1 Introduction, Motivation

Cardiovascular Disease often used interchangeably with “heart disease”, generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves, or rhythm, also are considered forms of heart disease.

The purpose of this project is to predict the effects of different parameters recorded in the data to predict risk parameters of the patient. By predicting so the physicians can determine high risk patients and can take better care of them thus helping them survive. This risk assessment is crucial for many existing treatment guidelines. For my case study, I am planning to apply the Logistic Regression and Decision Tree Classifier algorithms to predict risk parameters. The input into my algorithms will be age, bmi, PhysActivity, education, diabetes, HvyAlcoholConsump, CholCheck, GenHlth, and smoking (if the customer has any). My algorithm will use this information to output a prediction on whether a person will have heart disease 1 or not 0. Researchers can then target these people who are at risk of having heart disease and how can it be avoided.

2 Related Work, Literature Review

Heart disease has been identified as one of the largest causes of death even in developed countries. One of the reasons for fatality due to heart disease is due to the fact that the risks are either not identified, or they are identified only at a later stage.

I did not have to deal with problem of unbalanced datasets for my study which is a common issue in machine learning datasets. One general solution to this problem is simply under-sampling the larger class in order to balance out the dataset.

I also reviewed the quality of data which is very important for the accuracy of any machine learning model. Identified and removed duplicates.

Numerical values needs to stored correctly if represented as strings i.e. age was stored as days so converted it correctly for better analysis.

I also reviewed dealing with many dimensions, Dimensionality reduction ML method used to identify patterns in data and deal with computationally extensive problems. This method includes a set of algorithms aimed to reduce the number of input variables in a dataset by projecting the original high-dimensional input data to a low-dimensional space.

Dimensionality reduction is helpful when simplifying datasets in order to better fit a predictive model. There are many examples of dimensionality reduction algorithms, but I'd like to emphasize UMAP, which we found useful when working on one of the machine learning projects.

3 Dataset and Features

The dataset I used is titled "heart_disease_health_indicators_BRFSS2015" from Kaggle. It contains 253,680 survey responses data as a result of Behavioral Risk Factor Surveillance System (BRFSS) that is a health-related telephone survey that is collected annually by the CDC.of various patients. For each patient, there are 21 features and a [0,1] label of whether they have a Presence or absence of cardiovascular disease. The specific features for each patient are described in detail below:

```
#   Column                Non-Null Count   Dtype
---  ------               --------------   -----
 0   HeartDiseaseorAttack  253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   Diabetes              253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
```

I have opted to use 70% of our dataset for training, and the remaining 30% for testing. The dataset was very easy to use and not much preprocessing was needed.

In terms of normalization, I tried to work with categorical variables, break each categorical column into dummy columns with 1s and 0s also to scale the dataset used the StandardScaler.

4 Methods

4.1 Decision Tree Classifier

The first method used 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. I range features from 1 to 13.Also counting the dummy variables for gender.

4.2 Logistic Regression

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. For a given set of training data, each marked as belonging to one of two categories, SVM training algorithm develops a model by finding a hyperplane, which classifies the given data as correctly as possible by maximizing the distance between two data clusters.
The model showing the best accuracy seven features (age, gender, bp, chol, Smoking).

4.3 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. Then, I plot a line graph of the number of neighbors and the test score achieved in each case.

Another algorithm I experimented with was the K-Means Algorithm. I only used two centroids because an individual can either have heart decease, or no. To start, the centroids will be randomly initialized to two of the data points in the dataset. Then, data points will be assigned to a centroid depending on which one is closer. Next, the

centroids will re-center to be the mean of all points in their respective clusters. This process repeats until convergence is reached. For K-Means, convergence is reached when the two clusters created are the same as the clusters in the last iteration of the algorithm.

5 Experiments/Results/Discussion

A comparative analysis of various classification algorithms on the dataset has been performed. Some algorithms show good accuracy whereas some other algorithms perform average.

