IOE 491: Advanced Data Analytics
25 April 2022; Ann Arbor, MI

Analysis of Personal Key Indicators of Heart Disease

Harsh Hegde
Department of Industrial and Operations Engineering
Graduate Student
University of Michigan
Ann Arbor, MI
hegde@umich.edu

Charlie Hickman
Department of Industrial and Operations Engineering
Undergraduate Student
University of Michigan
Ann Arbor, MI
chickman@umich.edu

Daniel Korach
Department of Industrial and Operations Engineering
Undergraduate Student
University of Michigan
Ann Arbor, MI
dkorach@umich.edu

***Abstract*—Heart disease is widely acknowledged as one of the top causes of death around the world. The CDC's data collected through telephonic interviews with around 400k people would be valuable to gain an insight into the key predictors of heart diseases and their importance in estimating the chances of heart disease. To increase prediction accuracy, it's crucial to understand the precursors associated with heart disease and look at the different quantitative and qualitative indicators, the data was analyzed carefully and the correlation between the key predictors and the heart disease indicators was found. The performance of models generated using classification algorithms and relevant features was evaluated experimentally, this procedure is discussed in detail below. Three classification algorithms, namely Support Vector Machine, K-nearest neighbor, and Logistic Regression, were applied to the heart disease dataset as a consequence of the exploratory investigation. Each of the three algorithms was analyzed separately and it was found that each of the models was a capable predictor of those without heart disease, only undersampled(balanced) logistic regression provided the specificity worthy of a dual-directional predictive tool.**

I. Introduction

Heart disease is responsible for at least half of the deaths in the United States. Practitioners generally look at a vast combination of factors to determine a patient's susceptibility to heart disease. While our collective understanding of the causes of heart disease is growing, making precedent-based prognoses is still an immense challenge. Understanding how each factor contributes to an individual's health will be crucial as doctors look to make more justified and predictive diagnoses.

Considering that cardiovascular diseases have impacted large swathes of the population, over the years there have been many attempts at disease prediction. It is inappropriate for a person to frequently undergo costly tests like the ECG and thus there needs to be a system in place which is handy and at the same time reliable, in predicting the chances of heart disease. Features such as smoking, diabetes, general health condition, body mass indicators should generally be key indicators of heart diseases as found by several studies like Prabhakaran et al (2017) over the years Aggarwal et al. (2020) introduced a sequential feature selection strategy for identifying deaths in heart disease patients during therapy and finding the most critical aspects using LDA, KNN, and SVM. Sequential feature selection algorithms can be validated using the F-Method Score, precision, and recall rate. Al-Adhaileh et al. (2021) designed a detection model for kidney disease detection for 400 patients with 24 features. The k-nearest neighbors (KNN), support vector machine (SVM), decision tree, and random forest classification methods were used in this work achieving 100% accuracy.

Several studies suggest KNN is useful for predicting the presence of disease based on certain risk factors. Wang (2022) proposed the use of KNN for public health emergency decisions related to COVID-19. Kalita et al. (2022) used KNN to generate a model for the early identification of hypertension. KNN has also already been used to prevent cardiovascular disease by identifying risk levels (Li et al. 2022).

For medical applications, Huang, et. al. (2006) describe the substantial ability of SVM to take high-dimensional data and extract meaningful connections that improve prediction accuracy. In the application of heart disease, SVM will seek to maximize the distance of support vectors corresponding to subjects predicted to have heart disease and those predicted to not.

This study seeks to identify the most significant predictors of cardiovascular diseases and also analyze the predictions using different machine learning algorithms and identify the most useful algorithm. However, we haven't done a sensitivity analysis of each algorithm and in the future, there are more opportunities to explore the algorithms in depth. Using larger datasets and more powerful systems could have better results in the analysis.

II. Data

The data originated with the CDC and is a key component of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to collect information on Americans' health. The CDC-sourced data includes an array of binary (yes/no) and integer values of both scaled discrete and continuous types. Each is a question or measurement presented to or performed on the patient to create each anonymized row. Each year, the BRFSS conducts over 400,000 adult interviews, making it the world's biggest continually conducted health survey system. Data from 2020 is included in the most recent dataset (as of February 15, 2022). There are 401,958 rows and 279 columns in it. The data set contains 18 factors such as BMI, age, and weight are not patient-determined whereas the majority of the binary variables and a subset of discrete columns are comparatively subjective and patient-provided. A summary of some of the different factors can be found in Figures 1-7.

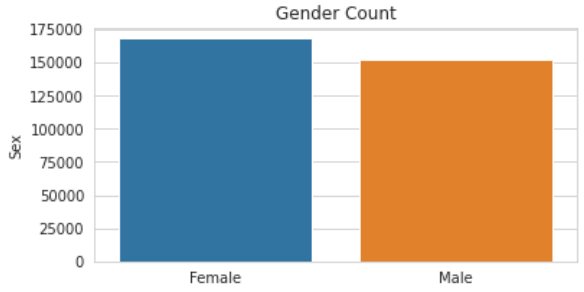

Figure 1: Gender count. There are more females than males in the data collected by the CDC

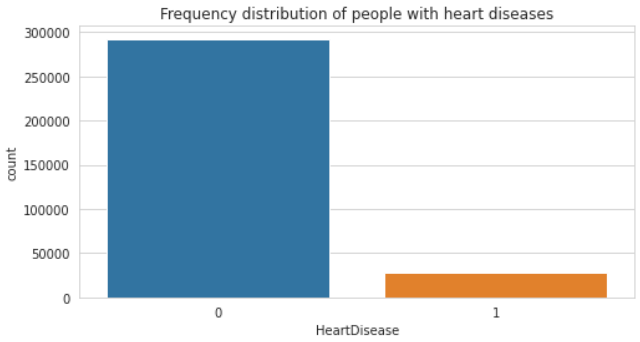

Figure 2: Frequency distribution of heart disease. Few people in the data set have heart disease.

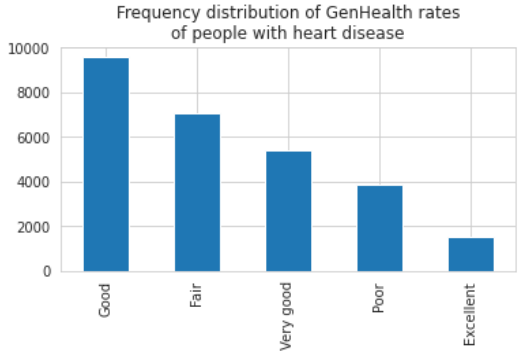

Figure 3: General health rates. Good health has the greatest frequency.

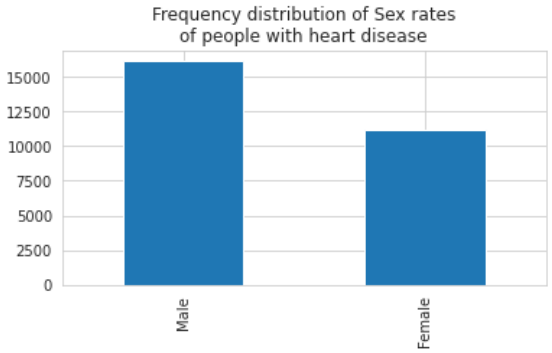

Figure 4: Sex distribution with heart disease; males with greater rates of disease.

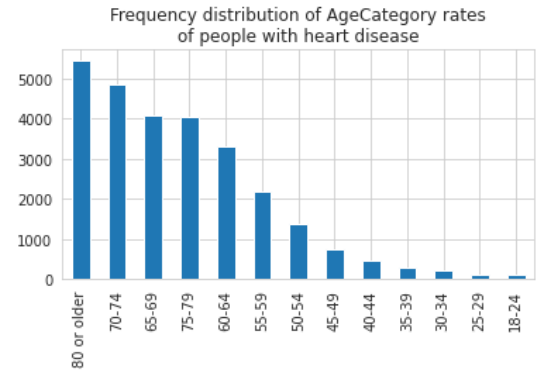

Figure 5: Age distribution with heart disease. Highest frequency above 80 years

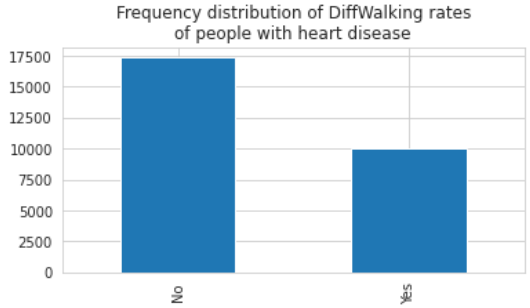

Figure 6: Walking distribution with heart disease. No difficulty walking did not prevent the high frequency of disease.

