**NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY**

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA



**Internship Project Report**

### on

Drug Classification and Deployment of the Model

*Submitted in partial fulfilment of the requirement for the award of Degree of*

*Bachelor of Engineering*

*in*

*Computer Science and Engineering*

*Submitted by:*

|  |  |
| --- | --- |
| M PRAVEEN KUMAR | 1NT21CS096 |



Department of Computer Science and Engineering

# (Accredited by NBA Tier-1)

2022-2023

**NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY**

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA

Department of Computer Science and Engineering

(Accredited by NBA Tier-1)



**CERTIFICATE**

This is to certify that the Internship Project Report on Drug Classification and Deployment of the Model is an authentic work carried out by **M PRAVEEN KUMAR (1NT21CS096)** bonafide student of **Nitte Meenakshi Institute of Technology**, Bangalore in partial fulfilment for the award of the degree of ***Bachelor of Engineering*** in COMPUTER SCIENCE AND ENGINEERING of Visveswaraya Technological University, Belagavi during the academic year ***2022-2023.*** It is certified that all corrections and suggestions indicated during the internal assessment has been incorporated in the report.

**Signature of the HoD**

HoD, Dr. Vijaya Shetty.S

**ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. I express my sincere gratitude to Mohammed Azar sir , IndoSkill AQMENZ AUTOMATION Pvt Ltd for providing facilities.

We wish to thank our HoD, Dr. Vijaya Shetty.S for the excellent environment created to further educational growth in our college. We also thank for the invaluable guidance provided which has helped in the creation of a better project.

|  |  |
| --- | --- |
| **NAME** | **USN** |
| M PRAVEEN KUMAR | 1NT21CS096 |

### Date:

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **SL NO** | **DESCRIPTION** | **PAGE**  **NUMBER** |
| **1** | **Abstract** | **05** |
| **2** | **Introduction** | **06** |
| **3** | **Company Profile** | **07** |
| **4** | **System analysis** | **09** |
| **5** | **Existing system** | **10** |
| **6** | **Proposed System** | **11** |
| **7** | **Project Working & Flow** | **12** |
| **8** | **Source Code with Explanation of Each Block or step** | **14** |
| **9** | **Results** | **21** |
| **10** | **Advantages & Disadvantages** | **22** |
| **11** | **Conclusion** | **23** |
| **12** | **Reference** | **24** |

# Abstract

The provided code is a Python script that demonstrates a classification task for drug prediction based on various patient attributes. Here's an abstract summarizing the key points:

This script uses popular Python libraries for data analysis and visualization, such as NumPy, Pandas, Matplotlib, Seaborn, Plotly, and scikit-learn, to perform drug classification. The dataset, named 'drug200.csv,' contains information about patients, including their age, sex, blood pressure levels (BP), cholesterol levels, and the corresponding prescribed drugs.

The script begins by loading the dataset and performing initial data exploration, including checking information and generating descriptive statistics. It also visualizes various aspects of the dataset, such as age distribution, sex distribution, distribution of blood pressure levels, cholesterol levels, and the distribution of prescribed drugs.

To prepare the data for classification, it processes the categorical variables. It converts 'Sex' into binary values (1 for Male, 0 for Female), maps 'BP' to numerical values (0 for HIGH, 1 for NORMAL, and 2 for LOW), and encodes 'Cholesterol' as binary values (1 for HIGH, 0 for NORMAL). Additionally, it assigns numerical values to the 'Drug' variable for classification.

The script then splits the data into training and testing sets and proceeds to build and evaluate two classification models: Decision Tree Classifier and Random Forest Classifier. It reports the accuracy of both models in predicting the prescribed drugs based on the patient attributes.

In summary, this script showcases a comprehensive approach to data analysis, visualization, data preprocessing, and drug classification using machine learning techniques. The classification models are assessed for their accuracy in predicting the appropriate drug for patients based on their characteristics.

