THYROID DISEASE DETECTION

Detailed Project Report

Ashwini Kakde
Data Science Intern at Ineuron.ai

INTRODUCTION

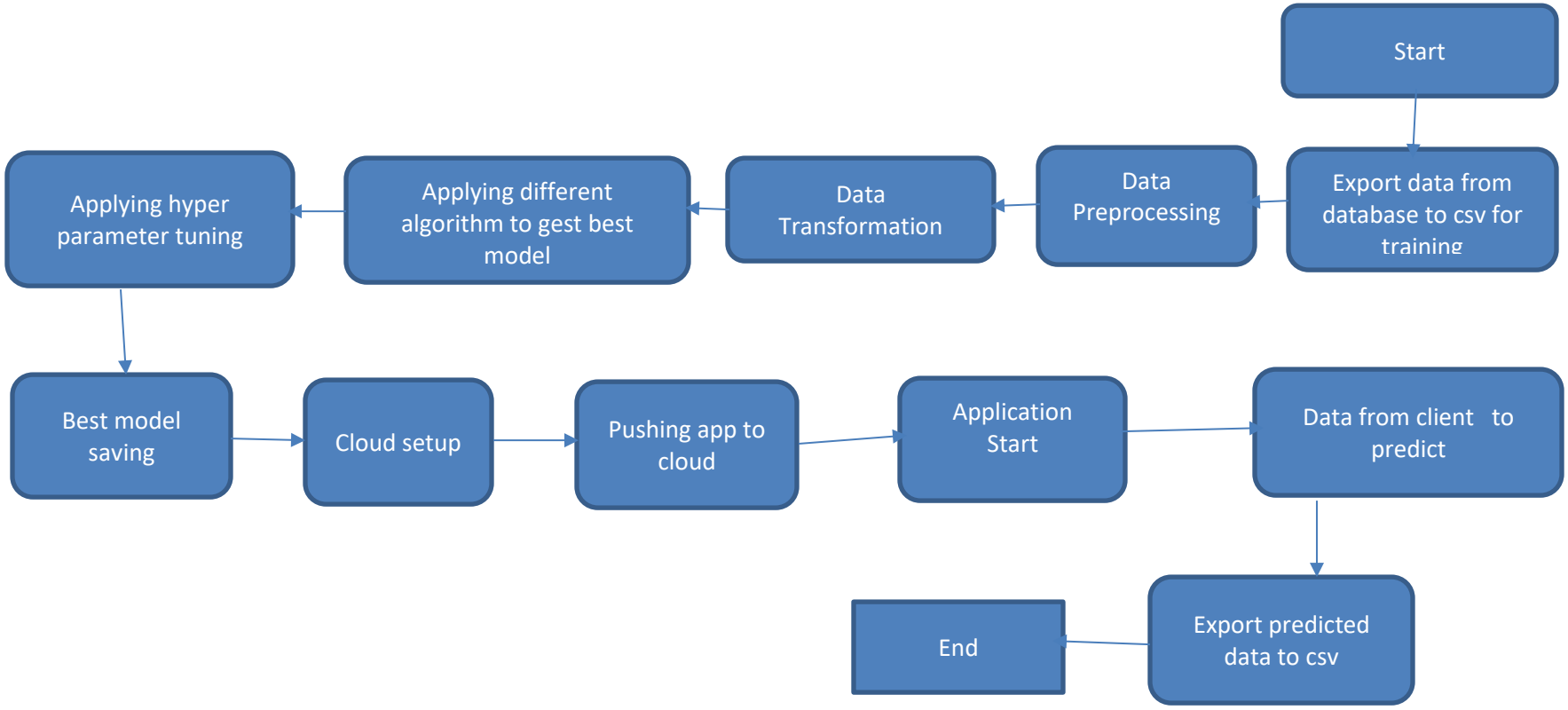
At least a person out of ten is suffered from thyroid disease in India. The disorder of thyroid disease primarily happens in the women having the age of 17–54. The extreme stage of thyroid results in cardiovascular complications, increase in blood pressure, maximizes the cholesterol level, depression and decreased fertility. The hormones, **total serum thyroxin (T4)** and **total serum triiodothyronine (T3)** are the two active thyroid hormones produced by the thyroid gland to control the metabolism of body. For the functioning of each cell and each tissue and organ in a right way, in overall energy yield and regulation and to generate proteins in the ordnance of body temperature, these hormones are necessary.

