Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Our primary goal is to understand the topic and get through all the possible insights and analyzed in every possible way. First, we need to analyze the topic then each and every variable in the dataset, after knowing the variables get through them analyzed and then used machine learning algorithms to find out the predicted analysis and compared with the actual one's to know about the models which worked more with accuracy.

As in cardiovascular risk prediction, heart disease is the major cause of morbidity and mortality globally: it accounts for more deaths annually than any other cause.

Of all heart diseases, coronary heart disease (aka heart attack) is by far the most common and the most fatal. Doctors and scientists alike have turned to machine learning (ML) techniques to develop screening tools and this is because of their superiority in pattern recognition and classification as compared to other traditional statistical approaches.

In this project, we will be giving you a walk through on the development of a screening tool for predicting whether a patient has a 10-year risk of developing coronary heart disease (CHD) using different Machine Learning techniques.

The most important evaluation metric that we go with, to address this problem statement, is the recall. Since we want to decrease the number of false negatives and predict all the true cases for 10-year risk of having CHD correctly, the focus of our ML models is to improve the recall.

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

The project workflow as Splitting to Train and Test sets to avoid Data bleeding, Simultaneously Data Cleaning of Train and Test sets after that EDA on features and Feature cleaning and engineering then Solving Class Imbalanced Problem with the use of Base Model and Candidate Models and at the end Hypertuning of parameters then our Final Predictions.

First, we split the data in 80-20 ratio before doing any sort of data cleaning or EDA so as to avoid data bleeding. Then, null values in the categorical values are imputed using a simple imputer which replaces the null values with the value that is most frequent in the particular feature. While the numerical features are imputed with a KNN imputer that imputes the null values with values that are close to the values of K nearest neighbors of that particular sample. Imputation like these ensure that none of the data is significantly lost and data is safely preserved in the data set.

Then in the exploratory data analysis that we performed on our train dataset helped us to realize how different features in our dataset influence the target variable. We used a chi-square test to find whether a certain categorical feature was dependent on our binary target variable. This helped us to decide whether certain features need to be included in our final model or not.

From the above EDA we try to establish some patterns which influence the cause of heart disease. We have tried to engineer some new features based on existing features by bucketing some of the continuous variables. We have created age buckets of population e.g., 18-25 -> **20s**, 25-40 -> **Mid30s** etc. In this way we might be able to target a particular age group which has a high risk of coronary heart disease. We also tried to reduce the multicollinearity from the dataset by removing features that are highly correlated, such as diabetes and glucose level, smoking and number of cigarettes per day, hypertension and systolic blood pressure.

In this problem we have a dataset of patients where we have to find out whether the given features or symptom a person has, he/she has a cardiovascular disease in future.

But here's the catch… the risk rate is relatively rare, only 15% of the people have this disease.

One of the major issues when dealing with unbalanced datasets relates to the metrics used to evaluate our model. Using simpler metrics like accuracy score can be misleading. In a dataset with highly unbalanced classes, the classifier will always

"predict" the most common class without performing any analysis of the features and it will have a high accuracy rate, obviously not the correct one.

There are various candidate models that are used to train our datasets like logistic regression, SVM, decision tree, KNN, naïve bayes and SGD After training each model and tuning their hyper-parameters using grid search, we evaluated and compared their performance using the following metrics:

The accuracy score: which is the ratio of the number of correct predictions to the total number of input samples.

The F1 Score: Which is defined as the weighted harmonic mean of the test's precision and recall. By using both precision and recall it gives a more realistic measure of a test's performance. (Precision, also called the positive predictive value, is the proportion of positive results that truly are positive. Recall, also called sensitivity, is the ability of a test to correctly identify positive results to get the true positive rate).

The Recall: Which provides an aggregate measure of performance across all possible classification thresholds. It gives the probability that the model ranks a random positive

After comparing all the parameters in all the models and based on our observations, the Support vector classifier seems to have performed better with a recall of 74%. Based on the recall metrics, the model performance is really good, which was our objective from the very beginning i.e., we wanted to correctly predict all the positive cases of high-risk CHD.

However, we sometimes also don't want to flag somebody with no risk of CHD as positive, which might eventually increase the operational cost. We need to ensure that our precision is not too low as well that's where our best model still lags. The current precision of the SVC model is around ~0.26. Further work needs to be done that might possibly improve the precision of our model on minority class as well.

Contributor's Role:

1. **Kanika Singh**
    - Introduction
    - Data handling
    - EDA
        Continuous Variables
    - Correlation
        Correlation of all continuous variables
    - Correlation and multicollinearity of all variables
    - Model Performance
        Decision Tree
        KNN
        Stochastic Gradient Descent (SGD)
    - Metrics comparison
    - Conclusion

1. **Tanjul Gohar**
    - Introduction
    - Data handling
    - EDA
        Categorical Variables
    - Correlation
        Correlation of all categorical variables
    - Correlation and multicollinearity of all variables
    - Model Performance
        Logistic Regression
        Support Vector Classifier (SVM)
        Naïve Bayes Classifier
    - Metrics comparison
    - Conclusion

Please paste the GitHub Repo link.

GitHub Link: - https://github.com/Kanika211/Cardiovascular_risk_prediction.git

Please paste the drive link to your deliverables folder, ensure that this folder consist of the project Colab notebook, project presentation and video

Drive Link –

https://drive.google.com/drive/folders/1xMNpLNTEkiNQTJ_lPNB16EOVfrvGikCM?usp=sharing