Introduction & Problem description - Predicting the Risk parameters for Cardiovascular patients

In this project, the datasets of interest is the Heart health datasets.

Using this Kaggle Dataset I am going to analyze the data to get more insights about heart health Indicators.

The features like HighBP, HighChol, age, sex, BMI & Diabetes play important role in calculating Heart health.

I am going to use all the features which ever one are important and drop the ones which are not as critical.

I like to analyze the factors impacts the heart health of an individual, it could be age, sex, smoking, BMI & other health related indicators. We want to make few visualizations to see which one has the most impact.

Why heart health analysis is important to stakeholders -

About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare.

Originally, the dataset come from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents.

As described the dataset was reduced to just about 20 variables.
We will identify the Heart disease using various Model classification methods.

We will use various plots and charts to see which variables has the most impact on heart health i.e. diabetes, smoking, lake of physical activity

DataSets & Sources-

Kaggle https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Overall, I like to clean the data as needed by dropping off columns which may not be required in further data analysis / visualizations, adding new columns / updating the data elements in selected columns to maintain consistent relationships between various data sources and then create the planned visualizations.

I like to analyze the factors impacts the heart health of an individual, it could be age, sex or other health indicators. We want to make few visualizations to see which one has the most impact.

In other milestones of the project & , I will explore pull the data from Kaggle apply data mining techniques that I have learned throughout the course.

Methods & Algorithms:

- EDA; include any visuals you think are important to your project
- Data preparation
- Model building and evaluation

I will use accuracy matrix method to analyze the data. The visualization of features importance allows us to understand more the effect of some features that the model consider more important in its classification. Thus, more process can be done to help the model reach high performance level. We can also continue to fine tune the hyper-parameters of the model to gain some % in the accuracy measure.

I will also use pipeline in this process as it helps to enforce desired order of application steps, creating a convenient work-flow. But, there is something more to pipeline, as we will used\ StandardScaler for cross validation, we can understand data bit better. .

I will also use Onehotencoder for categorical variables which used to turn categorical features into binary features that are “one-hot” encoded, meaning that if a feature is represented by that column, it receives a 1. Otherwise, it receives a 0.

Confusion Matrix also show lot of information.

Algorithms:

Decision Tree Classifier

The first method is decision tree which creates a decision tree for predicting categorical data which, it assigns the class values to each data point. Entropy/Information Gain and Gini Impurity are 2 key metrics used. Here, we can vary the maximum number of features to be considered while creating the model.

Support Vector Classifier

The second classifier aims at forming a hyperplane that can separate the classes as much as possible by adjusting the distance between the data points and the hyperplane. SVM is one of the most well-known supervised machine learning algorithms for classification.

K Neighbors Classifier

This classifier looks for the classes of K nearest neighbors of a given data point and based on the majority class, it assigns a class to this data point. However, the number of neighbors can be varied. I varied them from 1 to 20 neighbors and calculated the test score in each case.

Important Variables:

HeartDiseaseorAttack float64
HighBP float64
HighChol float64
CholCheck float64

BMI	float64
Smoker	float64
Stroke	float64
Diabetes	float64
PhysActivity	float64
Fruits	float64
Veggies	float64
HvyAlcoholConsump	float64
AnyHealthcare	float64
NoDocbcCost	float64
GenHlth	float64
MentHlth	float64
PhysHlth	float64
DiffWalk	float64
Sex	float64
Age	float64
Education	float64
Income	float64

Ethical Considerations

During our Analysis, we haven't used any PII data. All Data Sets are extracted from Public Websites. Used Datasets have no restrictions for Academic usage All used Datasets are Shared by respective government bodies for public benefits.

Challenges/Issues

I think some challenges around the Hypothesis Testing and choice of the test statistic. I chose the difference between the means of the two groups as my main test statistics, however I still think could I use some other comparisons too as test statistics such as standard deviation or chi squared based tests.

I feel still as a novice and learning as I read and practice with different datasets.

References

Shraddha Chauhan, Bani T. Aeri “The rising incidence of cardiovascular diseases in India: assessing its economic impact” J. Prev. Cardiol., 4 (4) (2015), pp. 735-740

Bao, Lei, et al. "Boosted near-miss under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets." Neurocomputing 172 (2016): 198-206.

Data, Svetlana Ulianova. "Cardiovascular Disease dataset." Kaggle, 2020. https:// www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

L. Verma, S. Srivastava, P.C. Negi A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data J Med Syst, 40 (7) (2016), pp. 1-7

Drummond, Chris, and Robert C. Holte. "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling." Workshop on learning from imbalanced datasets. Vol. 11. Washington DC: Citeseer, 2003.

Eitrich, Tatjana, and Bruno Lang. "Efficient optimization of support vector machine learning parameters for unbalanced datasets." Journal of computational and applied mathematics 196.2 (2006): 425-436.

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.