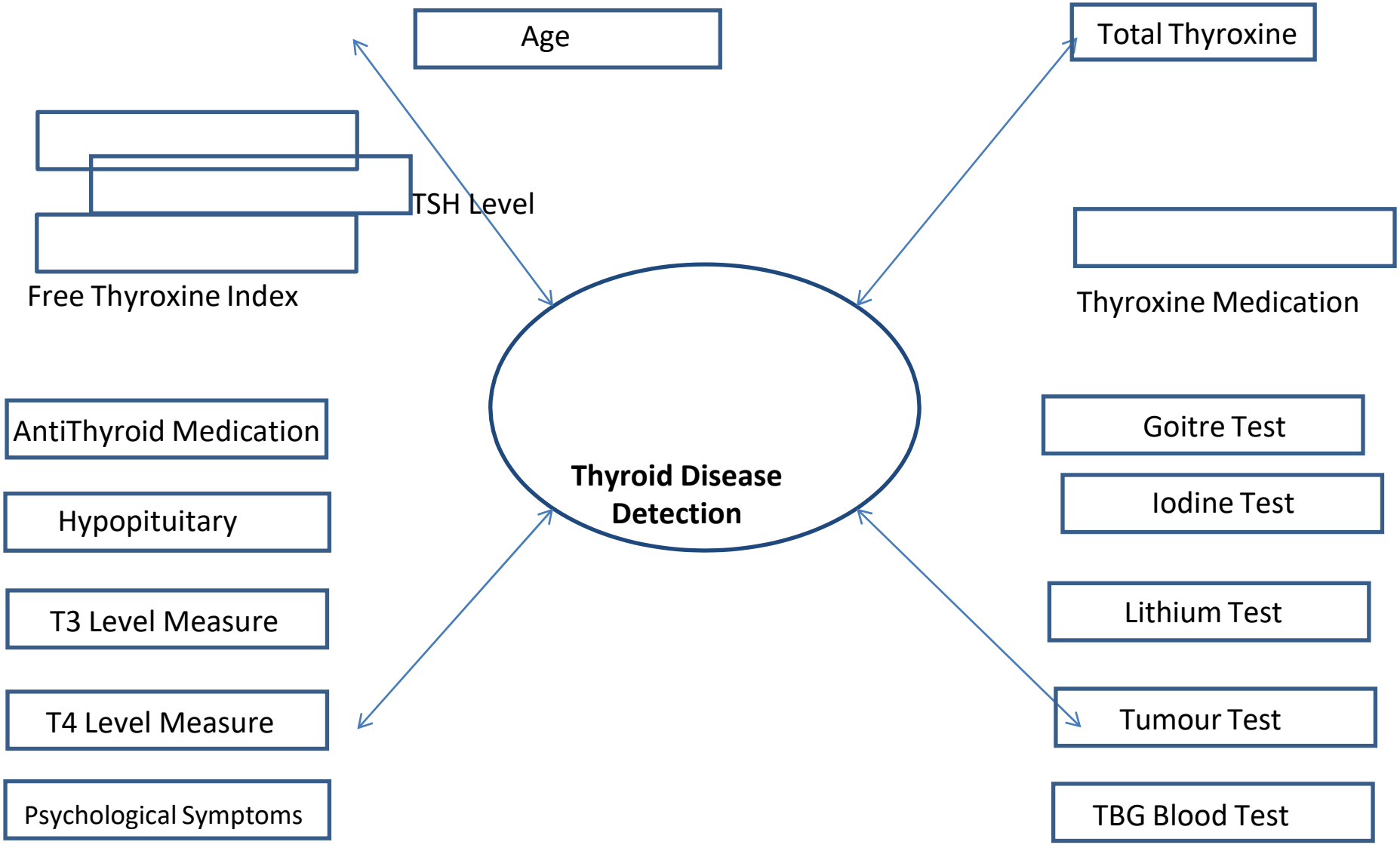
Hyperthyroidism and **Hypothyroidism** are the most two common diseases caused by irregular function of thyroid gland. Thyroid disorder can speed up or slow down the metabolism of the body. In the world of rising new technology and innovation, health care industry is advancing with the role of Artificial Intelligence. Machine learning algorithms can help to early detection of the disease and to improve the quality of the life. This study demonstrates the how different classification algorithms can forecasts the presence of the disease. Different classification algorithms such as Logistic regression, Random Forest, Decision Tree, Naïve Bayes, Support Vector Machine, XG Boost, KNN have been tested and compared to predict the better outcome of the model.

