THYROID DISEASE DETECTION

Detailed Project Report

Upendra Kumar
Data Science Intern at Ineuron.ai
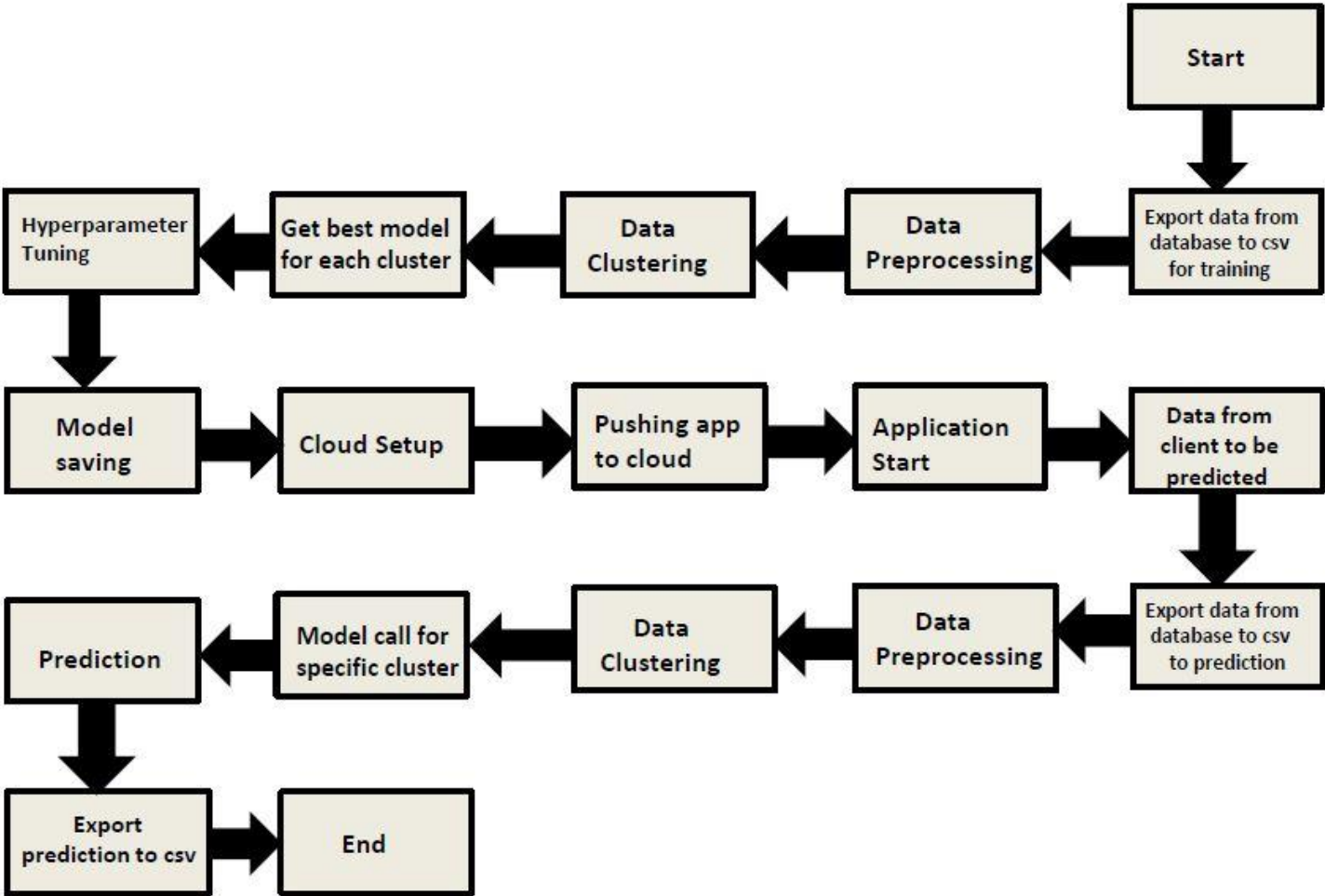
INTRODUCTION

