# Springboard Data Science Career Track

# Capstone Project I Milestone Report

# Thyroid Classification.

Rahul Ambati

May 2018

1. Introduction.
2. Data Acquisition and Wrangling.
3. Data Exploration and Inferential Statistics.
4. Machine Learning Techniques.
5. Summary and Further Steps.

# Introduction.

Definition:

Thyroid disease is a medical condition affecting the function of the thyroid gland. The symptoms of the disease vary depending on the type of thyroid disease.

Intention:

A physician needs to know the demographics suffering from thyroid disease and find what sector of people can be focused on so that they get admitted and get prior treatment.

Client:

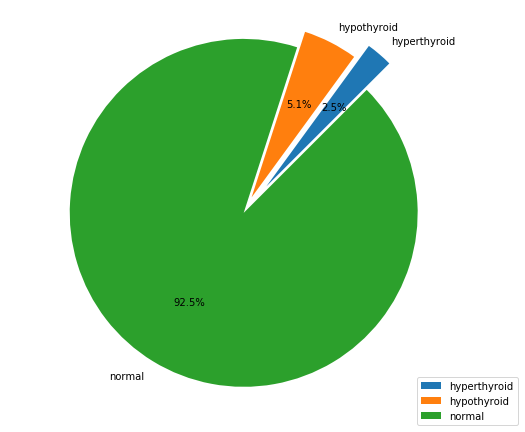
Physicians who want to understand what kind of demographics, medication, etc. to consider while treating patients with thyroid disease so that the right group of people can get the proper care and treatment.

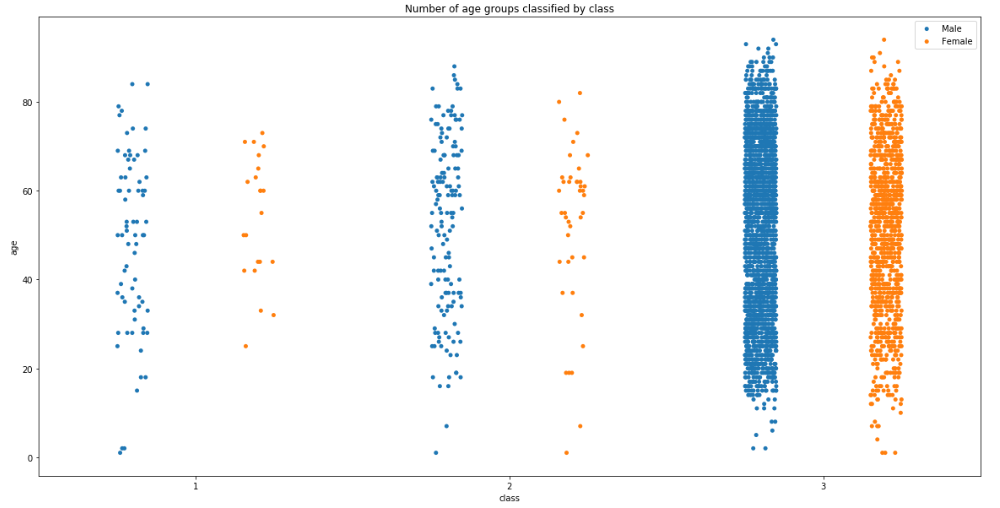
# Data Acquisition and Wrangling.

The data has been acquired from [UCI ML dataset Thyroid disease](https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease) and, in particular, this project will be focusing on the [ANN](https://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/ann-Readme). ANN dataset, which provides the information for characterizing thyroid disease.  
  
The dataset consists of demographics (age, sex), about medication, current conditions (sick, tumor, goiter, etc.) and some relevant measurements (TSH, T3, TT4, T4U, FTI) and category.

The dataset from the repository is clean, and not much cleaning or wrangling had to be done.

# Data Exploration and Inferential Statistics.

The dataset contains information about three categories of the disease ***hyperthyroid*** *(class 1),* ***hypothyroid*** *(class 2) and* ***normal*** *(class 3)*. ***Normal*** category means person does not suffer from the disease.  
  
Exploring the data, much of the demographics does not suffer from the disease, as one would expect (see Figure 1).  
  
Figure 1. Proportion of classes in the ANN dataset

From the dataset a strip plot (see Figure 2) is made to view the dataset clustered according to the classes, age group and gender.Figure 2. Strip plot showing the distribution of classes per age group and gender

Density plots (see Figure 3) show TSH, T3, TT4, T4U, FTI measurements and age.

