatalities highlights an urgent need for proactive and data-driven intervention strategies. The current systems for monitoring and responding to overdose trends are often reactive, relying on data

that is incomplete or reported with significant delays. This makes it difficult for public health officials to detect emerging overdose hotspots, understand the drivers behind them, and allocate resources effectively. While several machine learning models have been developed, they often suffer from disadvantages such as slow learning rates, long execution times, and an inability to detect emerging trends at an early stage. The practical use of collected data remains a time-consuming challenge. Therefore, there is a need for a more accurate and efficient methodology to predict medicine overdoses.

1.4 Scope of the Project

This project aims to create a decision support system that can serve as a vital tool for physicians, public health analysts, and policymakers. By providing accurate and timely predictions, the system can be applied in several key areas:

- **Clinical Diagnosis:** Assisting doctors in analyzing patient data to identify at-risk individuals.
- **Medicinal Combination Testing:** Analyzing how combinations of different drugs contribute to overdose risk.
- **Public Health Analysis:** Allowing officials to monitor, predict, and respond to overdose trends across different locations.

1.5 Limitations of the Project

The performance of the proposed prediction model is fundamentally dependent on the quality, volume, and accessibility of the input data from disparate sources. Challenges include:

- Variations in data formats across different sources
- Data-sharing restrictions due to privacy concerns
- Potential lack of real-time access to critical information

Additionally, the model's predictive accuracy is influenced by the algorithms used and may require regular updates based on new trends or emerging substances.

Chapter 2: Literature Survey

2.1 About Selected Research Papers and Area

Recent studies have applied various machine learning algorithms to predict opioid prescription patterns and overdose risk using demographic, clinical, and behavioral data.

One of the most recent and impactful works is by **IEEE 2023** DOI: 10.1109/TBDATA.2023.10207592. The researchers developed a deep learning framework utilizing multimodal healthcare data, including electronic health records (EHRs), lab results, and prescription histories. Their model demonstrated high predictive accuracy and interpretability using feature importance scores. However, its reliance on structured EHR data and deep architectures made it less scalable for public health deployment in real-time scenarios.
Several other researchers have proposed alternative machine learning approaches:

- **Arti Gupta and Maneesh Shreevastava (IEEE 2021)** used a **feed-forward back-propagation neural network**. Their method showed convergence issues on larger datasets due to small learning rates and lacked robust preprocessing.
- **Shraddha Subhash Shirsath and Prof. Shubhangi Patil (IEEE 2019)** implemented **Naive Bayes and CNN-MDRP**. These models failed to capture complex disease correlations and performed poorly with numerical and sparse data.
- **Nikita Kamble et al. (IEEE 2019)** also utilized **Naive Bayes**, but their dataset was limited in size and lacked comprehensive cleaning, affecting model generalization.
- **Nilesh Borisagar et al. (IEEE 2019)** explored an **ensemble of algorithms**, including Levenberg-Marquardt and Bayesian regularization, which were computationally expensive and slow to train.
- **Sellappan Palaniappan and Rafiah Awang (IEEE 2019)** proposed a **Multilayer Perceptron (MLP)** model that underperformed in classification accuracy and was not suitable for real-time decisions.
- **M.A. Nishara Banu and B. Gomathy (IEEE 2020)** applied **C4.5 and MAFIA** algorithms but lacked validation on large-scale or real-time data.
- **Unsupervised models** such as **K-Means** and **DBSCAN** were also explored by some studies, but they showed limited usefulness in early detection and individual risk classification due to their generic clustering behavior.

These prior works underscore the need for scalable, interpretable, and well-preprocessed models that work with mixed data formats (numerical + categorical) and can be trained and deployed efficiently.

2.2 Summary and Research Gap

From the reviewed literature, several common challenges emerge:

- **Slow learning** and **long execution times** hinder practical deployment.
- **Low interpretability** makes clinical trust and integration difficult.
- **Dependence on clean EHR data** limits model usability in broader public datasets.
- **Minimal feature engineering** results in reduced predictive power.
- **Lack of real-time performance** in many deep learning models.

Chapter 3: Feature Engineering

3.1 Data Collection

The dataset used for this project was acquired from **Kaggle**, containing records related to **opioid prescription behavior and prescriber attributes** across the United States. The dataset is in **CSV format** and links to publicly available health data, including sources like the **CDC Drug Overdose Dashboard** (as referenced in related literature).

- **Data snapshot date:** July 6, 2025
- **Number of records:** ~25,000 rows
- **Number of features:** 256 columns

The dataset includes the following information:

- **Prescriber demographics** (Gender, State, Credentials)
- **Specialty and prescription patterns**
- **Opioid prescription details and flags**
- **Drug names and refill counts**
- **Number of claims and cost information**

This dataset offers a detailed perspective on prescriber behavior and opioid prescription trends across a 12-month period in the U.S.

