Drug Type Classification Project Report

1. Introduction

This project analyzes the Drug Type Classification Dataset to predict the type of drug prescribed based on patient characteristics, including age, sex, blood pressure (BP), cholesterol levels, and sodium-to-potassium ratio (Na_to_K). The objective is to classify the drug type (Drug Y, drug A, drug B, drug C, drug X) and evaluate the predictive performance of machine learning models. The dataset, sourced from Kaggle, contains 200 records and 6 columns, with ”Drug” as the categorical target variable. The analysis involves data cleaning, exploratory data analysis (EDA), feature engineering, and classification using Logistic Regression and Decision Tree Classifier.

2. Methodology

2.1 Data Loading and Initial Exploration

• **Dataset:** Loaded using pandas from ”drug200.csv” with 200 records and 6 columns: Age (int64), Sex (object), BP (object), Cholesterol (object), Na_to_K (float64), and Drug (object).

• **Initial Checks:**

- ✓ No missing values were found (df.isnull().sum()).
- ✓ One duplicate record was identified and removed, reducing the dataset to 192 records (df.drop_duplicates()).
- ✓ Data types: Numerical (Age, Na_to_K) and categorical (Sex, BP, Cholesterol, Drug).
- ✓ Summary statistics (df.describe()) revealed:
  - ▪ Age: Mean ~44.32 years, range 15–74 years.
  - ▪ Na_to_K: Mean ~16.08, range 6.27–38.25.

2.2 Data Preprocessing

• **Handling Duplicates:** Remove done duplicate record to ensure data integrity.

• **Categorical Encoding:** Converted categorical variables (Sex, BP, Cholesterol) to numerical values using Label Encoding.

• **Feature Scaling:** Standardized numerical features (Age, Na_to_K, Sex, BP, Cholesterol) using StandardScaler to ensure uniformity for modeling.

2.3 Exploratory Data Analysis (EDA)

• **Univariate Analysis:**

- ✓ Categorical Variables: Analyzed Sex, BP, Cholesterol, and Drug distributions.
- ✓ Numerical Variables: Examined Age and Na_to_K distributions using histograms and box plots.

• **Bivariate Analysis:** Explored relationships between features (e.g., Age vs. Drug, Na_to_K vs. Drug) to identify patterns influencing drug type.

2.4 Modeling

• **Data Splitting:** Split data into training (80%, 153 records) and testing (20%,39 records) sets using train_test_split with random_state=40.

• **Models Used:**

Logistic Regression:

- ✓ Trained on standardized features to predict drug type.
- ✓ Evaluation metrics:
  - o Accuracy: 92.31%.

- Classification report: High precision and recall for most classes , with Drug Y(0.89 Precision, 0.94 recall), drug A (1.00,1.00),drug B(0.67,100),drug C(1.00,0.83), and drug X(1.00,0.90).

Decision Tree Classifier:

- ✓ Trained on the same dataset.
- ✓ Evaluation metric:
  - Accuracy: 100%

3. Results

• Data Characteristics:

- ✓ The dataset is clean with no missing values and minimal duplicates.
- ✓ Key features influencing drug type include Na_to_K, BP, and Cholesterol ,with Age and Sex showing less impact.

• Model Performance:

- ✓ **Logistic Regression:** Achieved an accuracy of 92.31%, with strong performance across most drug classes but slightly lower precision for drug B(0.67).
- ✓ **Decision Tree Classifier:** Achieved perfect accuracy (100%), indicating potential over fitting or a highly separable dataset.

4.Discussion

• Key Insights:

- ✓ Na_to_K ratio, BP, and Cholesterol are critical predictors of drug type, with higher Na_to_K ratios often associated with Drug Y.
- ✓ Age and Sex have minimal impact on drug classification, suggesting physiological factors dominate.

• Model Comparison:

- ✓ Decision Tree Classifier out performed Logistic Regression,achieving100% accuracy, likely due to its ability to capture non-linear decision boundaries.
- ✓ Logistic Regression's 92.31% accuracy is robust but less effective for complex patterns, as seen with drug B's lower precision.
- ✓ The perfect accuracy of the Decision Tree suggests potential over fitting, which could be validated with cross-validation or a larger test set.

• Data Quality:

Pre processing(duplicate removal, encoding, standardization)ensured reliable model inputs.

5. Conclusion

The analysis successfully classified drug types using patient characteristics, with the Decision Tree Classifier achieving perfect accuracy (100%) and Logistic Regression achieving 92.31%.Na_to_K , BP, and Cholesterol emerged as key predictors, while Age and Sex had less influence. The Decision Tree's performance suggests a highly separable dataset, but caution is needed to avoid over fitting.

Future work could include cross-validation, additional models (e.g., Random Forest), or feature engineering to enhance generalization. These findings can support medical decision-making by predicting appropriate drug types based on patient profiles.

6. References

• Dataset: ”drug200.csv” from Kaggle.

• Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn