THYROID DISEASE DETECTION

Detailed Project Report

Oviyashri B
Data Science Intern at Ineuron.ai

INTRODUCTION

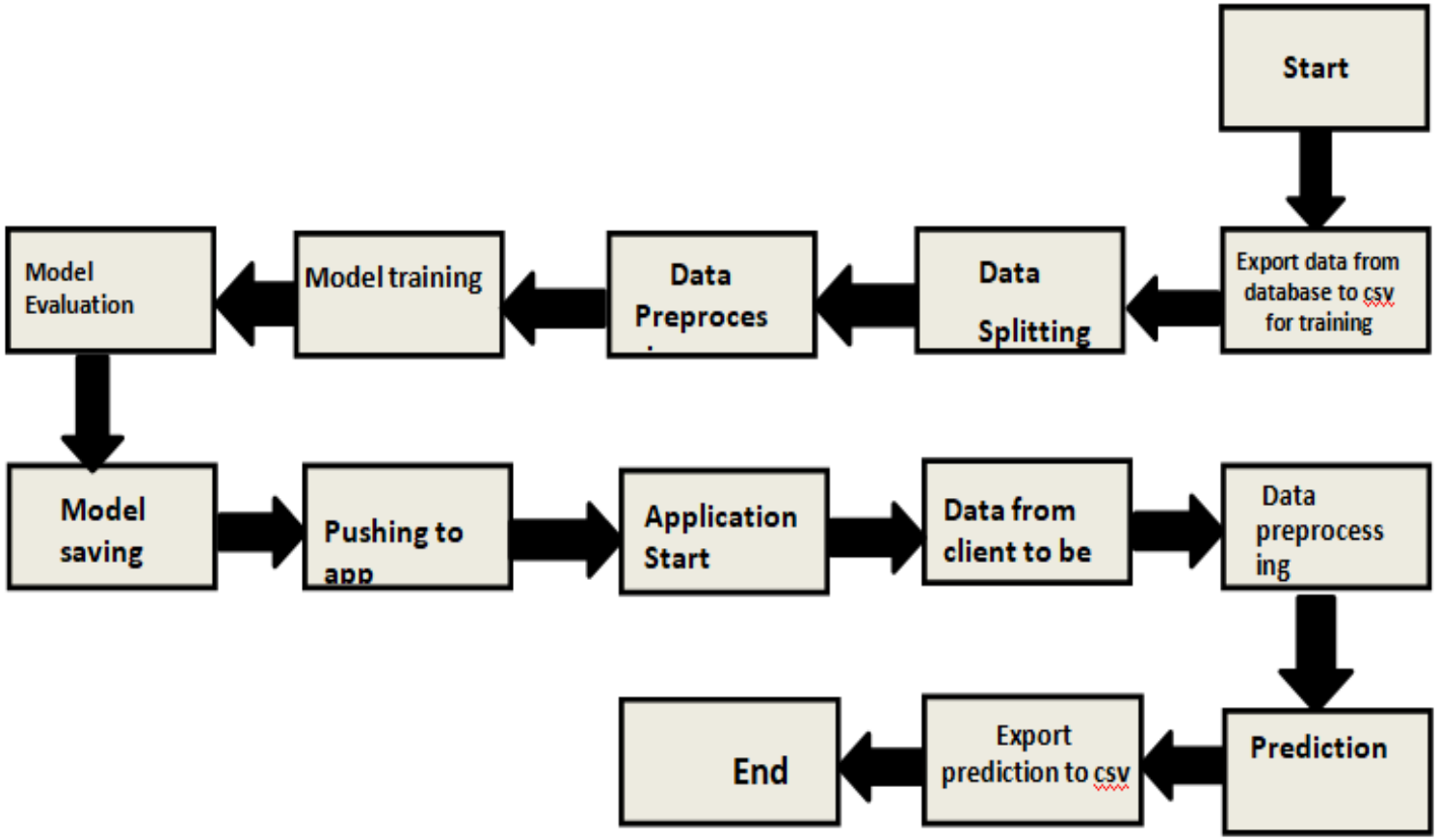
At least a person out of ten is suffered from thyroid disease in India. The disorder of thyroid disease primarily happens in the women having the age of 17–54. The extreme stage of thyroid results in cardiovascular complications, increase in blood pressure, maximizes the cholesterol level, depression 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. Cure of disease is a regular concern for the health care practitioners, and the errorless diagnostic at the right time for a patient is very important. Recently, by some advanced diagnosis methods, the common medical report can be generated with an additional report based on symptoms. We can find on implementing machine learning methods on Health care data. Health care data can be processed and after implementing with certain methodologies; it can provide information that can be used in diagnosis and treatment of diseases more efficiently and accurately with better decision making and minimizing the death risk.

