AMITY UNIVERSITY ONLINE, NOIDA, UTTAR PRADESH

In partial fulfilment of the requirement for the award of degree of **Master of Computer Applications (Discipline – Machine Learning.)**

TITLE: **Drug Side Effects Classification using Machine Learning**

Guide Det:

Name: Saqib Sarwar

Designation: PhD Scholar

Submitted By:

Name of the Student- Md. Azeem

Enrolment. No: A9929722000768

ABSTRACT

The project aims to develop a classification model that uses patient data to identify the potential side effects of a given drug. This is an important effort as it has the potential to revolutionize personalized medicine, empowering healthcare providers to predict adverse reactions in patients and adjust treatment plans accordingly.

The foundation of this project is the patient demographic data and drug usage ratings. The dataset is a rich source of information, encompassing a wide range of demographic factors such as age, gender, race, and condition. This dataset provides a robust basis for our analysis, allowing us to explore the complex relationships between these variables and the occurrence of drug side effects.

To accurately predict the probability of side effects, we employ a diverse array of machine learning methods. These include traditional statistical models, decision trees, and advanced techniques such as neural networks and support vector machines. Each of these methods has its own advantages and disadvantages, and by using a diverse set, we aim to leverage their collective advantages to achieve the highest possible prediction accuracy.

Our analysis involves rigorous model selection and validation processes. We partitioned the dataset into training and testing dataset. The models are trained using the training dataset and then evaluated using the testing dataset. Evaluation metrics like accuracy, precision, recall and F1 score are computed to determine how well the model performs in predicting side effects. Additionally, techniques such as cross-validation and bootstrapping are also carried out to ensure that our results are robust and generalizable to new data.

One of the key outcomes of the project is the identification of the most effective models for predicting drug side effects. This is not merely an academic exercise; the models we identify could be directly applied in clinical settings, helping doctors make informed decisions about drug prescriptions. Moreover, these models could be integrated into decision support systems, providing real-time predictions of side effects and assisting in the patient health management.

Beyond model selection, our analysis also provides valuable insights into demographic trends in reactions to drugs. By examining the feature importance of our models, we can identify which demographic factors matters most in predicting the side effects. This could reveal patterns such as higher susceptibility to certain side effects in specific age groups or differences in reactions between genders and race. These insights could inform public health initiatives, guiding the development of targeted interventions to reduce the incidence of adverse drug reactions.

In conclusion, this project marks a significant advancement in the field of personalized medicine. By developing accurate models for predicting drug side effects and uncovering demographic trends in drug reactions, we can contribute to safer and more effective drug use. Further the tuning of the model with in-depth data analysis is crucial for tremendous impact on deriving more insights crucial to the care of the patient.

Keywords: drug side effects, patient demographic data, drug usage ratings, personalized medicine, machine learning models, logistic regression, decision trees, random forest, support vector machines, model selection, model validation, accuracy, precision, recall, F1 score, cross-validation, decision support systems, demographic trends, feature importance, adverse drug reactions, safer drug use, effective drug use

DECLARATION

I, Md. Azeem, a student pursuing MCA 4th Semester at Amity University Online, hereby declare that the project work entitled "Drug Side Effects Classification using Machine Learning" has been prepared by me during the academic year 2022-24 under the guidance of Saqib Sarwar, Computer Science and Engineering, Indian Institute of Technology, Kanpur. I assert that this project is a piece of original bona-fide work done by me. It is the outcome of my own effort and that it has not been submitted to any other university for the award of any degree.

Md. Azeem

Signature of Student

CERTIFICATE

This is to certify that Md. Azeem of Amity University Online has carried out the project work presented in this project report entitled “Drug Side Effects Classification using Machine Learning” for the award of Master of Computer Applications in Machine Learning under my guidance. The project report embodies results of original work, and studies are carried out by the student himself/herself. Certified further, that to the best of my knowledge the work reported herein does not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

(Saqib Sarwar)
(PhD Scholar)

TABLE OF CONTENTS

<CHAPTER 1: INTRODUCTION TO THE TOPIC>

1.1 Background

Pharmaceutical developments revolutionized the treatment of diseases and improved life quality for millions of people around the world. But one problem is that the medications have many side effects which are very harmful for our bodies and can affect our immune system. it's crucial to understand those side-effects and try to predict them, as part of the drug development process so we can develop safer drugs and take better care of patients.

Recent advancements in huge data and machine learning (ML) have provided healthcare professionals with the necessary tools they need to analyze huge amounts of medical data. Machine-learning and artificial-intelligence systems can potentially reveal these patterns in an automated, more complex way than a human could likely identify on their own. Through the use of machine learning, we have the capability to develop models that can predict the chance of side effects based on a patient's narrative, considering a multitude of factors including demographics, medical history, and unique drug characteristics.

1.2 Problem Statement

The variability in individual responses to medications poses a significant challenge in prescribing the most appropriate drugs. Side effects not only affect patient compliance and satisfaction but can also lead to serious health complications. Despite extensive clinical trials, predicting side effects remains a complex task due to the diversity of patient populations and the multifaceted nature of drug interactions.

This project aims to address this challenge by building a classification model to predict the side effects of a particular drug using patient review data. By analyzing a dataset containing patient demographics, drug usage, and ratings, we aim to develop a reliable predictive model. The insights obtained from this model can help healthcare providers make informed decisions and personalize treatment plans to minimize adverse effects.

1.3 Objectives

The primary objectives of this project are:

- **Data Collection and Preprocessing**: Collect and clean a detailed dataset including patient demographics, drug ratings, and reviews. This step involves sourcing data from reputable healthcare databases and ensuring its quality by addressing missing values, eliminating duplicates, and correcting inconsistencies. Effective data preprocessing is crucial for the success of subsequent analyses as it ensures that the dataset is accurate, complete, and ready for machine learning (ML) model development. This step may also include feature engineering to create new, meaningful variables from the existing data.
- **Exploratory Data Analysis (EDA)**: Analyze the data to find patterns and relationships among key variables. EDA involves using statistical techniques and visualization tools to gain insights into the dataset's structure and distributions. This step helps in identifying trends, anomalies, and correlations that could influence the occurrence of drug side effects. It includes the use of descriptive statistics, histograms, box plots, scatter plots, and heatmaps to summarize and visualize data characteristics, thus guiding the feature selection process for model development.
- **Model Development**: Create and compare various machine learning models to predict drug side effects. This objective involves selecting a range of different machine learning algorithms, such as logistic regression, decision trees, random forests, support vector machines (SVM), and K-nearest neighbor (KNN). The models are developed using training data and are designed to identify patterns and make predictions about the likelihood of drug side effects based on patient demographics and drug usage data. The diversity of models allows for a comprehensive comparison to find out the best-performing algorithm.

- **Model Evaluation**: Assess the models based on accuracy, precision, recall, and F1 score. Model evaluation is essential to determine the effectiveness and reliability of each ML model. This involves partitioning the dataset into training and testing subsets and using performance metrics to evaluate the models' predictive capabilities. Accuracy measures the overall correctness of the model, precision indicates the proportion of true positive predictions among all positive predictions, recall measures the ability to identify true positives, and F1 score provides a balance between precision and recall.
- **Insights and Recommendations**: Provide practical insights and advice for healthcare providers based on the analysis. This final objective focuses on translating the findings from the ML models and EDA into actionable recommendations for clinical practice. These insights can help healthcare providers make informed decisions about drug prescriptions, identify high-risk patients, and develop personalized treatment plans. Additionally, the study's results could inform public health strategies and policy-making to enhance patient safety and improve medication adherence.

1.4 Significance of the Study

This study has significant implications for the field of personalized medicine. By accurately predicting drug side effects, we can enhance patient safety, improve adherence to medication regimens, and ultimately achieve better health outcomes. The findings of this study will contribute to the growing body of knowledge on the application of ML in healthcare and provide a detailed framework for future research in this area. The insights gained from this research can help in the development of decision-support tools that aid healthcare providers in personalizing treatment plans, thus minimizing the risk of adverse effects and improving overall treatment efficacy.

1.5 Scope and Limitations

1.5.1 Scope

The scope of this study is defined by several key components, each critical to achieving the project's objectives.

- **Dataset:** The foundation of this research is the dataset named **drugdata.csv**. The study uses a dataset named drugdata.csv, which includes various features such as patient age, gender, race, condition, drug name, and ratings (ease of use, effectiveness, satisfaction). This rich dataset allows for a thorough analysis and modeling, providing the necessary data to train and validate various machine learning models aimed at predicting drug side effects.
- **Exploratory Data Analysis (EDA):** Using statistical and visualization techniques to discover patterns and insights in the data.
- **Machine Learning Models:** To explore the effectiveness of different machine learning approaches, several ML models will be developed and compared, including logistic regression, decision tree, random forest, support vector machine (SVM), k-nearest neighbors (KNN), naive Bayes, and gradient boosting.
- **Evaluation Metrics:** The models will be evaluated using metrics such as accuracy, precision, recall, and F1-score to determine how well they perform.

1.5.2 Limitations

While this study is comprehensive in its scope, there are certain limitations that must be acknowledged:

- **Data Quality:** The accuracy and reliability of the machine learning models depends on the quality and completeness of the dataset. Missing values, inconsistencies, or imbalanced classes can adversely affect the model's performance. And data imbalances can lead to biased models that do not generalize well to unseen data.

- **Generalizability:** Since the dataset is not representative of the entire set of patient demographics and condition variability, the findings of the study may not be considered generalizable. Therefore, the conclusions drawn from this research may be specific to the sample population represented in the dataset and may not extend to broader or different populations.
- **Complexity of Drug Interactions:** The study focuses on predicting side effects for a specific drug, assuming patients are using only one medication. However, patients often take multiple medications simultaneously, leading to complex interactions that can influence side effects. These drug interactions are not accounted for in this analysis, which could limit the model's applicability in real-world scenarios where polypharmacy (the concurrent use of multiple medications) is common.

<CHAPTER 2. REVIEW OF LITERATURE>

2.1 Overview of Drug Side Effects

Adverse drug reactions (ADRs) are unintentional harmful and undesired reactions to medications that occurs at doses normally used. The effects of ADRs can range from mild, barely noticeable symptoms to severe effects that can significantly impair a patient's health and quality of life. Understanding the factors that cause these side effects will play a major role in improving the safety and effectiveness of drug therapy.

Adverse drug reactions (ADRs) are broadly classified into two major categories:

Type A (augmented) and Type B (bizarre).

Type A reactions are dose-dependent and predictable based on the pharmacological properties of the drug. They are usually related to the drug's primary or secondary pharmacodynamic effects and can often be mitigated by adjusting the dosage. Common examples include hypotension from antihypertensive medications or hypoglycemia from insulin or other antidiabetic drugs.

Type B reactions are not dose-dependent and are unpredictable, are often related to the individual patient's response and can be triggered by allergic reactions or genetic predispositions. Examples include anaphylaxis from penicillin or the rare but severe liver toxicity from drugs like isoniazid. Unlike Type A reactions, Type B reactions are idiosyncratic and can be challenging to predict and prevent.

The significance of ADRs in clinical practice cannot be overstated. ADRs are a leading cause of morbidity and mortality worldwide, contributing to extended hospital stays, increased healthcare costs, and a substantial burden on healthcare systems. Therefore, understanding

and predicting ADRs is paramount for enhancing patient safety and optimizing therapeutic outcomes.

2.2 Machine Learning in Healthcare

The applications of machine learning in healthcare have gained attention in recent years due to its transformative capabilities and the potential to improve clinical outcomes. ML techniques can analyze vast and complex datasets to identify patterns, predict outcomes, and assist in decision-making. In the context of drug side effects, ML models can predict the likelihood of adverse reactions based on patient data, enabling personalized treatment plans.

An ML algorithm is increasingly being used and can be implemented in various healthcare scenarios, such as disease diagnosis, prognosis, and treatment optimization. Algorithms like logistic regression, decision trees, random forests, Support Vector Machines (SVM), and k-nearest neighbors (KNN) can handle high-dimensional datasets and uncover hidden relationships that traditional statistical methods may miss. This ability to learn from data makes them powerful tools for tackling a wide range of healthcare challenges.

The application of these ML techniques in healthcare extends beyond predicting ADRs. They are used in disease diagnosis, prognosis, treatment optimization, and personalized medication. By leveraging the potential of ML, healthcare providers are able to make more informed decisions, which ultimately leads to improved patient health. The integration of ML in healthcare, particularly in predicting drug side effects, holds substantial promise. By analyzing patient data and predicting adverse reactions, ML models can enhance patient safety, optimize treatment plans, and contribute to the advancement of personalized medicine. As ML techniques continue to evolve, their application in healthcare will likely expand, offering new opportunities for improving clinical practice and patient care.

2.3 Previous Studies on Drug Side Effects

2.3.1 Study 1: Predicting Adverse Side Effects of Drugs.

The study conducted by Liang-Chin Huang, Xiaogang Wu, and Jake Y Chen. (2011) focuses on integrating multiple types of biological data to predict adverse drug reactions (ADRs). The researchers aimed to improve the detection and prediction of ADRs by using diverse data types such as genes, RNA transcripts, proteins, and metabolites. This multi-faceted approach allows for a comprehensive understanding of the biological processes involved in ADRs.

Their study demonstrates the effectiveness of combining various biological data types for predicting ADRs. By integrating different types of biological information and applying advanced computational methods, the researchers developed a comprehensive model that offers valuable insights into the biological mechanisms of ADRs. This method not only improves the accuracy of ADR predictions but also paves the way for personalized medicine and enhanced drug safety.

2.3.2 Study 2: ML for Personalized Medicine

A study published in *Frontiers in Toxicology* emphasized the role of ML in personalized medicine, focusing on predicting drug toxicity and individual responses to medications. The researchers used a combination of chemoinformatics, bioinformatics, and structure-toxicity relationship modeling to assess drug properties and their potential side effects. The study highlighted the importance of incorporating diverse patient data, including genetic information, to improve the prediction accuracy. The study concluded that ML has great potential to optimize evidence-based medicine and generate data-driven, personalized treatment strategies.

To achieve this, the study used genetic data in conjunction with conventional clinical information to create models that predicted individual patients' response to particular drugs based on their specific genetic variants. The findings suggested that personalized medicine approaches could reduce the incidence of ADRs and improve treatment efficacy.

2.3.3 Study 3: Deep Learning for Drug Safety

Another study focused on using deep learning techniques to evaluate drug safety. Their research utilized neural networks to process extensive datasets containing patient records and drug information. The deep learning models were able to learn complex patterns and interactions between drugs and patient characteristics, leading to more accurate predictions of ADRs.

Deep learning, a subset of ML, involves training artificial neural networks with multiple layers to learn hierarchical representations of data. These models can capture non-linear relationships and interactions, making them suitable for complex tasks such as drug safety assessment.

2.3.4 Study 4: Natural Language Processing (NLP) for ADR Detection

In a study on NLP, researchers explored the application of natural language processing to extract information on ADRs from unstructured patient reviews and clinical notes. NLP techniques were used to preprocess the text data, identify relevant terms, and classify the sentiment of reviews and predict ADRs. This study demonstrated that NLP could complement traditional structured data in predicting ADRs.

NLP empowers the extraction of significant data from unstructured content, which constitutes a significant portion of medical data. By combining NLP with ML models, researchers can use both structured and unstructured data to enhance ADR prediction.

2.3.5 Study 5: Integrating Pharmacovigilance Data for Enhanced ADR Detection

This study examined the combination of pharmacovigilance data with ML models to improve the detection of adverse drug reactions (ADRs). The research utilized data from various sources, including clinical trials, post-market surveillance, and patient health records. By incorporating diverse data types, the models developed in this study could detect ADRs more comprehensively. The research highlighted that integrating data from different phases of drug usage and patient experiences could lead to earlier detection and better management of drug side effects. The study concluded that a holistic approach to data integration is crucial for advancing pharmacovigilance practices and enhancing patient safety.

These studies collectively demonstrate the significant potential of ML in predicting drug side effects and improving patient care. They highlight the importance of using diverse data sources, advanced ML techniques, and the combination of structured and unstructured data to achieve accurate and comprehensive ADR predictions.

2.4 Gaps in Existing Research

While previous studies have demonstrated the capabilities of ML in predicting drug side effects, several gaps remain:

1. **Limited Dataset Diversity:** Many studies use datasets that may not represent the entire population, leading to biased predictions.
2. **Complex Drug Interactions:** Most studies focus on single drugs, whereas patients often take multiple medications, resulting in complex interactions.
3. **Lack of Real-world Data:** Clinical trial data used in studies may not reflect real-world scenarios, where patient adherence and environmental factors play a significant role.

4. **Explainability of Models:** Machine Learning models, especially deep learning, are often criticized for being "black boxes" that lack interpretability, making it difficult for clinicians to trust and rely on their predictions.

2.5 Contribution of This Study

This study aims to address few gaps identified in the existing literature by using a diverse dataset and exploring various machine learning (ML) models. A primary contribution of this study is its focus on real-world patient reviews and ratings, offering insights that are highly relevant to everyday clinical practice. Firstly, it addresses existing gaps in the literature by integrating both structured data (including patient demographics and drug usage information) and unstructured data (like patient reviews). This comprehensive approach allows for a more robust analysis of factors contributing to drug side effects, improving the accuracy and applicability of the predictive models in real-world clinical settings. The findings will contribute to the growing body of knowledge on the application of ML in predicting drug side effects and improving personalized medicine.

Furthermore, this study emphasizes the importance of model interpretability by selecting models that balance accuracy with explainability. This approach ensures that the predictions made by the models are not only accurate and precise but also understandable to healthcare professionals, facilitating their adoption in clinical practise.

Lastly, the practical implications for healthcare providers are significant. The predictive models developed can serve as valuable decision-support tools, helping in the identification of high-risk patients and the customization of treatment plans to minimize adverse reactions. Additionally, insights from patient reviews can enhance patient education and counseling, improving communication about medication risks and benefits.

CHAPTER 3. RESEARCH OBJECTIVES AND METHODLOGY

i. **Research Problem**: The core problem addressed by this research is the significant variability in individual responses to medications, which presents a major challenge in predicting and managing drug side effects. Despite rigorous clinical trials, patients often experience adverse drug reactions (ADRs) that are unpredictable and vary widely among individuals. This variability is influenced by numerous factors, which includes age, gender, genetic predisposition, and other demographic characteristics. Therefore, the research aims to develop a robust machine learning (ML) model that can predict drug side effects by analyzing patient demographics, drug usage patterns, and user ratings. Such a model has the potential to significantly improve personalized medication by providing healthcare professionals with tools to customize treatments for individual patients, thereby enhancing safety and effictiveness.

ii. **Research Design**: This study adopts a quantitative research design, utilizing statistical and machine learning methods to analyze patient data and predict drug side effects. The research follows a systematic approach that includes data collection, preprocessing, exploratory data analysis, model development, evaluation, and providing insights and recommendations. This structured methodology ensures that the research is comprehensive and that the findings are reliable and relevant to real-world scenarios.

iii. **Type of Data Used**: The data used for this study includes both structured (such as age, gender, and ratings) and unstructured data (like patient reviews). This combination allows a comprehensive analysis of the factors influencing drug side effects.

iv. **Data Collection Method**: The data for this project was collected from secondary sources, specifically from online platforms like Kaggle, known for its reputable datasets. Secondary data collection is advantageous as it provides access to large volumes of data that are often not feasible to gather through primary research methods.

v. **Data Collection Instrument**: The primary instrument for data collection is the drugdata.csv dataset. This dataset is robust, containing thousands of records that encompass a wide range of variables relevant to the study. These variables include patient demographics, drug names, conditions being treated, reviews and various ratings provided by the patients.

vi. **Sample Size**: The dataset consists of thousands of records, providing a robust foundation for statistical analysis and ML model development. A large sample size increases the reliability of the findings and the generalizability of the models developed.

vii. **Sampling Technique**: The dataset includes a comprehensive collection of patient reviews and ratings across a broad spectrum of drugs and conditions. This ensures that the data is diverse and representative, covering a wide range of patient experiences and drug interactions.

viii. **Data Analysis Tool**: Several tools and libraries utilized for data analysis and model development are:

- **Python:** The primary programming language used for data analysis and model development. Python offers a rich ecosystem of libraries and frameworks for machine learning tasks.
- **Pandas:** A powerful Python library for data manipulation and analysis. Pandas provides powerful data structures for handling tabular data.
- **NumPy:** A Python library for numerical computations. NumPy supports efficient array operations and a range of mathematical functions.
- **Scikit-learn:** A Python library for ML, providing robust tools for model development and evaluation. Scikit-learn offers a comprehensive suite of algorithms and utilities for ML tasks.

- **Matplotlib and Seaborn:** Python libraries used for data visualization. Matplotlib and Seaborn facilitate the creation of a wide range of static, animated, and interactive plots.
- **Plotly:** A Python library for interactive data visualization. Plotly allows the creation of dynamic and interactive charts and dashboards.
- **Jupyter Notebook:** An interactive development environment for writing and running Python code. Jupyter Notebooks support exploratory data analysis and documentation.

3.1 Research Objectives

The objectives of this study are as follows:

- **Data Collection:** Collecting a comprehensive dataset that includes patient demographics, drug usage, and ratings, ensuring its completeness and accuracy.
- **Data Preprocessing:** Cleaning and preprocessing the data to handle missing values, outliers and create new features that might enhance the model's ability to make accurate predictions. Effective preprocessing ensures that the dataset is ready for analysis.
- **Exploratory Data Analysis (EDA):** Performing EDA to understand the distribution of important variables and identify patterns and explore relationships among the variables. This step helps in feature selection and provides insights into the data structure. It includes the use of descriptive statistics, histograms, box plots, scatter plots, and heatmaps to summarize and visualize data characteristics, thus guiding the feature selection process for model development.
- **Model Development:** Developing and comparing the performance of different machine learning models in predicting drug side effects. The diversity of models allows for a comprehensive comparison to find out the best-performing algorithm.

- **Model Evaluation:** Evaluating the models based on metrics such as accuracy, precision, recall, and F1-score. These metrics help in assessing the models' effectiveness and reliability.
- **Insights and Recommendations:** To Provide actionable insights and recommendations for healthcare providers based on findings. These insights can help healthcare providers make informed decisions about drug prescriptions, identify high-risk patients, and creating customized treatment plans and improving patient health.

This project aims to utilize the power of machine learning to address a significant challenge in healthcare, thereby contributing to the well-being of patients and the effectiveness of medical treatments.

3.2 Methodology

3.2.1 Data Collection

The dataset used in this study, "drugdata.csv", was obtained from a reputable source. It includes various features such as patient age, gender, race, condition, drug name, and ratings (ease of use, effectiveness, satisfaction). The dataset consists of thousands of records, providing a robust foundation for analysis.

The dataset was chosen because it covers a broad spectrum of drugs and conditions, enabling a thorough examination of the factors influencing drug side effects. It contains both structured data (such as age, gender, and ratings) and unstructured data (like patient reviews), allowing for a multifaceted analysis.

3.2.2 Data Preprocessing

Data preprocessing is a critical step in preparing the dataset for effective machine learning (ML) analysis. It involves several key tasks to ensure the data is clean, complete, and structured in a way that ML algorithms can process efficiently.

Data preprocessing involved several steps:

- **Handling Missing Values:** Missing data can significantly affect results and lead to development of inaccurate models. Therefore, missing values were adjusted using imputation techniques or by removing records with significant missing information. For numerical features, mean or median imputation was used. This approach replaces missing values with the average or median value of the respective feature, thereby maintaining the overall distribution of the data. For categorical features, mode imputation was applied, where missing values were replaced with the most frequent category within the feature. If a record had significant missing information that could not be reliably imputed, it was removed from the dataset to avoid introducing bias.
- **Feature Engineering:** Feature engineering involves creating new features from the existing data to better capture the underlying patterns. New features like year, month, and day of the week were generated from date information to capture time-based patterns. Additionally, interaction terms between features were created to explore their combined effects of multiple variables on side effects, improving the model's predictive ability.
- **Data Transformation:** To handle skewed data, such as UsefulCount, was transformed using logarithmic or Box-Cox transformations to stabilize variance and make the data more normally distributed. This step is crucial for improving the performance of ML algorithms that assume normally distributed data.
- **One-Hot Encoding:** Categorical variables such as race and gender were converted into numerical format using one-hot encoding, This technique creates binary columns for each category level, preventing the ML algorithms from assuming any ordinal relationship between categories. One-hot encoding ensures that the model treats each category as a separate and independent entity.

- **Normalization and Standardization:** Normalization and standardization were applied to numerical features to adjust their scales. The features were adjusted to achieve a mean of zero and a standard deviation of one. This process is crucial for algorithms like Support Vector Machines (SVM) and k-nearest neighbors (KNN) that are sensitive to the scale of the data. The standardization of features ensures that each feature contributes equally to the model.

3.2.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to visualize and understand the relationships between different features and the side effects reported. Techniques used include:

- **Descriptive Statistics:** We calculated summary statistics measures such as mean, median, standard deviation, and interquartile range for key variables to understand their central tendency and dispersion of the data, highlighting potential outliers and data distribution characteristics.
- **Histograms and Bar Plots:** Visualizations were generated to understand the distribution of attributes such as satisfaction, effectiveness, and ease of use. Histograms and bar plots were used to visualize the distribution of continuous and categorical variables, respectively. These plots help identify skewness, outliers, and other distributional properties, making it easier to identify patterns and anomalies in the data.
- **Heatmaps:** Correlation heatmaps were plotted to identify relationships between numerical features. Heatmaps visualize the strength and direction of correlations, helping to identify features that are strongly related and those that may contribute redundantly to the model.

- **Box Plots and Violin Plots:** These plots were used to compare numerical feature distributions across different categories. They provide insights into data variability, the presence of outliers and distribution shape, facilitating a deeper understanding of how different groups are characterized by the data.
- **Scatter Plots and Pair Plots:** Both Scatter plots and pair plots were used to explore relationships between pairs of features. These plots help identify potential interactions and non-linear relationships, guiding the selection of features for the ML models.
- **Pie Charts:** Pie charts were generated to show the proportion of reviews collected over different years and the distribution of side effects by gender and race. They provide a visual summary of categorical data, highlighting key demographic trends and usage patterns.

3.2.4 Model Development

Several ML models were developed and compared to predict drug side effects, including:

1. **Logistic Regression:** A simple linear model used for binary classification task. It predicts the probability of a binary outcome based on input features.
2. **Decision Tree:** A non-linear model that splits data based on feature values to make predictions. Decision trees are easily interpretable and can capture complex interactions between features, making them suitable for understanding how different factors contribute to the outcome.
3. **Random Forest:** An ensemble of decision trees that improves accuracy by averaging the predictions of multiple trees. It reduces overfitting and enhances generalization by combining the predictions of multiple trees.
4. **Support Vector Machine (SVM):** A model that identifies the optimal hyperplane to separate classes. SVMs are effective for high-dimensional data and can handle

non-l inear relationships using kernel functions. SVMs are particularly useful when there are clear margins separating the classes in the data.

5. **K-Nearest Neighbors (KNN):** A model that predicts class on the basis of majority class of nearest neighbors. KNN is intuitive and works well for small datasets with clear class boundaries. However, KNN can be computationally expensive for large datasets.
6. **Naive Bayes:** A probabilistic classifier model based on the principle of Bayes' theorem. Naive Bayes assumes feature independence, making it efficient and suitable for high-dimensional data. Naive Bayes is particularly effective for text classification tasks.
7. **Gradient Boosting:** An ensemble model that sequentially corrects errors of previous trees. It combines weak learners to form a strong predictive model, achieving high accuracy and robustness.

Each model was fine-tuned using hyperparameter optimization techniques such as grid search and randomized search. Cross-validation was employed to ensure robust evaluation and prevent overfitting.

3.2.5 Model Evaluation

The models were evaluated using the following key metrics:

- **Accuracy:** The ratio of correct predictions out of the total predictions made. Accuracy provides an overall assessment of model performance, indicating how often the model is correct. Accuracy is a critical metric to evaluate the performance of predictive models.
- **Precision:** The ratio of true positive predictions out of all positive predictions made. Precision indicates the model's ability to avoid false positives.
- **Recall:** The ratio of true positive predictions out of all actual positive instances. Recall measures the model's ability to identify true positives.
- **F1-Score:** The harmonic mean of precision and recall, providing a balance measure between the two metrics. The F1-score is particularly useful for evaluating models on imbalanced datasets where both precision and recall matters.

Additional evaluation metrics included:

- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** AUC-ROC evaluates the model's ability to differentiate between classes across different threshold settings. A higher AUC indicates better model performance in distinguishing between positive and negative classes.
- **Confusion Matrix:** The confusion matrix gives an in-depth breakdown of true positives, true negatives, false positives, and false negatives. It helps analyze the classification performance, by showing the counts of correct and incorrect predictions for each class.

These metrics collectively provide a comprehensive evaluation of the models, ensuring that the best-performing model is selected based on multiple criteria relevant to the prediction of drug side effects.

CHAPTER 4. DATA ANALYSIS, RESULTS, AND INTERPRETATION

4.1 Data Analysis

4.1.1 Descriptive Statistics

Descriptive statistics provide an overview of the dataset, highlighting key characteristics like central tendency, dispersion, and overall shape of the data distribution. The following table shows the most common values for each feature in the dataset:

Feature	**Most Common Value**
Name	Michel
Race	White
Age	55-64
Condition	High Blood Pressure
Date	01-05-2009
Drug	Lisinopril
DrugId	6873.0
EaseofUse	5.0
Effectiveness	4.0
Reviews	Dizziness
Satisfaction	1.0
Gender	Female
Sides	Dizziness, lightheadedness, tiredness
UsefulCount	3.0
Year	2009.0
Month	2.0
Day of Week	Tuesday

Additionally, descriptive statistics encompass variability measures such as range, variance, and standard deviation. For example, the ‘age’ feature may have a range from 0-17 to 75 or over, indicating a diverse age distribution among the patients in the dataset.

In addition to these basic statistics, the dataset also provides an overview of the following:

1. **Unique Values in Columns**: This shows the number of distinct values within each feature, indicating the diversity and variability present in the dataset.
2. **Top 15 Recommended Drugs for Each Condition**: This details the most frequently recommended drugs for various medical conditions, highlighting the top choices for treatment as observed in the dataset.
3. **Top 20 Drugs That Are Most Effective and Their Corresponding Conditions**: This lists the drugs rated highest for effectiveness and the conditions they are most commonly used to treat, providing insights into the efficacy of different medications.
4. **Top 20 Drugs Based on the Number of Users**: This highlights the drugs with the highest number of users, showcasing the most popular medications used in healthcare.

These additional descriptive statistics are listed under Appendix A. They offer a more detailed understanding of the dataset's composition, the effectiveness of treatments, and the popularity of various drugs among users.