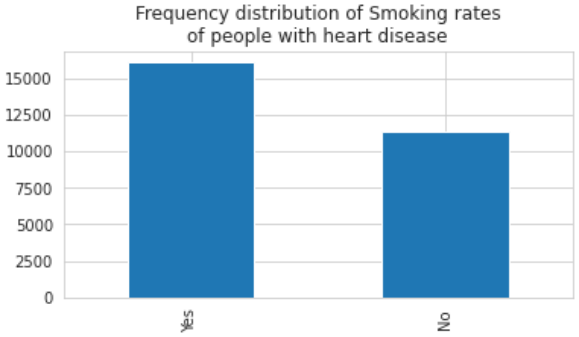

Figure 7: Smoking distribution with heart disease. Smokers were a greater proportion of patients.

We also analyzed the correlation for each parameter with heart disease. We found that difficulty walking, strokes, diabetes, and poor general health are all strong indicators of heart disease with values of 0.201, 0.197, 0.175, and 0.175 respectively. A full plot of the correlation values can be found in Appendix X.

In the pre-processing of the data, we try to find if there are any missing values and if every column has the appropriate data type, any outliers, or inappropriate responses. We then dropped the unnecessary data or encoded and scaled it as necessary. Data organization and cleaning were applied using pandas, sklearn, and numpy in Python. Firstly, the original dataset was converted from SAS to CSV format. Variables with a direct or indirect effect on heart disease were selected. The values of the categorical variables were converted from numeric type to text type to facilitate its analysis. Rows with missing records were removed.

III. Analysis

Before implementing the classification algorithms, we implemented a 5-fold method using logistic regression. This produced an accuracy of 0.913, 0.916, 0.918, 0.914, and 0.917 for each fold. The average accuracy was 0.916. Following classification, we utilized the standard scalar preprocessing tool to ensure our data is properly scaled and, combined with the encoders for ordinal variables, created a consistent dataset between positive and negative one. To split the preprocessed data, we used a 20/80 percent split between test/training with randomization using a seed of 42.

For classification problems in prediction, logistic regression, KNN, and SVM are effective tools. It's important to know when to use each of them to save money and time. First examining logistic regression, the binary nature of many health indicators renders logistic regression a natural fit for a similarly binary dependent variable.

A. Logistic Regression

A Machine Learning classification approach called logistic regression is used to predict the likelihood of a categorical dependent variable. It's a classification problem extension of the linear regression model. Unlike linear regression, which produces continuous numerical values, logistic regression produces a probability value that may be mapped to two or more discrete classes using the logistic sigmoid function. LR takes a probabilistic approach that explains each feature's prevalence in the construction of decisions. While an extremely effective tool in this and other contexts, it must be understood in diagnostic settings that LR assumes little variable interaction in that the correlation between variables will generally be below. Logistic regression typically produces low variance and high bias (overfitting).

$$P = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Figure 8: Logistic Regression Formula

Undersampling was also explored for the logistic regression classifier. Undersampling is the random omission of data similarly based on the frequency of the output class's two values. In the context of the heart disease data, this would allow for a more even mixture of those with and without heart disease. Figure 8 represents the probability proportion created with each additional data point within logistic regression. The exponential value is comprised of an intercept value and coefficients related to the impact of individual variables on the probability of a given outcome that, when added to 1, provides a clear value of outcome likelihood.