MultiCollinearity¶

we will have to check and remove multicollinearity from the data to get reliable coefficients and p-values. There are different ways of detecting (or testing) multi-collinearity, one such way is the Variation Inflation Factor. General Rule of thumb: If VIF is 1 then there is no correlation among the predictor and the remaining predictor variables. Whereas if VIF exceeds 5, we say it shows signs of high multi-collinearity. But the purpose of the analysis should dictate which threshold to use.

The blood pressure variables have a very interesting relationship. From the correlation chart it was shown that they have very low correlation(relationship) with the other numeric variables in the dataset. I went ahead to plot them together to see their relationship besides the correlation number. I also discovered relationships between the physical activity and HighBP negative relationship, that physical activity increases as HighBP decreases.

I also discovered another interesting relationship between the cholesterol and the cardiovascular disease variable.

	Model	Train_accuracy	Test_accuracy	Train_Recall	Test_Recall
0	Logisitic Regression with Optimal Threshold 0.104	0.89	0.89	0.35	0.34
1	Initial decision tree model	0.99	0.85	0.99	0.26
2	Decision treee with hyperparameter tuning	0.90	0.90	0.08	0.08
3	Decision tree with post-pruning	0.82	0.82	0.56	0.55

Observation:

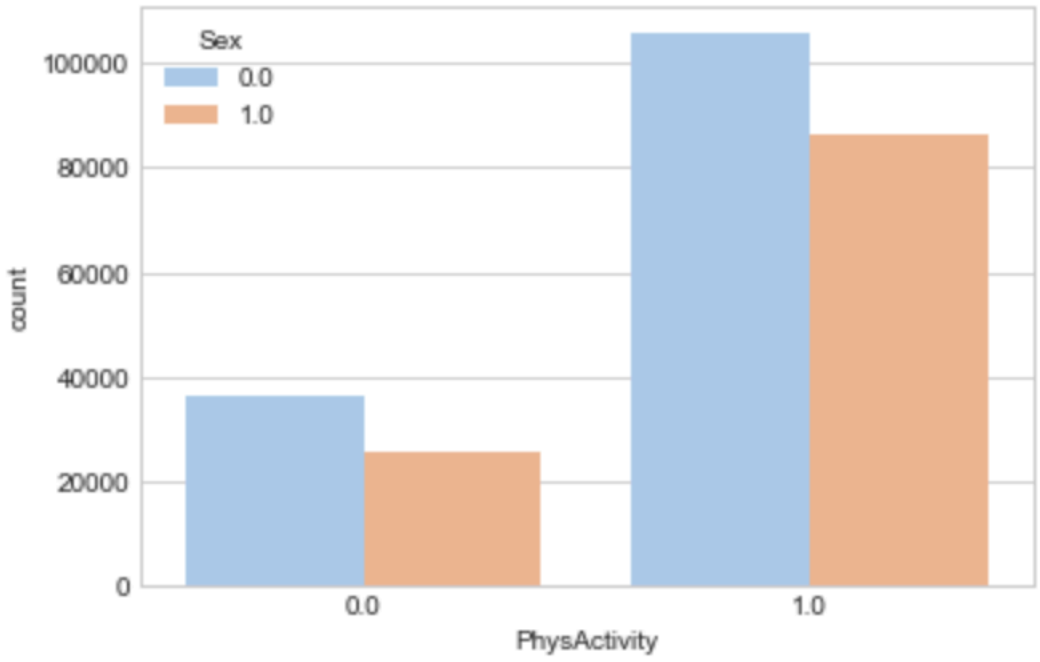
Decision tree model post pruning has given us best recall scores on data with 82% accuracy . Exploratory data analysis also suggested Age and GenHealth were important features in deciding if person will have heart issues. so choosing Decision Tree with post-pruning for our prediction.

Decision trees doesn't require to much data preparation or handling of outliers like logistic regression. They are easy to understand. Decision tress can easily overfit , so we have to be careful using decision tree.