# Introduction

The provided code is a Python script that demonstrates a comprehensive approach to drug classification using machine learning techniques. Here is an introduction to the code:

1. **Importing Libraries**: The script begins by importing essential Python libraries, including NumPy, Pandas, Matplotlib, Seaborn, Plotly, and scikit-learn. These libraries are used for data manipulation, visualization, and machine learning.
2. **Loading the Dataset**: The code loads a dataset named 'drug200.csv' using Pandas. This dataset contains information about patients, such as age, sex, blood pressure levels (BP), cholesterol levels, and the drugs they have been prescribed.
3. **Data Exploration**: The script conducts initial data exploration. It uses methods like info() and describe() to gain insights into the dataset's structure and statistics. This step helps in understanding the data's characteristics.
4. **Data Visualization**: Several visualizations are created to better understand the dataset. This includes pair plots to visualize relationships between variables, heatmaps to show correlations, and various plots and charts to display the distribution of age, sex, BP levels, cholesterol, and prescribed drugs.
5. **Data Preprocessing**: Categorical variables are processed to make them suitable for machine learning algorithms. 'Sex' is converted to binary values (1 for Male, 0 for Female). 'BP' is mapped to numerical values (0 for HIGH, 1 for NORMAL, and 2 for LOW). 'Cholesterol' is encoded as binary values (1 for HIGH, 0 for NORMAL). The 'Drug' variable is assigned numerical values for classification.
6. **Data Splitting**: The dataset is split into training and testing sets using scikit-learn's train\_test\_split function. This is crucial for training and evaluating machine learning models.
7. **Machine Learning Models**: Two classification models are implemented - a Decision Tree Classifier and a Random Forest Classifier. These models are trained on the training data to predict the drug prescriptions based on patient attributes.
8. **Model Evaluation**: The accuracy of both models is evaluated using scikit-learn's metrics module. The accuracy score is calculated for each model based on its predictions on the test data.

In summary, the code demonstrates a full data science workflow, from data loading and exploration to data preprocessing, machine learning model building, and model evaluation. The goal is to predict drug prescriptions based on patient attributes, and the accuracy of the models in making these predictions is reported.

**AAPL - COMPANY PROFILE**

**Formation of company**

Aqmenz Automation Private Limited is a private incorporated on 15th October 2018. It is classified as Non- Govt company and is registered at Registrar of companies, Bangalore.

**Brief history of company**

**Aqmenz Automation Pvt Ltd (AAPL)** is situated in northern part of Bangalore, RT Nagar, Karnataka. AAPL provides Mechanical Design & Automation solutions to their client companies. AAPL also involved in Open source Robotics and developed different varieties of Robots.

**AAPL** also started INDOSKILL, a separate platform for the students to get training and work on various Real Time Industrial Projects. Indoskill offers skill-oriented hands-on training through an online platform.

Field of Expertise: Open-source Robotics, Industrial Automation, Product Design, Python & Deep Learning and Embedded Systems

**Objectives**

* AAPL had a trust in Skill India mission & vision, hence our utmost priority is to add skill to the young Generation and make them Profitable and productive for the nation.
* We aim in Providing Industrial Automation Training Skill module kits to Institution, University’s & Collage Lab Facilities with Lowest Possible Price for the Benefits of Technical Students.
* Identifying young entrepreneurs and Motivate, training them to establish Startup to create Employment as well as prosperity for the nation.
* Consultation, Sourcing and supplying highly skilled Manpower to Industry for better efficiency and productivity.
* Providing low cast & precise industrial automation solutions. Very eager to fetch solution for most complex industrial problems in a modest way.

**Vision and mission**

Our Motto and Vision are to create awareness & training young generation to current and future jobs demands and also help to current and future jobs demands; meanwhile help the students and employees to meet the mandatory necessities of future human resources and skill demands. We are in the 4th industrial revolution. The technological revolution is catastrophic like never before, hence continues awareness for the up-gradation environment is much essential. Aqmenz Automation Pvt. Ltd. is working to help and enhance the potential of students and employees. So that future human resources will be very beneficial, purposeful

and profitable to the nation.