B. K-Nearest-Neighbors

K-nearest neighbors is a similarly powerful and efficient learning algorithm with significant practical application to diagnostic problems. The model provides a needed, non-linear decision boundary that can understand how feature interaction contributes to final outcomes. While assumptions and costs are limited, KNN requires the iterative updating of the number of

neighbors to examine. While generally an attainable figure, incorrectly choosing an initial K-value can lead to extensive misrepresentation down the line. Similar assumptions are required in Random Forest learning that, while more powerful, do not justify the computational costs necessary to create a non-overfit model on such a large dataset. K-Nearest-Neighbors (KNN) is a supervised classification algorithm. An input is compared to the K closest training data points by Euclidean distance, seen in Figure 9, and classified based on the majority.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Figure 9: Euclidean Distance for KNN

Small values of K can lead to a high variance and low bias (overfitting) while large values of K can create low variance and high bias (underfitting). Thus, it is often necessary to test multiple different K values to determine the optimal solution.

C. Support Vector Machine

Moving to Support Vector Machines, SVM provides advantages over onerous Bayesian alternatives like Naive Bayes Classification. While SVM may lack the power of NBCs to handle large datasets, the computational efficiency and similar ability to handle non-linear classifications render SVM an effective estimator in the heart disease space. SVM is used regularly in medical/healthcare settings due to its ability to notice complex patterns whose correlations are highly relevant to, especially rarer, disease diagnoses. Support Vector Machine is a classification method that seeks to group the binary outcome variables as distinctly far from the other in a decision boundary known as the hyperplane. Figure 10 highlights the desire to maximize the distance between the cluster groups denoted by the beta values described.

$$\begin{aligned} &\max_{\beta_0,\beta,\gamma} \quad \gamma \\ &\text{s.t.} \quad y_i(\beta^\top x_i + \beta_0) \geq \gamma \quad \forall i = 1, \ldots, N \end{aligned}$$

Figure 10: Formula for Support Vector Machine

In using the SVC algorithm, the authors implemented balanced class weighing for the SVM and logistic regression classifiers. Balanced class weighting is the process of assigning a weight to each data point based on the inverse proportion of its frequency amongst outputs. The process was applied to both the SVM and logistic regression model sections and was most effective in adding specificity in conjunction with the undersampled iteration of logistic regression as stated. SVM is generally a high variance, low-bias (overfitting) algorithm with the ability to rotate its bias and variance values by increasing the parameter shown in the maximization above, decreasing variance, and increasing bias.

Between the discussed methods, a comparative grid was created to highlight each model's performance across a litany of validation metrics. From this visual and quantitative aid, the team drew conclusions about the best classification method(s) for the data set. Bachem et al. (2018) suggest that a one-dimensional score may not be sufficient to capture the entire model. Thus, we compared the three algorithms based upon the following metrics:

- Accuracy
- Precision
- Recall
- F-Measure
- Kappa

Accuracy is the combined total of correctly predicted individuals with heart disease (true positives) and correctly predicted individuals without heart disease (true negatives) divided by the total number of predictions made. Precision is a measure of how many positive predictions were correct. Recall compares the number of correctly predicted positives to the number of actual positives in the sample. Precision and recall can be related using an F-Measure or F-1 score which is the harmonic mean of precision and recall. Lastly, a Kappa Score compares the prediction rate of the model against random chance. We want to give equal weight to people with and without heart disease so we compared the precision, recall, and F-Measures using macro averages. Because some of these metrics may have biases or flaws, we believe a comprehensive approach better captures the success of each model.

The models discussed in this project were developed in Python using a combination of Jupyter Notebook and Google Colab. The final results were generated using Google Colab. We chose to shift over to Google Colab once we began combining the individual contributions for easier group work.

IV. Results

For the KNN classifier, we tested different values for K from two to eight. This produced similar results for all of the comparison metrics except area under the curve (AUC) which increased with the value of K. Because the model is so complex, the optimal K value may be outside the range we tested. This would imply we are still on the side of overfitting. Thus, we will be using the results for K equals eight for the rest of the analysis. We also noticed a slight increase in run time for higher K values. This makes sense as each prediction must consider more nearby members. The comparison metrics by K value can be seen in Figure 11.