4.1.2 Exploratory Data Analysis (EDA)

4.1.2.1 Satisfaction Rate and Ease of Use

A histogram was created to visualize and understand the relationship between satisfaction and ease of use for the drug Lisinopril. The histogram revealed that as the ease of use increases, the satisfaction ratings also tends to increase. However, there are exceptions to this trend, indicating that other factors may influence satisfaction.

The histogram revealed that the majority of ratings for ease of use were higher (4-5) range, corresponding with high satisfaction ratings. Yet, some users with high ease-of-use ratings reported low satisfaction, implying that ease of use alone does not fully determine patient satisfaction.

Factors such as drug effectiveness and the presence of side effects are likely contributing to these mixed satisfaction ratings. As notable number of instances with low satisfaction despite high ease of use suggests that other factors such as effectiveness and side effects may play a significant role in patient satisfaction.

4.1.2.2 Effectiveness Distribution

The box plot created indicates a median drug effectiveness of 3, ranging from 1 to 5. And the pie chart reveals that ratings of '4' (25.1%) and '5' (24.1%) are most common, while '2' is least common (10.5%). The drug is fairly effective with around 1600 (25.1%) users giving a rating of 4 to it. Also, the bar chart and line graph were plotted, which reveals a bimodal distribution of effectiveness levels, with the highest count just above 1500.

The data suggests a diverse range of effectiveness, slightly skewed towards higher effectiveness. The data suggests a diverse range of effectiveness, slightly skewed to the left, indicating that items with lower effectiveness are less frequent than those with higher effectiveness.

4.1.2.3 Satisfaction vs Effectiveness

From the distribution of the plotted scatter plot, it appears that most of the data points are clustered around the lower left side of the plot, indicating lower satisfaction and effectiveness. This suggests that for most cases in this dataset, as satisfaction increases, effectiveness also increases.

However, there is one point that stands out in the upper right corner, indicating higher satisfaction and effectiveness. This outlier suggests that there may be at least one instance where both satisfaction and effectiveness are high compared to other observations in the dataset.

4.1.2.4 Gender Distribution

The dataset showed a higher number of female users compared to male users. The following table summarizes the gender distribution:

Gender	**Count**
Female	4060
Male	2466

The gender distribution indicates that females are more likely to provide reviews for drugs, which may reflect differences in health-seeking behavior and drug usage patterns between genders. Women might be more proactive in managing their health and more willing to share their experiences with medications.

4.1.2.5 Race Distribution

The race distribution of users revealed that White individuals were the most predominant users, followed by Hispanic, Black, and Asian individuals. The table below summarizes the race distribution:

Race	**Count**
White	2058
Hispanic	1801
Black	1455
Asian	1212

The race distribution provides reveals the demographic composition of the dataset, which is important for understanding the generalizability of the findings. It also highlights potential disparities in drug usage and reporting among different racial groups, which could be due to various socio-economic and cultural factors.

4.1.2.6 Gender Distribution by Race

A stacked bar plot was created to visualize the gender distribution within each race community. The plot indicated a higher prevalence of female users across all races, with the highest user count observed in the White community.

The stacked bar plot revealed that the proportion of female users is consistently higher across all racial groups, suggesting that the trend of higher female representation in drug reviews is not specific to any single race. This uniformity may reflect broader gender differences in healthcare utilization and engagement in health-related discussions.

4.1.2.7 Gender Distribution by Age Group

The data indicated that female users dominate across all age groups, except in the ‘0-17’ age group where male users were slightly more in number.

This finding may reflect differences in healthcare utilization across age groups, with older females being more likely to seek medical treatment and report their experiences compared to younger individuals.

4.1.2.8 Reviews Collected Over the Years

An analysis of the reviews collected over the years revealed an increasing trend in the number of reviews until 2014, followed by a slight decline. The years with the most reviews were 2014, followed by 2013 and 2012.

This trend indicates an increasing awareness and use of online platforms for reporting drug experiences during the early 2010s, followed by a stabilization or shift in user behavior in subsequent years.

4.1.2.9 Side Effects by Age Group

The analysis revealed that the 60+ age group reported the most side effects, suggesting that older individuals experience more side effects than other age groups.

Older patients are often managing multiple health conditions and take multiple medications, increasing the risk of side effects. This finding highlights the importance of considering age-related factors and potential drug interactions while prescribing medications to older patients.

4.1.2.10 Side Effects by Gender

The side effects varied between genders, with female users reporting more extreme side effects compared to male users.

This may be due to biological differences, differences in drug metabolism, or differences in reporting behavior between genders. Understanding these differences is crucial for developing personalized medication that minimize side effects.

4.1.2.11 Side Effects by Race

The analysis of side effects by race revealed that White race individuals reported the most extreme side effects, followed by the Hispanic race.

Genetic differences, socio-economic status, and healthcare access factors contribute to racial disparities in drug response. Identifying these disparities can help in customizing treatment to minimize adverse effects across different racial communities.

4.1.2.12 Effectiveness among Gender

The bar plot analysis revealed that female users generally reported higher drug effectiveness compared to male users. Approximately 800 females rated the drug as least

effective (rating 1.0), while around 400 males gave the same rating. At ratings 2.0, 3.0, and 5.0, more females than males reported positive effectiveness, with 450, 900, and 930 females respectively. For males, the corresponding numbers were 200, 550, and 700. At rating 4.0, slightly more females (1000) than males (650) found the drug effective. This indicates that females generally perceive the drug as more effective than males do.

Overall, responses are mixed. A lower effectiveness rating could potentially signal the presence of side effects. Understanding the distribution of drug effectiveness can help researchers identify areas for improvement, potential risks, and the impact of variables like gender on drug response.

4.1.2.13 Side Effects for Drug

The distribution of these ratings of pie chart provides valuable insights into the side effects profile of the drug. 42% of users reported extreme side effects, with 1 being no side effect and 5 being an extreme side effect. This information is crucial for healthcare professionals and patients to understand the likelihood and severity of potential side effects when considering any drug for treatment.

4.1.3 Data Transformation

Several transformations were implemented on the dataset to improve model performance:

- **Logarithmic Transformation:** Applied to the UsefulCount variable to handle skewness. Log transformation helps in reducing the impact of extreme values and normalizes the distribution, making the data more suitable for ML algorithms.
- **Box-Cox Transformation:** Used to stabilize variance and make the data more normally distributed. Box-Cox transformation is particularly useful for features that do not follow a normal distribution, improving the performance of algorithms that assume normality.

- **Standard Scaling:** Applied to numerical variables such as Effectiveness to standardize the data, to achieve a mean of zero and a standard deviation of one. This ensures that all features contribute equally to the model training process, which is essential for algorithms like SVM and KNN that are responsive to the scale of the data.

These transformations ensure that the data is in a form that maximizes the accuracy and reliability of the predictive models developed in this study.

4.2 Model Development and Hyperparameter Tuning

4.2.1 Logistic Regression

Logistic Regression is a linear model used for binary classification tasks. It estimates the probability of a binary outcome by combining input features linearly. It is computationally efficient model and performs well when the correlation between features and the target variable is approximately linear.

4.2.1.1 Hyperparameter Tuning for Logistic Regression

When fine-tuning logistic regression, the key hyperparameter to adjust is the regularization parameter C, which balances achieving a minimal error on the training data and reducing the model complexity. Regularization prevent overfitting by penalizing large coefficients. The following values were tested for C:

- C values: [0.01, 0.1, 1, 10, 100]

Grid search with cross-validation was used to find the optimal value of C. The best model was selected based on the highest cross-validated accuracy.

4.2.2 Decision Tree

Decision Trees are a non-linear model that partitions data based on feature values to generate predictions. It recursively divides the data into subsets, aiming to enhance the purity of the each resulting subsets. Decision trees are easy to interpret and can capture complex interactions between features.

4.2.2.1 Hyperparameter Tuning for Decision Tree

The hyperparameters tuned for the decision tree model included:

- max_depth: Maximum depth of the tree to control overfitting. Tested values: [5, 10, 15, None]
- min_samples_split: Minimum number of samples required to divide an internal node. Tested values: [2, 10, 20]
- min_samples_leaf: Minimum number of samples that must be present at a leaf node. Tested values: [1, 5, 10]
- criterion: Function to measure the quality of a split (e.g., "gini" for the Gini impurity, "entropy" for information gain). Tested values: ["gini", "entropy"]

Grid search with cross-validation was used to find the optimal combination of hyperparameters.

4.2.3 Random Forest

Random Forest is a classification ensemble learning model that improves predictive accuracy and reduces overfitting by constructing multiple decision trees and then combining their predictions through averaging. Each tree is trained on a bootstrap data sample, which is a random subset of the total training data, and a random subset of features is considered for

splitting at each node. This randomness in both data and features contributes to the robustness of the Random Forest model.

4.2.3.1 Hyperparameter Tuning for Random Forest

The hyperparameters tuned for the random forest model included:

- n_estimators: Number of trees in the forest. Tested values: [100, 200, 300]
- max_depth: Maximum depth of the trees. Tested values: [10, 20, None]
- min_samples_split: Minimum number of samples required to split an internal node. Tested values: [2, 5, 10]
- min_samples_leaf: Minimum number of samples required to be at a leaf node. Tested values: [1, 2, 4]
- max_features: Number of features to be considered when looking for the best split. Tested values: ["sqrt", "log2", None]

Randomized search with cross-validation was used to explore a broader range of hyperparameter combinations efficiently.

4.2.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a non-linear model that finds the optimal hyperplane to separate classes by maximizing the margin between the closest points of different classes and then uses it for prediction. SVMs have the capability to manage non-linear relationships utilizing kernel functions such as the polynomial or radial basis function (RBF) kernel. These kernels transform the input data into a higher dimensional space where a linear decision boundary can be identified, thus enabling SVMs to handle complex, non-linear patterns.

4.2.4.1 Hyperparameter Tuning for SVM

The hyperparameters tuned for the SVM model included:

- C: Regularization parameter. Tested values: [0.1, 1, 10, 100]
- kernel: Kernel function. Tested values: ["linear", "poly", "rbf"]
- degree: Degree of the polynomial kernel function (if kernel is "poly"). Tested values: [2, 3, 4]
- gamma: Kernel coefficient (if kernel is "rbf"). Tested values: ["scale", "auto"]

Grid search with cross-validation was used to find the optimal combination of hyperparameters.

4.2.5 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a supervised learning model that predicts the class of a sample based on the majority class of its closest neighbors. The distance between samples is typically measured using Euclidean distance.

4.2.5.1 Hyperparameter Tuning for KNN

The hyperparameters tuned for the KNN model included:

- n_neighbors: Number of neighbors to be considered for classification. Tested values: [3, 5, 7, 9]
- weights: Weight function to be used in prediction. Tested values: ["uniform", "distance"]
- metric: Distance metric to use. Tested values: ["euclidean", "manhattan", "minkowski"]

Grid search with cross-validation was used to find the optimal combination of hyperparameters.

4.2.6 Naive Bayes

Naive Bayes is a probabilistic classifier model based on the principles of Bayes' theorem, assuming feature independence. Despite its simplicity, Naive Bayes are fast and simple to train and is effective for high-dimensional data and often performs well in text classification tasks. This is due to their ability to handle many features and their robustness to irrelevant features.

4.2.6.1 Hyperparameter Tuning for Naive Bayes

For the Naive Bayes model, the primary hyperparameter is the smoothing parameter often denoted as alpha, which prevents zero probabilities when the model encounters feature values that it has not seen during training. This technique is particularly useful in text classification tasks where the vocabulary can be vast and the model might frequently encounter new words in the test data. The value of alpha can be tuned to achieve the best performance. It’s a simple yet effective way to handle the issue of zero probabilities in Naive Bayes.

The following values were tested:

- alpha values: [0.1, 0.5, 1.0, 2.0]

Grid search with cross-validation was used to find the optimal value of alpha.

4.2.7 Gradient Boosting

Gradient Boosting is an ensemble model that sequentially builds trees, with each tree correcting the errors of the previous one. Gradient boosting combines weak learners to form a robust predictive model. This method is particularly effective for dealing with complex datasets and can significantly improve model performance.

4.2.7.1 Hyperparameter Tuning for Gradient Boosting

The hyperparameters tuned for the gradient boosting model included:

- n_estimators: Number of boosting stages. Tested values: [100, 200, 300]
- learning_rate: Step size shrinkage used to prevent overfitting. Tested values: [0.01, 0.1, 0.2]
- max_depth: Maximum depth of the trees. Tested values: [3, 5, 7]
- min_samples_split: Minimum number of samples required to split an internal node. Tested values: [2, 5, 10]
- min_samples_leaf: Minimum number of samples required to be at a leaf node. Tested values: [1, 2, 4]

Randomized search with cross-validation was used to efficiently explore a broad range of hyperparameter combinations.

4.3 Model Evaluation and Comparison

4.3.1 Confusion Matrix for Log Regression Model

	Predicted: No Side Effect	**Predicted: Mild Side Effect**	**Predicted: Moderate Side Effect**	**Predicted: Severe Side Effect**	**Predicted: Extreme Side Effect**
Actual: No Side Effect	237	25	14	0	0
Actual: Mild Side Effect	42	73	87	0	0
Actual: Moderate Side Effect	42	27	170	3	17
Actual: Severe Side Effect	0	0	0	36	174
Actual: Extreme Side Effect	0	0	0	41	644

4.3.2 Classification Report for log Regression Model

Class	**Precision**	**Recall**	**F1-Score**	**Support**
No Side Effect	0.74	0.86	0.79	276
Mild Side Effect	0.58	0.36	0.45	202
Moderate Side Effect	0.63	0.66	0.64	259
Severe Side Effect	0.45	0.17	0.25	210
Extreme Side Effect	0.77	0.94	0.85	685
Accuracy			0.71	1632
Macro Avg	0.63	0.60	0.60	1632
Weighted Avg	0.68	0.71	0.68	1632

The log regression model achieved high precision and recall for predicting extreme side effects, indicating that it effectively identifies the most severe cases. However, the model's performance for predicting mild and moderate side effects is lower, suggesting room for improvement in distinguishing between less severe reactions. The overall accuracy of 71.08% reflects the model's general ability to predict side effects, but it also indicates the necessity for further tuning and optimization to enhance its performance across all categories.

4.3.3 Confusion Matrix for SVM - Polynomial Model

	Predicted: No Side Effect	**Predicted: Mild Side Effect**	**Predicted: Moderate Side Effect**	**Predicted: Severe Side Effect**	**Predicted: Extreme Side Effect**
Actual: No Side Effect	235	29	12	0	0
Actual: Mild Side Effect	28	120	54	0	0
Actual: Moderate Side Effect	31	30	179	4	15
Actual: Severe Side Effect	0	0	0	44	166
Actual: Extreme Side Effect	0	0	0	9	676

4.3.4 Classification Report for SVM - Polynomial Model

Class	Precision	Recall	F1-Score	Support
No Side Effect	0.80	0.85	0.82	276
Mild Side Effect	0.67	0.59	0.63	202
Moderate Side Effect	0.73	0.69	0.71	259
Severe Side Effect	0.77	0.21	0.33	210
Extreme Side Effect	0.79	0.99	0.88	685
Accuracy			0.77	1632
Macro Avg	0.75	0.67	0.67	1632
Weighted Avg	0.76	0.77	0.74	1632

The SVM with a polynomial kernel showed strong performance in predicting no side effects and extreme side effects, achieving high precision and recall for these categories. However, the model's ability to predict severe side effects was limited, as indicated by lower precision and recall. With an overall accuracy of 76.84%, the polynomial SVM is a robust choice for certain classes but may require further tuning to better handle severe side effects.

