# Detail Project Report

## INTRODUCTION

The Thyroid gland is a vascular gland and one of the most important organs of a human body. This gland secretes two hormones which help in controlling the metabolism of the body. The two types of Thyroid disorders are Hyperthyroidism and Hypothyroidism. When this disorder occurs in the body, they release certain type of hormones into the body which imbalances the body's metabolism. Thyroid related Blood test is used to detect this disease but it is often blurred and noise will be present. Data cleansing methods were used to make the data primitive enough for the analytics to show the risk of patients getting this disease. Machine Learning plays a very deciding role in the disease prediction. Machine Learning algorithms, Random Forest Classifier, logistic regression, are used to predict the patient's risk of getting thyroid disease. Web app is created to get data from users to predict the type of disease.

Keywords— Random Forest, Logistic Regression, Flask.

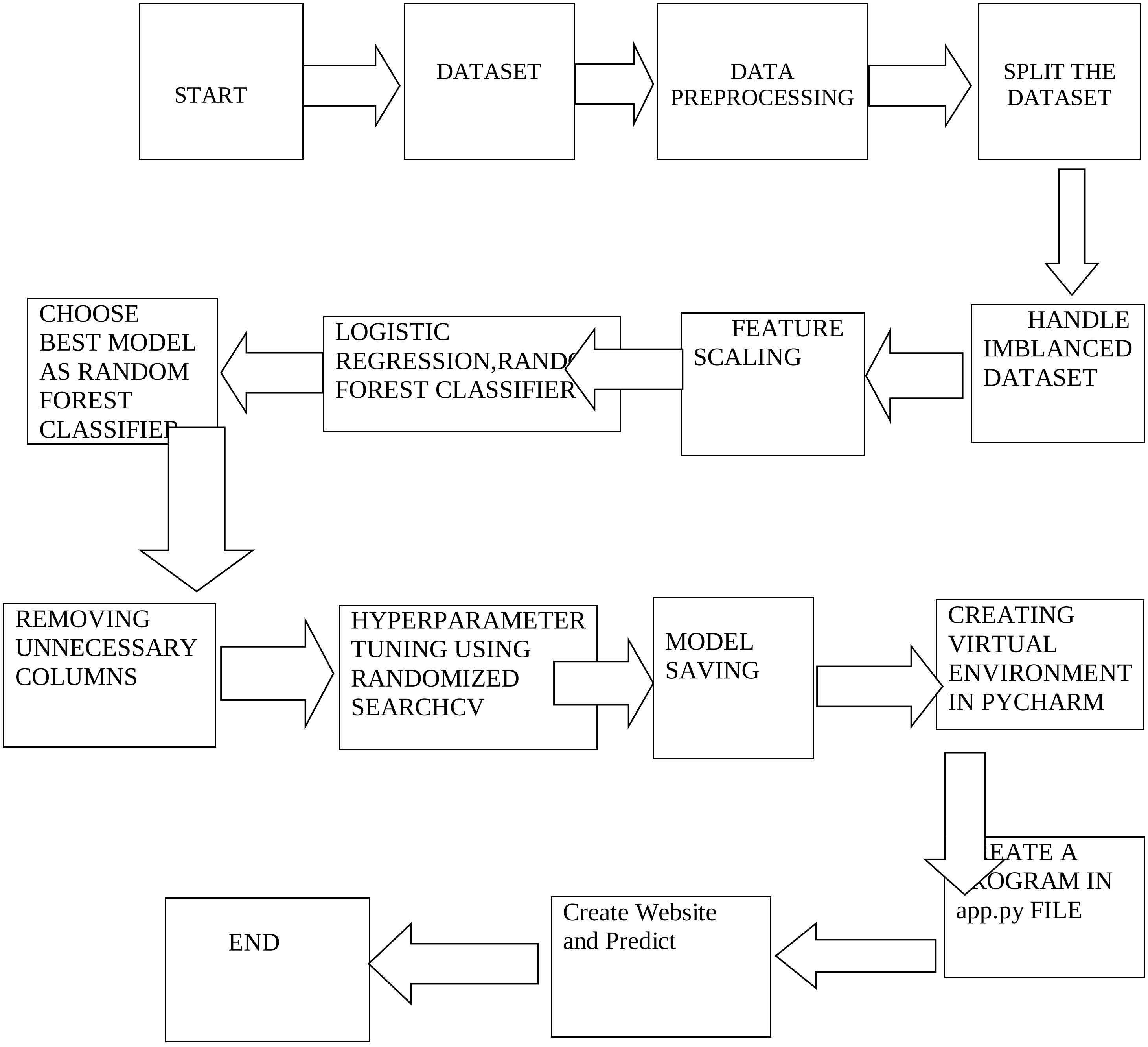
Objective:

- main objective is to develop a system which can predict of thyroid disease that patient is affected from.
- To predict thyroid disease with usage of minimum number of parameters.
- To predict all possible types of Thyroid diseases.

I used Logistics Regression and Random Forest classifier machine learning Technique to analyze Thyroid Dataset. Comparison was made between these two algorithms based on Precision, Recall, F measure, error. Random Forest classifier turned out has best classifier.

Detail Project Report

ARCHITECTURE


START
DATASET
DATA PREPROCESSING
SPLIT THE DATASET
HANDLE IMBLANCED DATASET
FEATURE SCALING
LOGISTIC REGRESSION,RANDOM FOREST CLASSIFIER
CHOOSE BEST MODEL AS RANDOM FOREST CLASSIFIER
REMOVING UNNECESSARY COLUMNS
HYPERPARAMETER TUNING USING RANDOMIZED SEARCHCV
MODEL SAVING
CREATING VIRTUAL ENVIRONMENT IN PYCHARM
CREATE A PROGRAM IN app.py FILE
Create Website and Predict
END

# Data Pre-processing

- Missing values handling by Simple imputation.
- Categorical features handling by ordinal encoding and label encoding.
- Feature scaling done by Min Max scaler method.
- Imbalanced dataset handled by SMOTE.
- Drop unnecessary columns.

DESCRIPTION:

1. Import the dataset and doing DataPreprocessing .
2. Second step is Expolatory Data Analysis using pie chart , heatmap.
3. Split the dataset into train test split.
4. AS the dataset is imblanced therefore use smote to overcome imblanced dataset.
5. Apply feature scaling using min max scaler.
6. Now applying Logistic Regression using grid searchcv and k fold cross validation and find accuracy, precision, recall.
7. Apply random forest classifier and find accuracy, precision,recall.
8. Using barplot and visualizing important features for Random forest classifier.
9. Remove unnecessary columns from the dataset.
10. Again using random forest classifier using Randomized searchcv and find accuracy, precision, recall.
11. Now import pickle and dump the model of high accuracy, precision, recall.
12. Make a python program using flask, in virtual environment in pycharm.