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.
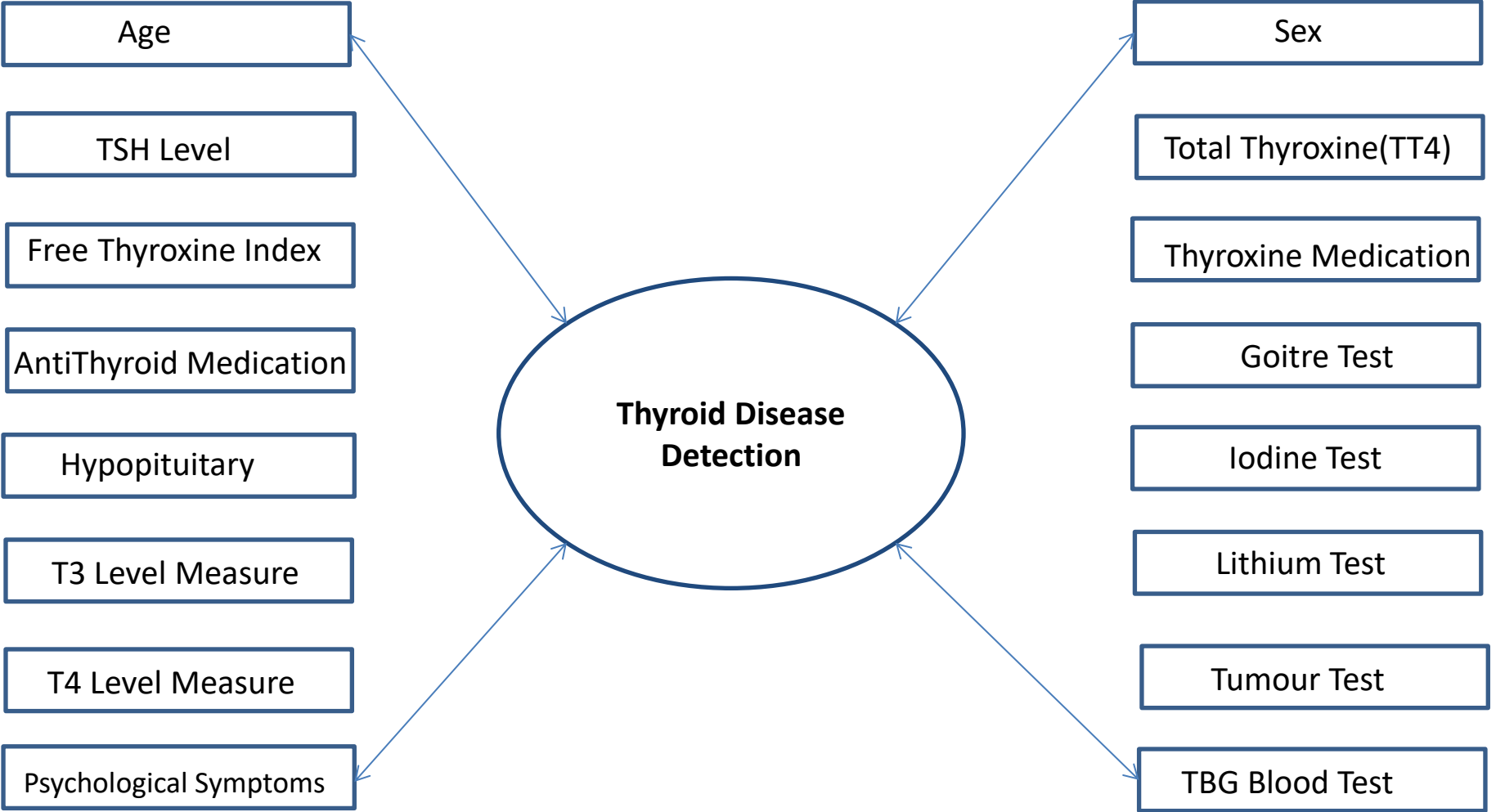
OBJECTIVE