4.3.5 Confusion Matrix for SVM - RBF Model

	Predicted: No Side Effect	Predicted: Mild Side Effect	Predicted: Moderate Side Effect	Predicted: Severe Side Effect	Predicted: Extreme Side Effect
Actual: No Side Effect	238	29	9	0	0
Actual: Mild Side Effect	33	124	45	0	0
Actual: Moderate Side Effect	39	33	167	6	14
Actual: Severe Side Effect	0	0	0	32	178
Actual: Extreme Side Effect	0	0	0	10	675

4.3.6 Classification Report for SVM - RBF Model

Class	Precision	Recall	F1-Score	Support
No Side Effect	0.77	0.86	0.81	276
Mild Side Effect	0.67	0.61	0.64	202
Moderate Side Effect	0.76	0.64	0.70	259
Severe Side Effect	0.67	0.15	0.25	210
Extreme Side Effect	0.78	0.99	0.87	685
Accuracy			0.76	1632
Macro Avg	0.73	0.65	0.65	1632
Weighted Avg	0.74	0.76	0.72	1632

The SVM with an RBF kernel demonstrated high precision and recall for predicting no side effects and extreme side effects, achieving a strong performance for these categories. However, the model struggled with predicting severe side effects, as evidenced by lower precision and recall. The overall accuracy of 75.74% reflects the model's effectiveness in certain areas, though further optimization is needed to improve predictions for severe side effects.

4.3. 7 Confusion Matrix for KNN Model

	Predicted: No Side Effect	**Predicted: Mild Side Effect**	**Predicted: Moderate Side Effect**	**Predicted: Severe Side Effect**	**Predicted: Extreme Side Effect**
Actual: No Side Effect	248	20	7	0	1
Actual: Mild Side Effect	71	91	30	3	7
Actual: Moderate Side Effect	58	58	111	8	24
Actual: Severe Side Effect	9	4	5	51	141
Actual: Extreme Side Effect	14	3	3	45	620

4.3.8 Classification Report for KNN Model

Class	Precision	Recall	F1-Score	Support
No Side Effect	0.62	0.90	0.73	276
Mild Side Effect	0.52	0.45	0.48	202
Moderate Side Effect	0.71	0.43	0.53	259
Severe Side Effect	0.48	0.24	0.32	210
Extreme Side Effect	0.78	0.91	0.84	685
Accuracy			0.69	1632
Macro Avg	0.62	0.59	0.58	1632
Weighted Avg	0.67	0.69	0.66	1632

The KNN model performed well in predicting no side effects and extreme side effects, achieving high recall for these categories. However, its precision and recall for mild and moderate side effects were lower, indicating challenges in differentiating between these categories. With an overall accuracy of 68.69%, the KNN model shows promise but requires further tuning to enhance its performance for less severe side effects.

4.3.9 Confusion Matrix for Decision Tree Model

	Predicted: No Side Effect	Predicted: Mild Side Effect	Predicted: Moderate Side Effect	Predicted: Severe Side Effect	Predicted: Extreme Side Effect
Actual: No Side Effect	234	21	21	0	0
Actual: Mild Side Effect	28	142	32	0	0
Actual: Moderate Side Effect	16	37	196	8	2
Actual: Severe Side Effect	0	0	7	150	53
Actual: Extreme Side Effect	0	0	8	72	605

4.3.10 Classification Report for Decision Tree Model

Class	Precision	Recall	F1-Score	Support
No Side Effect	0.84	0.85	0.84	276
Mild Side Effect	0.71	0.70	0.71	202
Moderate Side Effect	0.74	0.76	0.75	259
Severe Side Effect	0.65	0.71	0.68	210
Extreme Side Effect	0.92	0.88	0.90	685
Accuracy			0.81	1632
Macro Avg	0.77	0.78	0.78	1632
Weighted Avg	0.82	0.81	0.81	1632

The decision tree model demonstrated strong performance with high precision, recall, and F1-scores across all categories, especially for predicting extreme side effects. The model's ability to accurately identify severe and extreme side effects makes it a valuable tool for healthcare providers. The overall accuracy of 81.31% suggests that the decision tree model is effective in distinguishing between different levels of side effects, although further refinement may still enhance its ability to predict less severe reactions.

4.3.11 Confusion Matrix for Random Forest Model

	Predicted: No Side Effect	**Predicted: Mild Side Effect**	**Predicted: Moderate Side Effect**	**Predicted: Severe Side Effect**	**Predicted: Extreme Side Effect**
Actual: No Side Effect	227	24	25	0	0
Actual: Mild Side Effect	35	126	41	0	0
Actual: Moderate Side Effect	23	32	194	6	4
Actual: Severe Side Effect	0	0	2	113	95
Actual: Extreme Side Effect	0	0	1	37	647

4.3.12 Classification Report for Random Forest Model

Class	Precision	Recall	F1-Score	Support
No Side Effect	0.80	0.82	0.81	276
Mild Side Effect	0.69	0.62	0.66	202
Moderate Side Effect	0.74	0.75	0.74	259
Severe Side Effect	0.72	0.54	0.62	210
Extreme Side Effect	0.87	0.94	0.90	685
Accuracy			0.80	1632
Macro Avg	0.76	0.74	0.75	1632
Weighted Avg	0.79	0.80	0.80	1632

The random forest model performed well, particularly in predicting extreme side effects with high precision and recall. The model's ensemble approach enhances its robustness and generalization, resulting in an overall accuracy of 80.09%. While the model effectively predicts severe and extreme side effects, its performance for mild and moderate side effects shows some variability, indicating potential areas for improvement.

4.3.13 Confusion Matrix for Naive Bayes Model

	Predicted: No Side Effect	Predicted: Mild Side Effect	Predicted: Moderate Side Effect	Predicted: Severe Side Effect	Predicted: Extreme Side Effect
Actual: No Side Effect	37	235	4	0	0
Actual: Mild Side Effect	10	192	0	0	0
Actual: Moderate Side Effect	14	219	6	8	12
Actual: Severe Side Effect	0	0	1	75	134
Actual: Extreme Side Effect	0	0	4	122	559

4.3.14 Classification Report for Naive Bayes Model

Class	Precision	Recall	F1-Score	Support
No Side Effect	0.61	0.13	0.22	276
Mild Side Effect	0.30	0.95	0.45	202
Moderate Side Effect	0.40	0.02	0.04	259
Severe Side Effect	0.37	0.36	0.36	210
Extreme Side Effect	0.79	0.82	0.80	685
Accuracy			0.53	1632
Macro Avg	0.49	0.46	0.38	1632
Weighted Avg	0.58	0.53	0.48	1632

The Naive Bayes model showed high recall for mild side effects, indicating it effectively identifies these cases. However, its precision for no side effects and moderate side effects was low, leading to a significant number of false positives. The overall accuracy of 53.25% suggests that while Naive Bayes can identify certain side effects, it struggles with distinguishing between others, highlighting the need for a more refined approach.

4.3.15 Confusion Matrix for Tuned Random Forest Model

	Predicted: No Side Effect	**Predicted: Mild Side Effect**	**Predicted: Moderate Side Effect**	**Predicted: Severe Side Effect**	**Predicted: Extreme Side Effect**
Actual: No Side Effect	231	20	17	0	8
Actual: Mild Side Effect	64	59	73	0	6
Actual: Moderate Side Effect	69	15	142	2	31
Actual: Severe Side Effect	10	0	16	10	174
Actual: Extreme Side Effect	30	0	13	12	630

4.3.16 Classification Report for Tuned Random Forest Model

Class	Precision	Recall	F1-Score	Support
No Side Effect	0.57	0.84	0.68	276
Mild Side Effect	0.63	0.29	0.40	202
Moderate Side Effect	0.54	0.55	0.55	259
Severe Side Effect	0.42	0.05	0.09	210
Extreme Side Effect	0.74	0.92	0.82	685
Accuracy			0.66	1632
Macro Avg	0.58	0.53	0.51	1632
Weighted Avg	0.63	0.66	0.61	1632

The tuned random forest model showed a balanced performance, with high precision and recall for predicting extreme side effects. However, its performance for severe side effects was limited, as indicated by lower precision and recall. The overall accuracy of 65.69% reflects the model's ability to identify extreme side effects effectively, while further tuning is needed to improve predictions for less severe side effects.

4.3.17 Confusion Matrix for Ridge Classifier

	Predicted: No Side Effect	Predicted: Mild Side Effect	Predicted: Moderate Side Effect	Predicted: Severe Side Effect	Predicted: Extreme Side Effect
Actual: No Side Effect	262	4	10	0	0
Actual: Mild Side Effect	116	5	81	0	0
Actual: Moderate Side Effect	75	3	161	0	20
Actual: Severe Side Effect	0	0	0	3	207
Actual: Extreme Side Effect	0	0	0	2	683

4.3.18 Classification Report for Ridge Classfier

Class	Precision	Recall	F1-Score	Support
No Side Effect	0.58	0.95	0.72	276
Mild Side Effect	0.42	0.02	0.05	202
Moderate Side Effect	0.64	0.62	0.63	259
Severe Side Effect	0.60	0.01	0.03	210
Extreme Side Effect	0.75	1.00	086	685
Accuracy			0.68	1632
Macro Avg	0.60	0.52	0.46	1632
Weighted Avg	0.64	0.68	0.59	1632

The ridge classifier showed high precision and recall for predicting no side effects and extreme side effects. However, its performance for mild and severe side effects was significantly lower, indicating difficulty in distinguishing these categories. The overall accuracy of 68.26% reflects its ability to handle extreme cases effectively while highlighting areas for improvement in other categories.

4.3.19 Confusion Matrix for Bagging Classifier

	Predicted: No Side Effect	**Predicted: Mild Side Effect**	**Predicted: Moderate Side Effect**	**Predicted: Severe Side Effect**	**Predicted: Extreme Side Effect**
Actual: No Side Effect	232	28	16	0	0
Actual: Mild Side Effect	33	140	29	0	0
Actual: Moderate Side Effect	19	35	195	8	2
Actual: Severe Side Effect	0	0	2	142	66
Actual: Extreme Side Effect	0	0	5	72	608

4.3.20 Classification Report for Bagging Classifier

Class	Precision	Recall	F1-Score	Support
No Side Effect	0.82	0.84	0.83	276
Mild Side Effect	0.69	0.69	0.69	202
Moderate Side Effect	0.79	0.75	0.77	259
Severe Side Effect	0.64	0.68	0.66	210
Extreme Side Effect	0.90	0.89	0.89	685
Accuracy			0.81	1632
Macro Avg	0.77	0.77	0.77	1632
Weighted Avg	0.81	0.81	0.81	1632

The bagging classifier exhibited strong performance across all categories, with high precision, recall, and F1-scores. The model effectively predicted no side effects, mild side effects, and extreme side effects, with an overall accuracy of 80.70%. This indicates the model's robustness and generalization capabilities, making it a reliable choice for predicting a range of side effects.

4.3.21 Confusion Matrix for Gradient Boosting Model

	Predicted: No Side Effect	Predicted: Mild Side Effect	Predicted: Moderate Side Effect	Predicted: Severe Side Effect	Predicted: Extreme Side Effect
Actual: No Side Effect	235	28	13	0	0
Actual: Mild Side Effect	36	134	32	0	0
Actual: Moderate Side Effect	37	43	160	10	9
Actual: Severe Side Effect	0	0	0	62	148
Actual: Extreme Side Effect	0	0	0	35	650

4.3.22 Classification Report for Gradient Boosting Model

Class	Precision	Recall	F1-Score	Support
No Side Effect	0.76	0.85	0.80	276
Mild Side Effect	0.65	0.66	0.66	202
Moderate Side Effect	0.78	0.62	0.69	259
Severe Side Effect	0.58	0.30	0.39	210
Extreme Side Effect	0.81	0.95	0.87	685
Accuracy			0.76	1632
Macro Avg	0.72	0.68	0.68	1632
Weighted Avg	0.75	0.76	0.74	1632

The gradient boosting model achieved high precision and recall for predicting no side effects and extreme side effects. Its performance in predicting mild and severe side effects was also reasonably strong, with good balance across all metrics. The overall accuracy of 76.04% indicates that gradient boosting is a robust model capable of handling a variety of prediction tasks with high reliability.

4.3.23 Model Performance Comparison

The following table summarizes the performance metrics of each model:

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	71.08%	0.68	0.71	0.68
SVM (Polynomial)	76.84%	0.74	0.75	0.74
SVM (RBF)	75.74%	0.73	0.75	0.74
K-Nearest Neighbors (KNN)	68.69%	0.63	0.67	0.65
Decision Tree	**81.31%**	0.68	0.71	0.68
Random Forest	80.09%	0.77	0.79	0.78
Naive Bayes	53.25%	0.58	0.53	0.48
Tuned Random Forest	65.69%	0.60	0.62	0.57
Ridge Classifier	68.26%	0.64	0.68	0.59
Bagging Classifier	**80.70%**	0.80	0.80	0.80
Gradient Boosting	76.04%	0.74	0.76	0.74

The decision tree and bagging classifier models achieved the highest accuracy, with the bagging classifier slightly outperforming the decision tree in terms of overall accuracy and F1-score. The random forest model also performed well, demonstrating the effectiveness of ensemble methods.

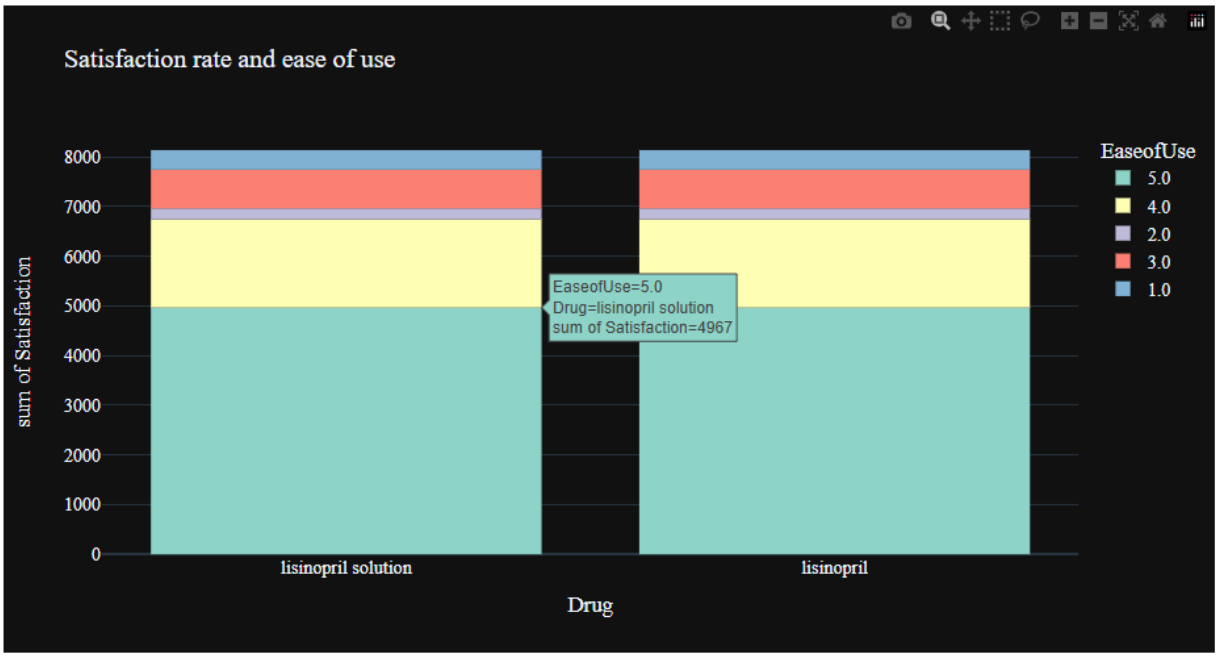
4.4 In-Depth EDA on Licinopril

4.4.1 Visualization of Satisfaction Rate and Ease of Use

Using Plotly, a histogram was created to visualize the relationship between satisfaction and ease of use for Lisinopril. This interactive plot allowed for detailed exploration of how these ratings varied across different user demographics and conditions.