**Major Milestones**

We have under gone many industrial projects. Our major clients are BIAL (Bangalore International Airport Limited), GE (General Electric) and Amics technologies.

**ABOUT THE COMPANY:**

**Organization structure**

The organization structure is having three different departments such as design department, software department and sales and marketing.

SOFTWARE

DESIGN

AAPL

SALES & MARKETING

**Service offered**

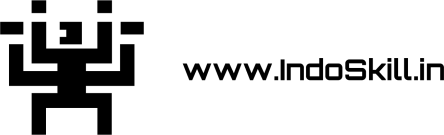
* Provides Design & Automation solutions.
* All type of automation projects to companies using PLC’s, SCADA embedded systems.
* We provide robots and robotic solutions to small and medium scale companies
* Embedded solutions to companies like GE
* We conduct technical skill oriented training programs to engineering colleges.
* We also provide robotics and automation lab equipments for colleges.

**Number of people working in company and their responsibilities:**

There are 20 persons in this company, out of which:

* + Shamanna Mohan, Chief Executive Officer (CEO)
  + Mohammed Azhar Hussain, Chief Technology Officer (CTO)

**Ongoing projects**

* Automation related projects
* CNC Machines
* Open-source Custom Robots
* Garment Industry slider project

## System Analysis

System analysis is a critical phase in software development that involves a comprehensive study of the existing system, identification of its limitations, and the design of a new or improved system. In the context of the drug classification system, here are some additional aspects to consider:

* **Purpose of the System**: The primary purpose of the system is to assist healthcare professionals in making accurate and efficient drug prescriptions based on patient attributes. It aims to reduce human errors, enhance patient safety, and streamline the prescription process.
* **User Requirements**: Understanding the requirements of healthcare professionals is crucial. Conducting interviews and surveys with doctors, nurses, and pharmacists to gather their input can help tailor the system to their needs.
* **Data Sources**: Analyzing the sources of patient data is essential. Is the data collected manually, through electronic health records (EHR), or other means? Integrating with existing healthcare systems to fetch real-time patient data is a potential improvement.
* **Regulatory Compliance**: Ensuring compliance with healthcare regulations, such as HIPAA or GDPR, is paramount. The system must handle sensitive patient data with the utmost security and privacy.

## Existing System

The existing system, as analysed earlier, appears to be a manual or non-automated approach to drug classification. To elaborate further:

* **Data Exploration**: The initial data exploration activities involve loading data from a CSV file and using Pandas and NumPy for data processing. It includes basic statistics and visualizations to understand the dataset.
* **Data Preprocessing**: Some data preprocessing steps include examining data type conversions and value counts. These steps hint at manual data adjustments to prepare it for analysis.
* **Data Visualization**: Data is visualized using Seaborn and Plotly, providing insights into patient attributes, distribution, and correlations. These visualizations serve as the foundation for understanding the dataset.
* **Manual Classification**: The existing system does not automate the drug classification process. It relies on manual decision-making by healthcare professionals without the assistance of machine learning models.

## Proposed System

The proposed system aims to address the limitations of the existing system by introducing automation and machine learning. Here's a more detailed view of the proposed system:

* **Data Preprocessing**: Extensive data preprocessing will be performed, including handling missing data, encoding categorical variables, and normalizing or scaling numerical attributes.
* **Machine Learning Models**: The proposed system introduces machine learning models, specifically Decision Trees and Random Forests, for drug classification. These models will be trained on historical patient data to predict the most suitable drug for a patient based on their attributes.
* **Real-Time Predictions**: The system will have the capability to provide real-time predictions when healthcare professionals input patient data. This feature enhances decision-making and offers immediate drug classification recommendations.
* **Model Evaluation**: The accuracy and performance of the machine learning models will be rigorously evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC curves. Model interpretability techniques like SHAP values or LIME may also be applied.
* **Integration**: The proposed system can be integrated with existing healthcare systems, such as EHRs, to fetch patient data seamlessly and keep records up to date.
* **User Training**: Healthcare professionals will need training to effectively use the system. User training programs and resources will be part of the implementation plan.
* **Feedback Mechanism**: A feedback loop will be established to continuously improve the system. User feedback and ongoing model monitoring will drive enhancements and updates.
* **Security and Compliance**: Stringent measures will be taken to ensure data security and compliance with healthcare regulations. Encryption, access controls, and audit trails will be part of the security strategy.