3.2 Data Preprocessing

Before training the model, the dataset was cleaned and preprocessed using **Pandas, NumPy, and Scikit-learn** libraries in Python.

The key steps included:

- **Missing Value Treatment**
    - Categorical fields such as Gender, Credentials, and Specialty were cleaned. Missing values were handled using imputation techniques or dropped if deemed insignificant.

- **Categorical Encoding**
    - Categorical variables like Gender, State, Credentials, and Specialty were encoded using **Label Encoding** to convert them into numeric format suitable for the Random Forest model.
- **Feature Selection**
    - A combination of **domain knowledge**, a **correlation matrix**, and **feature importance scores** from the trained Random Forest model were used to select impactful features for prediction.
- **Dropping Irrelevant Columns**
    - Columns such as Id (identifier) were removed as they don't contribute to model prediction.

3.3 Feature Extraction

Feature extraction was carried out to enhance the predictive power of the dataset. This involved identifying the most relevant features and creating new ones that could reveal underlying patterns.

- **Domain Knowledge Selection:**
  Based on medical and healthcare insights, the following were chosen as key predictors:
    - Total Claim Count
    - Opioid Claim Count
    - Opioid Cost
    - Days' Supply
    - Specialty and Credentials

- **Derived Features:**
  New features were engineered to enrich the dataset:

- **Opioid Claim Ratio** = Opioid Claim Count / Total Claim Count
  → Helps highlight prescribers with a high tendency toward opioids.
- **Average Opioid Cost per Claim** = Opioid Cost / Opioid Claim Count
  → Indicates the economic impact of individual prescriptions.
-

- **Feature Importance Analysis:**
  A **Random Forest Classifier** was trained on the preprocessed dataset to compute **feature importances**. The top contributing features were retained, and those with negligible impact were dropped.

This structured feature engineering pipeline transformed the raw dataset into a refined input suitable for building robust machine learning models.

Chapter 4: Development and Coding

4.1 Technology Used

The system is built using the following software and hardware configurations:

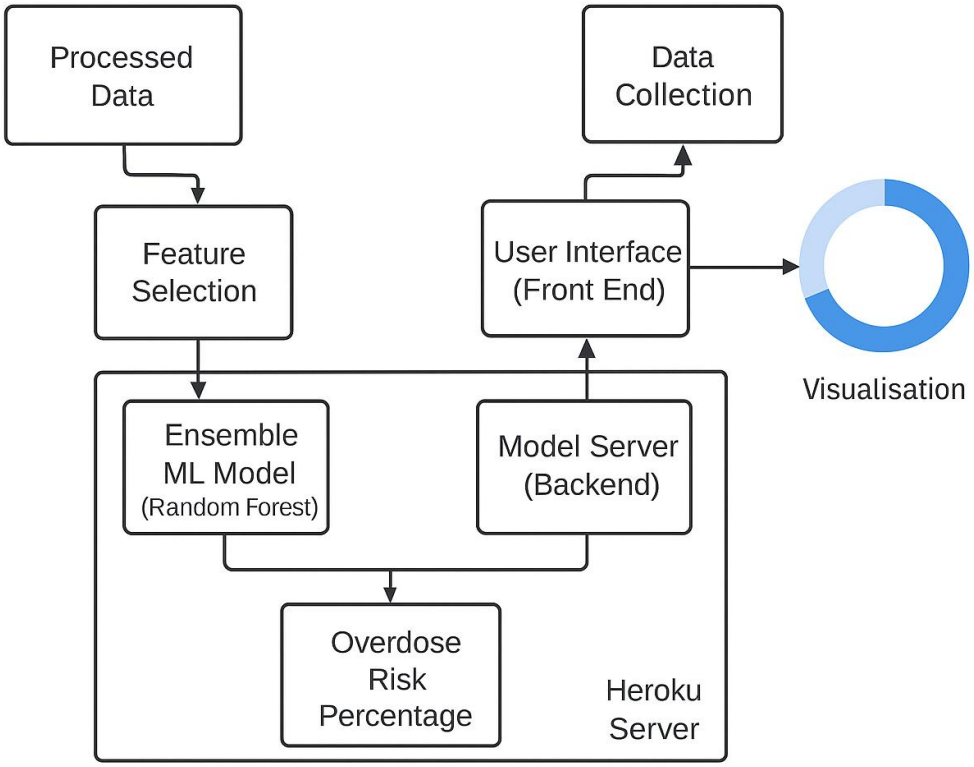
Software Requirements

- **IDE**: Anaconda Navigator with Jupyter Notebook
- **Programming Language**: Python 3.12.11
- **Libraries Used**:
  - **Pandas** and **NumPy** – for data preprocessing
  - **Scikit-learn** – for machine learning (Random Forest classifier)
  - **Matplotlib** and **Seaborn** – for visualization
  - **Pickle** – for model serialization