OBJECTIVE

The main goal is to predict the estimated risk on a patient's chance of obtaining thyroid disease or not.
The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing.

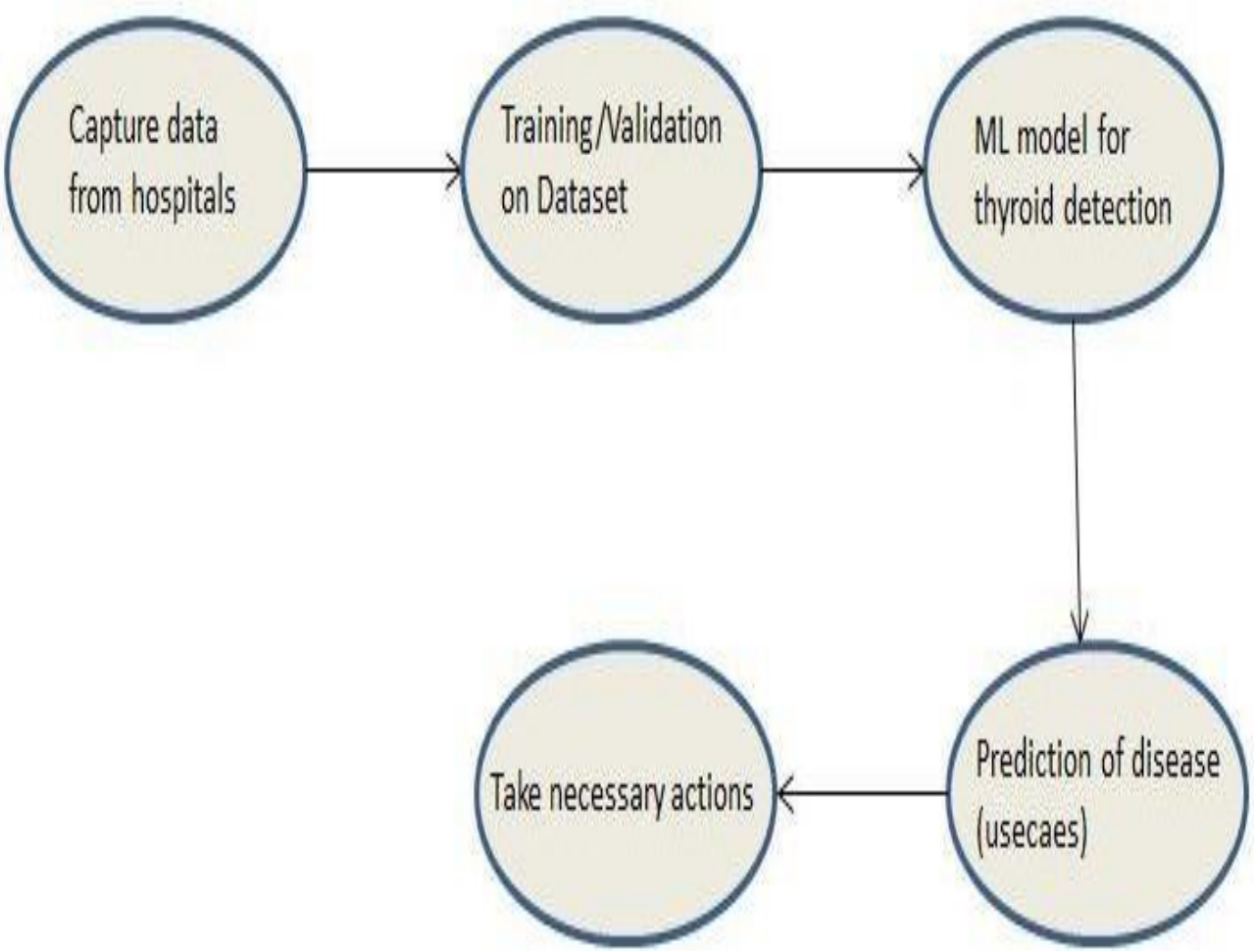
ARCHITECTURE



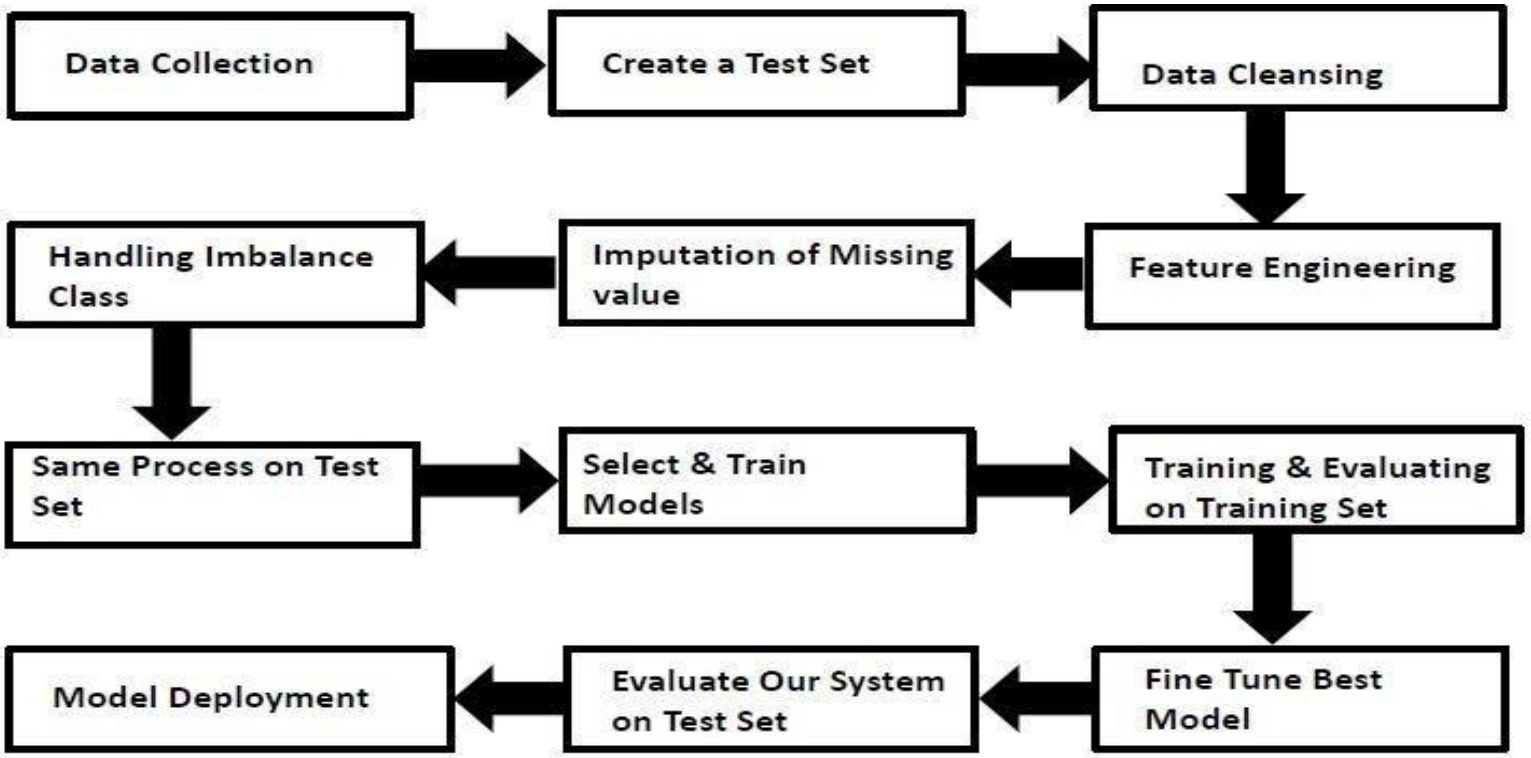
DATASET

age:	continuous,?.
sex:	M,F,?.
on_thyroxine:	f,t.
query_on_thyroxine:	f,t.
on_antithyroid_medication:	f,t.
thyroid_surgery:	f,t.
query_hypothyroid:	f,t.
query_hyperthyroid:	f,t.
pregnant:	f,t.
sick:	f,t.
tumor:	f,t.
lithium:	f,t.
goitre:	f,t.
TSH_measured:	f,t.
TSH:	continuous,?.
T3_measured:	f,t.
T3:	continuous,?.
TT4_measured:	f,t.
TT4:	continuous,?.
T4U_measured:	f,t.
T4U:	continuous,?.
FTI_measured:	f,t.
FTI:	continuous,?.
TBG_measured:	f,t.
TBG:	continuous,?.

PROCESS FLOW



MODEL TRAINING AND EVALUATION WORKFLOW



MODEL TRAINING AND EVALUATION

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
- Categorical features handling
- Imbalanced dataset handled by Random over sampling
- Drop unnecessary columns

MODEL TRAINING AND EVALUATION

Model Training and Evaluation

- Classification algorithm - Random Forest is tested since it given better result and was chosen for model training and testing.
- Model performance evaluated based on accuracy, classification report.

