THYROID DISEASE DETECTION

Detailed Project Report

Navanshu Khare
Data Scientist

INTRODUCTION

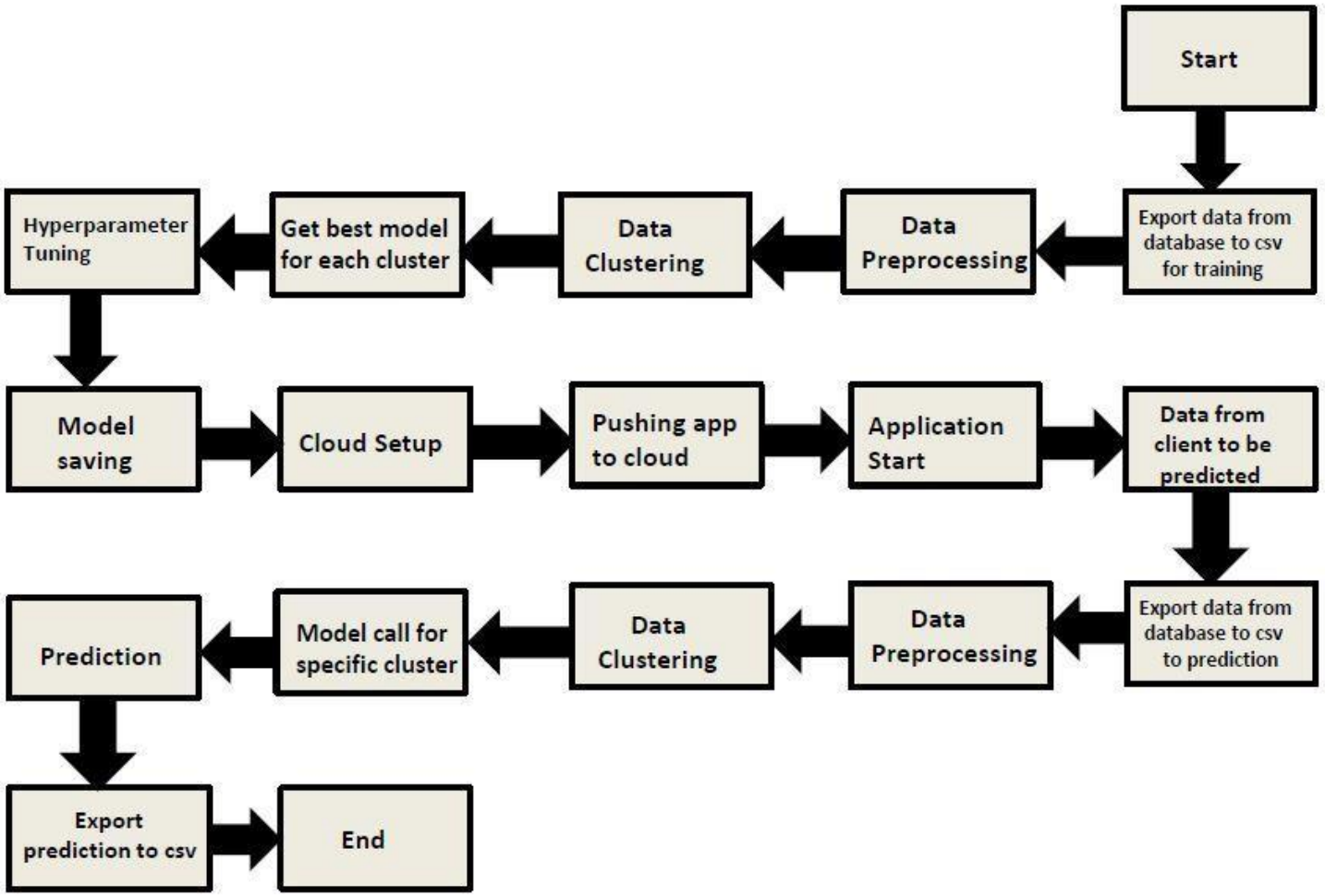
In India and USA, thyroid illness affects at least one person in ten. Thyroid illness primarily affects females between the ages of 18 to 60. Cardiovascular problems, a rise in blood pressure, a peak in cholesterol levels, depression, and reduced fertility are all brought on by the severe thyroid stage. The thyroid gland produces two active thyroid hormones, total serum thyroxin (T4) and total serum triiodothyronine (T3), to regulate the body's metabolism. These hormones are essential for the proper operation of every cell, tissue, and organ as well as for general energy production, control, and protein synthesis to maintain body temperature.