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 fore cast 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.
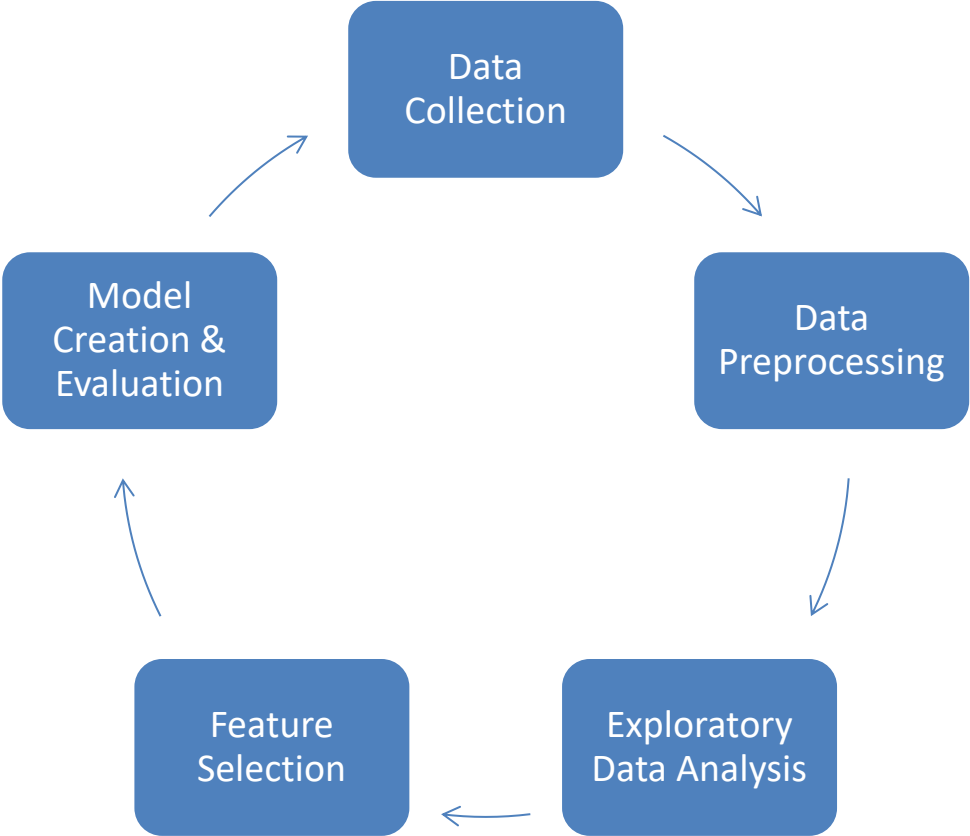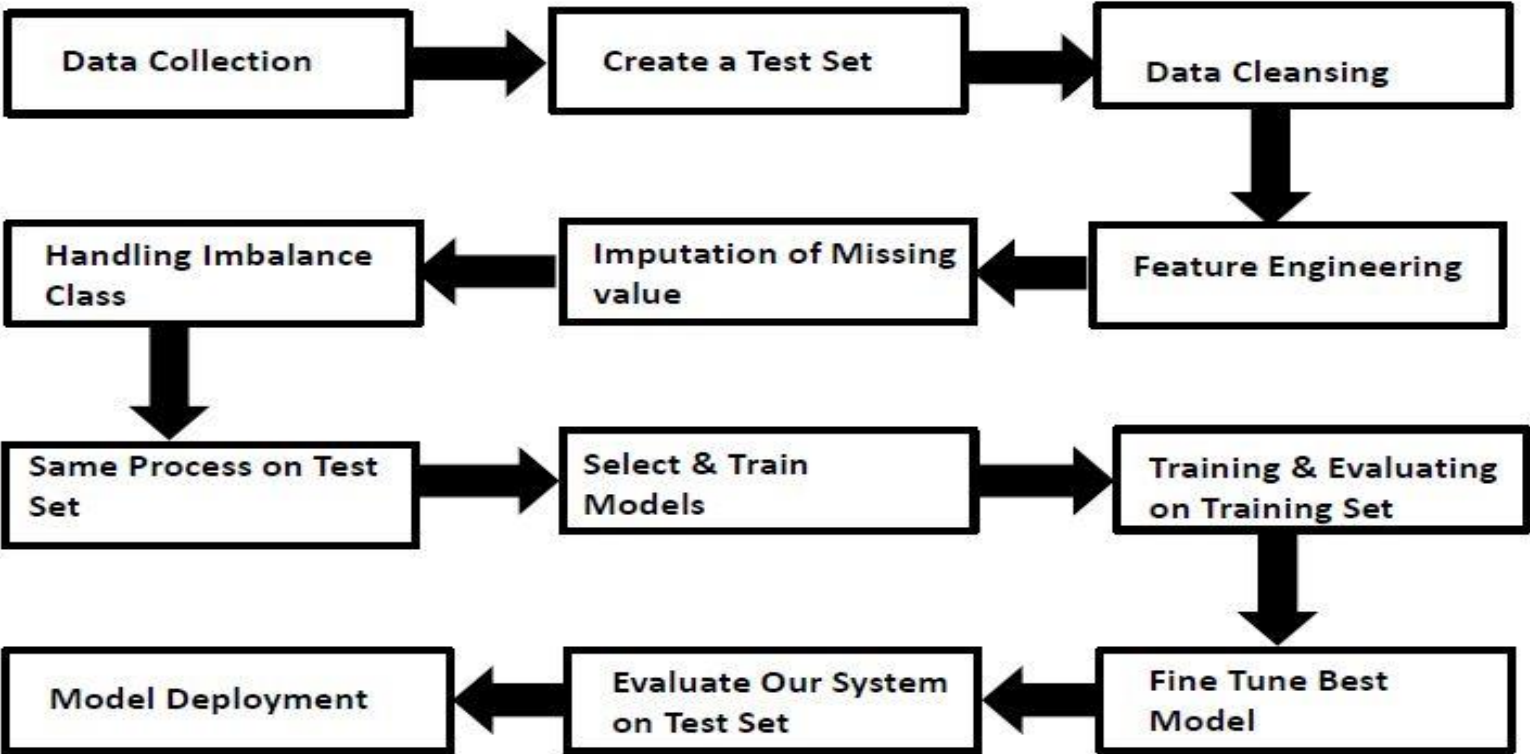
ARCHITECTURE