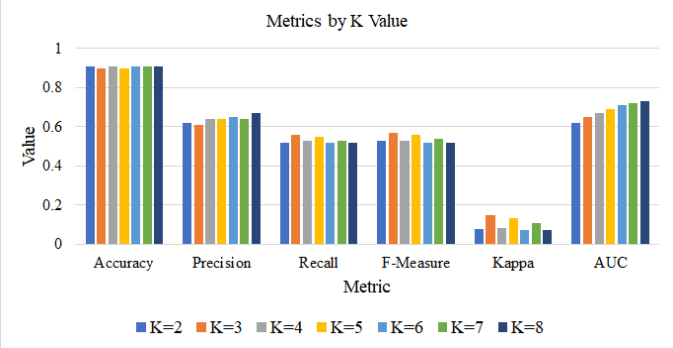

Figure 11: Metrics by K value

Next, we compared KNN with the other two classifiers: SVM and logistic regression. KNN and logistic regression had the highest accuracy values of 0.91. This was verified using a k-fold cross validation using a k =5, cross-validation is a resampling technique for evaluating machine learning models on a small sample of data. The process includes only one parameter, k, which specifies the number of groups into which a given data sample should be divided. Logistic regression with undersampling had the best precision, recall, F-Measure, and Kappa Score with values of 0.76, 0.76, 0.76, and 0.53 respectively. The full comparison of metrics can be seen in Figure 12 below.

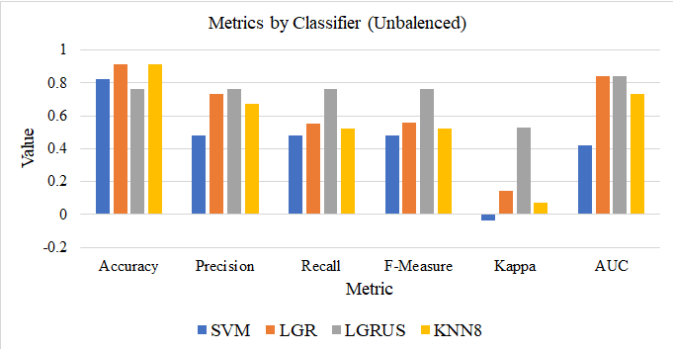

Figure 12: Comparison of metrics by classifier (unbalanced)

In terms of the area under the curve, both logistic regression methods performed the best with values of 0.84. KNN performed slightly worse with a value of 0.73. SVM had an AUC of 0.42. The AUC graphs can be seen in Figure 13.

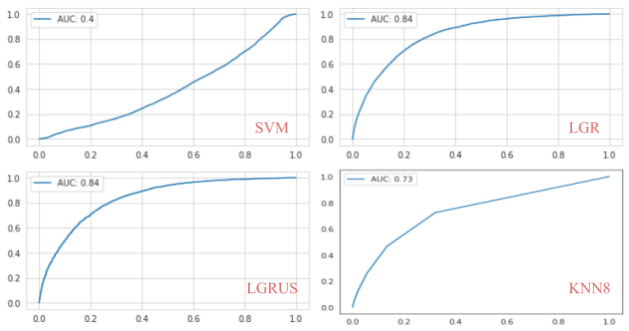

Figure 13: Area under the curve by classifier

When examining the confusion matrices for the different classifiers, we noticed the classifiers were performing well in identifying true negatives. However, they were also producing a large false-negative rate as well. The confusion matrices can be seen in Figure 14. One possible explanation for this is the imbalance of people without heart disease in the data set. Because most of the people in the data set do not have heart disease, the classifier can achieve a high accuracy by correctly predicting true negatives from false positives. True positives and false positives do not have as much of an impact on the weighted metrics such as accuracy or area under the curve. It is also possible that the classification algorithms could be heavily biased towards certain factors and thus underfitting the data.

