Predictive Classifications Models   
Using Key Indicators for Heart Disease

Jessie Browne, Sindhu Sai Sangani, Rama Sai Arun Varma Penmasta, & Murali Krishna Vattikunta, Department of BioHealth Informatics, SOIC

# INTRODUCTION

Cardiovascular disease (CVD) is a progressive disorder that moves from the presence of known risk factors to progressive disease that causes organ damage, organ failure, and death (Dahlöf, 2010). The National Center for Health Statistics summarized data collected from 1999 - 2010 in a brief issued in August 2012, and they looked at population stratification for the three known risk factors: uncontrolled high blood pressure, uncontrolled high LDL cholesterol, and current cigarette smoking (Fryar, Chen, & Li, 2012). They reported that about 47% of American adults had at least one of these three risk factors for CVD.

The CDC stresses that it is important to reduce the risk factors that are in the patient’s control, such as eating a high fat diet, not getting enough physical exercise, smoking, and drinking alcohol (CDC, 2019). Physical health and environmental effects are not the only covariates to consider - mental health has been shown to affect a patient's risk for cardiovascular disease. According to Ariyo et al. (2000), depressive symptoms are an independent predictor of heart disease in elderly patients. Dahlöf goes on to indicate too much or too little sleep as a risk factor (2010).

**Dataset.** Many of these risk factors are included in the 2020 CDC dataset published on Kaggle (CDC, 2020). The dataset, *Personal Key Indicators of Heart Disease*, has 18 variables (9 booleans, 5 strings, and 4 decimals), and the first of these indicates a self-reported diagnosis of heart disease. The other columns are BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, and AgeCategory. This dataset is further described in the methodology section, where exploratory data analysis and cleaning are detailed.

**Decision Tree.** One of the models used for extracting, predicting, and understanding the data is decision trees, which are built on a dataset having multiple features and use a target variable as the outcome variable. The decision tree model works on both classification and regression methods. The difference between classification and regression trees is the use of a categorical variable as target variable as opposed to a continuous variable.

Decision trees can handle large datasets without having any mandatory constraints of having data prepared. It takes its name from its tree-like structure, which has a target or root node, decision nodes, and terminal nodes. If the dataset is substantially large, we can divide it into train and test data, which is the validation set approach. In practice, we fit the model on the training data and use the testing data set to make predictions. The accuracy of these predictions is a good indicator of whether the model is optimal for the task at hand.

Fitting a decision model involves three steps: splitting, stopping, and pruning. The splitting of the nodes will happen according to the importance of variables to this model. It is important to be aware of the tradeoff between bias and variance – a very complex model may high purity in the nodes, but it could easily lead to overfitting. A small depth between parent node and end node means a fewer number of nodes in the model and subsequently fewer covariates. When the stopping rule does not work well, pruning may offer further improvements. The two types of pruning are pre-pruning, done by using chi square test, and post-pruning, in which a deep decision tree is built and then the branches are removed accordingly to improve the accuracy. One approach to pruning is error prediction, taking the test data set and making predictions on the fitted model. Cross validation can improve this approach.

**Linear Discriminant Analysis.** LDA is one of the popular machine learning techniques, which is mainly used to tackle classification problems rather than the supervised classification problems. It's fundamentally a dimensionality decrease procedure. Utilizing the linear blends of indicators, LDA attempts to foresee the class of the offered perspectives.

LDA algorithm attempts to observe the headings that can amplify the partition among the classes. Then, at that point, it involves these bearings for foreseeing the class of every person. These headings are known as linear discriminants and are a linear combination of the predictor variables. The LDA model can also be used in data preprocessing to reduce the number of features just as PCA (Principal Component Analysis) which reduces the computing cost significantly.

**Support Vector Machine.** The support vector machine (SVM) is a powerful model for classification. It is an extension of the support vector classifier, which uses a maximal margin classifier to separate classes using a linear boundary (Hastie, Tibshirani, & Friedman, 2009). The primary concepts comprising the SVM are a separating hyperplane, a margin with tunable permeability (C, slack or cost variable), and a kernel function. For this project, we performed both slack variable tuning and explored five kernels (linear, radial, quadratic, cubic, and sigmoid). The kernel function projects data from a two-dimensional space to a higher dimension (Rüping, 2001). Equation 1 gives the general form.

(1)

One major drawback of the SVM is low interpretability. Since it is a highly flexible model, it’s function is complex, which makes the relationship between variables and the output difficult to understand (Hastie et al., 2009). The SVM is also limited to problems with two classes, and it is computationally expensive.

# METHODOLOGY

​​**Exploratory Data Analysis.** The Python programming language was used for data wrangling on this dataset. First, we checked for null values and data types of all the covariates and the target variable in this dataset. Second, we identified the numerical columns and categorical columns. We plotted the numerical columns We generated several figures, using Python libraries Seaborn and Matplotlib, to visualize the data distribution and evaluate variable collinearity.

**Data preprocessing.** Steps in preprocessing the data included converting categorical attributes to numerical by using label\_encoder from sklearn.preprocessing(Scikit-learn) and the standardization of numerical data by using StandardScaler from sklearn.preprocessing. We split the data into 80-20 train-test sets and oversampled the train and test to achieve balance to avoid overfitting and underfitting.

**Decision Trees.** After data wrangling, we implemented a decision tree classification model on the train dataset. We used rpart and caret R libraries to fit, prune, and plot the decision tree. Firstly, we have fit the model with train data using all the features as variables and heart disease as target variable. We chose the degree of pruning with cross-validation. Finally, we generated a summary of the model, observed and recorded the covariate importance.