Hardware Requirements

- **Processor**: Intel Core i3 or higher (Core 2 Duo may be outdated)
- **RAM**: Minimum 4 GB (8 GB recommended for faster computation)
- **Storage**: At least 500 GB HDD or 128 GB SSD

4.2 Architecture Diagram



Model Design and Flow

The system aims to predict whether a medical prescriber is likely to issue opioid prescriptions. It follows a modular machine learning pipeline comprising the following key stages:

1. Data Ingestion

- The primary dataset (prescriber-info.csv) is imported using the Pandas library.
- Supplementary datasets such as overdoses.csv, opioids.csv, and Dataset_Upload.csv are optionally integrated to enrich the feature set and provide additional context.

2. Data Preprocessing

- **Missing Values:** Null values in critical columns like Gender and Credentials are treated using imputation (e.g., mode or mean).

- **Encoding Categorical Variables:** Features such as Gender, State, Credentials, and Specialty are encoded using Label Encoder for model compatibility.
- **Column Dropping:** Non-predictive or identifier columns such as Id are removed.
- **Normalization:** Numerical features (like claim counts, costs, and supply duration) are normalized to ensure uniformity and better model performance.

3. Feature Engineering

- **Feature Selection:** Random Forest's built-in feature importance and domain expertise are used to select the most impactful features.
- **Feature Extraction:** Derived features are created from existing ones to highlight potential indicators of opioid prescribing behaviour.

4. Model Training

- A **Random Forest Classifier** is employed due to its robustness, ability to handle high-dimensional data, and resistance to overfitting.
- The dataset is split into training and testing sets using stratified sampling.
- The model is trained and then evaluated on test data using metrics like Accuracy, Precision, Recall, and F1-Score.

5. Model Evaluation and Storage

- The model's performance is analyzed using classification reports and confusion matrices.
- Once validated, the trained model is serialized using pickle and saved as finalized_model.sav for future deployment and use.

6. Visualization and Insights

- Feature importance scores are visualized using bar charts to explain which features influence predictions the most.
- Confusion matrices are plotted using heatmaps for a better understanding of prediction accuracy across classes.

Chapter 5: Visualisation and Testing

5.1 Model Evaluation

After training the RandomForestClassifier on the preprocessed dataset, the model was evaluated using the **test set** (20% split). The following evaluation metrics were calculated using sklearn.metrics:

- **Accuracy Score:** Indicates the proportion of correct predictions.
- **Precision Score:** Proportion of true positive predictions among all predicted positives.
- **Recall Score:** Measures the proportion of actual positives correctly identified.
- **F1-Score:** Harmonic mean of precision and recall, used for imbalanced classes.

These metrics were printed using the classification_report() function for detailed insights.

5.2 Confusion Matrix

A **confusion matrix** was generated using confusion_matrix(y_test, y_pred), giving a breakdown of:

- **True Positives (TP)** – correctly predicted opioid prescribers
- **True Negatives (TN)** – correctly predicted non-opioid prescribers
- **False Positives (FP)** – non-prescribers incorrectly labeled as prescribers
- **False Negatives (FN)** – actual prescribers incorrectly labeled as non-prescribers

To improve readability, a **heatmap** of the confusion matrix was plotted using seaborn.heatmap().

5.3 Visualisation Tools Used

Tool	Purpose
Matplotlib	Used for plotting bar graphs of feature importance.
Seaborn	Used for confusion matrix heatmap visualization.

5.5 Final Testing Insights

- The Random Forest model demonstrated **high classification accuracy**, especially in distinguishing opioid prescribers from non-prescribers.
- Visualizations helped verify the model's performance and interpret results.
- These tools support the explainability and trustworthiness of the system, which is essential in healthcare-related use cases.

Chapter 6: Conclusion

Drug overdose, especially due to opioids, continues to be one of the most serious public health problems across the globe. Through this project, we aimed to create a data-driven solution that could help predict whether a prescriber is likely to issue opioid prescriptions, using machine learning techniques.

By using Random Forest as our main classification algorithm, we were able to build a predictive model that performed well in terms of accuracy, precision, recall, and F1-score. The model worked on structured data, including prescriber details, credentials, drug claims, and other relevant attributes. Before training the model, we performed important preprocessing steps like handling missing values, encoding categorical data, and selecting meaningful features based on importance scores.

One of the strengths of this project is that the model not only makes accurate predictions but also provides insights into which features are most important for identifying potential opioid prescribers. This makes it easier for healthcare professionals and policymakers to understand the reasoning behind the predictions.

