KAUSHIKI

Kaushiki.bhardwaj10@gmail.com

Detailed Project report

Thyroid Disease Detection

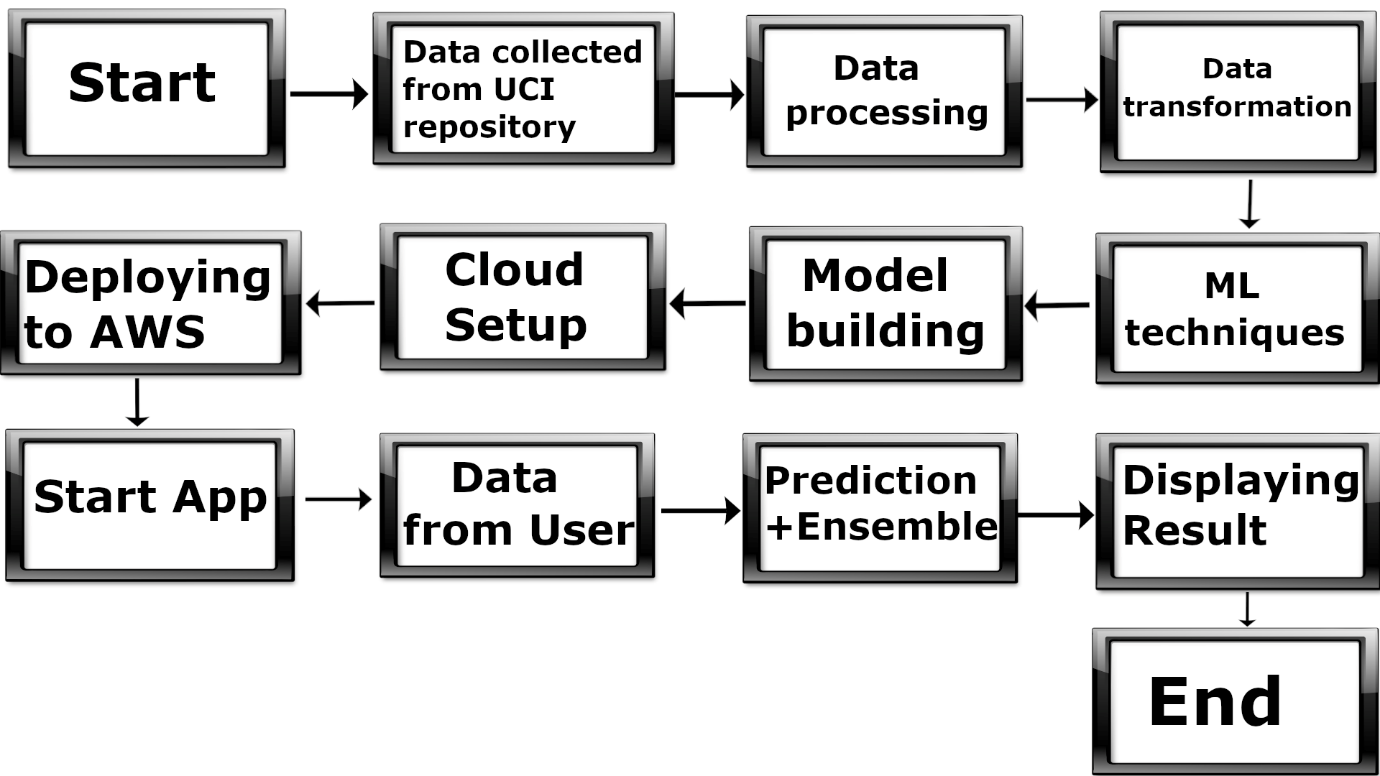
**Objective**

To build a classification methodology which predicts whether a person is thyroid free or not based on the given training data.

**Benefits**

* Detection of thyroid disease
* Eliminates human error
* Higher efficiency
* Time efficient

**Architecture**



**Data Transformation**

* Removing of outliers
* Deletion of unnecessary features
* Imputation of null values
* Feature Selection
* Re-scaling the data into a range of 0 and 1
* Fixing the imbalanced data

**Model Training**

After data processing, data is sampled into training, testing and validation. Then training data is fitted into 6 different models and they are tested on testing data. Finally, two best models are selected based on performance metrics.

**Model Selection**

Training data is used to train multiple models. Two models were selected based on performance metrics. They are Random Forest and XGBoost.

**Prediction**

The validation data is used as a final test for our models. The output of both the models is ensembled and the final values are tested.

**Q & A**

Q1) What’s the source of the data?

The entire data is collected from the UCI machine learning repository.

Q2) What was the type of data?

The data was a combination of numerical, categorical and null values.

Q3) How logs are managed?

Once the code is executed a log file will be created in the same repository.

Q4) What techniques were used for data processing?

* Removing features/columns that are not much useful
* Visualizing relation of independent variable with each other through heatmaps.
* Removing outliers
* Imputing null/missing values
* Encoding the categorical features
* Fixing the imbalanced data
* Scaling of the data

Q5) How training was done and what models were used?

After data processing, the training data was fitted into 6 different models. These included Logistic Regression, K-nearest Neighbours, Support vector machine, XGBoost, Decision Tree and Random Forest. Finally based on performance metrics Random Forest and XGBoost were selected.

Q6) How prediction was done?

The final evaluation was made with the validation dataset. Same data processing steps were applied and prediction of each model was ensembled and tested.

Q7) What are the different stages of deployment?

* Once the models were tuned, they were saved as a pickle file.
* AWS account was created and EC2 instance on Linux environment was created.
* Flask was used to integrate the models and the simple HTML frontend.
* Requirement file was created which contained all the necessary libraries needed for the code to execute.
* Putty Gen and Putty was used to create the private key and access the EC2 instance.
* WinSCP was used to transfer all the files to the EC2 instance.
* Run the required files in the command prompt.