MedTrackPro

Exploring Healthcare Data for Precision Medicine

DONE BY

Ashwin K

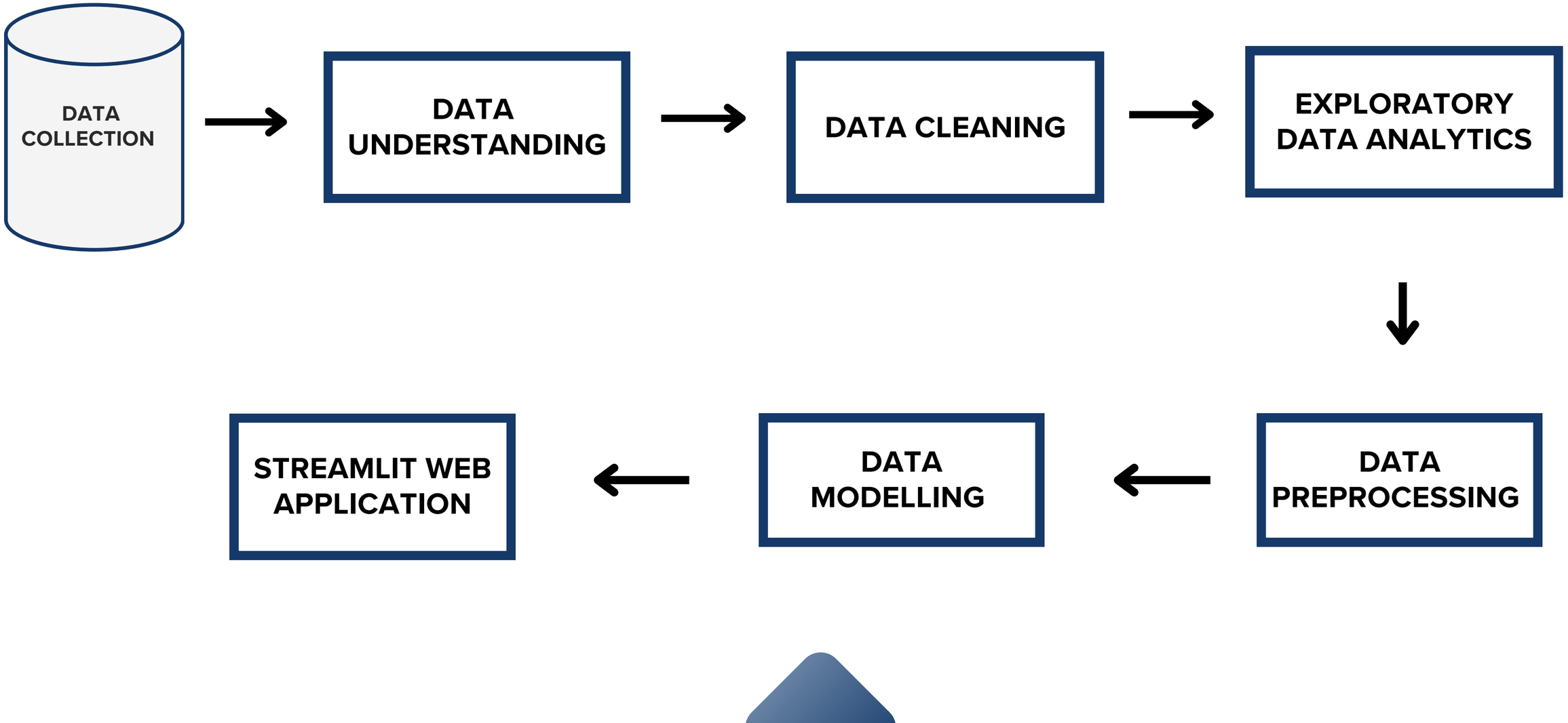
Agenda

1. PROJECT DESCRIPTION
2. PROJECT APPROACH
3. TECH - STACK USED
4. MODULES
5. PROJECT OUTCOMES
6. RESULTS

PROJECT DESCRIPTION

- The Project aims to develop a predictive model for medication prescription patterns using healthcare data.
- Conducting thorough data analysis to understand distributions, correlations, and patterns and preprocess data to create relevant features for medication modeling.
- Utilizing machine learning techniques to build predictive models and evaluate their performance in predicting medication prescriptions accurately.

PROJECT APPROACH



TECH - STACK USED

JUPYTER NOTEBOOK

PYTHON

STREAMLIT

MODULES

1. DATA COLLECTION AND UNDERSTANDING
2. DATA CLEANING
3. EXPLORATORY DATA ANALYSIS (EDA)
4. DATA PREPROCESSING
5. DATA MODELLING
6. FINAL IMPLEMENTATION

DATA COLLECTION AND UNDESTANDING

- At first, the dataset is fetched from Kaggle and then imported into a variable named df using the Python library, pandas.
- The dimensionality of the dataframe is determined using Python code, revealing that it comprises 5000 rows and 7 columns.
- The information regarding the column name, number of non-null values, count, and data types of the data frame is presented.

- Printing the columns of the dataframe reveals that it comprises the following: Name, Age, Blood Type, Gender, Test Result, Disease, and Medication.

- The count, mean, standard deviation, minimum, 25th percentile, 50th percentile, 75th percentile, and maximum values of the numerical column (age) within the dataframe are displayed.

- The count, number of unique values, most common value, and its frequency for categorical columns (Name, Blood Type, Gender, Test Result, Disease, and Medication) are displayed.

- A list of all unique values within each column of the dataframe is listed.

DATA CLEANING

- Upon inspection, it is determined that the dataframe contains unnecessary columns, with the "Name" column identified as non-essential. Consequently, the "name" column is removed from the dataframe.
- As the name column is deemed unnecessary for our model, it is excluded from the dataframe, while all other columns deemed essential are retained without alteration.
- The dataframe is inspected for null values, and it is confirmed that no null values are present within it.

- Upon examination, it was discovered that the dataframe contains 178 duplicate values. Subsequently, these duplicates were eliminated from the dataframe.

BEFORE DATA CLEANING	AFTER DATA CLEANING
ROWS : 5000 COLUMNS : 7	ROWS : 4822 COLUMNS : 6

- The age column in the dataframe is examined for outliers, and it is determined that no outliers are present.

EXPLORATORY DATA ANALYTICS

- The most commonly occurring value in each column of the dataframe is identified.
- The p-values were derived by conducting chi-squared tests to assess the independence between each predictor variable (Age, Blood Type, Gender, Test Result, Disease) and the target variable (Medication).
- Age, Test Result, and Disease exhibit statistically significant associations as independent variables, whereas blood type and gender show non-statistically significant associations.
- A histogram illustrating age and frequency indicates that individuals aged 68 are more susceptible to illness compared to other age groups.

- A horizontal bar chart depicting gender and count reveals a higher number of males affected by illness compared to females.

- A pie chart illustrates the percentage of individuals affected by illness based on blood groups. It is apparent that individuals with AB+ blood type are predominantly affected by the virus, while those with B- blood type are least affected.

- A horizontal bar chart displaying the count of people affected by various diseases indicates that Impetigo affects the highest number of individuals, while Hepatitis A affects the least.

- A bar chart depicting the test result and the count of disease reveals that the majority of test results are normal, while a smaller proportion are abnormal.

DATA PREPROCESSING

- Two variables are instantiated: X, which stores the categorical features (Disease, Test Result) and numerical feature (age), and y, which stores the target variable (Medication).
- The categorical column values are encoded into binary format using one-hot encoding.
- The dataframe is partitioned into training and testing datasets, denoted as X_train, y_train, and X_test, y_test, respectively.
- The training data and testing data are typically split in a ratio of 70% for training and 30% for testing.

DATA MODELLING

- For this task, we have chosen four machine learning algorithms: decision tree classifier, random forest classifier, K-nearest neighbor, and support vector machines.

- The training accuracy, cross-validation score, mean cross-validation score, and standard deviation of cross-validation scores for each algorithm are computed and presented.

MODEL SCORE COMPARISION

ALGORITHM	TRAIN ACCURACY	CROSS VALIDATION SCORES
DESISION TREE	97.86	0.97037037 0.97333333 0.98666667 0.98962963 0.97777778
RANDOM FORESRT	97.86	0.97333333 0.97481481 0.98814815 0.98814815 0.96888889

