Report on a Machine Learning Model using Regression Algorithm

Problem Statement:

- Machine Learning algorithms always helps in providing insights / predictive results, but based on the dataset and algorithm using, the results will vary. In this study, we are trying to implement a regression-based dataset in supervised machine learning model.
- The dataset is related to the indicators of the heart diseases. Some of the variables present in the dataset are mentioned in here:
    - Age
    - Gender
    - Blood Pressure
    - Smoker
    - Heart Stroke
    - BMI
    - Diabetes

Solution Pipeline:

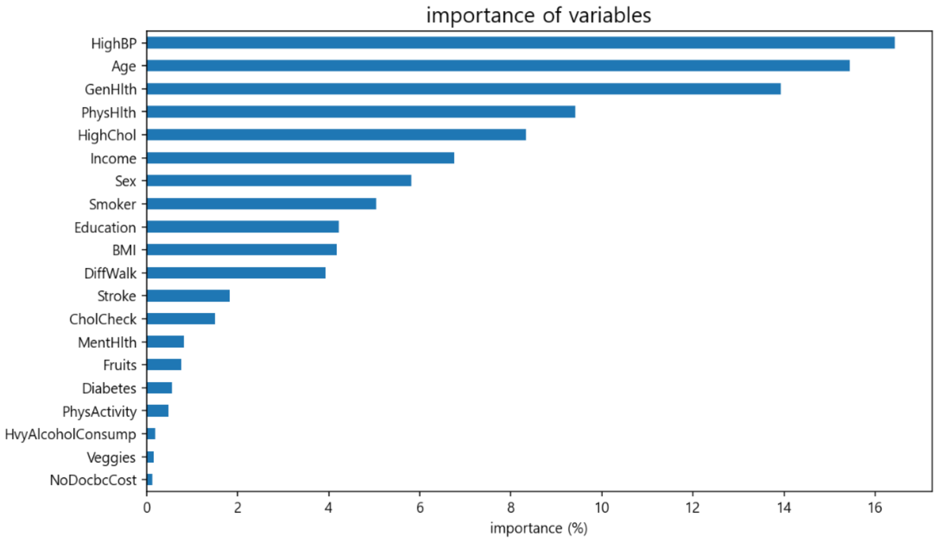
- For building a machine learning model we need to first understand the data, for providing the best algorithm. And with the algorithm selected we will be having many evaluation metrics to identify the best algorithm and use that one for prediction. So here is the pipeline for completing the entire process mentioned here:
    1. First, we identified the problem we are focusing that is “to build a model for predicting the possibility of the heart attacks”
    2. Selecting the correct dataset with correct variables – in this case “Heart Indicators Dataset”
    3. Started building the program for prediction

4. Importation of packages required – for processing, cleaning, visualizing, modeling, evaluating, algorithms, and prediction.
5. Performing the data preprocessing steps like removing null values, understanding the data, and finding the key features.
6. Checked the data for null values – no null values found in this scenario.
7. Checked the data for unique values – to identify the key attributes.
8. Created the confusion matrix to identify the correlation between the features.
9. Performed the data visualization to get any insights on the data.
10. Used some basic python functions to look at the data like head(), describe(), info(), dtypes(), tail() and others.
11. Now some insights are obtained like size of the dataset, unique values in the dataset, main features in the dataset and algorithms that can be used on the dataset.
12. Now the main part of the program building that is data modeling.
13. Since the dataset is ready to use, we will split the data into training data and testing data.
14. Using python functions, we will correct the data that is divided. Functions used are scaler and over sampling.
15. Now to understand the potential of the model, we have created a user-defined function to identify the performance. Some of the main results obtained from this function are accuracy, precession, mean absolute error, and f1 score.
16. Now as the basic functions are ready and the data is divided as required, we will be creating the algorithms and training them with the data. Here are the algorithms used:
    a. Logistic Regression
    b. XG Boost
    c. Light GBM
    d. Voting Classifier – A algorithm to combine the all the other algorithms.
17. After all the main steps are completed, we have concluded that the voting classifier has the better results compared to other individual results.

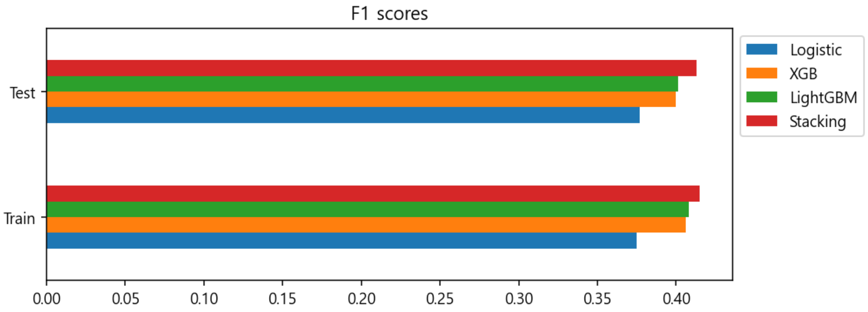
18. Now using the voting algorithm as the base, we will be taking inputs from the user – health related information. And the program will be able to predict the chances of person having heart attack.
19. To conclude we have performed the results on the entire dataset.
20. For better understanding, we have identified the main features to cross-check. And concluded the results with visualization graphs.