6 Conclusion/future works

Conclusions

1) Here we gonna see how physical activity is important to be good and avoid lot of things such as Diabetes and HeartDiseaseorAttack and etc



2) The amount of Males that have cardiovascular diseases is more than the females that have this cardiovascular disease.

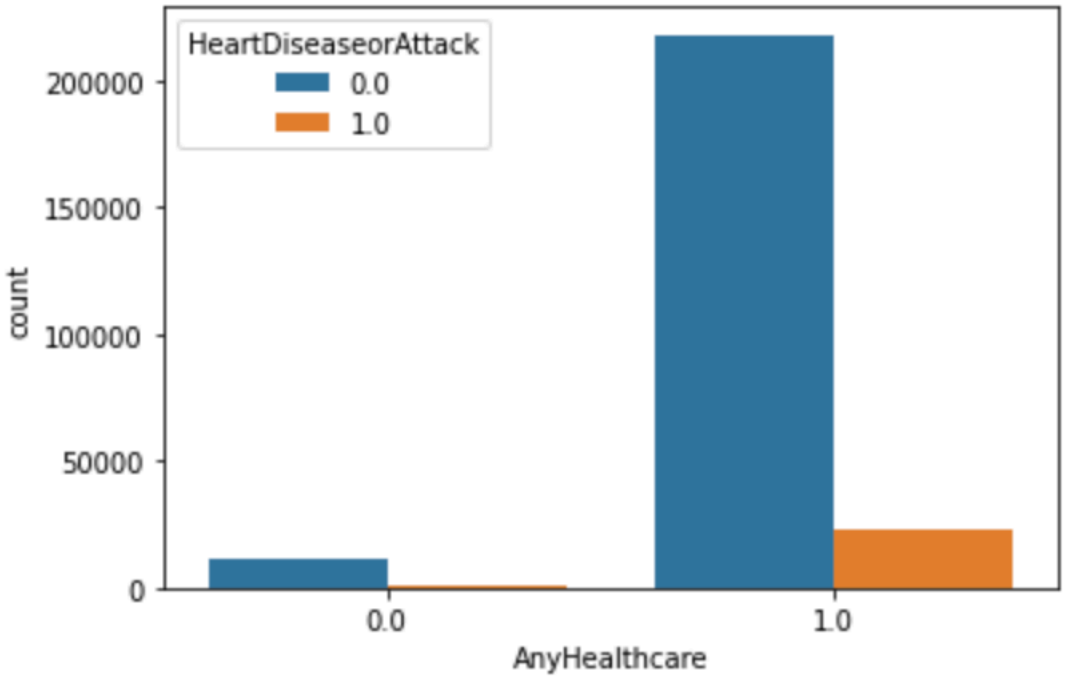
This was first explored by checking the general population plot where I discovered that half population did not have heart disease. I furthered explored the distribution with gender where I observed and concluded that more males have cardiovascular diseases

than the females. Increase in Age, number of cigarettes smoked show increasing odds of having heart disease.

3) All attributes selected after the elimination process show Pvalues lower than 5% and thereby suggesting significant role in the Heart disease prediction.

4) how healthcare is so important and how it affect on person and make him good more than people who don't have
;.

.



For future work, I would like to explore **Naïve Bayes** and also research more solutions for classification problems with datasets. As mentioned above, a specific area of improvement i would like to pursue is under-sampling and/or over-sampling in my dataset. Came across a literature recently using combination of under and over sampling techniques to increase effectiveness of data so would explore that and also will research if hyper parameter tuning could improve the prediction power of algorithms I am using.

was found to be the best algorithm, followed by neural networks and decision trees.

Multivariate Exploration Summary is also something I am really interested in.

7 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.

Karan, towardsdatascience, Predicting presence of Heart Diseases using Machine Learning

10 Questions from the Audience

1. What are the risk factors you identified which may lead to Heart Issues from this case study?
2. How well is your model accuracy defined?