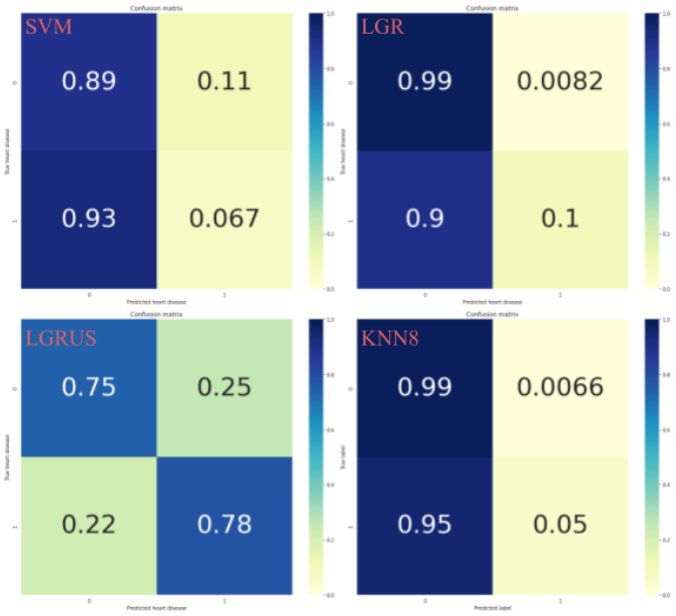

Figure 14: Confusion matrices (unbalanced)

While the majority of people may not have heart disease, we want to make sure the classifier can also identify individuals with heart disease. To improve the prediction rate for people with heart disease, we balanced the class sizes for the SVM and logistic regression classifiers. This improved all the metrics for the SVM classifier. For the logistic regression classifier, balancing the classes decreased accuracy but improved the other metrics. Logistic regression with under-sampling was unaffected. The metrics for the balanced classifiers can be seen in Figure 15.

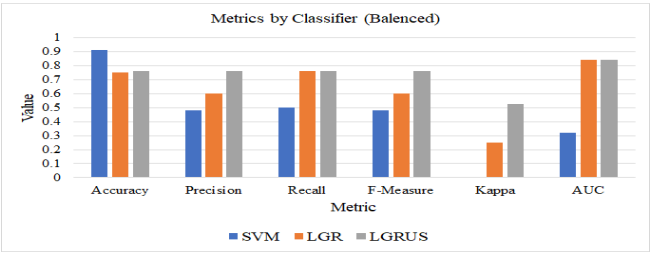

Figure 15: Comparison of metrics by the classifier (balanced)

After balancing the class sizes, the true negative false-negative time rates for SVM increased. The true positive rate did not improve. However, logistic regression's true positive rate increased significantly. The true positive and negative rates are now comparable. The confusion matrix for logistic regression with undersampling did not change. The balanced confusion matrices can be seen in Figure 16.

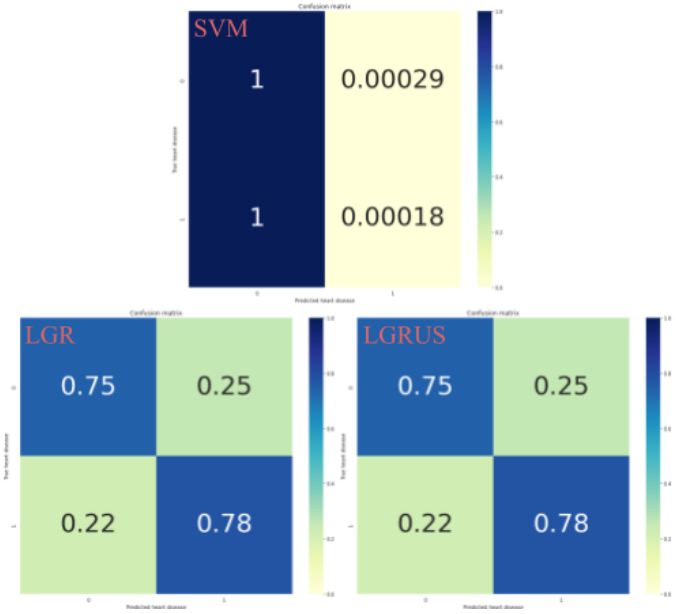

Figure 16: Confusion matrices (balanced)

V. Conclusions

When initially constructing analysis for this dataset, the goal of any resultant classification model was an accurate and precise diagnostic tool for heart disease. In an optimal setting, one model could be appropriately tuned to provide specificity in both the positive and negative directions, meaning doctors could use the tool to confidently tell patients that they likely do/will have heart disease or do not/will not have cardiovascular issues.

