TOPIC

A comparative study of the performance of various Machine Learning algorithms to classify if a patient has hypothyroid.

OBJECTIVE

There is a vast difference between the number of specialists to the total number of patients. For instance, the required number of cardiologists in India is as high as 88,000 whereas the actual number is merely 4000. Hyperthyroidism is a form of thyroid disease, causing the gland to produce too much thyroid hormone. Around 1 in every 100 Americans has hyperthyroid.  Through this project, I aim to bridge this gap and make health care more consumer centric, affordable and accessible by using Classification algorithms to aid the clinical decision making process.

DATASET DESCRIPTION

The dataset was sourced from UCI Machine Learning repository. It has 3772 rows/objects and 30 columns/features. It has 2 classes (0: no hypothyroid, 1:hypothyroid).

Patients with hypothyroid: 3481

Patients without hypothyroid: 291

|  |  |  |
| --- | --- | --- |
| Sr. No.- | Attribute | Details |
| 1 | Age | Patient’s age in yrs |
| 2 | Sex | F:female, M:male |
| 3 | On thyroxine | True/false |
| 4 | Query on thyroxine | True/false |
| 5 | On antithyroid mdication | True/false |
| 6 | Sick | True/false |
| 7 | Pregnant | True/false |
| 8 | Thyroid surgery | Tru/false |
| 9 | I131 treatment | True/false |
| 10 | TSH | Thyroid simulating hormone level |
| 11 | T3 | Triiodothyronine level |
| 12 | Lithium | True/false |
| 13 | Goitre | True/false |
| 14 | Tumor | True/false |
| 15 | Hypopituitary | True/false |
| 16 | Psych | True/false |
| ….…. | ….… | ….…. |
| 30 | Outcome | 0: patient does not have hypothyroid1: patient has hypothyroid |

DATA CLEANING

I converted columns to the required datatype. "age","TSH","T3","TT4","T4U","FTI" were converted to float. Null values were found in columns sex, T3, TT4, T4U, FTI and TBG. Since TBG has majority of values missing, I dropped the column.For the rest of the columns, I replaced NA values with the median of the particular column with respect to class label. For example, in case of T3, missing values were filled with 1.5 for class label 0 (Healthy patients) and 2.0 for class label 1(Hypothyroid patients).

df0["T3"].median() …where df0 is the subset of healthy patients without hypothyroid

#1.5

df1["T3"].median() …where df1 is subset of patients with hypothyroid

#2.0

I performed label encoding on all the categorical columns. I checked for outliers using the Inter Quartile Range criteria and found extreme values in certain columns but decided to keep the outliers since they show important variability in the data. It is possible that patients had extreme values for these hormone levels, they did not seem like data entry errors. Getting rid of the outliers may have increased model's accuracy but they are essential to show the natural variation in medical data. To reduce the effect of difference in units and magnitude of values, I scaled the data using StandardScaler().

DATA VISUALIZATIONS

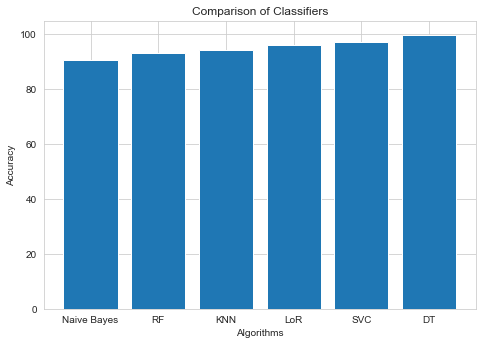
Analyzing hormone levels for patients with and without hyperthyroid based on if they are sick or not/ have a tumor not not. Here are some of the observations I got:

* T3 levels were higher in those who were not sick. Hypothyroid patients had comparatively higher T3 levels.
* Hypothyroid patients had higher TT4 levels.
* T4U levels were higher for patients who were not sick.
* For those who were not sick, patients without hypothyroid had higher T4U levels. For those who were sick, patients with hypothyroid had higher T4U levels.
* Hypothyroid patients had higher FTI levels.
* Healthy patients had higher TSH levels. These were highest for those who were not sick.
* Patients with tumor had higher T3 levels. Patients with hypothyroid had slightly higher T3 levels.
* Patients with a tumor had slightly higher T4U level.
* Non-hypothyroid patients with a tumor had larger variability in FTI levels. In general, hypothyroid patients had higher FTI levels.
* For patients without a tumor, median age was approximate the same, irrespective of if they had hypothyroid or not. For patients with a tumor, age was slightly higher for patients with hypothyroid.
* Highest number of patients were between 50-60 years old.
* I computed the pearson correlation for each pair of features and found that TT4 and FTI had strongest correlation, followed by TT4 and T3.

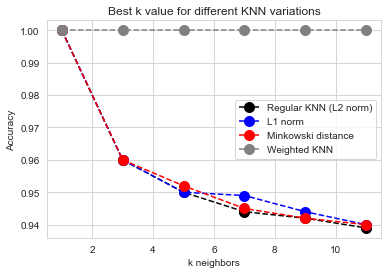
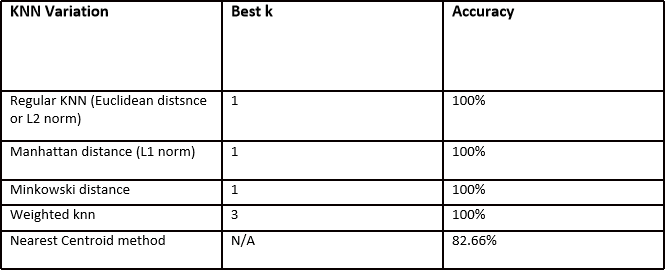
ALGORITHMS USED

* Logistic Regression
* KNearst Neighbors
* Support Vector Classifer
* Decision Tree
* Random Forest
* Naive Bayes

Here is their comparison based on different evaluation metrics:

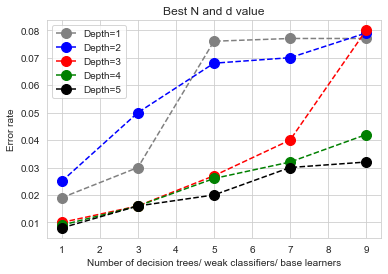


I also implement different variations of KNN with various values of k (number of neighbours).



I got best accuracy for weighted knn, k=3.

After implementing Random Forest for different combinations of Number of weak learners and maximum depth of each Decision Tree, best combination was found to be: N=3, d=5 i.e. 3 weak learners or deicision trees, with a maximum depth of 5 each.



After comparing the classifiers, I got highest accuracy for Decision Tree Classifier. True Positive Rate and True Negative Rate were also highest for Decision Tree.

HYPERPARAMETER TUNING AND STRATIFIED KFOLD CROSS VALIDATION

1. Randomized Search- It selects random combinations of hyperparameters to train the model.

Providing parameter grid:

random\_grid = {"max\_depth": [i for i in range(1,100)],

"max\_features": ['auto', 'sqrt','log2'],

"min\_samples\_leaf" : [j for j in range(1,11)],

"criterion": ["gini", "entropy"],

"splitter":["best", "random"]}

Using Startified KFold Cross Validation:

cv = RepeatedStratifiedKFold(n\_splits=10, n\_repeats=5, random\_state=3)

Specifying base estimator, parameter grid, number of iterations and cross validation.

dt\_after\_randomizedsearch=RandomizedSearchCV(estimator=dt2,param\_distributions=random\_grid,n\_iter=100,cv=cv)

Best parameters were found to be:

{'splitter': 'best',

'min\_samples\_leaf': 1,

'max\_features': 'sqrt',

'max\_depth': 9,

'criterion': 'gini'}

2.Grid Search- It tries every possible combination of hyperparameters. It is more exhaustive, but time consuming.