3. What do you suggest to make the model better?

4. What is your plan to explain AI/ML to patients and gain confidence in using this model?

5. How is this model different from other models available?

6. How reliable this model is?

7. What is the implementation time and effort?

8. What calibration methods are being used to ensure the model is working as planned?

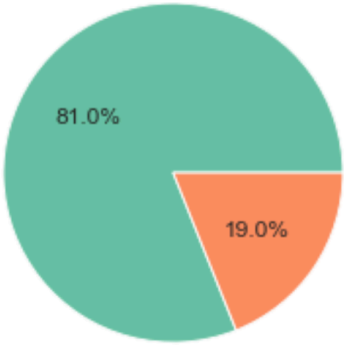
9. Any additional consumers due diligence expected for model in case of false negatives?

10. Should we be concerned that the model could potentially adversely impact certain demographic?

11 Ethical Considerations:

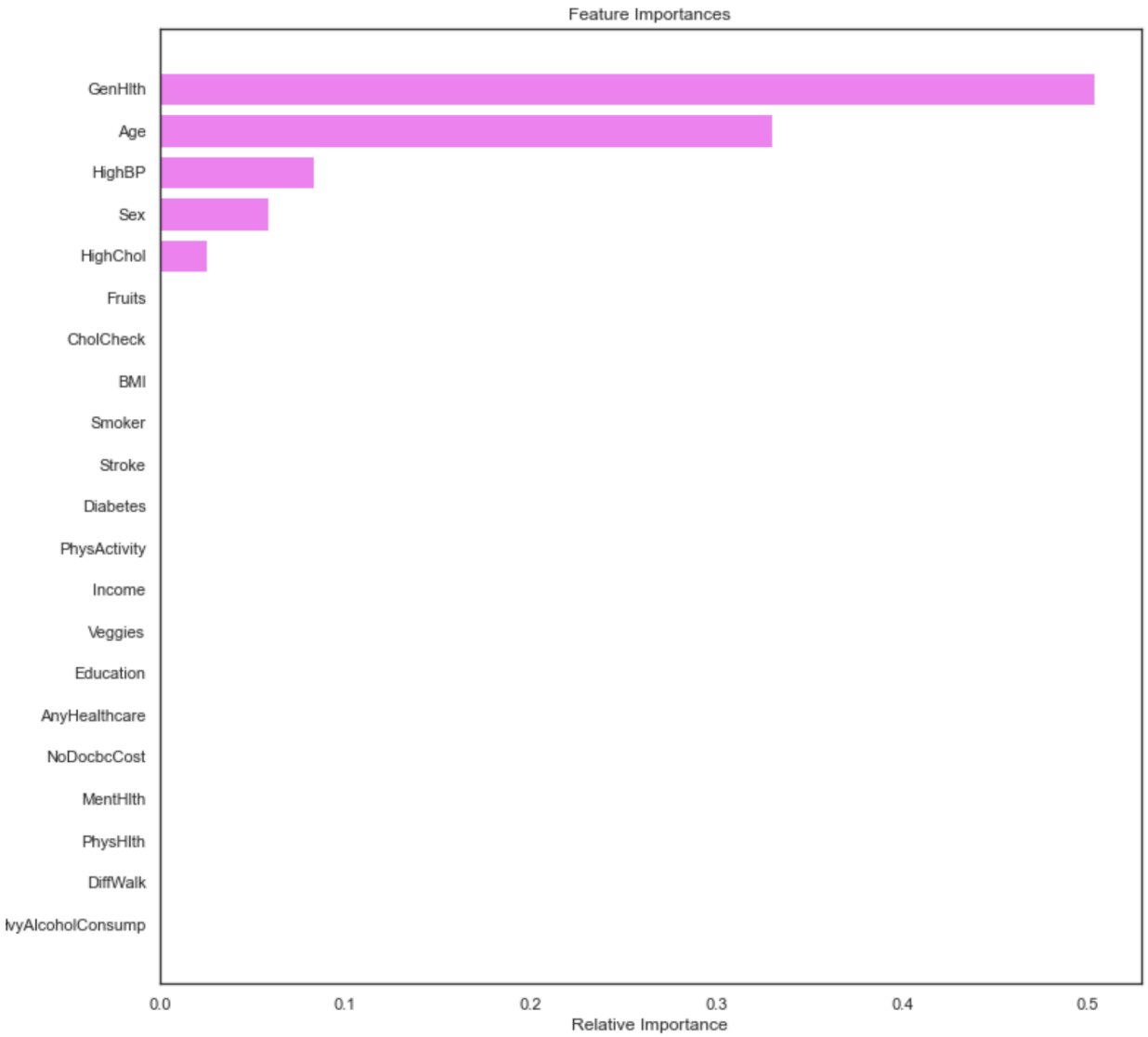
1. This Data contains processed physical images information related to multiple verities of rice and does not contains any PII-related information.
2. Datasets and information on data were extracted from the public websites kaggle machine learning repositories.
3. This data research is not going to harm any privacy.