Overall, this project shows how machine learning can be used to support public health decisions and reduce opioid-related risks. While our model is based on historical and structured data, it can be further improved by integrating real-time updates and larger datasets. With continuous development, it can be turned into a practical tool for monitoring and prevention of opioid misuse in communities.

Appendices

Appendix A: Dataset

The dataset used in this project (prescriber-info.csv) was sourced from a CDC-linked Kaggle dataset and contains 25,000 records with 256 features. Below is a sample of the first 100 rows used for model training and testing.

Dataset Highlights:

- **Format:** CSV
- **Source Link:** prescriber-info.csv
- **Features include:** Gender, State, Specialty, Opioid Prescriber, Total Claim Count, Opioid Cost, etc.
- **Target Variable:** Opioid.Prescriber (0 = No, 1 = Yes)

Id	Gender	State	Credential	Specialty	ABILIFY	ACETAMIN	ACYCLOVI	ADVAIR.DI	AGGRENO	ALENDROI	ALLOPURI	ALPRAZOL	AMIODAR(	AMITRIPTY	AMLODIPII	AMLODIPII	AMOXICILI	AMOX.TR.I	AMPHETAI	ATENOLOI	ATORVAST	AVOD
1710982582	M	TX	DDS	Dentist	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1245278100	F	AL	MD	General Surgery	0	0	0	0	0	0	0	134	0	0	15	0	0	0	0	15	0	
1427182161	F	NY	M.D.	General Practice	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1669567541	M	AZ	MD	Internal Medicine	0	43	0	0	0	21	0	0	0	0	58	0	0	0	0	0	13	
1679650949	M	NV	M.D.	Hematology/Oncology	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	
1548580897	M	PA	DO	General Surgery	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1437192002	M	NH	MD	Family Practice	0	0	0	25	0	0	47	54	0	0	191	12	0	29	0	43	87	
1407113988	F	PA	RN, MSN, /	Nurse Practitioner	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	0	29	
1023260569	M	TX	O.D.	Optometry	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1821106832	F	WI	MD	Internal Medicine	0	0	0	11	0	15	22	12	0	0	89	0	0	0	0	55	68	
1609931914	F	PR	M.D.	General Practice	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	
1659334472	M	TX	MD	General Surgery	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1144205303	M	CO	MD	Family Practice	0	14	0	0	0	31	29	30	0	20	128	28	0	15	0	0	223	
1548275050	M	OH	MD	Cardiology	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1952598419	F	TX	MD	Hematology/Oncology	0	0	16	0	0	0	0	22	0	0	0	0	0	0	0	0	0	
1780661793	M	OH	M.D.	General Surgery	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1356388011	F	MA	M.D.	Internal Medicine	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	
1588788921	M	CT		Hematology/Oncology	0	0	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1548238389	M	FL	M.D.	General Surgery	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1366582587	M	MN	D.D.S.	Dentist	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1083907059	M	UT	M.D.	Obstetrics/Gynecology	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1598748931	F	AL	M.D.	Family Practice	0	0	0	20	0	68	17	64	0	15	154	22	0	0	11	37	145	
1609121748	F	IA	ACNP	Nurse Practitioner	0	0	0	0	0	0	0	0	0	0	0	0	0	28	0	0	0	
1063566727	F	IL	DPM	Podiatry	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	
1487630919	M	CT	PAC	Physician Assistant	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	
1194795351	F	FL	A.R.N.P.	Nurse Practitioner	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	
1669671194	M	NY	M.D.	Diagnostic Radiology	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	
1629394754	M	NY	MD	Student in an Organized Health Care	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1174555080	F	FL	MD	Internal Medicine	0	0	63	16	0	12	11	0	0	0	17	0	0	0	0	0	0	
1508170895	F	CO	DDS	Dentist	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	
1871548537	F	MT	MD	Internal Medicine	0	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	
1841349677	F	IN	MSN, APRI	Nurse Practitioner	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	
1801816095	M	FL		Neurology	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	
1629038443	M	VA	MD	General Surgery	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1386607315	F	FL	D.O.	Neurology	0	0	0	0	0	13	0	0	0	163	0	0	0	0	0	0	0	
1811005366	M	CA	M.D,	Internal Medicine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1902018518	F	NV	APRN	Nurse Practitioner	294	0	0	0	0	0	0	13	0	15	0	0	0	0	0	0	0	
1043399322	M	TX	MD	Urology	0	23	0	0	0	0	0	12	0	16	0	0	0	25	0	0	0	
1255471140	M	IL	MD	Cardiac Electrophysiology	0	0	0	0	0	0	0	0	33	0	48	0	0	0	0	23	0	
1619975919	F	CT	PAC	Physician Assistant	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1023062759	M	MI	M.D.	Family Practice	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	
1376865949	F	PR	D.M.D.	Dentist	0	0	0	0	0	0	0	0	0	0	0	0	60	0	0	0	0	
1609867399	M	MO	M.D.	Internal Medicine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1225079817	F	IA	M.D.	Family Practice	0	0	0	0	0	14	0	20	0	0	37	0	0	0	0	11	42	
1093777229	F	PA	M.D.	Dermatology	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1134260607	F	NH	ARNP	Nurse Practitioner	0	0	0	0	0	0	0	26	0	0	24	0	0	0	0	22	42	
1164544045	M	AR	D.O.	Family Practice	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1538141569	F	IN	O.D.	Optometry	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1902916729	M	CA	M.D.	Internal Medicine	0	0	0	0	0	12	0	0	0	0	40	0	0	0	0	30	16	
1679647747	M	WI	D.D.S.	Dentist	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	
1154304293	F	TN	D.D.S	Dentist	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	
1659395069	M	TX	DPM MD	Podiatry	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1114977758	M	NE	MD	Emergency Medicine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1912339961	F	NC	FNP	Nurse Practitioner	0	0	0	0	0	0	0	16	0	0	21	0	0	0	0	0	32	
1699933614	F	IA	PA-C	Physician Assistant	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0	
1174601066	M	CA	MD	Internal Medicine	0	45	23	0	0	105	82	106	22	0	225	0	53	15	0	376	408	
1003893140	M	TX	MD	Psychiatry & Neurology	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	
1104843101	F	MN	MD	Infectious Disease	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	
1912114752	M	OH	MD	Psychiatry	25	0	0	0	0	0	0	23	0	38	0	0	0	0	0	0	0	
1326073214	M	NJ	PA-C	Physician Assistant	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