WORKFLOW

Data Description

We will be using Thyroid Disease Data Set present in UCI Machine LearningRepository. This Data set is satisfying our data requirement. Total 7200 instances present in different batches of data.

Export Data from database to CSV for Training

Here we will be exporting all batches of data from database into one csv filefor training.

Data Splitting

We split the data here for our train and test data for further uses.

Data Preprocessing

We will be exploring our data set here and perform data preprocessing depending on the data set. We first explore our data set in Jupyter Notebook and decide what pre-processing and validation we have to do such as imputation of null values, dropping some column, handling imbalanced data etc and then we have to write separate modules according to our analysis, so that we can implement that for training as well as prediction data.

Model Training

We trained various model in our notebook and Random Forest Classifier was good on it. We trained with our processed data.

Model Saving

We will save our models so that we can use them for prediction purpose.

Push to app

Here we will do cloud setup for model deployment. We also create our flask app and user interface and integrate our model with flask app and UI.

Data from client side for prediction purpose

Now our application on cloud is ready for doing prediction. The prediction data which we receive from client side.

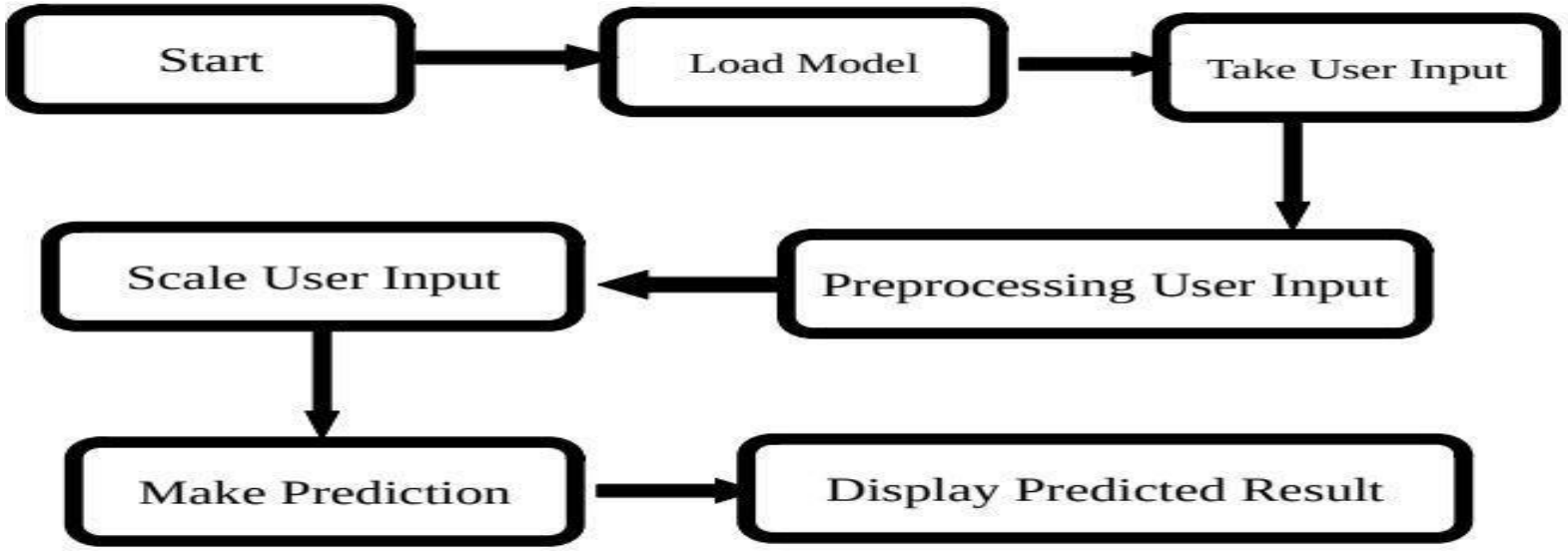
Data processing

Client data will also go along the same process Data pre-processing and according to that we will predict those data.

Export Prediction to CSV

Finally when we get all the prediction for client data, then our final task is toexport prediction to csv file and hand over it to client

Model Deployment



Model Deployment

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

HEROKU

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
- 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.
- Various model such as Decision Tree, Random Forest and XGBoost models are trained on all clusters and based on performance, 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 cloud platform Heroku.

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

- For this project I have made for one user input prediction.
- UI is very user friendly and easy to use.

THANK YOU