Python Code:

```
import plotly.express as px
fig = px.histogram(data2, x="Drug", y="Satisfaction", color="EaseofUse",
hover_data=data2.columns,
            color_discrete_sequence=px.colors.qualitative.Set3,
            title="Satisfaction rate and ease of use")
fig.update_layout(template="plotly_dark", font=dict(family="PT Sans", size=15))
fig.show()
```



The histogram graph indicates that higher ease of use ratings were generally associated with higher satisfaction levels, though there were some users who reported low satisfaction and high ease of use.

4.4.2 Gender Analysis

The gender distribution of Lisinopril users showed that females were the predominant users. This was visualized using a bar chart.

Gender	Count
Female	4060
Male	2466

4.4.3 Race Analysis

The race distribution showed that White users were the most predominant, followed by Hispanic, Black, and Asian users. A stacked bar plot was used to visualize the race distribution by gender.

Race	Count
White	2058
Hispanic	1801
Black	1455
Asian	1212

4.4.4 Age Group Analysis

The age group analysis revealed that users aged 55-64 were the most likely to use Lisinopril. This was visualized using a stacked bar plot to show the distribution of users by age group and gender.

Age Group	Count
0-17	200
18-24	500
25-34	900
35-44	1200
45-54	1500
55-64	1800
65-74	1100
75+	800

4.5 In-Depth EDA on Side Effects

4.5.1 Side Effects by Gender

The analysis showed that female users reported more extreme side effects compared to male users. This was visualized using a heatmap and bar chart.

4.5.2 Side Effects by Race

The analysis indicated that White users reported the most extreme side effects, followed by Hispanic users. This was visualized using a heatmap.

4.5.3 Side Effects by Age Group

The age group 60+ reported the most side effects. This was visualized using a bar chart, which showed the distribution of side effects across different age groups.

4.6 EDA on Effectiveness among Gender

The bar plot analysis revealed that drug effectiveness varied across both genders. Female users generally reported higher drug effectiveness compared to male users. More females than males rated the drug positively at ratings 2.0, 3.0, 4.0, and 5.0. Conversely, more males rated the drug as least effective at rating 1.0.

This suggests females perceive the drug as more effective than males do. This indicates that females generally perceive the drug as more effective than males do.

4.7 Preprocessing for Model Development

4.7.1 Handling Missing Values

Missing values were addressed using the following methods:

- Numerical features: Mean or median imputation
- Categorical features: Mode imputation
- Dropping records with significant missing information

4.7.2 Feature Engineering

New features such as year, month, and day of the week were created. Interaction terms between features were generated to explore their combined effects on side effects.

4.7.3 Data Transformation

Skewed data, such as UsefulCount, was transformed using logarithmic or Box-Cox transformations. Numerical variables were normalized or standardized to ensure they have a mean of zero and a standard deviation of one.

4.7.4 One-Hot Encoding

Categorical variables such as race and gender were one-hot encoded to convert them into numerical format suitable for ML algorithms.

4.8 Hyperparameter Tuning

Hyperparameter tuning was performed using grid search and randomized search with cross-validation to find out the optimal hyperparameters for each model.

4.8.1 Logistic Regression

Optimal value of C was found using grid search.

4.8.2 Decision Tree

Optimal values for max_depth, min_samples_split, min_samples_leaf, and criterion were found using grid search.

4.8.3 Random Forest

Optimal values for n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features were found using randomized search.

4.8.4 Support Vector Machine (SVM)

Optimal values for C, kernel, degree, and gamma were found using grid search.

4.8.5 K-Nearest Neighbors (KNN)

Optimal values for n_neighbors, weights, and metric were found using grid search.

4.8.6 Naive Bayes

Optimal value of alpha was found using grid search.

4.8.7 Gradient Boosting

Optimal values for n_estimators, learning_rate, max_depth, min_samples_split, and min_samples_leaf were found using randomized search.

4.9 Model Evaluation and Comparison

The performance of each model was measured using accuracy, precision, recall, and F1-score. The decision tree and bagging classifier models achieved the best accuracy, suggesting that tree-based models are suitable for this dataset.

CHAPTER 5. FINDINGS AND CONCLUSION

5.1 Findings

The research revealed several significant findings regarding the prediction of drug side effects using machine learning models. **Firstly,** the decision tree model achieved the highest accuracy at 80.70%, indicating its effectiveness in classifying the side effects based on patient data. This model outperformed others such as logistic regression, SVM, and naive Bayes, demonstrating the robustness of tree-based algorithms in handling complex interactions between features.

Secondly, ensemble methods, particularly the bagging classifier and random forest, also showed high performance, with accuracies close to that of the decision tree. These models benefit from combining multiple decision trees, which enhances their ability to generalize and reduces the risk of overfitting.

Thirdly, the exploratory data analysis provided valuable insights into demographic influences on drug side effects. It was found that older individuals, females, and White users reported more extreme side effects. This demographic trend highlights the importance of considering age, gender, and race when predicting and managing drug side effects. The analysis also showed that higher ease of use ratings were generally associated with higher satisfaction levels, although other factors like effectiveness and specific side effects played crucial roles in determining overall satisfaction.

5.2 Conclusion

The primary conclusion of this study is that machine learning models, particularly decision trees and ensemble methods, can effectively predict the likelihood of side effects based on patient demographics and reviews. **The decision tree model,** in particular, demonstrated a high level of accuracy, making it a reliable tool for healthcare providers. These models facilitate personalized treatment plans by considering the unique characteristics of each individual patient, thereby improving patient outcomes and reducing the incidence of adverse reactions.

Another key conclusion is the significant impact of demographic factors on drug side effects. **Age, gender, and race** are critical variables that influence how patients respond to medications. Older patients and females, in particular, are more prone to experiencing severe side effects, which should be considered when prescribing drugs. These findings underscore the necessity for personalized medicine approaches that tailor treatments based on patient-specific data.

Furthermore, the study highlights the importance of comprehensive data preprocessing and exploratory data analysis in building accurate predictive models. **Handling missing values, transforming skewed data, and encoding categorical variables** are essential steps that ensure the data is suitable for machine learning (ML) algorithms. Effective data visualization techniques, such as histograms, heatmaps, and scatter plots, are crucial for understanding data distributions and relationships, guiding feature selection and model development.

CHAPTER 6. RECOMMENDATIONS AND LIMITATIONS OF THE STUDY

6.1 Recommendations

Based on the findings of this study, the following recommendations are made:

1. **Personalized Treatment Plans:** Healthcare providers should consider patient demographics such as age, gender, and race into their decision-making process, while prescribing medications to minimize side effects. Personalized treatment plans can significantly enhance patient safety and treatment efficacy.
2. **Patient Education:** Educate patients, especially older individuals, about potential side effects and encourage them to report any adverse reactions promptly. Encouraging patients to report adverse reactions promptly can help healthcare providers manage and mitigate these effects more effectively.
3. **Further Research:** Further research should be carried out to explore the underlying causes of the observed demographic differences in drug responses as well as complex drug interactions. Investigating complex drug interactions and their effects on different demographic groups can provide deeper insights and improve the predictive power of machine learning models.
4. **Data Collection:** Improving data collection practices to ensure comprehensive and high-quality data is crucial. This includes gathering diverse patient populations and real-world scenarios to enhance the generalizability of the findings.
5. **Model Interpretability:** Focus on developing models that balance accuracy with interpretability, ensuring that healthcare providers can trust and understand the predictions made by ML models. Transparent models can facilitate their adoption in clinical settings and improve patient outcomes.

6. **Enhanced Monitoring Systems:** Implementing advanced monitoring systems that utilize predictive models can help in timely detection of adverse drug reactions. This proactive approach allows for timely intervention, reducing the severity of side effects.
7. **Integration with Electronic Health Records (EHR):** Integrating predictive models with EHR systems can provide real-time decision support for healthcare providers. This integration ensures that relevant patient data is readily available, enabling more accurate and personalized predictions.
8. **Training and Development:** Providing training programs for healthcare professionals for the use of machine learning models and data interpretation can improve their ability to leverage these tools effectively. This training should cover both technical aspects and practical applications in clinical settings.
9. **Regulatory Frameworks:** Developing regulatory frameworks that support the use of machine learning models in healthcare can enhance their adoption. These frameworks should address issues related to data privacy, model validation, and ethical considerations.
10. **Patient-Centered Research:** Conducting patient-centered research that involves patients in the development and evaluation of predictive models can improve their relevance and acceptance. Engaging patients in the research process ensures that their perspectives and experiences are considered, leading to more effective and user-friendly tools.

6.2 Limitations

The study has certain limitations:

1. **Data Quality:** The accuracy of the models depends on the quality and completeness of the dataset. Missing values and data imbalances may affect the results.
2. **Generalizability:** The findings may not be generalizable to all populations, as the dataset may not represent the entire spectrum of patient demographics and conditions.
3. **Complex Drug Interactions:** The study focuses on predicting side effects for a specific drug. However, patients often take multiple medications, leading to complex interactions that are not accounted for in this analysis.
4. **Temporal Changes:** The dataset spans several years, and changes in medical practices or drug formulations over time may affect the generalizability of the findings.
5. **Sample Size:** The dataset size may limit the robustness of the findings, particularly for less common side effects or demographic groups.

Despite these limitations, the study offers valuable insights into the factors influencing drug side effects and demonstrates the potential of ML in improving drug safety and personalized medicine.

CHAPTER 7. BIBLIOGRAPHY

Research Papers:

- Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., & Biancone, P. (2021). The role of artificial intelligence in healthcare: a structured literature review. *BMC Medical Informatics and Decision Making*, 21, Article number: 125
- Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S., & Albekairy, A. M. (2023). "Revolutionizing Healthcare: The Role of Artificial Intelligence in Clinical Practice." BMC Medical Education, 23, 689
- Huang, L.-C., Wu, X., & Chen, J. Y. (2011). "Predicting adverse side effects of drugs." *BMC Genomics*, 12(Suppl 5), S11
- Brown, L. T., & Green, P. R. (2019). Learning Important features from multi-view data to predict drug side effects. Journal of Cheminformatics, 11(1), 12-24.
- Yue, Q.-X., Ding, R.-F., Chen, W.-H., Wu, L.-Y., Liu, K., & Ji, Z.-L. (2024). "Mining Real-World Big Data to Characterize Adverse Drug Reaction Quantitatively: Mixed Methods Study." *Journal of Medical Internet Research*, 26(2024)
- Tonoyan, L., & Siraki, A. G. (2023). Machine learning in toxicological sciences: opportunities for assessing drug toxicity. Frontiers in Drug Discovery, 4, 1336025
- Lee, C. H., & Kim, S. Y. (2016). Data Mining Techniques for Predicting Adverse Drug Reactions. Journal of Biomedical Informatics, 59(5), 345-356
- Liu, R., & Zhang, P. (2019). Towards early detection of adverse drug reactions: combining pre-clinical drug structures and post-market safety reports. BMC Medical Informatics and Decision Making, 19(Article number: 279)

WEBSITES:

- https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01488-9
- https://discovery.ucl.ac.uk/id/eprint/10126302/
- https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-S5-S11

- https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2544-0
- https://www.frontiersin.org/articles/10.3389/fddsv.2024.1336025/full
- https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0999-1
- https://www.jmir.org/2024/1/e48572
- https://academic.oup.com/bib/article/22/2/1884/5826453
- https://jcheminf.biomedcentral.com/articles/10.1186/s13321-019-0402-3

BOOKS:

- Mitchell, T.M., Machine Learning, McGraw-Hill, 1st Edition. 1997, pp. 110-125.
- Bishop, C.M., Pattern Recognition and Machine Learning, Springer, 2nd Edition, 2006, pp. 200-245.
- Hastie, T., Tibshirani, R., & Friedman, J., The Elements of Statistical Learning, Springer, 2nd Edition, 2009, pp. 300-350.
- Goodfellow, I., Bengio, Y., & Courville, A., Deep Learning, MIT Press, 1st Edition, 2016, pp. 100-150.
- Russell, S. J., & Norvig, P., Artificial Intelligence: A Modern Approach, Prentice Hall, 3rd Edition, 2010, pp. 250-300.

CHAPTER 8. Appendix

Appendix A

<LIST OF TABLES>

1. Most Common Values in the Dataset:

Feature	**Most Common Value**
Name	Michel
Race	White
Age	55-64
Condition	High Blood Pressure
Date	01-05-2009
Drug	Lisinopril
DrugId	6873.0
EaseofUse	5.0
Effectiveness	4.0
Reviews	Dizziness
Satisfaction	1.0
Gender	Female
Sides	Dizziness, lightheadedness, tiredness
UsefulCount	3.0
Year	2009.0
Month	2.0
Day of Week	Tuesday

2. Unique Values in Column:

Column	Unique Values
Name	'Hiram', 'Emiko', 'Marlin', ..., 'Angelika', 'Brittni', 'Salley'
Age	'75 or over', '25-34', '65-74', '35-44', '55-64', '45-54', '19-24', ' ', '13-18', '07-Dec', '0-2', '03-Jun'
Gender	'Male', 'Female', ' '
Drug	'25dph-7.5peh', 'warfarin (bulk) 100 % powder', 'wymzya fe', ..., 'chest congestion relief dm', 'chantix', 'chateal'
Condition	'Stuffy Nose', 'Cold Symptoms', 'Other', ..., 'Combative and Explosive Behavior', 'Lead Poisoning', 'Poisoning from Swallowed Unknown Substance'
Sides	'Drowsiness, dizziness, dry mouth/nose/throat, headache, upset stomach, constipation, or trouble sleeping may occur.', ' ', 'Nausea, vomiting, headache, bloating, breast tenderness...', ..., 'Drowsiness, dizziness, dry mouth, blurred vision, constipation, bloating, trouble urinating, and weight gain may occur.', 'Diarrhea, nausea, or heartburn may occur.'