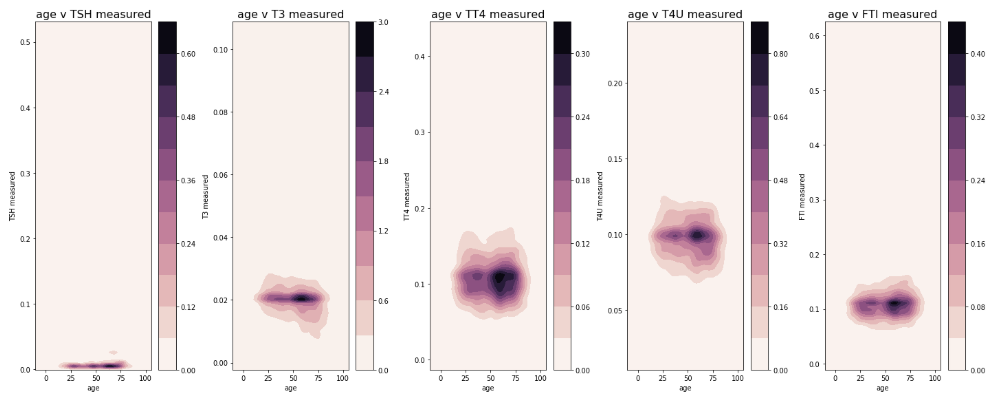
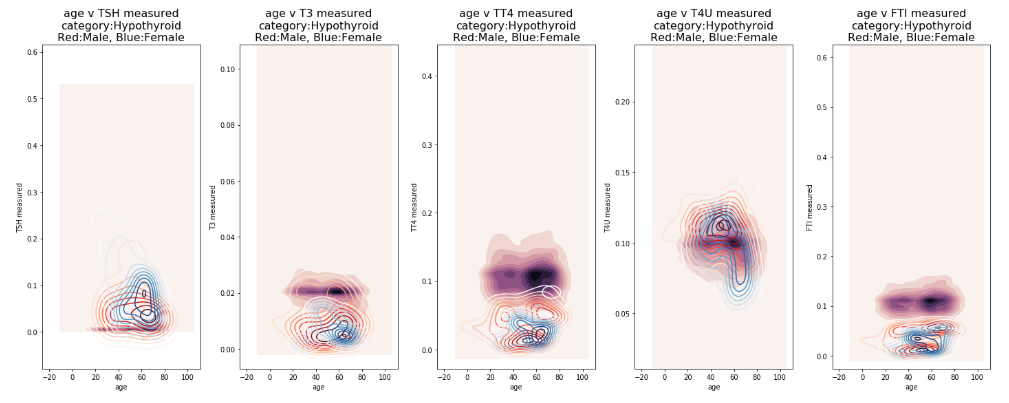


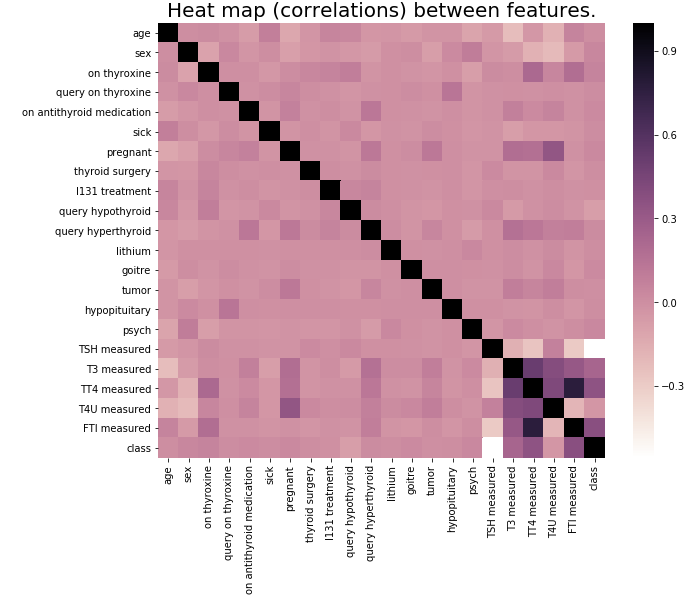
Figure 3. Density plots for various thyroid-related measurements and age

***Hyperthyroid*** class was extracted from the data set and overlapped on the measurements to see where it lies in the distribution (Figure 4).

Figure 4. Density plot for various thyroid-related measurements, age and classes

A clear outlier can be seen where the ***hyperthyroid*** lies and we can make some inferences that it might lie outside the normal distributions.

With many and different features each feature can affect differently, so a correlation heatmap is made to see how features are affecting the most and the least (Figure 5).

