THYROID PREDICTION PROJECT

CONTENTS:

BY: ANAM FATIMA

1. Introduction to our project

“Thyroid Prediction” – As the name suggests we’ll build a model that predicts the Thyroid Disease in a patient. Early detection and accurate diagnosis are crucial for effective treatment of Thyroid Disease. In this project I have used Machine learning algorithms and compare multiple algorithms like logistic regression, lasso, OneVsRestClassifier, elastic net. The goal is to establish a reliable and efficient predictive framework that can assist healthcare professionals in identifying individuals at risk and providing timely treatment, thereby enhancing patient outcomes and reducing the burden of thyroid-related health issues.

The goal is to establish a reliable and efficient predictive framework that can assist healthcare professionals in identifying individuals at risk and providing timely treatment, thereby enhancing patient outcomes and reducing the burden of thyroid-related health issues.

2. PROBLEM STATEMENT

Here is the dataset of thyroid disease. There are various components like age, values of TSH,TT4,T4U,T3,FTI,whether the patient is on thyroxine treatment, having goiter, tumor, pregnant, previously done thyroid surgery.

I have to train my machine learning model to predict whether the patient is having thyroid or not , by getting al the values of TT4,T4U,FTI,T3,TSH.

4. Data Preprocessing

Steps to follow in data preprocessing

- **Handling Missing Values**
- **Handling Categorical Values**
- **Checking the distributions of features**
- **Removing the outliers**

First find the columns which have missing values, Dropping the columns which have missing values, Dropping the rows which have missing values, Mean/Median replacement, Random Sample replacement, End of Distribution value replacement.

- **Easiest way is to delete (drop the column) which have the missing values but it is a loss of the data at large level. Like if we single-single missing values in each columns should we drop every column. Obviously not, so it's not the right approach**

- **Other way is dropping all the rows which have missing values and this is quite a bit good approach but if we have small dataset, dropping rows is also not a good approach**
- **Replacing the all the missing values with some values is although a very good approach because no loss of columns, no loss of rows in short no loss of data.**

- **<u>Handling categorical feature</u>**

Categorical features are those features which have values in form of string.

We can replace the missing values of categorical feature (column) with the most frequent value. This would be the good approach for imputing categorical features. Now all machine learning algorithms/models does work well with categorical features and it's not good approach to drop the columns as we discussed above. So, let's replace the value of categorical features with a number. We are doing onehotencoding with categorical features.

5. Feature Engineering

Checks the distribution and treat the outliers are those which are damaging our models and have an reverse relation with our model. For example: - we have age column in which almost values are above 12 and below 40, but some values are 70,80. So these values are considered as an outlier. We can see the outlier's using boxplot for every feature. Know more about outliers.

6. Feature Selection

Now that maybe the case that columns are not important for our dataset and model so should drop that column. We'll find the relation using corr() method of pandas to find out the relation between each column. We have two types of features i.e. Dependent Feature and Independent Features. Corr() gives a number which shows the strength of relation . we see a high relation between two independent features, we'll drop one of them. If we see a high relation between independent feature and dependent feature, we'll keep the relation because it is important for our model. We can see the relation by plotting it using heatmap.

I've used feature selection library of sklearn module to find the top 10 essential features that we can keep. We are using chi_2 method to find it.

7. Playing with different models and checks the accuracy

Now we build the model for prediction. I've splitted the data in training data and testing data using train_test_split and classifies the model KNeighborsClassifier, SVC, RandomForestClassifier, AdaBoostClassifier, GaussianNB, I've build the models one by one and tune it also using hyperparameter tuning with GridSearchCV library.

8. Pickle file

Now we have to import our model to other source,editor,codebase etc. We cant take our whole file ,so for this we'll make a model file (model.pkl).