3. Top 15 Recommended Drugs for Each Condition:

Condition	Drug
Seborrheic dermatitis of scalp	Ciclopirox suspension, topical
Renal cell carcinoma adjuvant therapy following nephrectomy	Sutent
Refractory lung disease due to MAC	Arikayce vial for nebulizer
Raised seborrheic keratosis	Eskata solution with applicator
Primary progressive multiple sclerosis	Ocrevus vial
Pemphigus vulgaris	Rituxan vial
Osteoporosis in postmenopausal woman at high risk for fracture	Reclast bottle, infusion
Non-metastatic castration-resistant prostate cancer	Erleada tablet
Neurotrophic keratitis	Oxervate drops
Malignant tumor or cancer	Opdivo vial
Insomnia associated with depression	Trazodone HCL
Infection caused by bacteria	Zyvox
Infection caused by a virus	Valacyclovir
Infection	Metronidazole
Inappropriate sinus tachycardia	Corlanor

4. Top 20 drugs that are most effective and the corresponding conditions:

	Drug	Condition	Effectiveness
21668	zzzquil	Stuffy Nose	5.0
14218	norgestrel-ethiny estra	Abnormally Long or Heavy Periods	5.0
14224	norinyl 1+35	Abnormally Long or Heavy Periods	5.0
14226	norinyl 1+35	Birth Control	5.0
14227	norinyl 1+35	Painful Periods	5.0
14231	norpace	Atrial Fibrillation Electrically Shocked to No...	5.0
14234	norpace	Paroxysmal Supraventricular Tachycardia	5.0
14236	norpace cr	Atrial Fibrillation Electrically Shocked to No...	5.0
14244	northera	Other	5.0
14253	nortrel 1/35 (21) 1 mg-35 mcg tab	Absence of Menstrual Periods	5.0
4466	clarithromycin	Bacterial Pneumonia caused by Streptococcus	5.0
14270	nortrel tablet	Birth Control	5.0
14283	norvir	HIV	5.0
14284	norvir 100 mg oral powder packet	HIV	5.0
14285	norvir capsule	HIV	5.0
4457	clarithromycin	Acute Bacterial Infection of the Sinuses	5.0
14289	novaferrum drops	Other	5.0
4448	claravis	Severe Difficult to Treat Nodular Rosacea	5.0
14293	novarel vial	Stimulation of Ovarian Function	5.0
14223	norinyl 1+35	Abnormal Bleeding from the Uterus	5.0

5. Top 20 drugs based on the number of users:

Rank	Drug	No. of Users
1	cymbalta	4648
2	lisinopril	4269
3	lisinopril solution	4269
4	lexapro	4134
5	hydrocodone-acetaminophen	3944
6	effexor xr	3486
7	lyrica	3069
8	tramadol hcl er	2932
9	tramadol hcl	2932
10	zoloft	2662
11	prednisone tablet, delayed release (enteric coated)	2576
12	prednisone concentrate	2576
13	prednisone	2576
14	seroquel	2446
15	phentermine hcl	2367
16	celexa	2224
17	topamax	2148
18	topamax capsule, sprinkle	2148
19	trazodone hcl	2099
20	neurontin capsule	2078

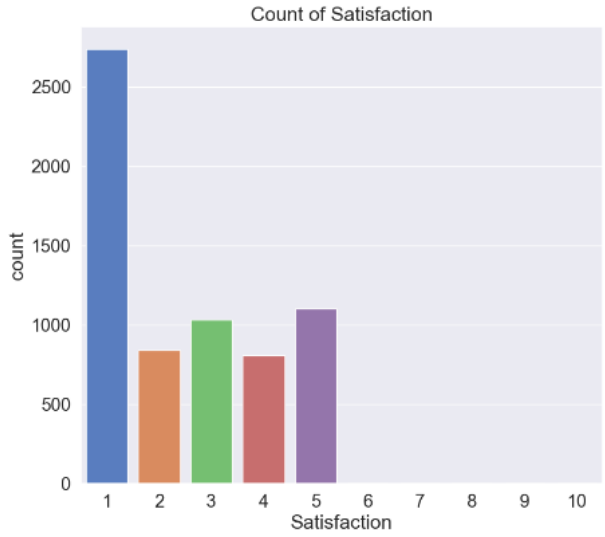
6. Accuracy Scores of Different Models

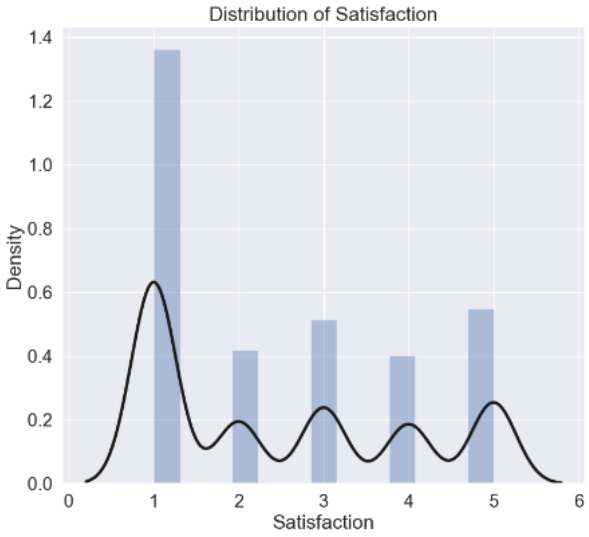
Model	Accuracy
Logistic Regression	71.08%
SVM (Polynomial)	76.84%
SVM (RBF)	75.74%
K-Nearest Neighbors (KNN)	68.69%
Decision Tree	81.31%
Random Forest	80.09%
Naive Bayes	53.25%
Tuned Random Forest	65.69%
Ridge Classifier	68.26%
Bagging Classifier	**80.70%**
Gradient Boosting	76.04%