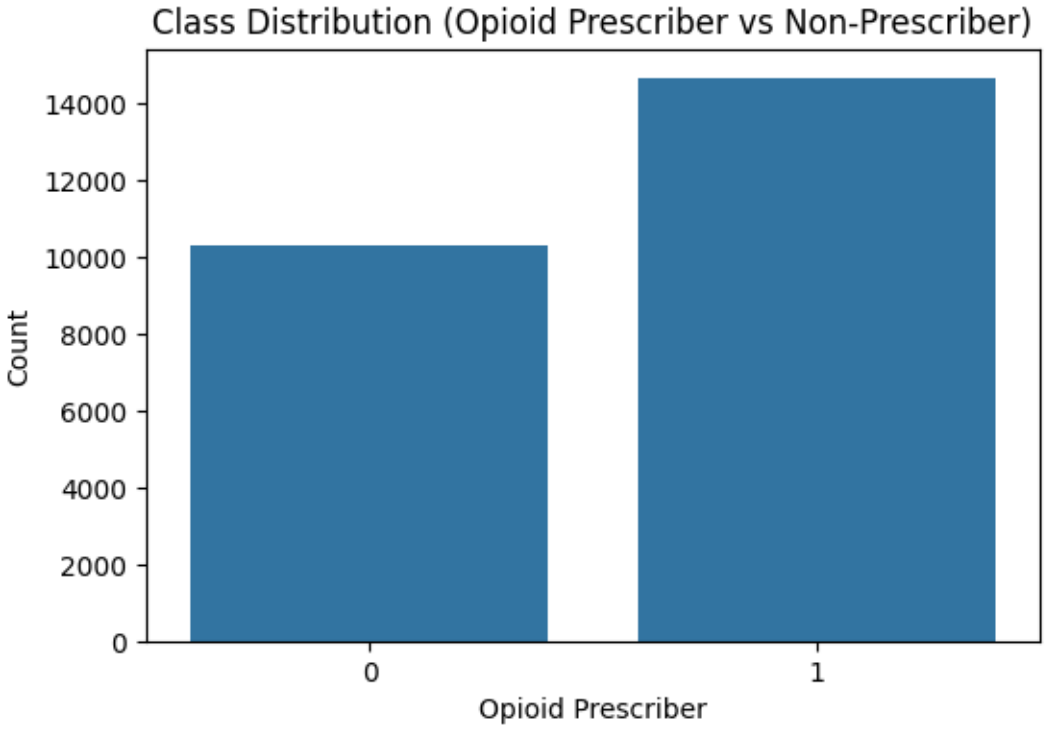
prescriber-info

Appendix B: Data Visualization

Data visualization was performed using **Matplotlib** and **Seaborn** to gain insights and interpret model performance. Below are key visuals included:

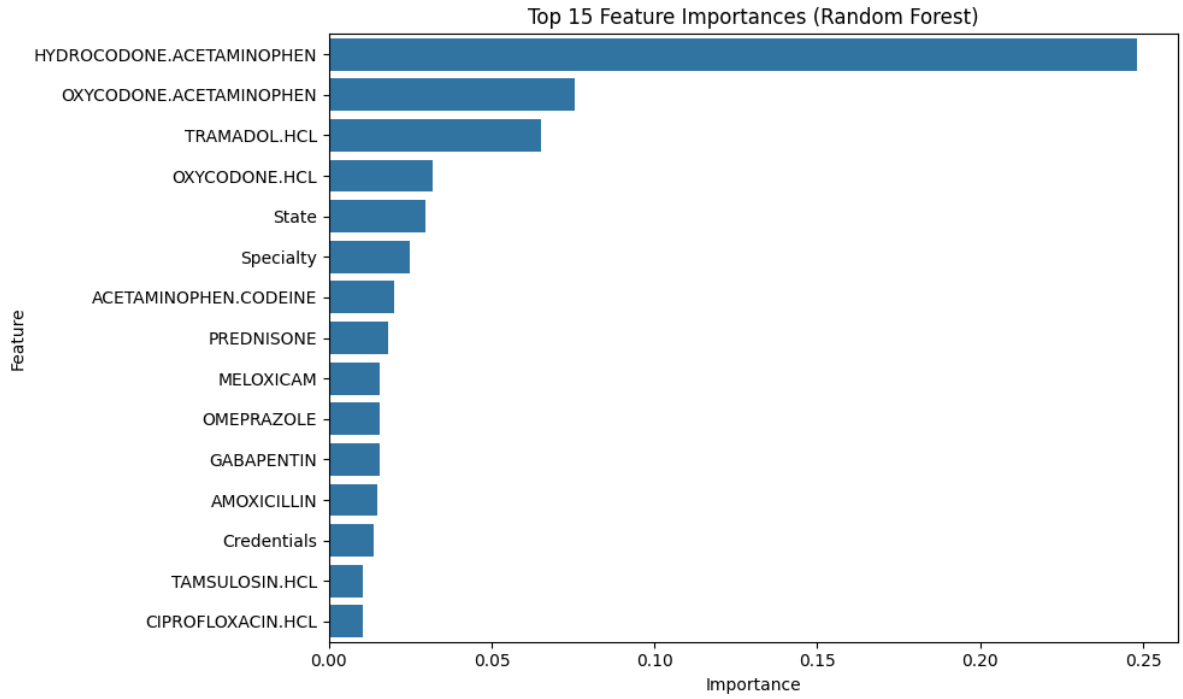
1. Class Distribution Plot

A bar chart depicting the balance of opioid prescriber (1) vs non-prescriber (0) in the dataset.