   
Figure 5. Correlation heatmap for some variables

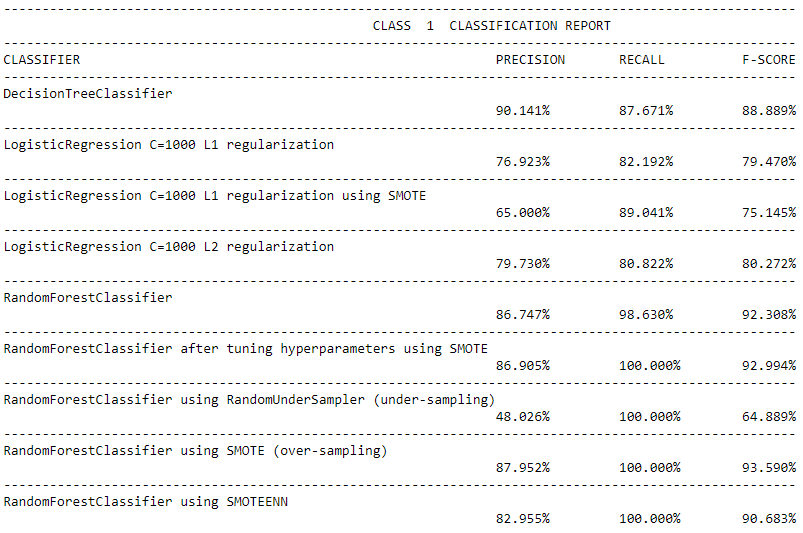
Looking at the correlation map (Figure 5) the measurements have a varied level of influence.

***Statistical inference*** was used to determine if correlation between TT4 and FTI measurements is statistically significant, by performing a t-test.

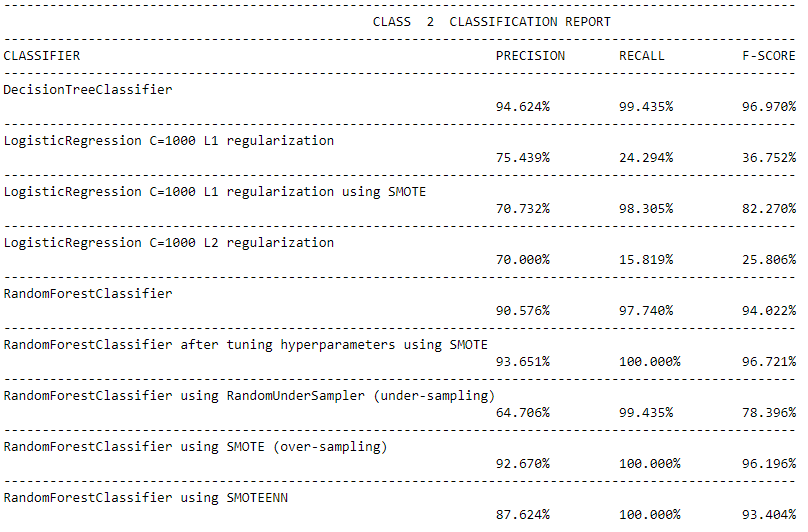
*Null hypothesis* mentions correlation between TT4 and FTI measurements is zero.

# Machine Learning Techniques.

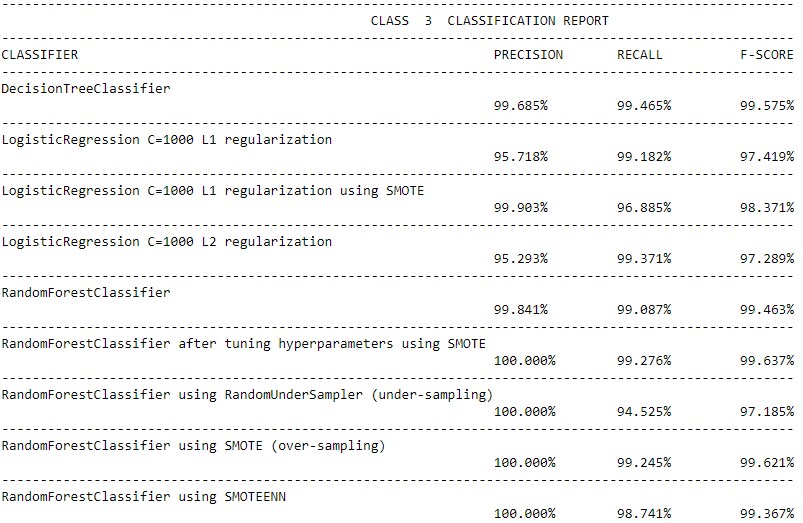
Summary table showing the algorithms applied (Machine learning techniques) and their test-set metrics for each class (Precision, Recall, F-score).



For Class 1 RandomForestClassifier using SMOTE has high precision amongst the classifiers, 100% recall with highest F-score.



For Class 2 RandomForestClassifier after tuning hyperparameters using SMOTE has highest precision and f-score with 100% recall.



For Class 3 RandomForestClassifier after tuning hyperparameters using SMOTE has highest recall amongst the classifiers with 100% precision and highest F-score.

# Summary.