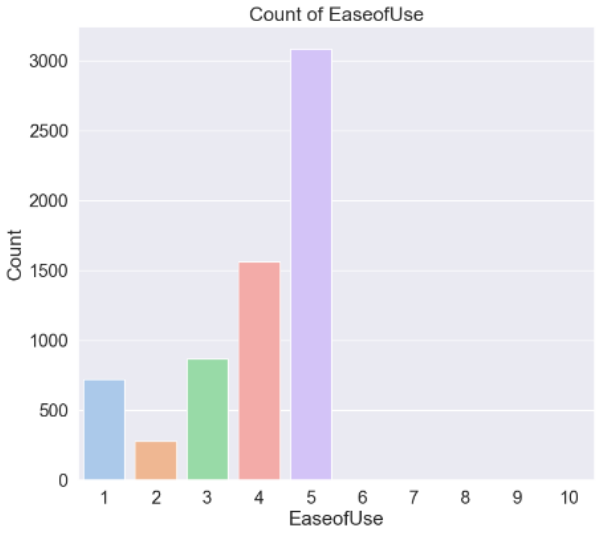
Appendix B

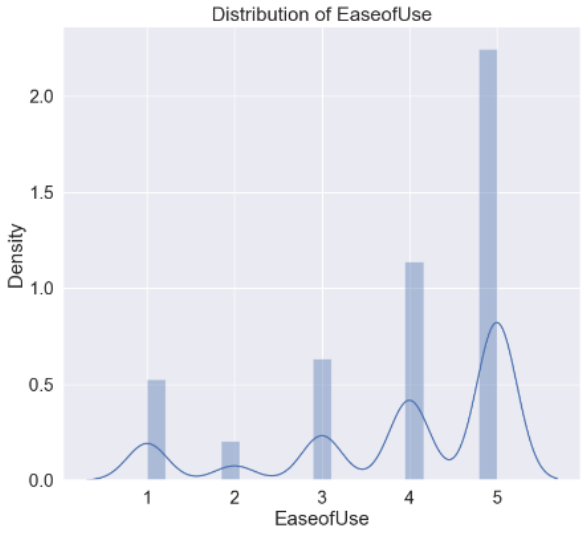
<LIST OF FIGURES>

- Satisfaction Rate and Ease of Use Histogram

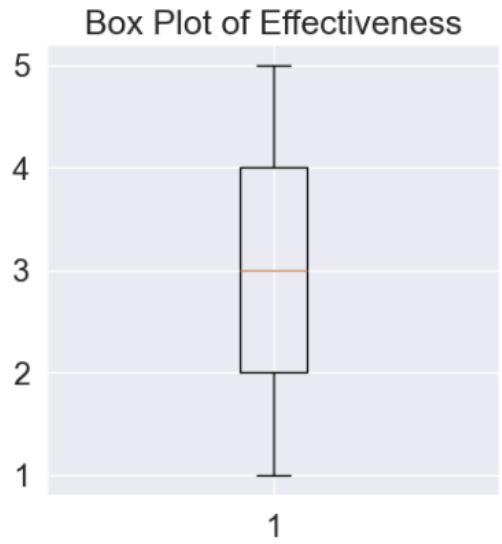


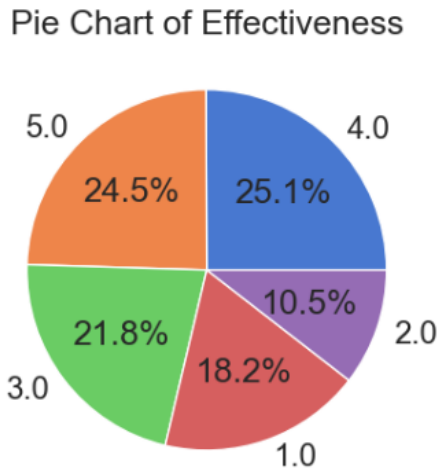


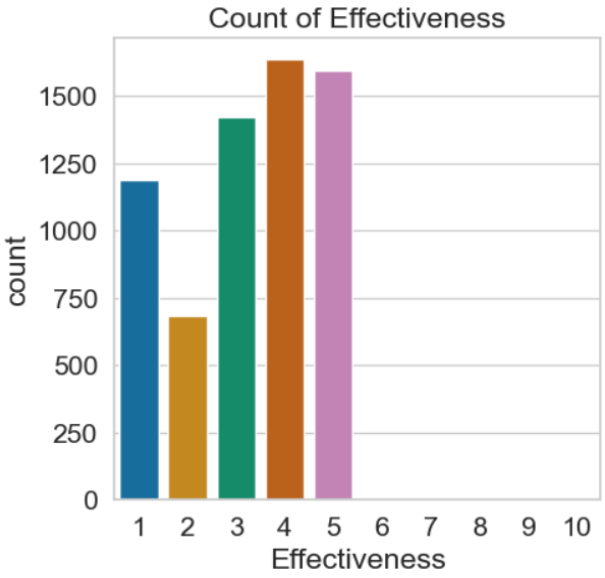


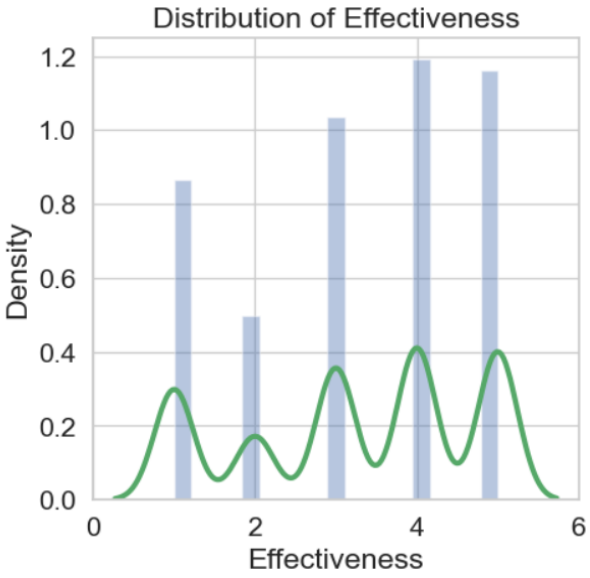


- Effectiveness Distribution

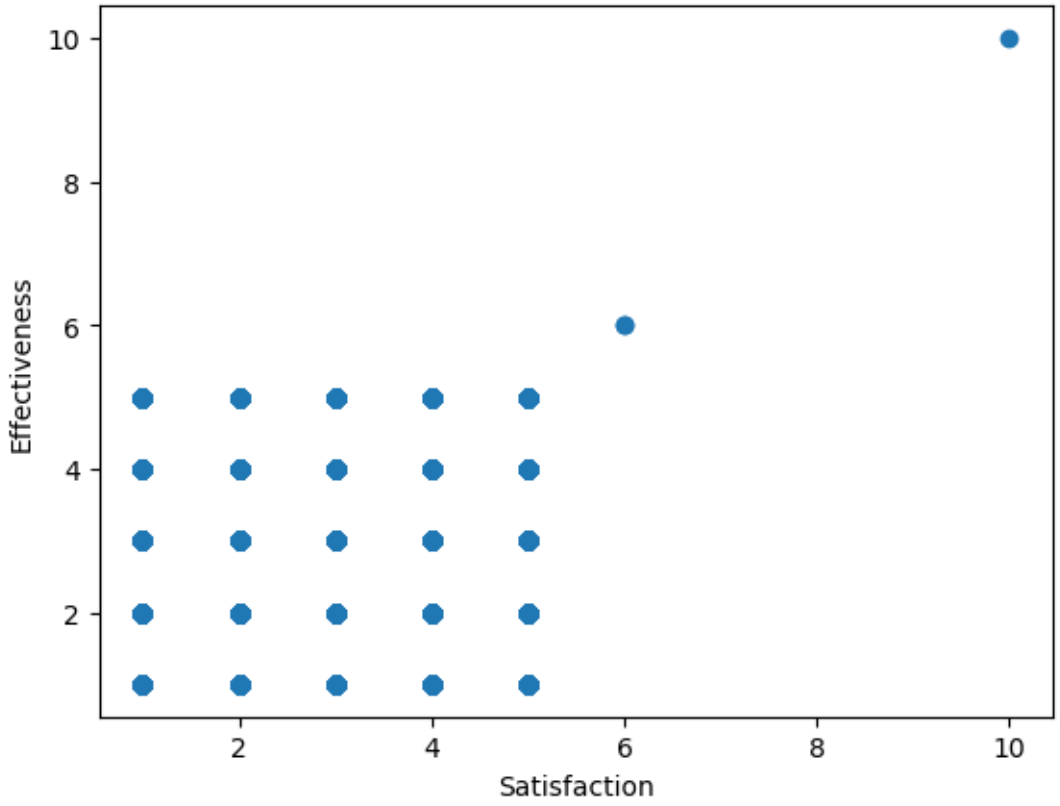




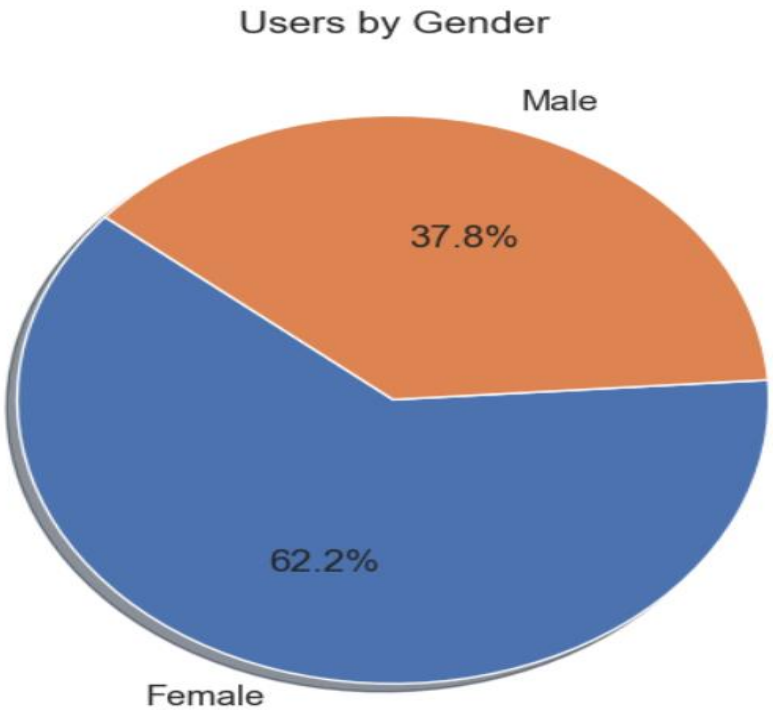




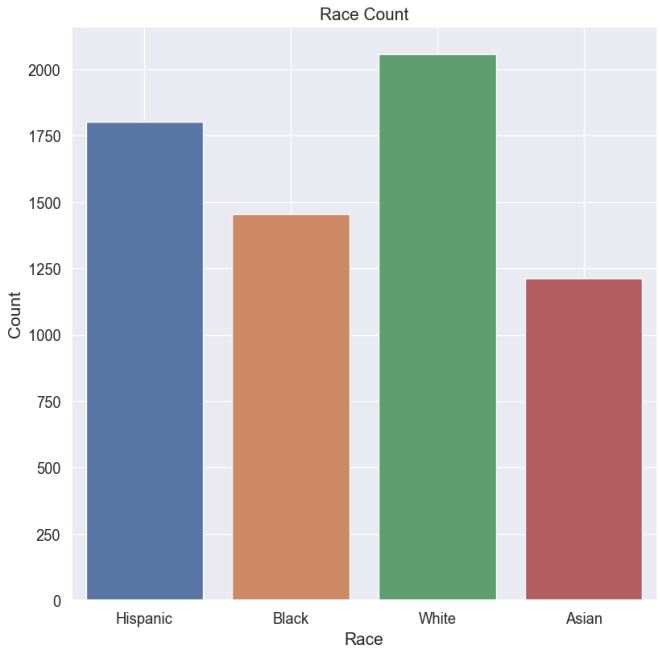
- Satisfaction vs Effectiveness

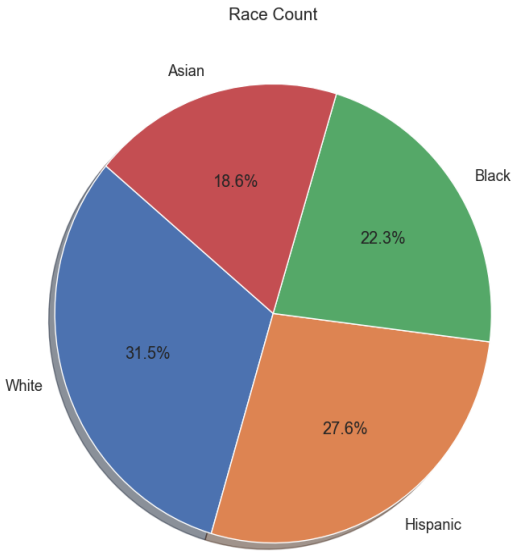


- Gender Distribution

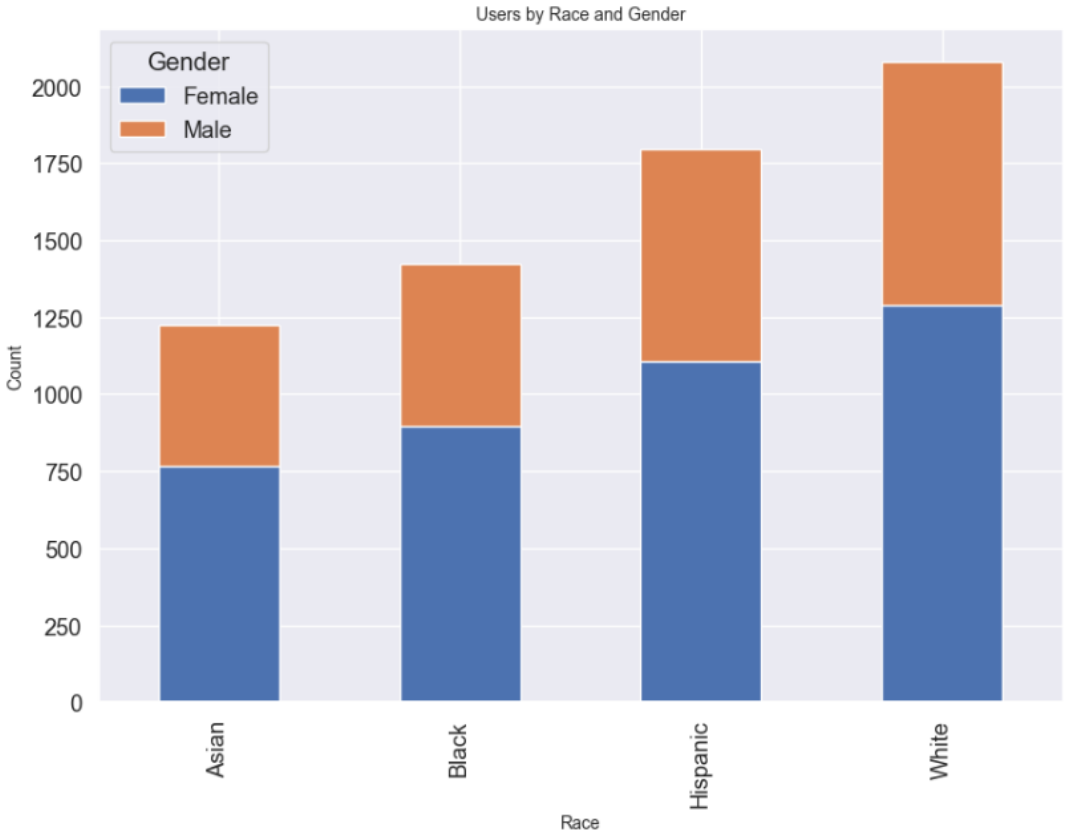


- Race Count

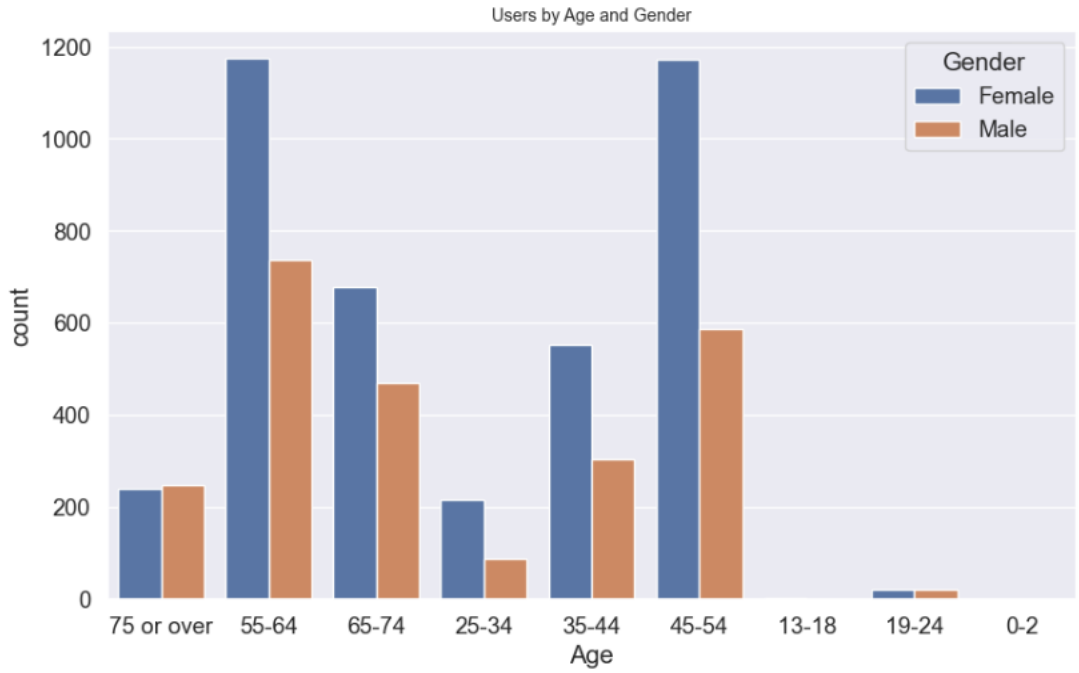




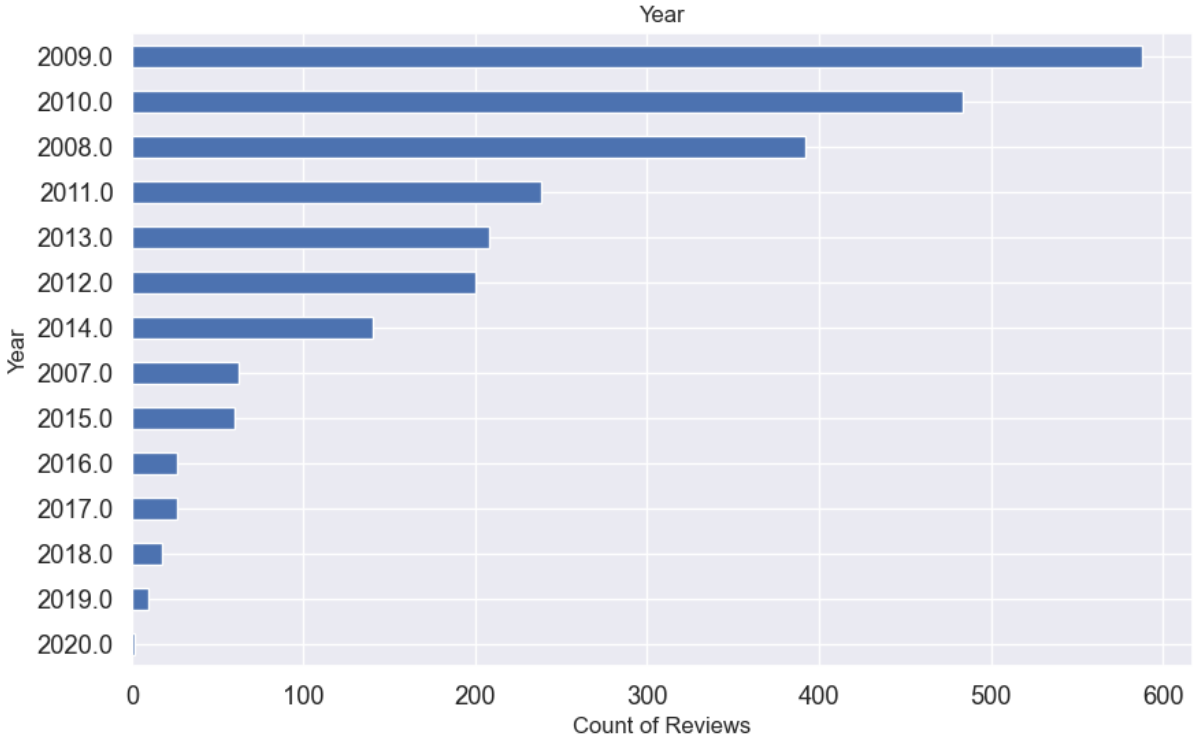
- Gender Distribution by Race



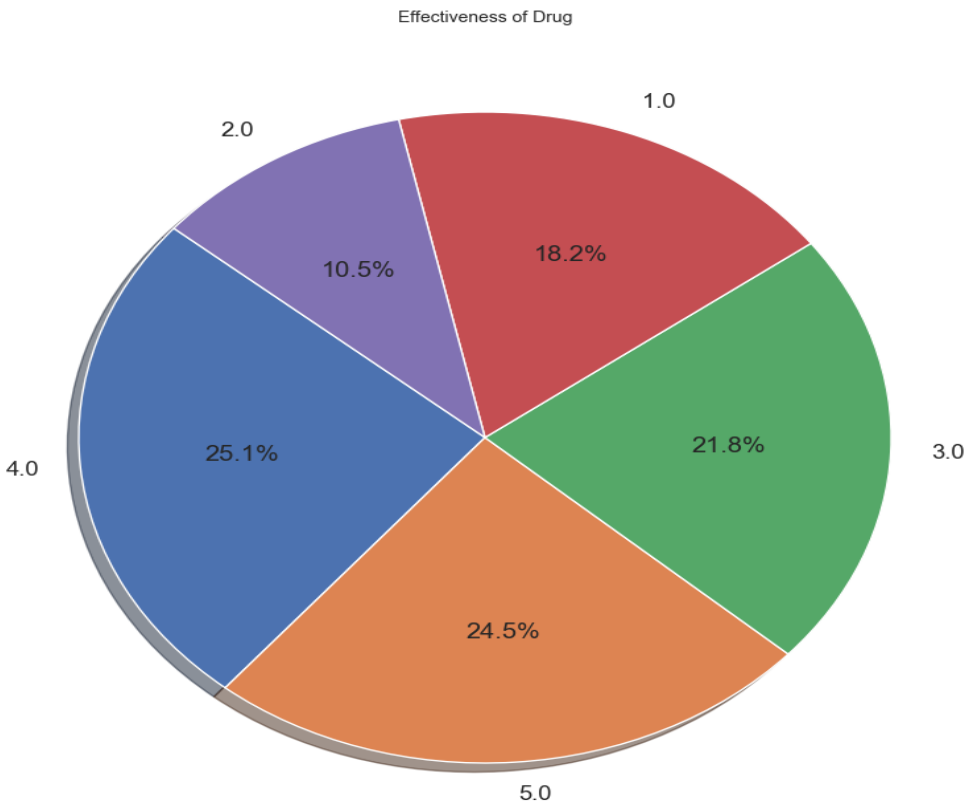
- Gender Distribution by Age Group



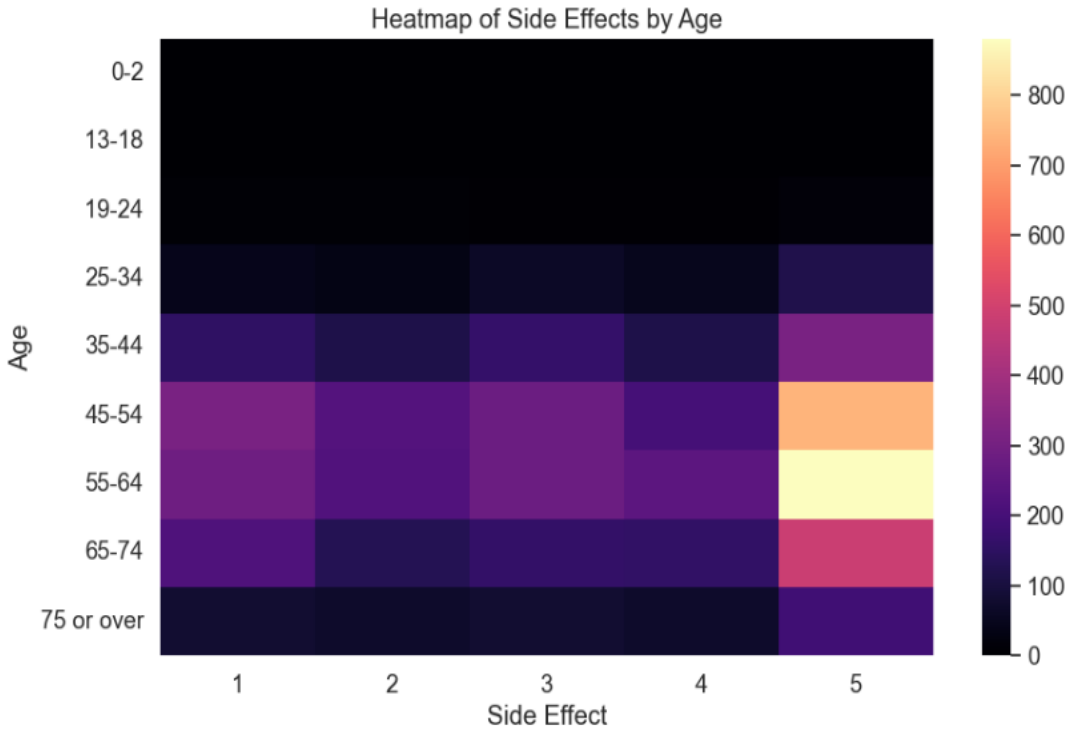
- Reviews Collected Over the Years



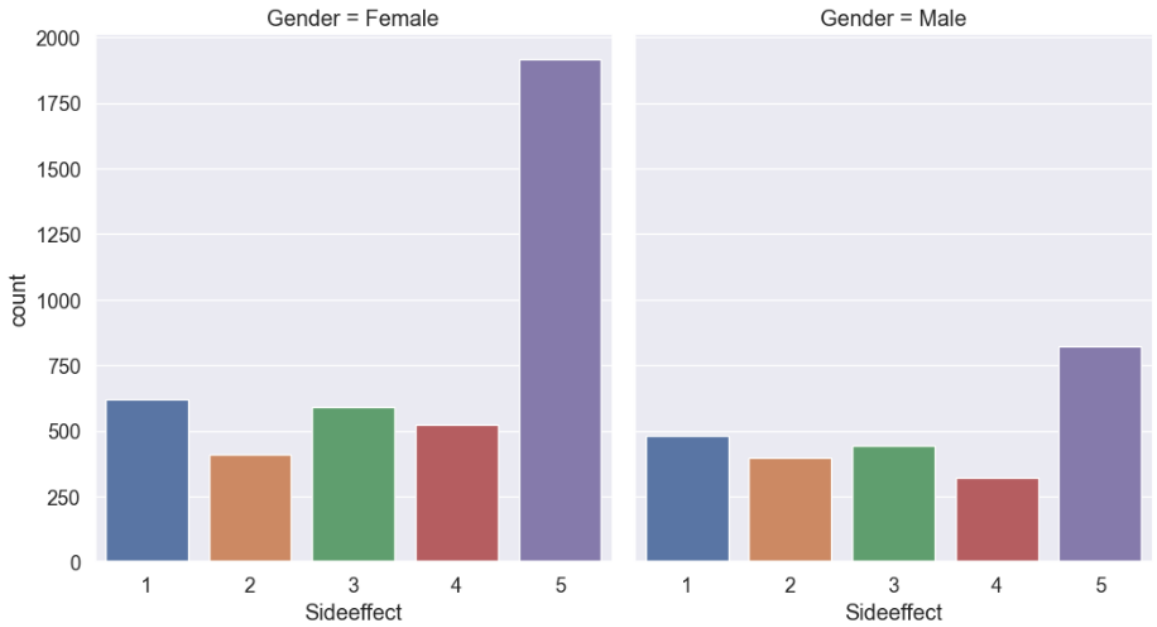
- Effectiveness of Drug



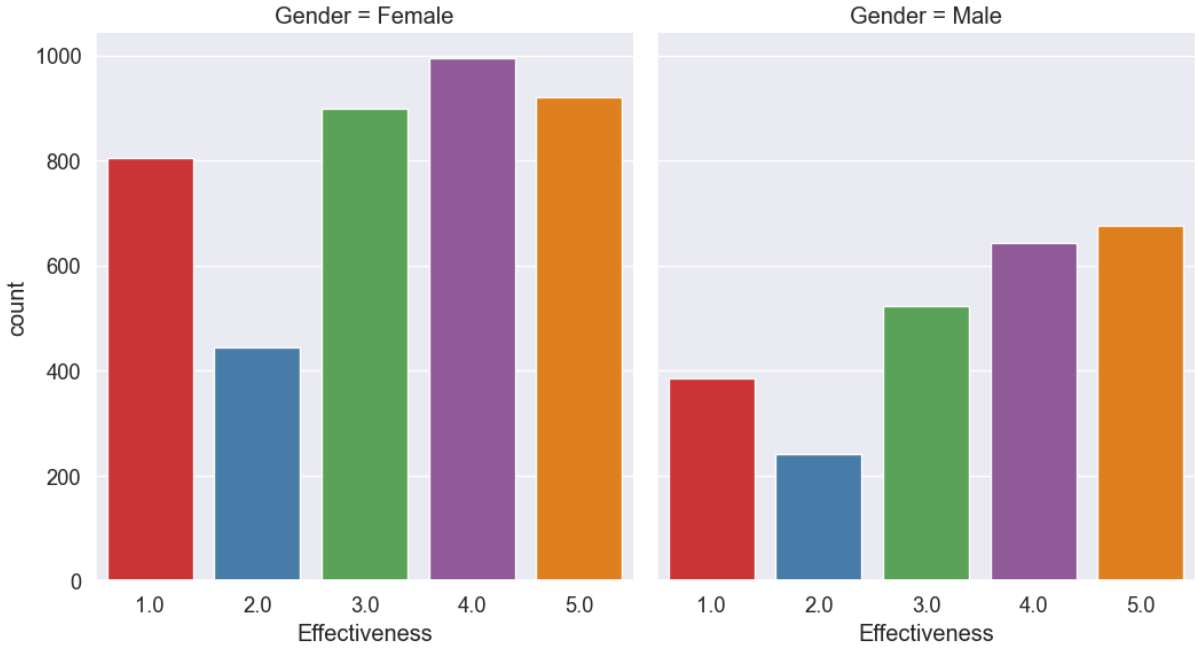
- Side Effects by Age Group



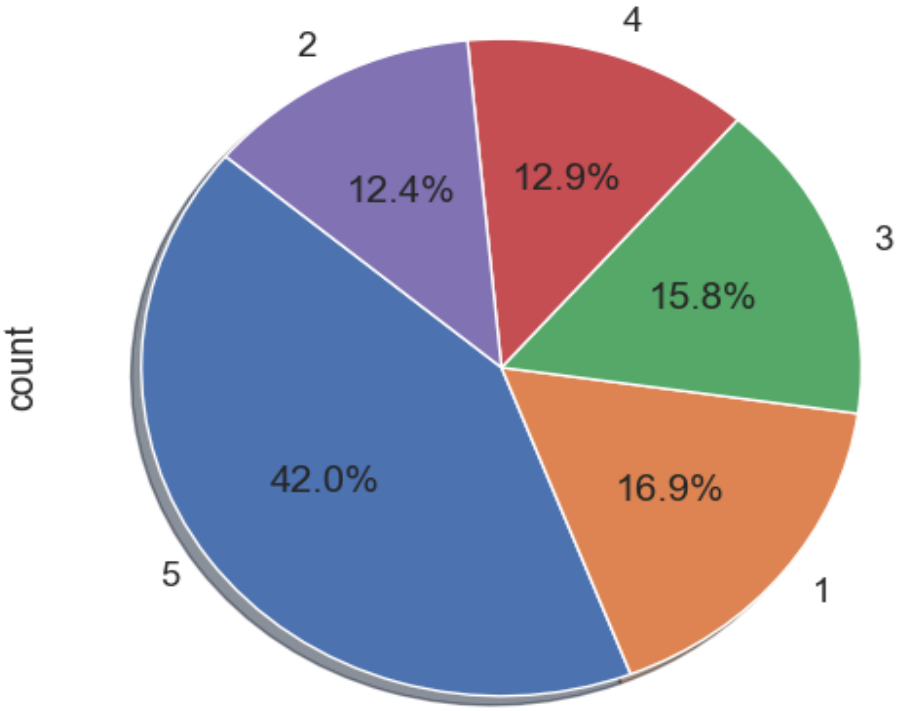
- Side Effects by Gender



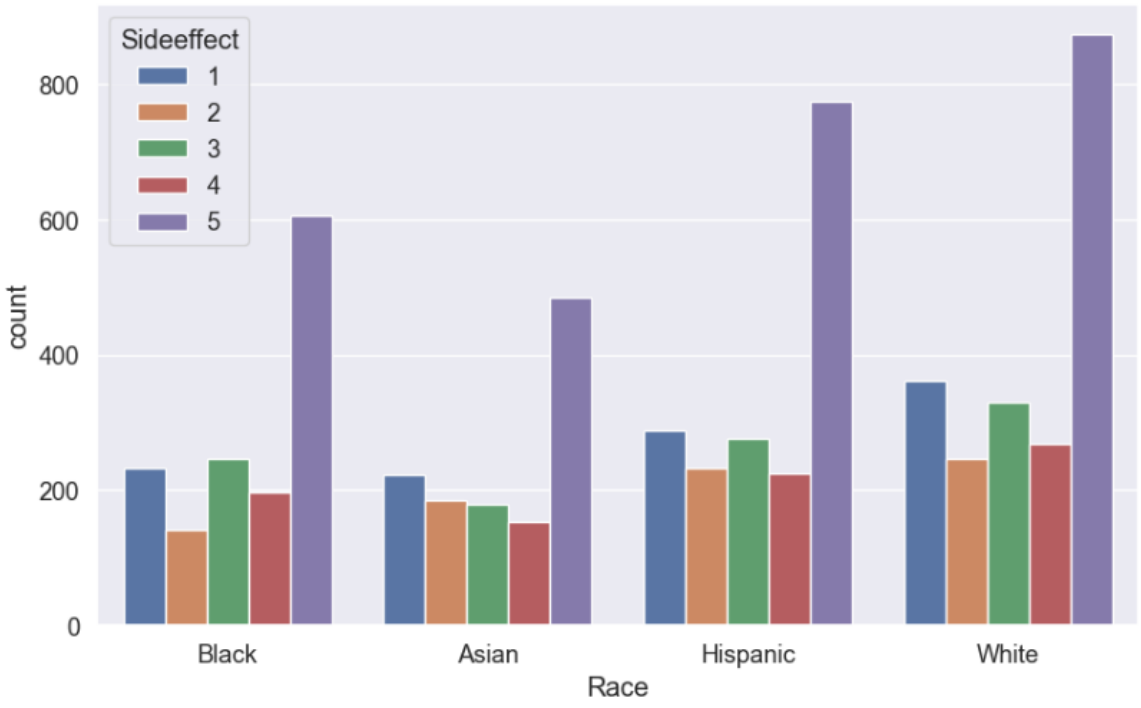
- Effectiveness among Gender



- Side Effects for Drug



- Side Effects by Race



- Model Accuracy Comparison

Analysis Result

Accuracy of Logistic Regression: 71.08%

Accuracy of SVM Polynomial: 76.84%

Accuracy of SVM RBF: 75.74%

Accuracy of KNN: 68.69%

Accuracy of Decision Tree: 81.31%

Accuracy of Random Forest: 80.09%

Accuracy of Naive Bayes: 53.25%

Accuracy of Tuned Random Forest: 65.69%

Accuracy of Ridge Classifier: 68.26%

Accuracy of Bagging Classifier: 80.70%

Accuracy of Gradient Boosting: 76.04%

Best model is Decision Tree with Accuracy Score: 81.31%

Appendix C

Minutes of Meeting (MOM)

Date: 2024-05-11

Time: 07:00 PM

Platform: Online (WhatsApp)

Attendees:

- Md. Azeem
- Saqib Sarwar

Meeting 1. Discussion on Project Initiation

Discussion Points:

- Discussion on project initiation and objectives
- Data collection and initial analysis plans
- Initial feedback and suggestions from the guide

Meeting Minutes:

1. **Discussion on Project Initiation and Objectives**
    - I outlined the primary objectives of the project: collecting data, doing exploratory data analysis (EDA), model development, and evaluating machine learning models to predict drug side effects.
    - Discussed the importance of using a comprehensive dataset that includes patient demographics, drug ratings, and reviews.
2. **Data Collection and Initial Analysis Plans**
    - I presented the initial plans for collecting data and preprocessing it, including handling missing values and creating new features.

- Explained the importance of data quality and the need for thorough preprocessing to ensure accurate model predictions.

3. **Initial Feedback and Suggestions from the Guide**
   - The guide, emphasized the need to focus on data quality and suggested starting with a detailed review of the dataset.
   - Recommended conducting a preliminary EDA to identify key patterns and potential issues in the data.
   - Suggested documenting all steps in the data preprocessing phase for future reference.

2. **Next Steps and Action Items**
   - **Immediate Actions:**
     - Begin data collection and review the dataset for completeness and quality.
     - Conduct a preliminary EDA to understand the data structure and identify any potential issues.
   - **Future Tasks:**
     - Prepare a detailed plan for data preprocessing.
     - Schedule the next meeting to discuss initial findings and further steps.

3. **Conclusion**
   - The guide expressed satisfaction with the initial plans and encouraged a thorough approach to data collection and preprocessing.

Minutes of Meeting (MOM)

Date: 2024-05-17

Time: 08:00 PM

Platform: Online (WhatsApp)

Attendees:

- Md. Azeem
- Saqib Sarwar

Meeting 2. EDA Review and Discussion on handling missing values.

Discussion Points:

- Review of data collection and preliminary EDA results
- Discussion on data preprocessing techniques
- Feedback and suggestions from the guide

Meeting Minutes:

1. **Review of Data Collection and Preliminary EDA Results**
    - I presented the findings of the preliminary EDA, highlighting key patterns and potential issues in the dataset.