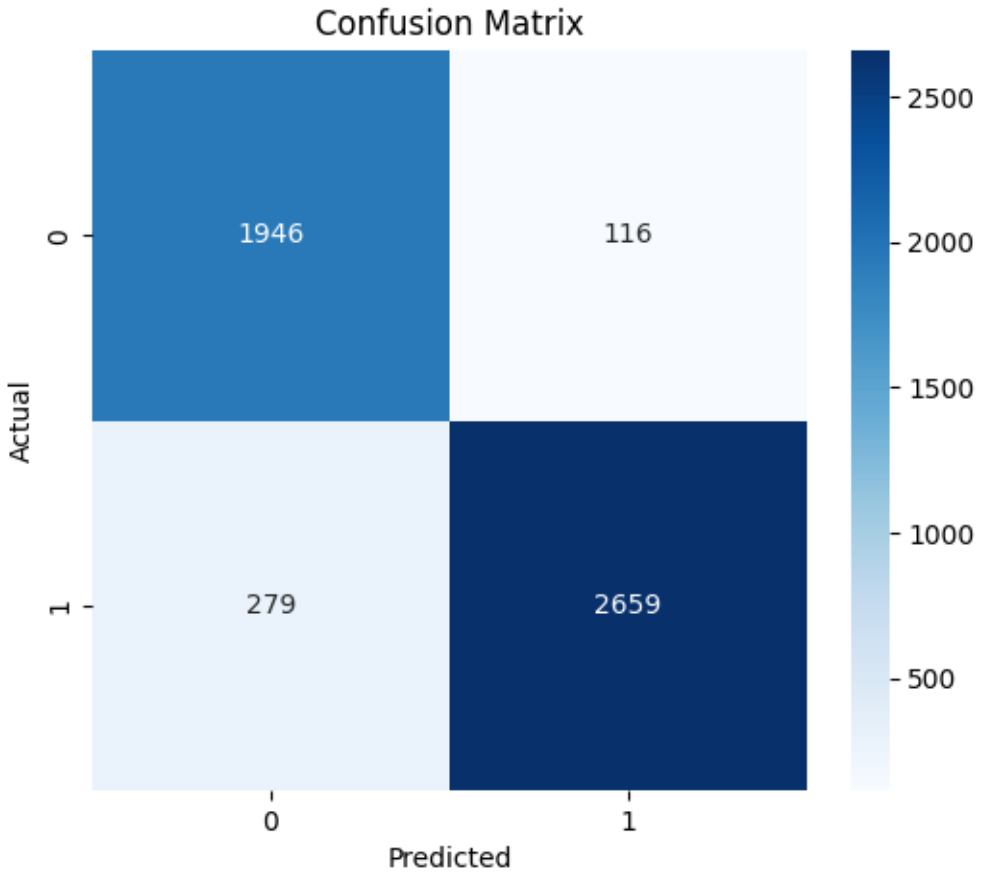
2. Feature Importance Plot

A ranked bar chart showing top features influencing opioid prescription prediction (generated using RandomForest feature importances).



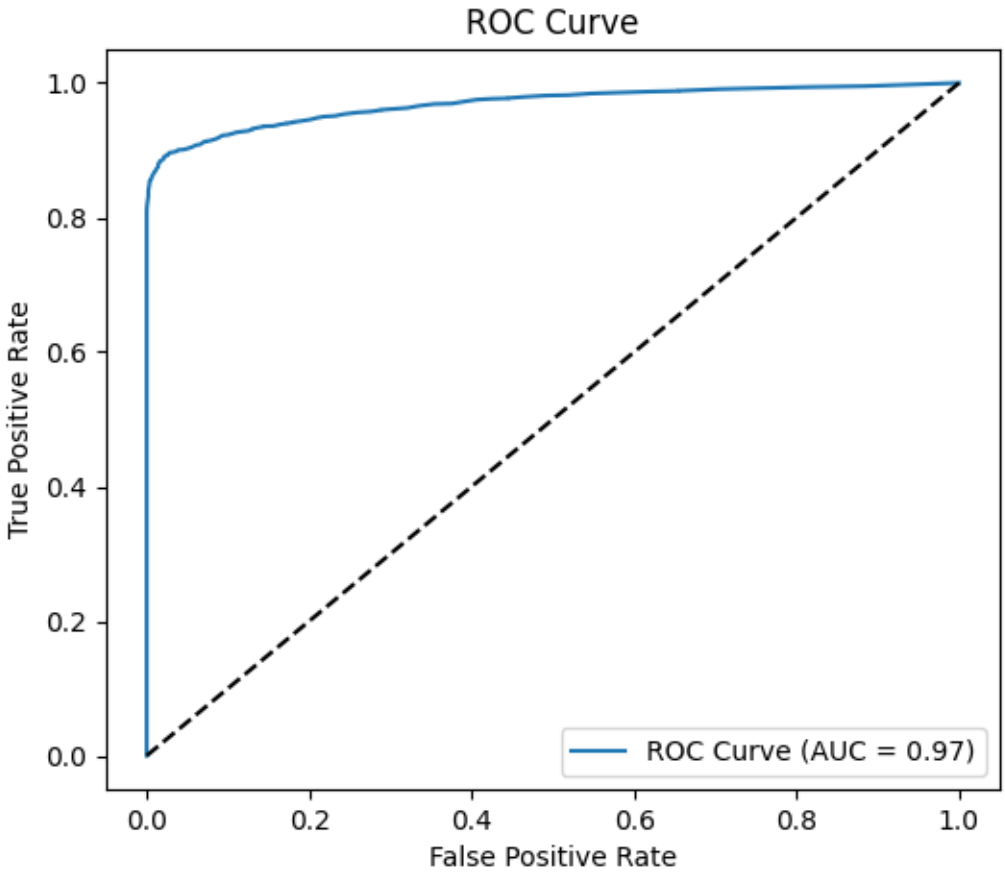
3. Confusion Matrix (Test Set)

A heatmap visualizing True Positives, False Positives, True Negatives, and False Negatives.



4. ROC Curve

Receiver Operating Characteristic (ROC) curve to evaluate the model's classification performance.



Appendices C: Coding Algorithm Implementation

This appendix presents key Python code used to implement the machine learning pipeline for predicting opioid prescribers. The code was developed using Python 3.12.11 in Anaconda IDE, leveraging libraries like pandas, scikit-learn, matplotlib, and seaborn.

