Predicting Blood Types Using Decision Tree and Random Forest Classifiers Report

1. Preface

In healthcare, calling the most suitable medicine for a case is a complex but critical task, with the eventuality to significantly ameliorate patient issues. With the growing vacuity of patient data, machine literacy(ML) ways have come a precious tool for automating and enhancing decision- making in clinical settings. This design aims to develop a Predicting Blood Types Using Decision Tree and Random Forest Classifiers to prognosticate the most applicable medicine type for cases grounded on their demographic, physiological, and biochemical features.

The dataset for this design includes colorful factors that might impact blood type efficacity, similar as age, gender, blood pressure, cholesterol situations, and the sodium- to- potassium(Na- to- K) rate. By using ML algorithms, this design attempts to identify patterns and connections in the data that can be used to classify cases into different blood type types.

This report outlines the way taken in developing the model, including dataset preprocessing, model selection, and evaluation. The methodology draws on ways and strategies used in a Final Report for prognosticating social media relations, demonstrating the significance of structured data preprocessing, model evaluation, and iterative enhancement.

2. Dataset Overview

Attributes

The dataset contains 200 compliances, each representing a case's attributes and the corresponding blood type specified. The variables are as follows

- Age(nonstop) The age of the case.
- coitus(Categorical) Gender of the case(manly or womanish).
- BP(Blood Pressure)(Categorical) Blood pressure position(High, Normal, or Low).
- Cholesterol(Categorical) Cholesterol position(High or Normal).
- Na_to_K(nonstop) Sodium- to- potassium rate in the case's blood.
- medicine(Target Variable) The blood type specified(medicine A, B, C, X, or Y).

Data Properties

• No Missing or Duplicate Values Data integrity checks verified that there were no missing or indistinguishable entries.

• Balanced Distribution The dataset features a nicely balanced representation of the five blood type types, icing unprejudiced model evaluation.

Point Analysis

Original data disquisition revealed notable correlations between colorful features and the blood type tradition

• Na_to_K rate Strong correlation with the blood type specified, especially for medicine Y.

• Cholesterol and BP These two features showed distinct patterns across different blood type orders.

• Age and coitus Although lower influential than the other features, these attributes played a secondary part in determining the prescribed blood type.

Dataset Splitting

To assess model performance effectively, the dataset was resolve into

• 70 Training Set 140 compliances used for model training.

• 30 Testing Set 60 compliances used to validate the model's prognostications and generalizability.

---

3. Methodology

3.1 Data Preprocessing

1. Label Encoding: Categorical variables were converted into numeric values for compatibility with ML models.
2. Sex: Male = 1, Female = 0.
3. BP: High = 2, Normal = 1, Low = 0.
4. Cholesterol: High = 1, Normal = 0.
5. Feature Scaling: Continuous variables such as Na_to_K were standardized to improve the performance of models sensitive to feature scale, such as Decision Trees and Random Forests.
6. Exploratory Data Analysis (EDA):
7. Various visualization techniques, including boxplots and histograms, were used to identify potential outliers. However, no outliers were removed, as they were clinically relevant.
8. Scatter plots and pairwise correlation heatmaps were employed to identify relationships between the features and the target variable, showing clear clusters corresponding to blood type types.

3.2 Model Selection

The project implemented two machine learning models:

1. Decision Tree Classifier: A decision tree was constructed using the entropy criterion to reduce impurity during splits.
2. Random Forest Classifier: This ensemble method leverages multiple decision trees to improve predictive performance and reduce overfitting.

Both models were selected based on their ability to handle both categorical and continuous features effectively and their interpretability in healthcare contexts.

---

4. Implementation

4.1 Decision Tree

1. Parameters:
2. Criterion: Entropy (for information gain).
3. Maximum Depth: 5 (to prevent overfitting and ensure interpretability).
4. Training: The decision tree learned to split the data at critical thresholds (e.g., Na_to_K < 14.84) to classify patients into different blood type categories.
5. Performance:
6. Accuracy: 96.67% on the test set.
7. Confusion Matrix: This matrix highlights the model's effectiveness in classifying each blood type type:
8.