OBJECTIVE

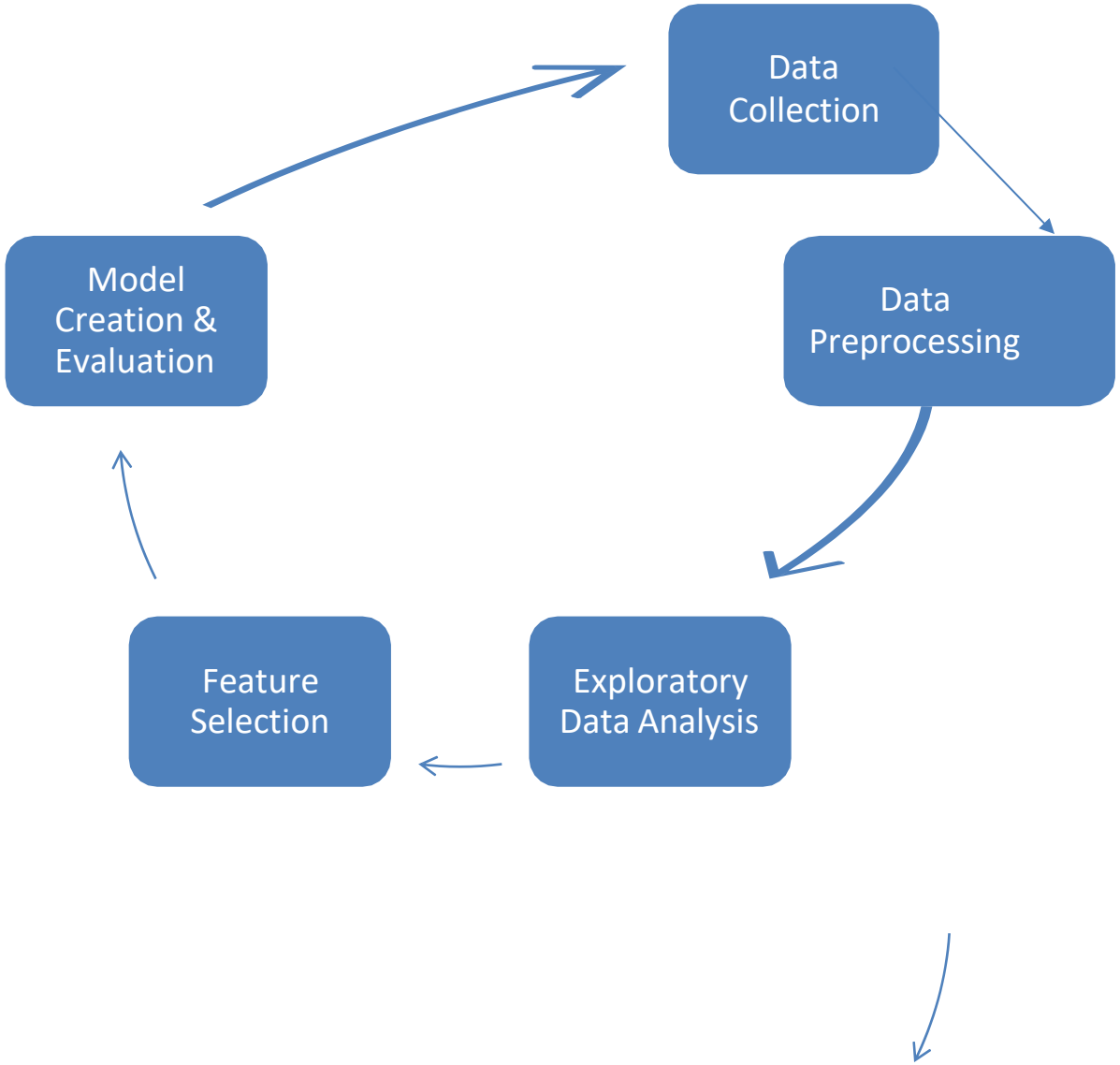
The main goal of this project is to predict the risk of hyperthyroid and hypothyroid based on various factors of individuals. Thyroid disease is a common cause of medical diagnosis and prediction, with an on set that is difficult to forecast in medical research. It will play a decisive role in order to early detection, accurate identification of the disease and helps the doctors to make proper decisions and better treatment.