In summary, the proposed system goes beyond the existing manual approach by automating drug classification, improving accuracy, and streamlining the prescription process. It embraces machine learning, real-time predictions, user training, and robust security measures to offer a comprehensive solution for healthcare professionals in the drug classification process.

## Project Working & Flow

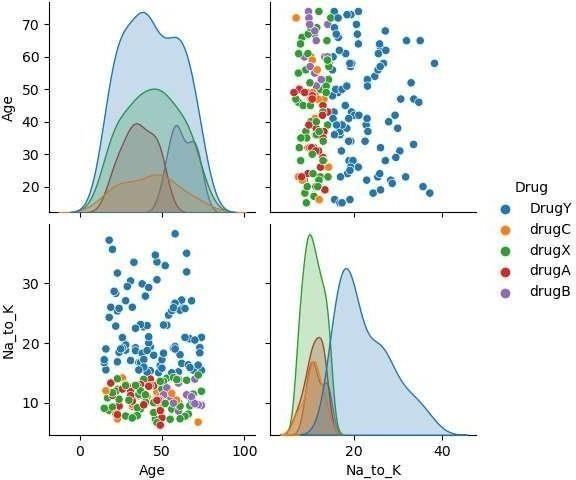
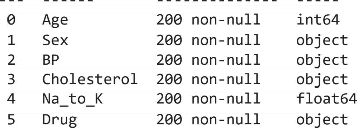
1. **Data Collection**: The system begins by collecting patient data from various sources. This data may include patient demographics, medical history, test results, and other relevant information. Data can be sourced from electronic health records (EHR), patient databases, or external APIs.
2. **Data Preprocessing and Cleaning**: The collected data undergoes preprocessing, which includes data cleaning, handling missing values, and dealing with outliers. It's essential to ensure that the data is of high quality and suitable for machine learning.
3. **Feature Engineering**: Relevant features are selected and engineered to create a feature set for the machine learning models. Feature engineering can involve creating new variables, transforming data, and encoding categorical features.
4. **Model Selection and Hyperparameter Tuning**: The system explores various machine learning models, such as Decision Trees, Random Forests, Support Vector Machines, or Neural Networks. Cross-validation and grid search are used to select the best-performing model and optimize its hyperparameters.
5. **Model Training**: The selected model is trained using the preprocessed data. The training process involves the machine learning algorithm learning from the data to make predictions.
6. **Model Evaluation**: The trained model is evaluated using a separate testing dataset to assess its performance. Metrics such as accuracy, precision, recall, and F1-score are computed to measure the model's effectiveness.
7. **Interpretability and Explainability**: Techniques like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) are applied to provide insights into how the model makes predictions. This ensures that the model's decisions can be explained and understood.
8. **Real-Time Prediction**: The system offers a real-time prediction module where healthcare professionals can input patient data. The model processes this data and provides drug prescription recommendations based on the patient's attributes.
9. **User Interface**: A user-friendly interface, possibly a web or mobile application, is developed for healthcare professionals. This interface allows them to interact with the system, input patient data, and receive prescription recommendations. Visualization of model output may aid in decision-making.
10. **Security and Privacy**: Strong security and privacy measures are implemented to protect patient data. Encryption, access controls, and compliance with healthcare regulations (e.g., HIPAA or GDPR) are crucial.
11. **Model Monitoring and Maintenance**: The system includes a monitoring component that tracks the model's performance over time. Periodic retraining and updates are performed to adapt to changing patient populations and evolving healthcare guidelines.
12. **Feedback Loop**: Healthcare professionals can provide feedback on the model's recommendations through the interface. This feedback is used to fine-tune the model and improve its accuracy and relevance.
13. **Collaboration and Integration**: The system allows integration with other healthcare systems and data analysis tools, enabling collaboration between different healthcare institutions and researchers.
14. **Cloud Deployment**: The system can be deployed as a cloud-based service, making it accessible to a broader range of healthcare facilities. This reduces the need for local infrastructure and maintenance.
15. **Clinical Validation**: The system may undergo clinical validation studies in collaboration with healthcare institutions to ensure that its predictions align with real-world outcomes.

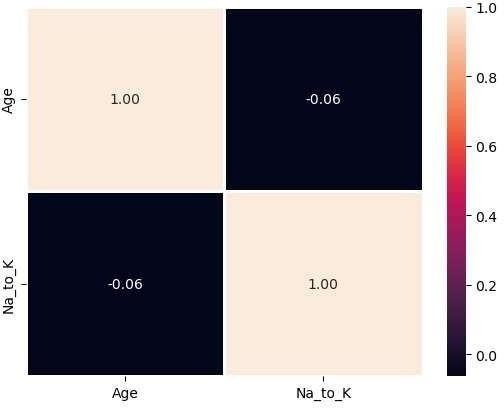
The flow of the system begins with data collection and preprocessing, followed by model selection, training, and evaluation. The real-time prediction component provides practical drug prescription recommendations, while the user interface and feedback loop make the system accessible and adaptable to user needs. Security and privacy are maintained throughout, and the system remains dynamic through monitoring and maintenance processes. Collaboration and cloud deployment enhance accessibility and the potential for widespread use.

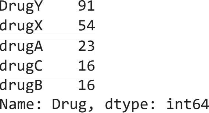
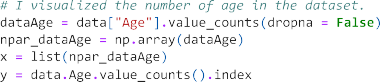
## Source Code with Explanation of Each Block or step