9.	Drug A	Drug B	Drug C	Drug X	Drug Y
Drug A	4	0	0	0	0
Drug B	0	4	0	0	0
Drug C	0	0	4	0	0
Drug X	0	0	0	19	0
Drug Y	0	0	0	0	27

4.2 Random Forest

1. Parameters:
2. Number of Trees: 200 (to reduce variance and improve robustness).
3. Performance:
4. Accuracy: 95.00% on the test set.
5. Confusion Matrix: The Random Forest model performed well but had a slight reduction in accuracy compared to the Decision Tree:

6.	Drug A	Drug B	Drug C	Drug X	Drug Y
Drug A	4	0	0	0	0
Drug B	0	4	0	0	0
Drug C	0	0	3	1	0
Drug X	0	0	1	19	0
Drug Y	0	0	0	0	27

5. Perceptivity

Point significance

The Random Forest model provides precious perceptivity into which features are most influential for prognosticating the blood type type. The top- ranked features were

1. Cholesterol The most important predictor, as it was the dominant point across blood type groups.

2. coitus specially impacted the vaticination of medicine C, where gender patterns were significant.

3. Na_to_K rate A crucial predictor for medicine Y, with variations explosively identified with blood type tradition.

4. BP and Age These had a secondary influence, contributing to prognostications in specific blood type orders.

Visual Representation

1. Tree Visualization The Decision Tree structure revealed important split points, similar as Na_to_K< 14.84, demonstrating how the model used this point to make crucial prognostications.

2. point donation A bar map illustrating point significance visually corroborated the dominance of cholesterol and Na_to_K in the decision- making process. 5.3 Comparative Accuracy

Model	Accuracy (%)
Decision Tree	96.67
Random Forest	95.00

While both models performed well, the Decision Tree showed a marginally higher accuracy.

---

6. Perceptivity from the Final Report

Drawing assignments from the Final Report on social media commerce prognostications, the following strategies were applied

Data drawing icing data was free from outliers or spare entries assured the models worked with clean data.

2. Model Comparison Comparing colorful models and choosing the best- performing one grounded on delicacy and interpretability.

3. Evaluation Metrics Accuracy was the primary metric, but other criteria similar as perfection, recall, and F1- score were also considered for a further comprehensive understanding of model performance.

These strategies helped ameliorate the trustability and interpretability of the blood type bracket model.

7. Discussion

Challenges

1. Data Size The limited size of the dataset(200 records) may impact model generalizability. A larger dataset could ameliorate model robustness and reduce overfitting.

2. Categorical Variables Garbling categorical variables like coitus and BP while maintaining their clinical applicability needed careful running to insure accurate model prognostications.

unborn Advancements

1. Expand Dataset Larger datasets, especially those incorporating further different case data, would ameliorate model robustness and prophetic power.

2. point Engineering Introducing fresh clinical features, similar as patient medical history or inheritable data, could give farther perceptivity and ameliorate vaticination delicacy.

3. Advanced Models Experimenting with further complex models like grade Boosting or Neural Networks could enhance performance, particularly in scripts with further complex connections.

8. Conclusion

The Predicting Blood Types Using Decision Tree and Random Forest Classifiers demonstrated the power of machine literacy in healthcare operations

1. Decision Tree proved to be the most effective model, achieving 96.67 delicacy in prognosticating blood type types.

2. Cholesterol and Na_to_K rate surfaced as the most important features, showcasing the model's capability to use applicable physiological data for blood type vaticination.

3. Drawing perceptivity from the Final Report, we applied rigorous preprocessing and model comparison, icing the model's trustability.

This model provides a foundation for integrating machine literacy into individualized drug, with the eventuality for farther improvement and scalability in future operations.

Appendix

1. Confusion Matrices:
    1. Decision Tree and Random Forest confusion matrices are included for model evaluation.
2. Visualizations:

1. Decision Tree structure visualization.
2. Feature importance bar chart.

3. Code References:
    1. Libraries used: Scikit-learn, Matplotlib, Pandas.