DATASET

Age		Sex
TSH Level		Total Thyroxine(TT4)
Free Thyroxine Index		Thyroxine Medication
AntiThyroid Medication	**Thyroid Disease Detection**	Goitre Test
Hypopituitary		Iodine Test
T3 Level Measure		Lithium Test
T4 Level Measure		Tumour Test
Psychological Symptoms		TBG Blood Test

Data Analysis Steps

- Data Collection
- Data Preprocessing
- Exploratory Data Analysis
- Feature Selection
- Model Creation & Evaluation

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 (Used KNN Imputer)
- Outliers' detection and removal by boxplot and percentile methods
- Categorical features handling by ordinal encoding and label encoding
- Feature scaling done by Standard Scalar method
- Imbalanced dataset handled by SMOTE -Over sampling
- 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. XG Boost was chosen for the final model training and testing.
- Hyper parameter tuning was performed.
- Model performance evaluated based on accuracy, confusion matrix, classification report.

XG Boost Classifier Model

INTRODUCTION

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

The XGBoost Classifier is a supervised learning algorithm which we can use for regression and classification problems. It is among the most popular machine learning algorithms comes under boosting ensemble technique.

XGBoost Classifier being ensemble algorithm tends to give more accurate result. This is because it works on the principle i.e. number of weak estimators when combined forms strong estimator. Even if one or few decision tree are prone to noise, overall results would tend to be correct.

Reason to use XGBoost Classifier model:
- It has high execution speed.
- It gives better model performance.

MODEL PREDICTION RESULTS
ON TEST DATASET

Classification Report

```
              precision    recall  f1-score   support

         0.0       0.94      1.00      0.97        33
         1.0       1.00      0.36      0.53        14
         2.0       0.99      1.00      1.00      1068
         3.0       1.00      1.00      1.00      1156

    accuracy                           1.00      2271
   macro avg       0.98      0.84      0.87      2271
weighted avg       1.00      1.00      1.00      2271
```

Confusion Matrix

```
[[  33    0    0    0]
 [   2    5    7    0]
 [   0    0 1068    0]
 [   0    0    0 1156]]
```

DATABASE CONNECTION & DEPLOYMENT

Database Connection

- Cassandra Database used for this project.

Model Deployment

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

THANK YOU