The two most prevalent conditions brought on by the thyroid gland's abnormal function are hyperthyroidism and hypothyroidism. The body's metabolism can be sped up or slowed down by a thyroid condition. Artificial intelligence is playing an increasingly important part in the advancement of the health care sector in the age of new technology and innovation. Machine learning algorithms can aid in early disease identification and enhance overall quality of life. The results of this study show how several categorization algorithms can predict the presence of a disease. To determine which classification algorithm will produce the best results for the model, various algorithms have been examined and compared, including Random forest, SVM and KNN

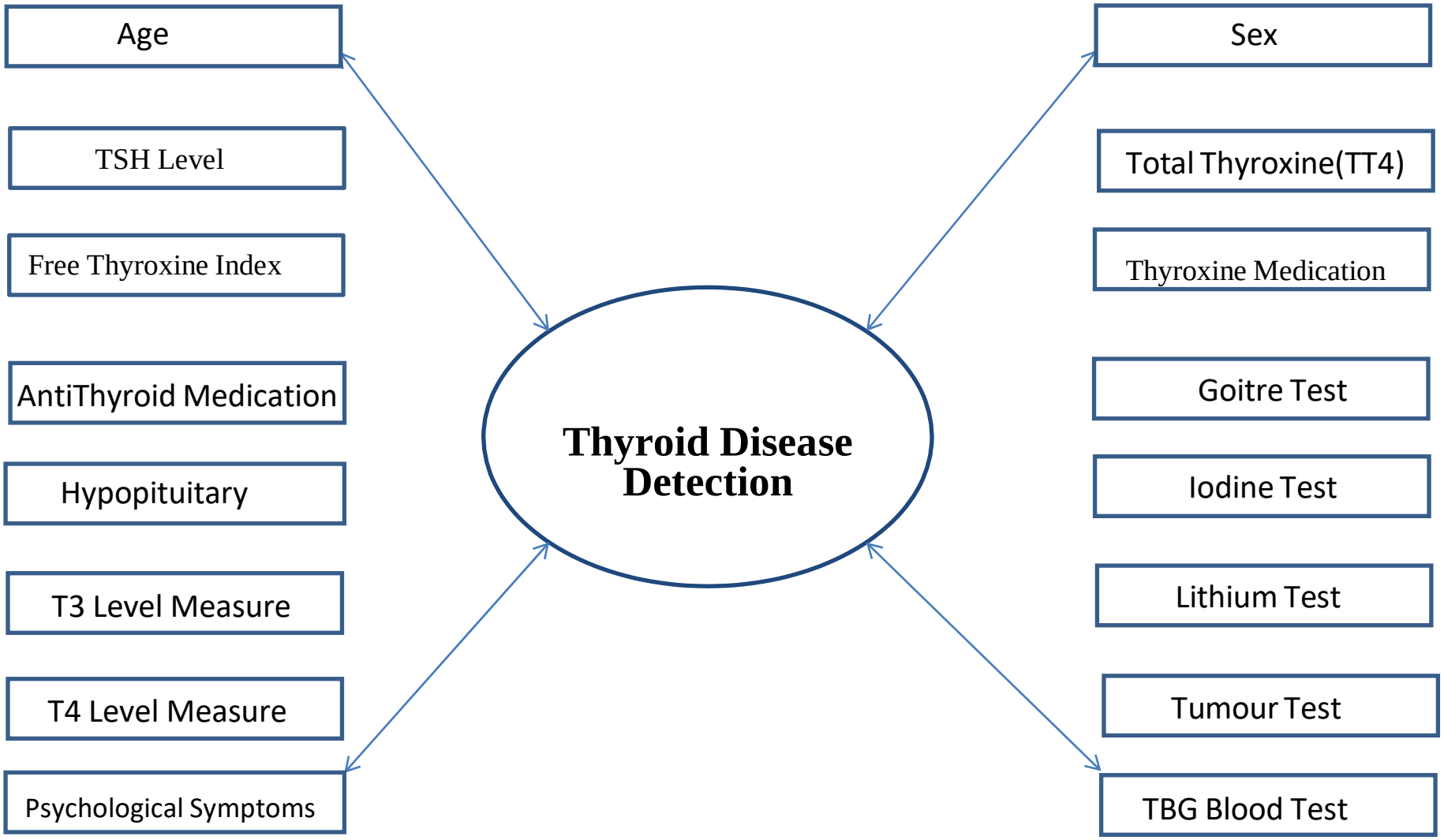
OBJECTIVE

The primary objective of this project is to forecast an individual's risk of hyperthyroidism and hypothyroidism using a variety of personal data. Medical research finds it challenging to predict the on-set of thyroid disease, a prevalent source of medical diagnosis and prediction. It will be crucial for early detection, accurate disease identification, and for assisting medical professionals in making wise choices and providing better care.

