Predicting Heart Disease from Health and Lifestyle Factors

Kavya K.

Introduction

- Heart disease is the leading cause of death.
- Around 697,000 people died from heart disease in the US.
- High blood pressure, high cholesterol and smoking are key risk factors for heart disease.
- According to the CDC, about 47% of Americans have at least one of the three risk factors.
- Several other conditions and lifestyle choices, such as diabetes and obesity, can also put people at a higher risk of heart disease.

Executive Summary

- In this project, I’ve used the data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS).
- The objective is to use 21 risk factors to build models for predicting risk of heart disease using supervised machine learning algorithms. This study can raise awareness for those who may have a potential high risk of heart disease.
- There is 90% class imbalance that is dealt with by using under-sampling.
- Using imbalanced data for training leads to the accuracy paradox. The accuracy is high, but other parameters are skewed.
- Overall, the best performing model is logistic regression followed by random forest.

Problem Statement

- Data from 2015 Behavioral Risk Factor Surveillance System (BRFSS) is used in this project.
- The questions are regarded as the features and the responses are the recorded observations.
- Main objective is to build and compare supervised learning models for predicting heart disease risk, based on the given health and lifestyle factors.
- Additionally, we will also address the problem of using highly imbalanced datasets for training models.

Related Work

- In 2020, a study compared results between MLP Neural Network and k-NN models, using the UCI heart disease data. MLP achieved 82% accuracy and k-NN was 74% accurate.
- Using 2014 BRFSS data, a study conducted by CDC researchers developed a predictive model to identify risk factors for Type II diabetes. Neural networks had the highest accuracy (82%), specificity (90%) and AUC score (79%).
- A 2023 study trained various machine learning models, using a real-world dataset to evaluate how accurately these models can predict risk of cardiovascular diseases. They concluded that the multilayer perceptron was the best performing model, having an accuracy of 87% and AUC score of 0.95.

Proposed Work: Dataset and Tools

- Dataset: 2015 BRFSS data containing 253,680 responses and 21 features. (Available on Kaggle)
- Tools: Implemented in Jupyter Notebook, using Python
  - Python libraries used
    - pandas
    - numpy
    - Scikit-learn
    - Seaborn
    - Imbalance-learn
    - Matplotlib

Dataset Features and Description

#	Features	Data Type	Description
1	HeartDiseaseorAttack	Binary	0 – No heart disease 1 – Has heart Disease
2	HighBP	Binary	0 – No high blood pressure 1 – Has high blood pressure
3	HighChol	Binary	0 – No high cholesterol 1 – Has high cholesterol
4	CholCheck	Binary	0 – Hasn't checked cholesterol in the last 5 years 1 – Has checked in the last 5 years
5	BMI	Numeric	Data Range – (12, 98)
6	Smoker	Binary	0 – Has not smoked at least 100 cigarettes in entire life 1 – Has smoked at least 100 cigarettes in entire life
7	Stroke	Binary	0 – Never suffered from a stroke 1 – Has suffered from a stroke
8	Diabetes	Numeric	0 – No diabetes or only during pregnancy 1 – Pre-diabetes or borderline 2 – Has diabetes
9	PhysActivity	Binary	0 – No physical activity in last 30 days 1 – Has physical activity in the last 30 days
10	Fruits	Binary	0 – No fruit consumption per day 1 – Consumes one or more fruits per day
11	Veggies	Binary	0 – No vegetable consumption per day 1 – Consumes one or more vegetables per day
12	HvyAlcoholConsump	Binary	0 – No heavy drinking 1 – Heavy drinking
13	AnyHealthCare	Binary	0 – No healthcare access 1 – Has healthcare access
14	NoDocbcCost	Binary	0 – Has met a doctor in last 12 months 1 – Has not met a doctor in last 12 months due to cost
15	GenHlth	Numeric	1 – Excellent, 2 – Very Good, 3 – Good 4 – Fair, 5 – Poor
16	MentHlth	Numeric	Number of bad mental health days in a month. Range – (0, 30)
17	PhysHlth	Numeric	Number of bad physical health days in a month. Range – (0, 30)
18	DiffWalk	Binary	0 – No difficulty while walking 1 – Has difficulty while walking
19	Sex	Binary	0 – Female 1 – Male
20	Age	Numeric	Age categories: 1 – 18-24, 2 – 25-29, 3 – 30-34, 4 – 35-39, 5 – 40-44, 6 – 45-49 7 – 50-54, 8 – 55-69, 9 – 60-64, 10 – 65-69, 11 – 70-74, 12 – 75-79, 13 – 80 and older
21	Education	Numeric	Categories: 1 – Never attended school or only kindergarten, 2 – Grades 1 to 8, 3 – Grades 9 to 11, 4 – High school graduate, 5 – College 1 to 3 years, 6 – College graduate or higher
22	Income	Numeric	1: <$10 K, 2: $10–$15 K, 3: $15–$20 K, 4: $20–$25 K, 5: $25–$35 K, 6: $35–$50 K, 7: $50–$75 K, 8: >$75 K

Proposed Work: Main Tasks

- Data Loading and Cleaning
- Exploratory Data Analysis
- Feature Selection, Balancing the Dataset
- Train Models using Supervised Learning Algorithms
  - Logistic Regression
  - Random Forest
  - K-Nearest Neighbors
  - Naïve Bayes
- Model Analysis using Evaluation Metrics

Evaluation: Experimental Setup

- Experimental Setup
  - Dataset will be split 80:20 (80% for training and 20% for testing)
  - Each model will be trained with the full set of 21 features and a set of uncorrelated features
  - Three datasets will be used
    - Imbalanced dataset
    - Balanced dataset with resampling ratio 50:50
    - Balanced dataset with resampling ratio 60:40

Evaluation: Evaluation Metrics

- Evaluation Metrics
  - *Confusion Matrix*: Defined by four parameters
    - True positive (TP)
    - True Negative (TN)
    - False Positive (FP)
    - False Negative (FN)
  - $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
  - $Precision = \frac{TP}{TP+FP}$
  - $Recall = \frac{TP}{TP+FN}$
  - *AUC Score*

Predicted Labels

True Labels		1	0
	1	TP	FN
	0	FP	TN

Confusion Matrix Table

Key Results

For Imbalanced Dataset

	Model	Accuracy	Precision	Recall	AUC Score
0	Logisitic Regression	0.900864	0.548474	0.131003	0.559411
1	Random Forest Classifier	0.892530	0.398073	0.115137	0.547736
2	K-NN Classifier	0.890136	0.335891	0.084477	0.532806
3	Gaussian NB Classifier	0.806254	0.270910	0.537521	0.687064

Table 1: Model Performance Using All Features

	Model	Accuracy	Precision	Recall	AUC Score
0	Logisitic Regression	0.898209	0.511161	0.048848	0.521771
1	Random Forest Classifier	0.897687	0.483871	0.044795	0.519684
2	K-NN Classifier	0.886481	0.245428	0.054394	0.517698
3	Gaussian NB Classifier	0.853450	0.305751	0.343643	0.627503

Table 2: Model Performance Using Uncorrelated Features

Key Results

For Balanced Dataset using Resampling Ratio 50:50

	Model	Accuracy	Precision	Recall	AUC Score
0	Logisitic Regression	0.767269	0.675088	0.568753	0.716926
1	Random Forest Classifier	0.756728	0.645275	0.584291	0.712999
2	K-NN Classifier	0.710351	0.615849	0.325883	0.612853
3	Gaussian NB Classifier	0.733539	0.588098	0.643678	0.710751