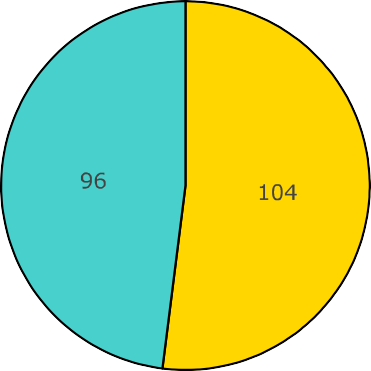


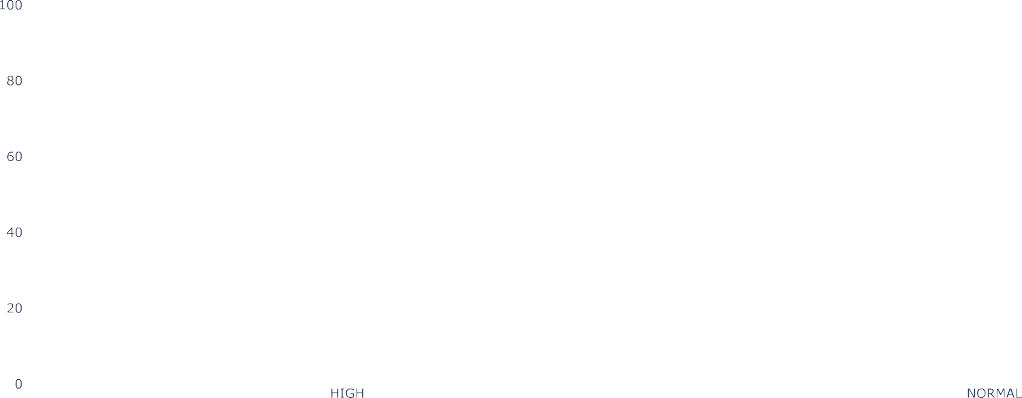


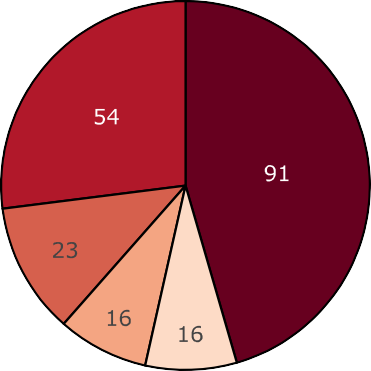






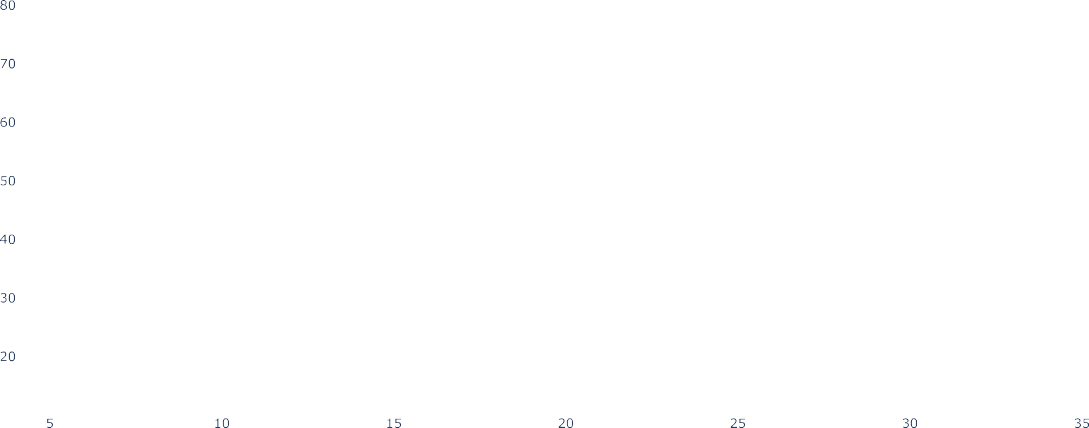
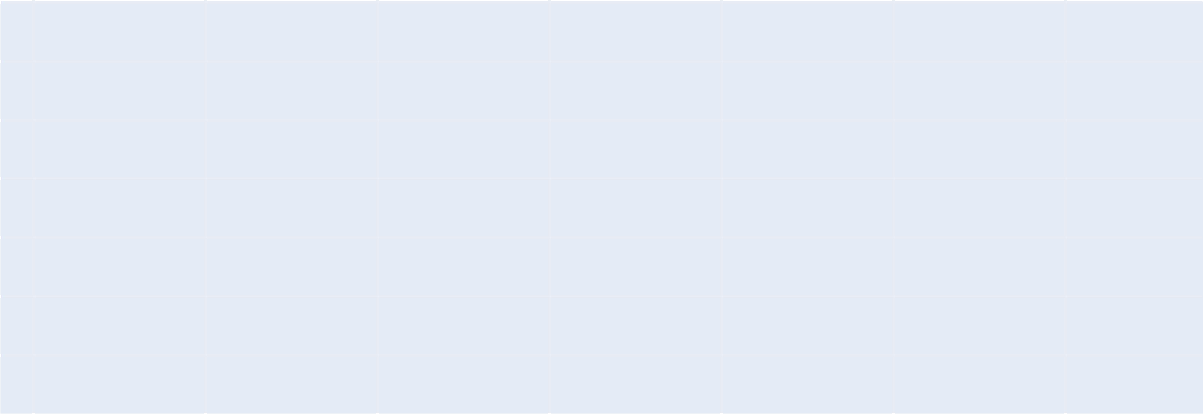
















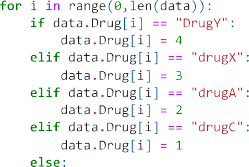




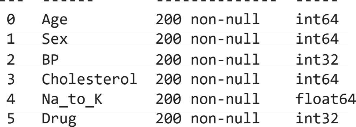
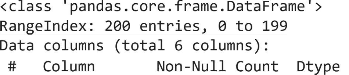
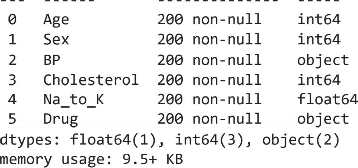