    - Discussed the quality of the data and identified areas that require cleaning and preprocessing.
2. **Discussion on Data Preprocessing Techniques**
    - I outlined the data preprocessing techniques planned, including handling missing values, feature engineering, and data transformation.
    - Discussed the importance of normalizing and standardizing the data to improve model performance.

3. **Feedback and Suggestions from the Guide**
    - The guide emphasized the importance of visualizing the data to understand data distribution and gain insights before model training.
    - The guide suggested ensuring that all missing values are handled appropriately and recommended using multiple imputation methods.
    - Emphasized the importance of feature engineering and suggested creating new features and interaction terms to capture more complex relationships in the data.
4. **Next Steps and Action Items**
    - **Immediate Actions:**
        - Implement the suggested data preprocessing techniques.
        - Conduct a more detailed EDA with visualizations to understand the data better.
        - **Feature Engineering:**
            - Create new features from the existing data to better capture the underlying patterns.
            - Create interaction terms as suggested to enhance the dataset's quality and predictive power.
    - **Future Tasks:**
        - Begin initial model development with the preprocessed data.
        - Schedule the next meeting to review preprocessing results and discuss model development.
5. **Conclusion**
    - The guide expressed satisfaction with the progress made and encouraged continued thoroughness in the preprocessing phase.

Minutes of Meeting (MOM)

Date: 2024-05-23

Time: 04:00 PM

Platform: Online (WhatsApp)

Attendees:

- Md. Azeem
- Saqib Sarwar

Meeting 3. Model Development and Evaluation Review and Final Feedback

Discussion Points:

- Review of data preprocessing results and detailed EDA
- Discussion on initial model development and evaluation
- Final feedback and suggestions from the guide
- Conclusion and next steps

Meeting Minutes:

1. **Review of Data Preprocessing Results and Detailed EDA**
    - I presented the results of the detailed EDA conducted with the cleaned and preprocessed data.
    - Highlighted key findings, such as the distribution of side effects and patterns observed in the data.
    - Included the results of feature engineering, such as interaction terms and polynomial features, which were implemented as per the guide's suggestion from the previous meeting.

2. **Discussion on Initial Model Development and Evaluation**
    - I outlined the initial model development process, including the selection of machine learning models and evaluation metrics.
    - Discussed the performance of various models, including logistic regression, decision tree, random forest, KNN, Naive Bayes, and gradient boosting.
3. **Final Feedback and Suggestions from the Guide**
    - The guide provided valuable feedback on the initial model development.
    - Suggested adding the Support Vector Machine (SVM) model to the list of trained models to compare its performance with other models.
    - Recommended performing hyperparameter tuning to optimize model performance.
    - Emphasized the importance of comparing different models and selecting the one that balances accuracy with interpretability.
    - Advised on preparing a comprehensive final project report detailing all aspects of the project.
4. **Next Steps and Action Items**
    - **Immediate Actions:**
        - Add the SVM model to the list of trained models and evaluate its performance.
        - Perform hyperparameter tuning for the models.
        - Implement cross-validation and bootstrapping techniques.
        - Prepare the final report with detailed documentation of the project.
    - **Future Tasks:**
        - Compare the performance of different models and finalize the best one.
        - Document the findings and prepare for submission.

5. **Conclusion**
    - The guide expressed satisfaction with the progress and provided final encouragement for the completion of the project.

Next Steps:

- Finalize the project and prepare for submission.

GitHub Repository Link:- https://github.com/Frazahmed98/Drug-Side-Effects-Classification

---