1. Data Loading and Initial Analysis

```
import pandas as pd

df = pd.read_csv('prescriber-info.csv')  # Primary dataset
overdoses_df = pd.read_csv('overdoses.csv')  # Supplementary
opioids_df = pd.read_csv('opioids.csv')  # Supplementary

print(df.head())  # Preview data
print(df.isnull().sum())  # Check for missing values
```

2. Data Preprocessing and Feature Engineering

```
from sklearn.preprocessing import LabelEncoder

df_processed = df.copy()
df_processed = df_processed.drop('Id', axis=1)

categorical_features = ['Gender', 'State', 'Credentials', 'Specialty']
for col in categorical_features:
    le = LabelEncoder()
    df_processed[col] = df_processed[col].astype(str).str.upper().str.strip()
    df_processed[col] = le.fit_transform(df_processed[col])
```

3. Train-Test Split

```
from sklearn.model_selection import train_test_split

X = df_processed.drop('Opioid.Prescriber', axis=1)
y = df_processed['Opioid.Prescriber']

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y)
```

4. Model Training (Random Forest)

```
from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

RandomForestClassifier

▼Parameters

n_estimators	100
criterion	'gini'
max_depth	None
min_samples_split	2

5. Evaluation Metrics

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix

y_pred = model.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
print("Precision:", precision_score(y_test, y_pred))
print("Recall:", recall_score(y_test, y_pred))
print("F1 Score:", f1_score(y_test, y_pred))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
```

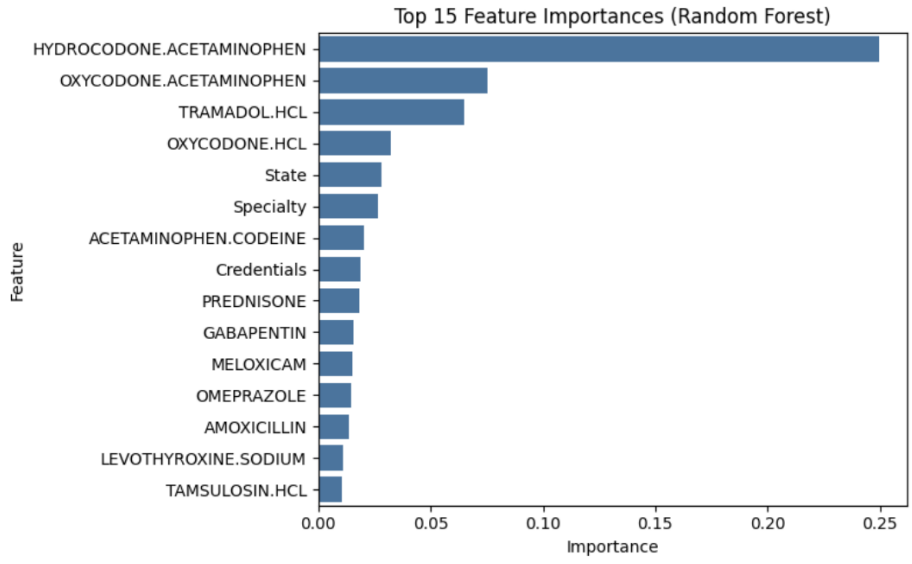
```
Accuracy: 0.9204
Precision: 0.9578226387887527
Recall: 0.904356705241661
F1 Score: 0.9303221288515406
Confusion Matrix:
 [[1945  117]
 [ 281 2657]]
```

6. Save Model & Visualize

```
import seaborn as sns

feature_importance = pd.DataFrame({
'Feature': X_test.columns,
'Importance': model.feature_importances_
}).sort_values(by="Importance", ascending=False)

plt.figure(figsize=(8, 5))
sns.barplot(x="Importance", y="Feature", data=feature_importance.head(15))
plt.title("Top 15 Feature Importances (Random Forest)")
plt.tight_layout()
plt.show()
```



References

[1] Centers for Disease Control and Prevention (CDC), Drug Overdose Deaths. Available at: https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm Accessed: July 6, 2025.

[2] Kaggle Dataset – U.S. Opioid Prescriptions Dataset, CDC Linked. Available at: https://www.kaggle.com/datasets/gaurav2020us/opioid-overdose-deaths

[3] IEEE, "Deep Learning for Opioid Risk Prediction Using Multimodal EHR Data,"
IEEE Transactions on Big Data, 2023. DOI: 10.1109/TBDATA.2023.10207592

[4] Arti Gupta, Maneesh Shreevastava, "Overdose Risk Prediction using Neural Networks,"
IEEE Conference Proceedings, 2021.

[5] Shraddha Subhash Shirsath & Prof. Shubhangi Patil, "CNN-MDRP for Drug Prediction,"
IEEE Conference Proceedings, 2019.

[6] Project Implementation by Aditya Mishra (Roll No. 250260825001) and Vidyuth Ramachandran (Roll No. 250260825014), Opioid Prescriber Prediction using Machine Learning — DBDA February 2025 Batch.