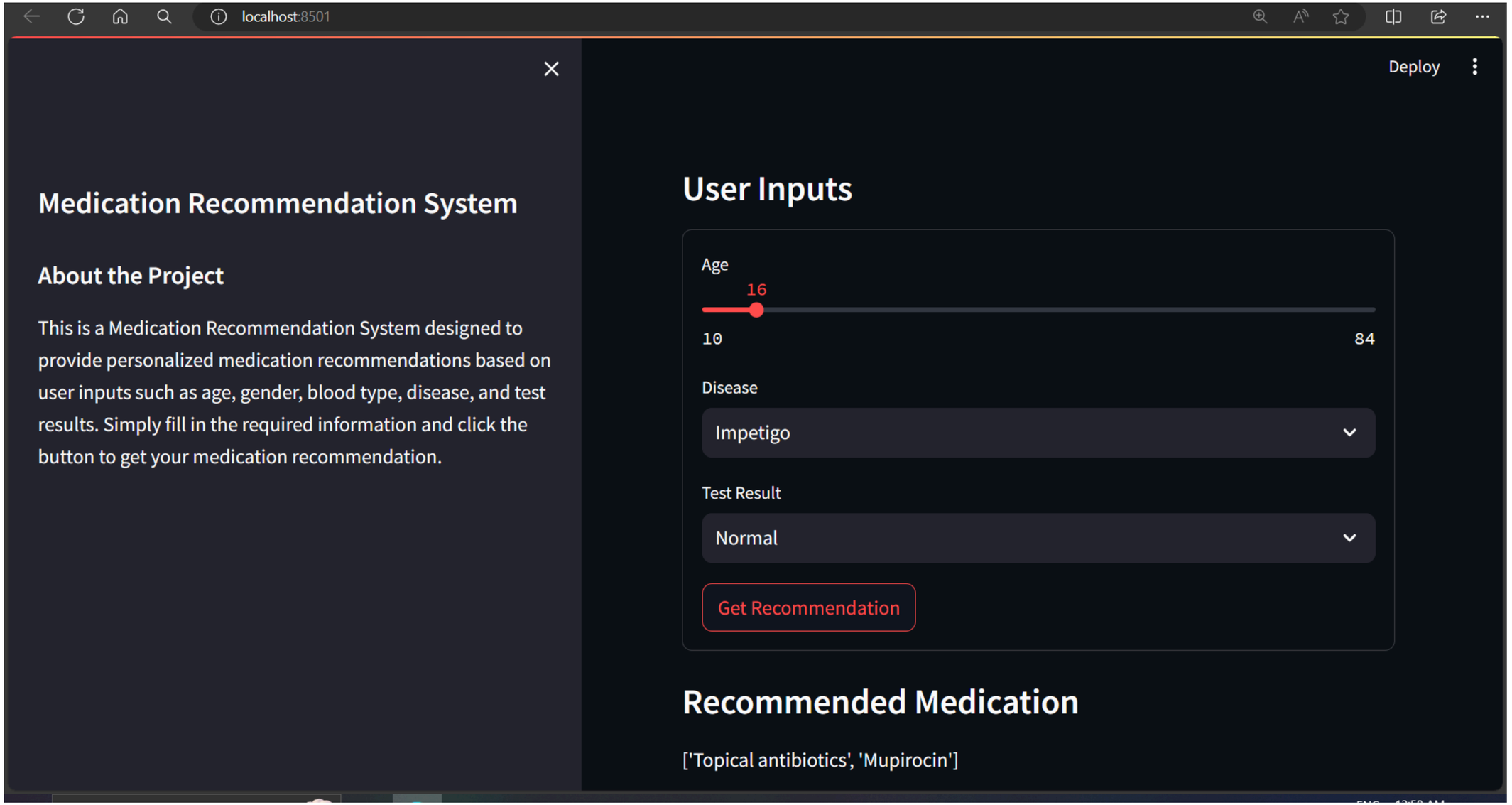
ALGORITHM	TRAIN ACCURACY	CROSS VALIDATION SCORE
K NEAREST NEIGHBOUR	0.48	0.48888889 0.45481481 0.47555556 0.49333333 0.50222222
SUPPORT VECTOR MACHINES	55.22	0.97037037 0.97333333 0.98666667 0.98962963 0.97777778
LOGISTIC REGRESSION	80.86	0.81777778 0.80444444 0.8237037 0.8237037 0.78518519

- Upon comparison of all four models, it is evident that the decision tree classifier exhibits higher accuracy and cross-validation scores compared to the other three algorithms. Therefore, it is selected for our project.

FINAL IMPLEMENTATION

- The code for the Streamlit web application is written in a manner that ensures it is easily comprehensible for the user.
- This web application is crafted to receive user input parameters such as age, disease, and test results, forwarding them to the decision tree classifier for prediction.
- Once the decision tree classifier predicts the medication, it is showcased to the user via the Streamlit web application.

PROJECT OUTCOMES



RESULT

- By considering age and blood type along with test results, our model can offer more targeted medication recommendations compared to a one-size-fits-all approach.
- This user-friendly system empowers individuals to take a more active role in their healthcare by providing personalized medication suggestions based on their specific characteristics.
- This user-friendly application can bridge the gap between patients and medication knowledge.

THANK YOU