Table 3: Model Performance Using All Features

	Model	Accuracy	Precision	Recall	AUC Score
0	Logisitic Regression	0.719064	0.640112	0.340358	0.623027
1	Random Forest Classifier	0.712810	0.588576	0.432099	0.641624
2	K-NN Classifier	0.675146	0.513446	0.304811	0.581231
3	Gaussian NB Classifier	0.710702	0.580094	0.447850	0.644045

Table 4: Model Performance Using Uncorrelated Features

Key Results

For Balanced Dataset using Resampling Ratio 60:40

	Model	Accuracy	Precision	Recall	AUC Score
0	Logisitic Regression	0.766543	0.711200	0.637416	0.740830
1	Random Forest Classifier	0.754368	0.678069	0.659091	0.735396
2	K-NN Classifier	0.725512	0.642159	0.608375	0.702187
3	Gaussian NB Classifier	0.727963	0.632451	0.658670	0.714165

Table 5: Model Performance Using All Features

	Model	Accuracy	Precision	Recall	AUC Score
0	Logisitic Regression	0.702585	0.653536	0.443392	0.650973
1	Random Forest Classifier	0.701399	0.605497	0.588805	0.678979
2	K-NN Classifier	0.677682	0.580223	0.513678	0.645025
3	Gaussian NB Classifier	0.696577	0.615813	0.511364	0.659696

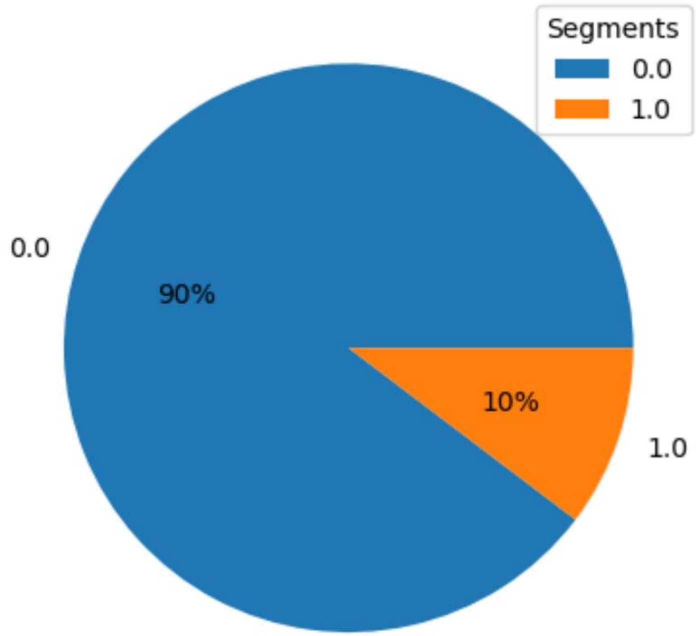
Table 6: Model Performance Using Uncorrelated Features

Timeline

	Timeline	Current Status
Data Loading, Cleaning, EDA	Week 1	Finished
Feature Selection, Data Preprocessing, Balancing Dataset	Week 2	Finished
Model Training	Week 2.5	Finished
Evaluation and Comparison	Week 3	Finished
Results and Conclusion	Week 3.5	Finished

Potential Challenges

- Data is heavily imbalanced.
- Cannot rely on accuracy as the only evaluation metric
- Alternate approach:
    - Create balanced datasets by resampling
    - I've used imblearn's RandomUnderSampling to resample the data into ratios 60:40 and 50:50



Key Findings

- When training on data with high class imbalances, it is necessary to use other metrics like precision, recall, f1 score and AUC to analyze the model performance.
- Models performed better when using the full set of features, instead of just set of uncorrelated features.
- Best performing model is Logistic Regression, followed by Random Forest.
- Gaussian NB model has a better recall than all the models.

Future Work

- A risk factor to be considered in later work is ethnicity as certain groups are more prone to heart disease.
- Use different feature selection methods such as Fisher score or Variance threshold
- Implement deep learning algorithms as they are efficient and self-adaptive. They are known to obtain better analyses and results.