ARCHITECTURE

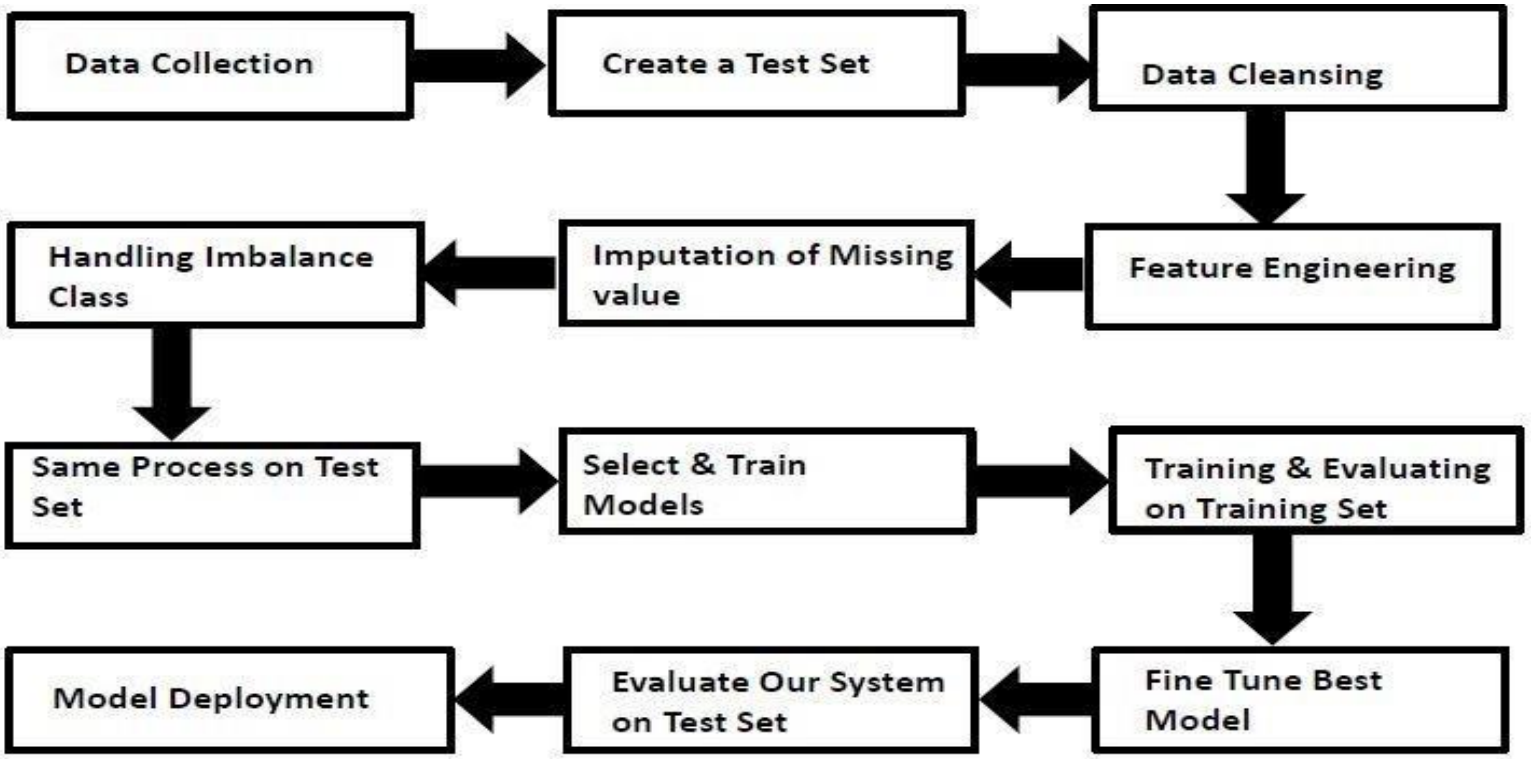



Age
Total Thyroxine
TSH Level
Free Thyroxine Index
Thyroxine Medication
Thyroid Disease Detection
AntiThyroid Medication
Goitre Test
Hypopituitary
Iodine Test
T3 Level Measure
Lithium Test
T4 Level Measure
Tumour Test
Psychological Symptoms
TBG Blood Test

Data Analysis Steps



MODEL TRAINING AND VALIDATION WORKFLOW



MODEL TRAINING AND VALIDATION WORKFLOW

Data Collection

- Thyroid Disease Data Set from UCI Machine Learning Repository
- For Data Set: https://archive.ics.uci.edu/ml/datasets/thyroid+disease

Data Pre-Processing

- Missing values handling by Simple imputation by most frequent values for categorical data and By median values for numerical data
- Categorical features handling by one hot encoding and label encoding for target values
- Feature scaling done by robust Scalar method
- Imbalanced dataset handled by randomoversampler
- Drop unnecessary columns

MODEL TRAINING AND VALIDATION WORKFLOW

Model Creation and Evaluation

- Various classification algorithms like Random Forest, XG Boost, KNN etc tested.
- Random Forest, XGBoost and KNN all were given better results. DecisionTreeClassifier was chosen for thefinal model training and testing.
- Hyper parameter tuning was performed.
- Model performance evaluated based on f1_score, recall score, confusion matrix, classification report.