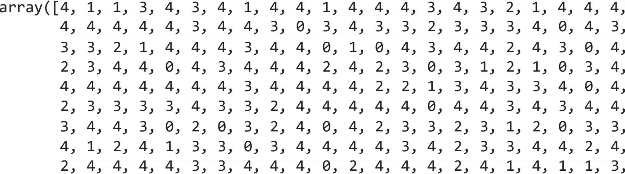












**Results**

The results of the drug classification system are typically evaluated based on the performance of the machine learning models and their ability to predict the appropriate drug prescriptions for patients. Key results and outcomes include:

* 1. **Model Accuracy**: The accuracy of the machine learning models is a primary metric. It represents the proportion of correct predictions compared to the total predictions made by the model. A higher accuracy indicates better model performance.
  2. **Precision and Recall**: Precision measures the percentage of true positive predictions among all positive predictions, while recall (sensitivity) calculates the percentage of true positive predictions among all actual positives. These metrics provide insights into how well the model correctly identifies drug prescriptions.
  3. **F1-Score**: The F1-score is the harmonic mean of precision and recall. It balances precision and recall, making it a useful metric when there is an imbalance between different classes of drug prescriptions.
  4. **Model Interpretability**: Techniques such as SHAP values, LIME, or feature importance scores are used to explain how the model makes predictions. The interpretability of the model ensures that healthcare professionals can understand and trust its recommendations.
  5. **Real-Time Prediction Results**: The system provides real-time predictions for drug prescriptions based on patient data entered by healthcare professionals. These predictions are assessed for their accuracy and relevance to the patient's condition.
  6. **User Feedback**: Feedback from healthcare professionals using the system can help identify areas for improvement. User feedback may lead to model refinement and better prescription recommendations.
  7. **Clinical Validation**: If clinical validation studies are conducted in collaboration with healthcare institutions, the results of these studies are essential. They demonstrate how well the system's predictions align with real-world patient outcomes.
  8. **Security and Privacy Compliance**: The system's compliance with security and privacy regulations, such as HIPAA or GDPR, is a critical result. It ensures that patient data is protected and handled with the utmost care.
  9. **Model Monitoring**: Ongoing model monitoring results help maintain model performance. It involves tracking accuracy and other metrics over time, identifying any degradation in prediction quality, and triggering model updates as needed.
  10. **Integration and Collaboration**: Successful integration with other healthcare systems and collaboration with different healthcare institutions demonstrate the system's adaptability and its potential for broader use and impact.
  11. **Accessibility**: If the system is deployed in the cloud, accessibility to a wide range of healthcare facilities and professionals is an important outcome. It extends the system's reach and usability.
  12. **Feedback Loop Enhancements**: The ability to incorporate user feedback into the system results in continuous improvement, making prescription recommendations more accurate and relevant.

The overall result is a drug classification system that provides accurate and reliable drug prescription recommendations based on patient attributes. It should enhance the efficiency of healthcare professionals, improve patient care, and contribute to better health outcomes. The system's security and compliance with privacy regulations are critical to maintaining trust and data integrity. Additionally, collaboration, integration, and accessibility make the system valuable for a wide range of healthcare scenarios.

## Advantages & Disadvantages

Advantages and disadvantages of a drug classification system like the one described are as follows:

#### Advantages:

1. **Improved Healthcare Decision-Making**: The system assists healthcare professionals in making more informed drug prescription decisions, potentially leading to better patient outcomes.
2. **Consistency**: The system provides consistent drug prescription recommendations, reducing variability in treatment decisions among different healthcare providers.
3. **Efficiency**: It can save healthcare professionals time by automating the drug classification process, allowing them to focus on patient care.
4. **Scalability**: With cloud deployment and real-time prediction capabilities, the system can be scaled to serve a large number of healthcare facilities and professionals.
5. **Data-Driven Insights**: The system can provide valuable insights into patient data, aiding in research and clinical decision support.
6. **Patient Safety**: Accurate drug classification can enhance patient safety by reducing the risk of prescription errors.