Results Obtained:

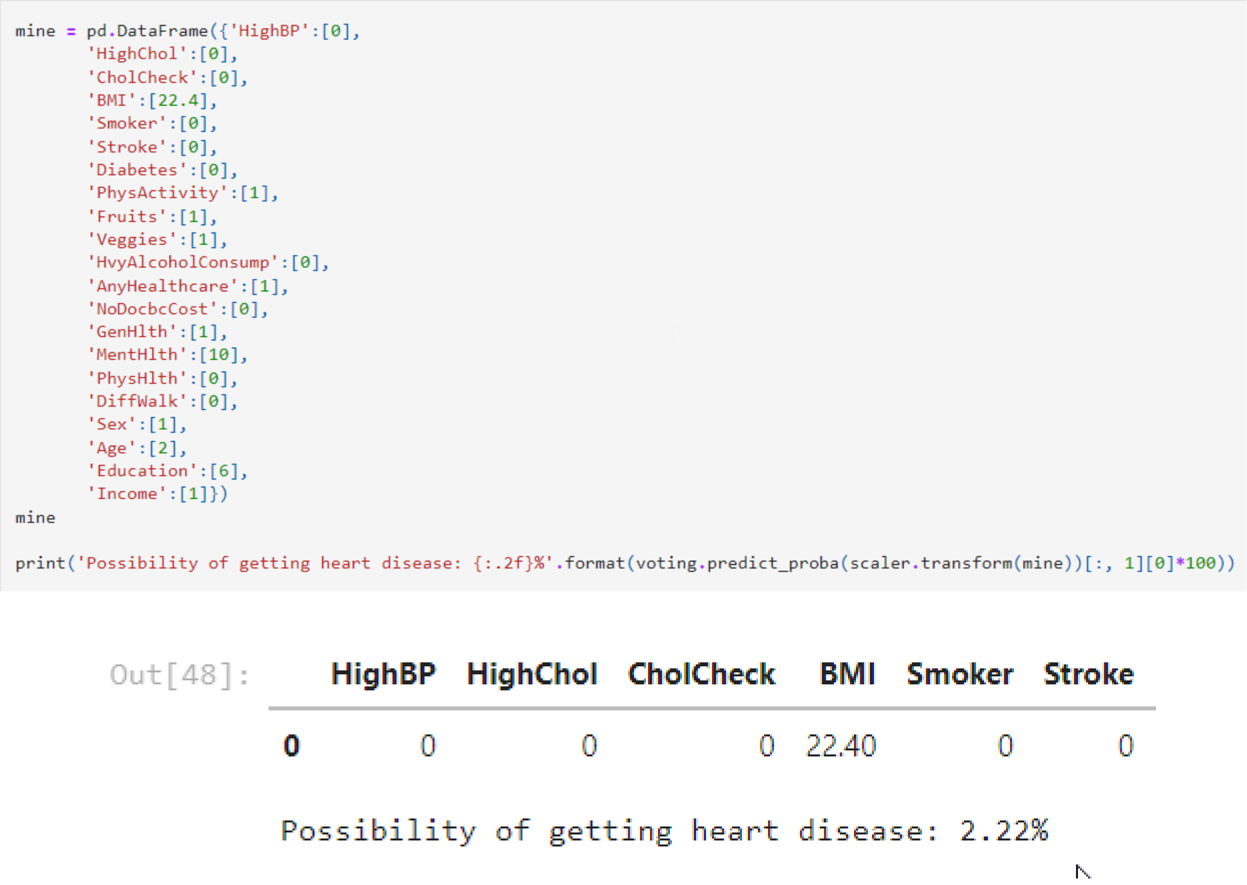
- From the program we have identified some of the main features that are responsible for any heart-disease are



- Here are the main results that we have obtained from this program.



Results of the performance of the Machine Learning Algorithms



```
mine = pd.DataFrame({'HighBP':[0],
        'HighChol':[0],
        'CholCheck':[0],
        'BMI':[22.4],
        'Smoker':[0],
        'Stroke':[0],
        'Diabetes':[0],
        'PhysActivity':[1],
        'Fruits':[1],
        'Veggies':[1],
        'HvyAlcoholConsump':[0],
        'AnyHealthcare':[1],
        'NoDocbcCost':[0],
        'GenHlth':[1],
        'MentHlth':[10],
        'PhysHlth':[0],
        'DiffWalk':[0],
        'Sex':[1],
        'Age':[2],
        'Education':[6],
        'Income':[1]})
mine

print('Possibility of getting heart disease: {:.2f}%'.format(voting.predict_proba(scaler.transform(mine))[:, 1][0]*100))
```

Out[48]:

	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke
0	0	0	0	22.40	0	0

```
Possibility of getting heart disease: 2.22%
```

Results on a real-time predection

References:

1. https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset
2. https://heartbeat.comet.ml/5-regression-loss-functions-all-machine-learners-should-know-4fb140e9d4b0
3. https://towardsdatascience.com/forecasting-house-prices-with-python-kaggle-competition-5cc791a642de
4. https://towardsdatascience.com/performance-comparison-catboost-vs-xgboost-and-catboost-vs-lightgbm-886c1c96db64