Supplying best parameters I got from random search + additional parameter options as well. Specifying base estimator (from randomized search), possible parameters and cross validation.

dt\_after\_gridsearch=GridSearchCV(estimator=dt3,param\_grid=param\_grid,cv=cv)

Fitting classifier on best parameters.

dt4 = DecisionTreeClassifier(splitter='best',

min\_samples\_leaf=1,

max\_features= 'auto',

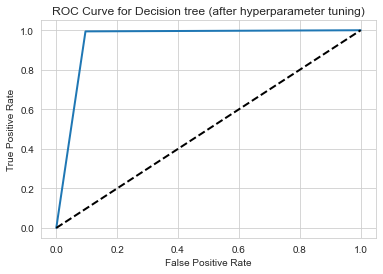
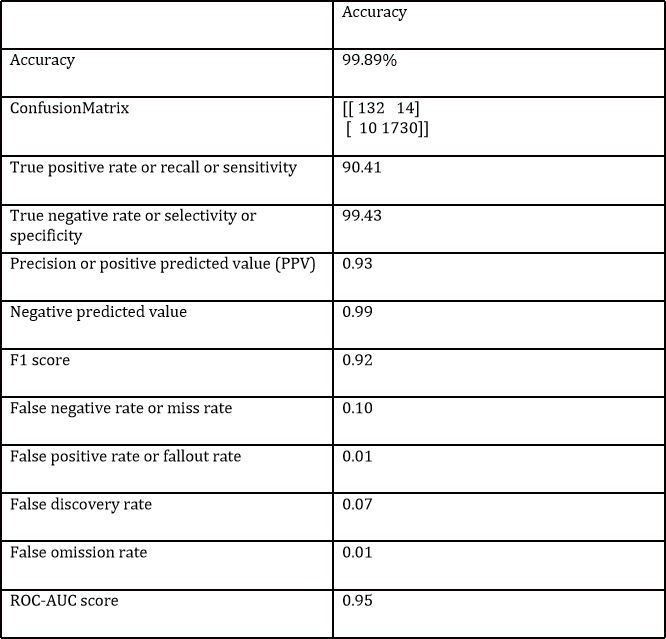
max\_depth= 200,

criterion= 'entropy')

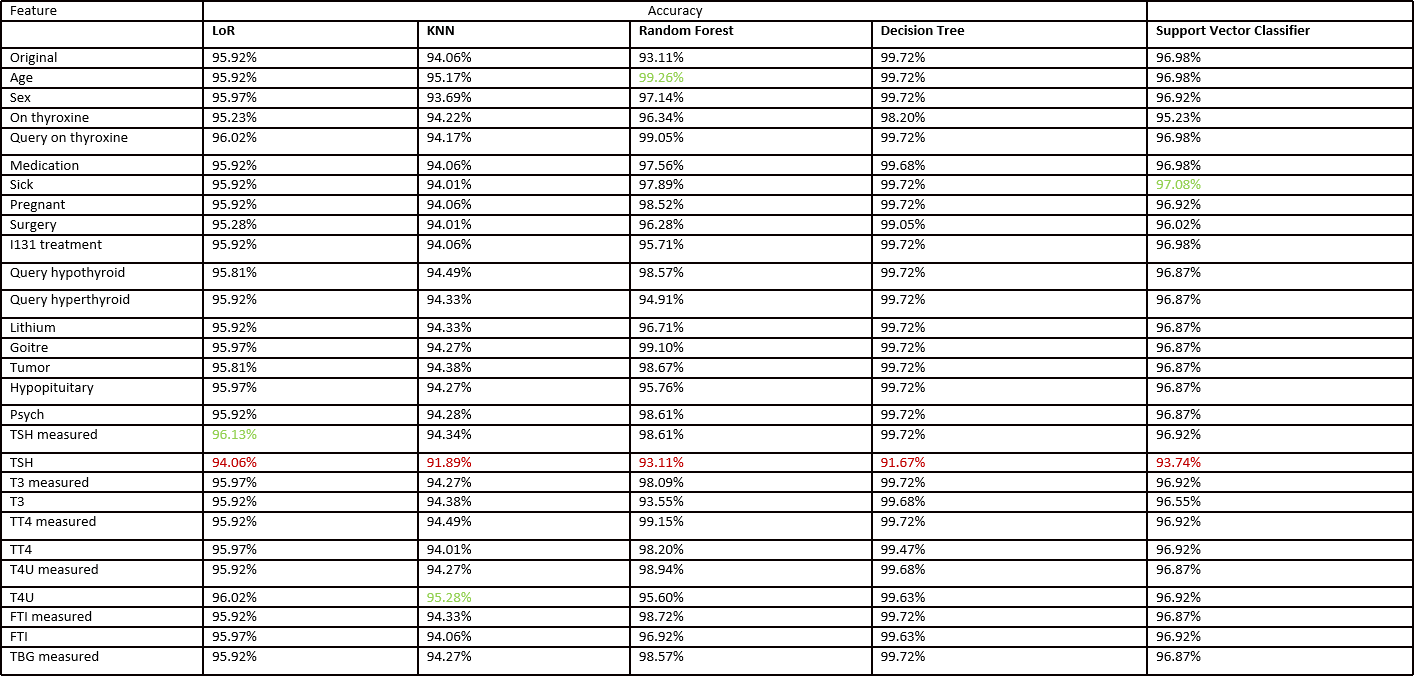
Final accuracy was found to be slightly higher (99.72% ---> 99.89%).