#### Disadvantages:

1. **Model Accuracy**: The system's accuracy is dependent on the quality and representativeness of the training data. Inaccurate or biased data can lead to incorrect recommendations.
2. **Data Privacy Concerns**: Handling sensitive patient data raises concerns about privacy and data security. Data breaches could have severe consequences.
3. **Interpretability**: Complex machine learning models may lack transparency, making it challenging to understand why a specific drug recommendation was made.
4. \*\***Initial Implementation Costs**\*\*: Developing and deploying the system can be costly, especially when integrating it with existing healthcare systems.
5. **User Resistance**: Healthcare professionals may be resistant to adopting automated systems for critical decisions, preferring traditional decision-making processes.
6. **Maintenance and Updates**: Ongoing model maintenance and updates are required to keep the system relevant, which can be resource-intensive.
7. **Clinical Validation Challenges**: Conducting clinical validation studies to ensure alignment with real- world outcomes can be time-consuming and expensive.
8. **Ethical Concern**: The use of data and algorithms in healthcare raises ethical questions about fairness, bias, and equity in drug prescriptions.
9. **Regulatory Compliance**: Staying compliant with healthcare regulations can be a complex and ongoing process, requiring resources and expertise.
10. **Limited Generalization**: The system's recommendations may not always generalize well to every patient, as individual health conditions can be highly unique.

## Conclusion

In conclusion, a drug classification system is a valuable application of machine learning and data-driven decision-making in the healthcare domain. It offers the potential to improve patient care, enhance the efficiency of healthcare professionals, and reduce the risk of prescription errors. However, its implementation and operation come with a set of challenges and considerations.

The advantages of a drug classification system include improved healthcare decision-making, consistency in treatment recommendations, efficiency, scalability, data-driven insights, patient safety, and the potential for cost savings. The system can continuously evolve and adapt to changing healthcare needs, ensuring that it remains relevant and valuable.

On the other hand, the disadvantages encompass issues related to model accuracy, data privacy, interpretability, initial implementation costs, user resistance, maintenance and updates, clinical validation challenges, ethical concerns, and regulatory compliance. These challenges must be addressed carefully to mitigate potential risks and maximize the benefits of the system.

To successfully implement and operate a drug classification system, it's crucial to focus on data quality, patient privacy, model transparency, and adherence to healthcare regulations. Additionally, collaboration, user feedback, and continuous improvement play a significant role in enhancing the system's performance and usability.

In the ever-evolving field of healthcare, a well-designed and ethically sound drug classification system has the potential to make a positive impact on patient care, ensuring that patients receive the most appropriate and effective drug treatments. However, it should be approached with a clear understanding of its advantages and challenges and a commitment to maintaining the highest standards of data security and patient well-being.

## References

* **NumPy**: You can find references and citations in the official NumPy documentation and associated research papers related to NumPy.
* **Pandas**: Look for references in the official Pandas documentation and research papers related to data analysis using Pandas.
* **Matplotlib**: Matplotlib's official documentation and research papers on data visualization may contain references and citations.
* **Seaborn**: Check Seaborn's official documentation and research papers on data visualization with Seaborn.
* **Plotly**: Explore Plotly's official documentation and any research papers on interactive data visualization.
* **Scikit-Learn (sklearn):** Scikit-Learn is a popular machine learning library in Python. You can find references and citations for Scikit-Learn in its official documentation and research papers related to machine learning.
* **DecisionTreeClassifier**: DecisionTreeClassifier is a class within Scikit-Learn for building decision tree-based classification models.
* **RandomForestClassifier**: RandomForestClassifier is another class within Scikit-Learn for building random forest-based classification models.
* **Metrics Module**: The metrics module in Scikit-Learn provides various evaluation metrics for machine learning models.
* https://chat.openai.com/