ARCHITECTURE

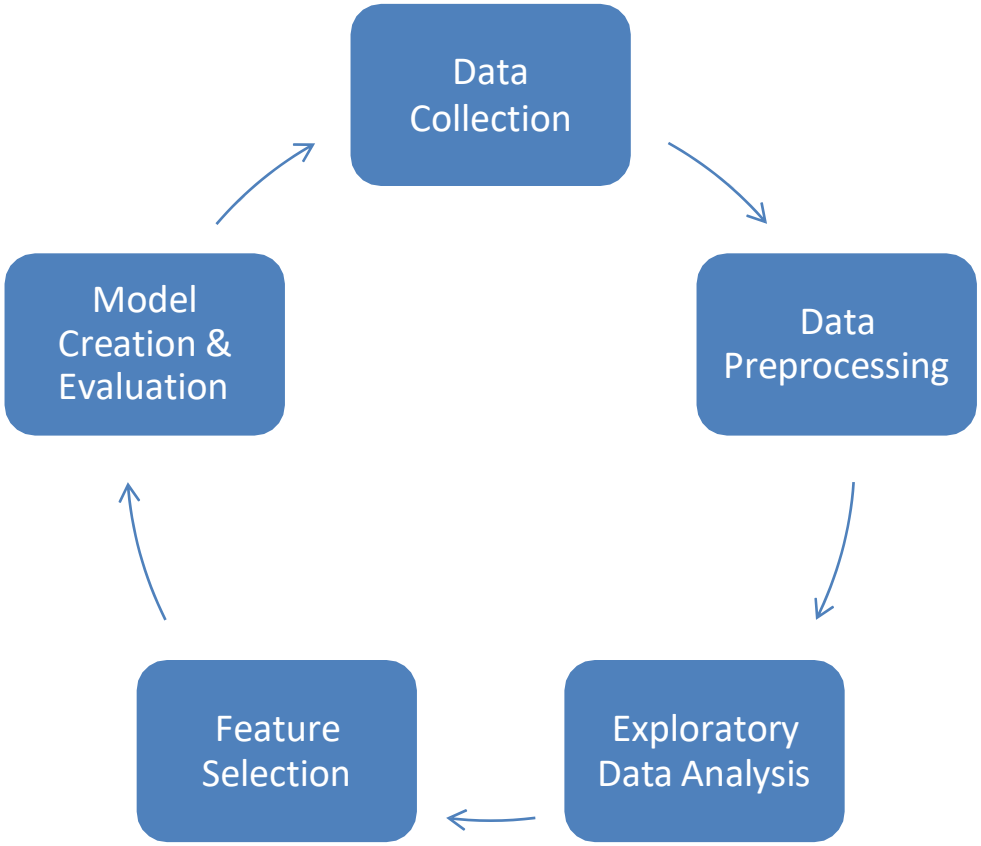


DATASET

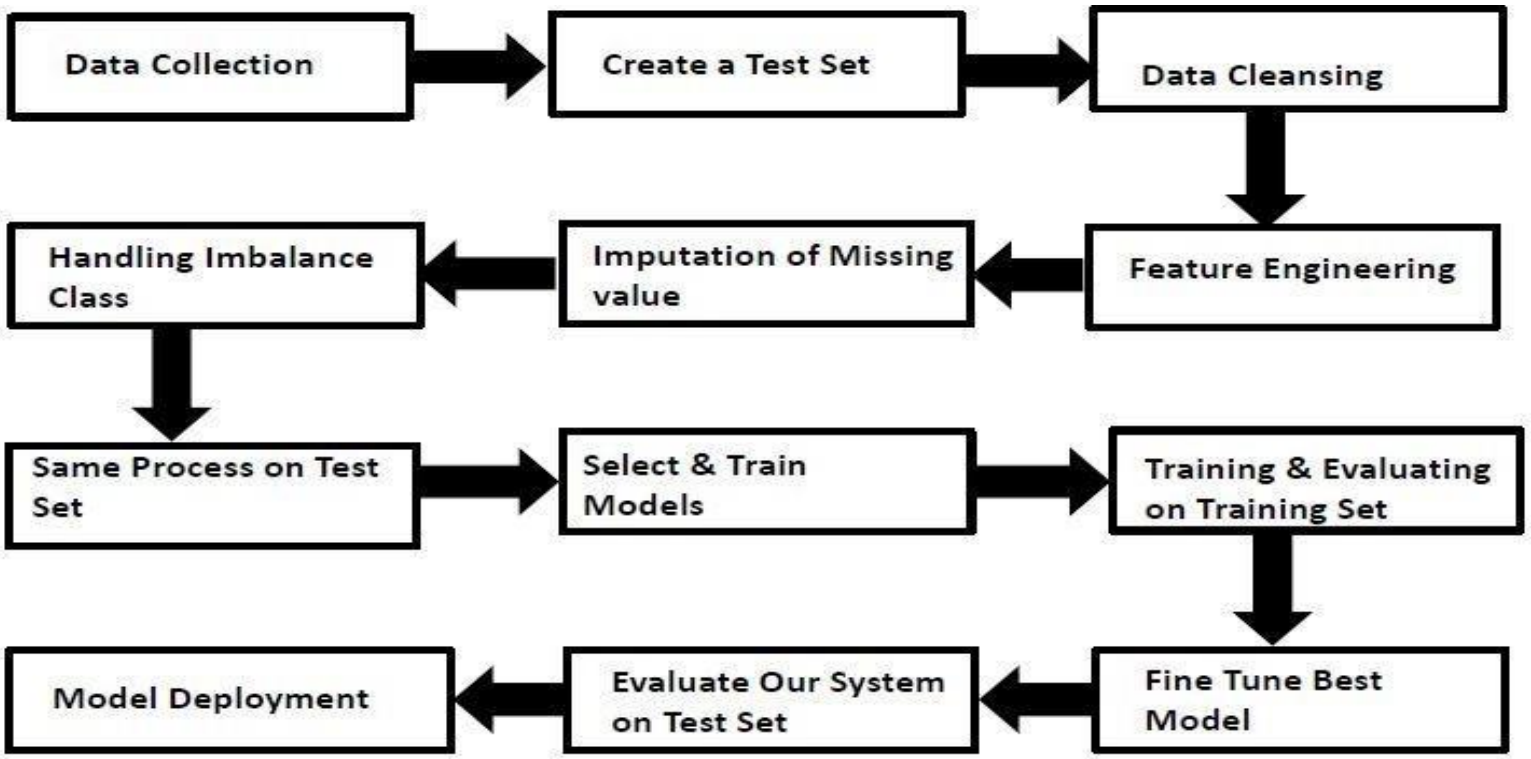

Age
TSH Level
Free Thyroxine Index
AntiThyroid Medication
Hypopituitary
T3 Level Measure
T4 Level Measure
Psychological Symptoms
Thyroid Disease Detection
Sex
Total Thyroxine(TT4)
Thyroxine Medication
Goitre Test
Iodine Test
Lithium Test
Tumour Test
TBG Blood Test

DPR: Thyroid Disease Detection

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 (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

- Data is divided in clusters
- Various classification algorithms like Random Forest, SVM, KNN are performed on clusters
- Each clustes gave best model and that model is saved after hyper parameter tuning
- Model performance evaluated based on accuracy, confusion matrix, classification report.

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]]
```

DEPLOYMENT

Model Deployment

- The final model is deployed on Heroku 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?

Please see the architecture of this project

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

Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removedthe 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 andprediction 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 KNN, Random Forest and SVM 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 theaccumulated 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.