EVALUATING THE MODEL

Final evluation metrics for Decision Tree after hyperparameter tuning and Cross Validation:



FEATURE IMPORTANCE USING SHAPELY ALGORITHM



* LoR: Accuracy decreased the mose on removing feature TSH. It increased the most on removing feature TSH measured.

Most import feature: TSH

Least important feature: TSH measured

* KNN:

Most import feature: TSH

Least important feature: T4U

* RF:

Most import feature: TSH

Least important feature: age

* DT:

Most import feature: TSH

Least important feature: Accuracy remained almost the same when other features were removed.

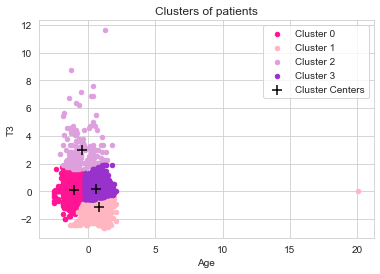
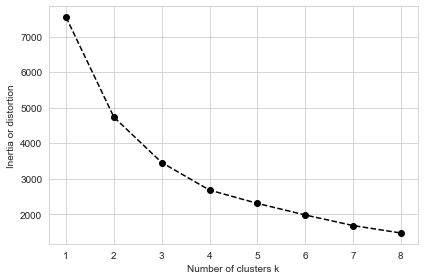
* SVC:

Most import feature: TSH

Least important feature: sick

KMEANS CLUSTERING (UNSUPERVISED LEARNING)

By plotting number of clusters against inertia, I got best k (number of clusters)=4.



Cluster centroids are: [[-1.00486819 0.08231764]

[ 0.81368309 -1.1619196 ]

[-0.45584153 2.9871567 ]

[ 0.56775931 0.13522548]]

Percentage of healthy and unhealthy labels in first cluster is 6.53 % and 93.47 % respectively.

Percentage of healthy and unhealthy labels in second cluster is 19.8 % and 80.2 % respectively. (Comparatively impure cluster). Percentage of healthy and unhealthy labels in third cluster is 2.38 % and 97.62% respectively.(Purest cluster-mostly hypothyroid patients). Percentage of healthy and unhealthy labels in fourth cluster is 3.78 % and 96.22 % respectively.

Cluster1: Patients with lower age and lower T3. Cluster2: Patients with comparatively higher age and lower T3

Cluster3: Patiets with higher age and slighty higher T3. Cluster4: Patientswith higher T3.

PROJECT SCOPE

The project can be extended to classifying which type of thyroid disease a ptient has (hyperthyroidism, hypothyroidism, thyroiditis, etc.) using multi label classification. Lime alorithm can be used to explain feature importance. Similar techniques can also be used for proactive detection and diagnosis of other diseases in order to aid the clinical decision making process.

SOME SCREENSHOTS OF MY CODE