11. Appendix



Percentage of value predicted by our model has been very close to the actual values. Lets find out False Negative and False Positive observations

Chart below describes which features are more important for heart health.



These are the various methods of Machine Learning algorithms I have used and analysis is performed.

	Model	Train_Accuracy	Test_Accuracy	Train Recall	Test Recall	Train Precision	Test Precision	Train F1	Test F1
0	Logistic Regression Model-Sklearn	0.61	0.61	0.92	0.92	0.18	0.18	0.31	0.31
1	Logistic Regression Model - Statsmodels	0.91	0.91	0.13	0.13	0.55	0.53	0.21	0.21
2	Logistic Regression - Optimal threshold = 0.092	0.72	0.72	0.83	0.84	0.23	0.23	0.36	0.36
3	Logistic Regression - Optimal threshold = 0.3	0.89	0.89	0.35	0.34	0.42	0.42	0.38	0.37
4	Logistic Regression - Sequential feature selec...	0.58	0.58	0.92	0.92	0.18	0.17	0.29	0.29

Since we want higher Recall with higher accuracy Optimal Threshold 0.3 seems to be a good choice. Lets explore a model with decison tree if this score can be improved further.

```
# Text report showing the rules of a decision tree -

print(tree.export_text(best_model,feature_names=feature_names,show_weights=True)
```

```
|--- GenHlth <= 3.50
|   |--- Age <= 9.50
|   |   |--- HighBP <= 0.50
|   |   |   |--- weights: [10346.25, 905.25] class: 0.0
|   |   |--- HighBP >  0.50
|   |   |   |--- HighChol <= 0.50
|   |   |   |   |--- weights: [1983.30, 368.90] class: 0.0
|   |   |   |--- HighChol >  0.50
|   |   |   |   |--- weights: [1877.10, 1105.00] class: 0.0
|   |--- Age >  9.50
|   |   |--- Sex <= 0.50
|   |   |   |--- GenHlth <= 2.50
|   |   |   |   |--- weights: [2382.30, 753.95] class: 0.0
|   |   |   |--- GenHlth >  2.50
|   |   |   |   |--- weights: [1484.70, 1276.70] class: 0.0
|   |   |--- Sex >  0.50
|   |   |   |--- GenHlth <= 2.50
|   |   |   |   |--- weights: [1642.80, 1409.30] class: 0.0
|   |   |   |--- GenHlth >  2.50
|   |   |   |   |--- weights: [969.60, 1929.50] class: 1.0
|--- GenHlth >  3.50
|   |--- Age <= 8.50
|   |   |--- HighBP <= 0.50
|   |   |   |--- weights: [799.05, 371.45] class: 0.0
|   |   |--- HighBP >  0.50
|   |   |   |--- weights: [863.55, 1263.10] class: 1.0
|   |--- Age >  8.50
```