Decision Tree Classifier Model

INTRODUCTION

A decision-tree-based ensemble Machine Learning algorithm that uses a Gradient boosting framework.

A decision tree is a flowchart-like tree structure where an internal node represents a feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome.

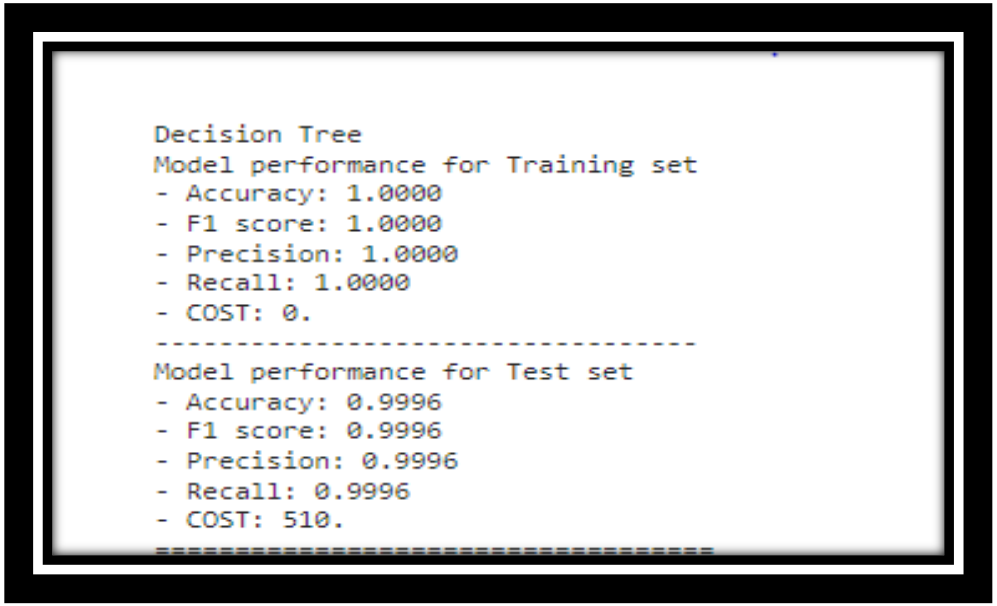
The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in a recursive manner called recursive partitioning. This flowchart-like structure helps you in decision-making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.

Reason to use Decision Tree Classifier model:

- High accuracy, robustness, feature importance, versatility, and scalability
- It gives better model performance.

MODEL PREDICTION RESULTS ON TEST DATASET

Classification Report



DATABASE CONNECTION & DEPLOYMENT

Database Connection

- Mongo dB Database is used

Model Deployment

- The final model is deployed on AWS using Flask framework.

FREQUENTLY ASKED QUESTIONS

Q1) what is the source of data?

The data for training is obtained from famous machine learning repository.
UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/thyroid+disease

Q2) what was the type of data?

The data was the combination of numerical and Categorical values.

Q3) what's the complete flow you followed in this Project?

Refer slide 7th, 8th and 9th for better understanding.

Q4) After the File validation what you do with incompatible file or files which didn't pass the validation?

Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.

Q5) how logs are managed?

We are using different logs as per the steps that we follow in training and prediction like model training log and prediction log etc. And then sub log are inside those folder.

Q 6) what techniques were you using for data pre-processing?

- Removing unwanted attributes
- Visualizing relation of independent variables with each other and output variables
- Checking and changing Distribution of continuous values
- Removing outliers
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values.

Q 7) how training was done or what models were used?

- First Data validation done on raw data and then good data insertion happen in DB.
- Then Data preprocessing done on final CSV file received from DB.
- We did clustering over the data to divide it on desired cluster based on elbow method.
- Various model such as Decision Tree, Random Forest and XGBoost models are trained on all clusters and based on performance, for each cluster different model is saved.

Q 8) How Prediction was done?

- The testing files are shared by the client .We Perform the same life cycle till the data is clustered.
- Then on the basis of cluster number model is loaded and perform prediction. In the end we get the accumulated data of predictions.

Q 9) what are the different stages of deployment?

- After model training and finalizing all models. We created required files for deployment.
- Finally deployed our model over various cloud platforms such as Heroku and AWS.

Q 10) how is the User Interface present for this project?

- For this project I have made two types of UI.
- First is for bulk prediction.
- Second is for one user input prediction.
- Both UI are very user friendly and easy to use.

THANK YOU