Upon the final model comparison, it was evident that without specific parameter tuning, results such as the confusion matrices shown in Figure 13 indicated an inability for the model, in the cases of initial iterations of all three of SVM, Logistic Regression, and KNN, to properly predict the occurrences of true positives. However, this inability to target true positives does not eliminate the diagnostic utility of this model selection process for optimizing patient outcomes.

A similarly prevalent diagnostic application comes from the ability to rule out disease, a feature often sought when attempting to label mysterious ailment causes. While the models with the default class weight setting of none were unable to nail in on true positives as desired, the incorporation of balanced class weights and, more potently, undersampling led to a more specific predictor at the expense of some underlying metrics.

Because of this greater equality of outcome combined with the balanced class data at a cell level, the undersampled logistic regression produced the most specific data with comparable metric returns in kappa, f-1, accuracy, and precision. Across each of the three models, bias/variance tendencies were different across SVM, LR, and KNN but yielded similar results when sampled properly and without balanced weights. It should be noted that the propensity of high-bias models like LR to overfit certain variables, requires the addition of undersampling and other techniques to eliminate these potential disruptors.

Physicians and data scientists alike could look to increase successful patient outcomes by using the combination of a robust, true negative indicator in conjunction with a slightly less accurate but significantly more specific tool in the undersampled regression algorithm. On an individual level, there was significant evidence that reflected the immense impact that smoking, walking, and poor general health can have on heart disease outcomes. Utilizing large datasets, often that are publically available, can and will give doctors a predictive edge never before possible in medicine. This report hopes that data-driven intervention drives down the lethality and "caught-too-late" cases of heart disease.

VI. Team Participation

Each member of the project worked on one method of classification however there was constant collaboration in the coding. The results and analysis were done by everyone together and the report writing was split mutually. Throughout the project, each team member showed a pervasive willingness to accommodate flexibility as well as help one another. This cohesion yielded time to explore curious feature selection methods while constantly improving the efficiency of the code.

Individually, Daniel Korach worked through the SVM model construction, code organization, and preprocessing/encoding. For the report, he focused on the project approach, SVM information, and conclusions. Harsh Hegde worked through logistic regression, function creation, graphics/display outputs, and feature selection while focusing on the data, code, and introduction sections of the report. Charlie Hickman was responsible for loop construction, preprocessing, the KNN information and coding, as well as the analysis and literature review in the report.

VII. References

[1] Bachem, O. et al. (2018). Assessing Generative Models via Precision and Recall. 32nd Conference on Neural

Information Processing Systems. doi: 10.48550/arXiv.1806.00035.

[2] Al-Adhaileh, M. H., et al. (2021). Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. Journal of Healthcare Engineering. doi: 10.1155/2021/1004767.

[3] Huang, S., et al. (2006). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer genomics & proteomics, 15(1), pp 41-51. doi:10.21873/cgp.20063.

[4] Kalita, U., et al. (2022). Signal-based automated hypertension detection using the Fourier decomposition method and cosine modulated filter banks. Biomedical Signal Processing and Control, 76. doi: 10.1016/j.bspc.2022.103629.

[5] Li, J., et al. (2022). Personalizing cholesterol treatment recommendations for primary cardiovascular disease prevention. Scientific Reports, 12(1). doi: 10.1038/s41598-021-03796-6

[6] Aggrawal, R. & Pal, S. (2020). Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease. SN Computer Science, 1(6). doi: 10.1007/s42979-020-00370-1.

[7] Wang, H. (2022). Public health emergency decision-making and management system sound research using rough set attribute reduction and blockchain. Scientific Reports, 12(1). doi: 10.1038/s41598-022-07493-w.

Appendix A: Open-Source Notebooks

Logistic Regression notebook: https://tinyurl.com/3bhan278

KNN notebook: https://tinyurl.com/2p88buhr

SVM notebook: https://tinyurl.com/3fx6ay7u

Appendix B: Correlation Plot