**LDA.** LDA can be computed in R using the lda() function which is present in the package MASS. In our LDA model, initially we trained with 10000 samples of train data and 2000 samples of test data and then we trained and tested our model with the whole dataset. We calculated the precision, recall and f1-score.

**SVM**. The support vector machine (SVM) models were built using RStudio and R version 4.1.3 (2022-03-10) by R Core Team (2019) and the R package e1071 by Meyer, Dimitriadou, Hornik, Weingessel, & Leisch (2021). The first step was to sub-sample the very large data set, which had n = 467702, in order to explore kernels and cost parameters. The training data set was randomly sampled with n = 1000 for parameter tuning and n = 10000 for training sub-models to explore five different kernels (linear, radial, quadratic, third-degree polynomial, and sigmoid). The test data set was also sub-sampled with n = 1000 for these exploratory models.

Parameter tuning was performed with the tune() function, which uses 10-fold cross-validation. The column ‘HeartDisease’ was the response variable, and all 17 predictor variables were used in the formulas for each of five models. To evaluate the SVM models, the subsampled test data was used to make predictions, and predictions > 0.5 were classified as ‘HeartDisease == 1.’ The accuracy score is the fraction of correct classifications with this criteria. Confusion matrices were populated for all five sub-models (not shown), and sensitivity (recall or TPR) and precision were calculated from these matrices.

Since the complexity of the SVM model is exponential, it wasn’t possible to fit the entire training data set, with 467702 observations, so an ensemble of SVM models with linear kernels and constant cost parameter (0.001) were created and their predictions averaged. The ensemble was composed of 104 models trained on 15000 observations each. The member predictions were weighted by accuracy and averaged to calculate a final result vector on 117142 test observations.

# RESULTS

**EDA.** The pair plot between the numerical columns is displayed as Figure 1and the correlation plot in Figure 2. We observed a weak correlation between MentalHealth and PhysicalHealth covariates (0.29), but there was no strong collinearity between numerical attributes. There were no null values present in the dataset. There were four numerical columns in the dataset: BMI, PhysicalHealth, MentalHealth, SleepTime.

Diagram

Description automatically generated

**Figure 1:**Pair plots for numerical features in the dataset:  
 Personal Key Indicators of Heart Disease

Chart

Description automatically generated

**Figure 2:** Correlation matrix to check for collinearity in the predictors

**Decision Tree Results and Discussion.** Examining the unpruned decision tree in Figure 3, at the apex we have the AgeCategory variable which is the most important variable in classifying the observations. The AgeCategory in the tree divided into two classifications, AgeCategory< 6.5 and more than 6.5. Now going with the next node which is AgeCategory < 6.5, we have reached the terminal node which says that 30 percent of people falling into AgeCategory less than 6.5 are having the heart disease probability of 20 percent. If we go with the node on the right side of the tree (AgeCategory>7), we come across the second most important variable in classifying the results further. That is the General Health variable which is now the decision node. It says that 70 percent of the people who falls under the AgeCategory>7 with a probability of 63 percent having heart conditions.

If we go with General Health > 4, we reach the terminal node which concludes that 45 percent of people who falls under this division (GenHealth > 4) with 74 percent probability of having heart disease. Now if we go further deep into the tree, we can see that we have the AgeCategory again as important variable with AgeCategory < 10 within GenHealth <3 we have equal division of 13 percent on both left and right nodes with probabilities of heart disease 33 percent and 55 percent respectively. If we look at the left side of the plot, we can see that 25 percent of people in this node have a 20 percent probability of having heart disease.

Finally, with AgeCategory > 10, then we have another important variable Sex being Female we have 6 percent of people having heart disease probability of 43 percent and on the other side, we have with Sex being male we have 6 percent again with probability of 66 percent with having heart disease. Now we can run our model against our test data by predicting on our test data and getting the performance evaluation of it as shown in Supplemental Figure 3.

Timeline

Description automatically generated

**Figure 3**: Unpruned decision tree

The accuracy of the model on the test data set is reported to be 72.99 percent. I found that the accuracy of the train data, we can see it is 73.4%. We can see there is slight overfitting of the data as train data performs better than test data. To balance this, we have performed post pruning on the data. For this we have used library caret. We have plotted a tree against the train data. Then we used cv.tree function to perform cross validation against this data and plotted it. We can see from Supplemental Figure 3 there is low deviation at the 5th sized tree. Now using prune misclass we have plotted the tree (Supplemental Figure 1). Now building the confusion matrix (Supplemental Figure 2), along with other performance evaluation metrics, we can see that the accuracy obtained is 71.6% which is actually good having the test observations classified validly. The above pruning process provided a better understandable tree at the cost of little classification accuracy rate.

**LDA and SVM.** The LDA reported separate statistics for each of the result variable groups, with 0 indicating no heart disease, and 1 indicating heart disease (Supplemental Table 1). Overall, the model achieved and accuracy of 0.760, precision of 0.774, and recall of 0.753. The results for the exploratory model SVM models, one for each of five kernels with their individually tuned slack variables, comprise Supplemental Table 2. The ensemble SVM achieved an accuracy of 0.744, precision of 0.702, and recall of 0.821. These results are summarized together with the decision tree results in Table 1.

**Table 1:** Summary of results for the three primary models developed

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** |
| LDA | 0.760 | 0.760 | 0.760 |
| Decision Tree | 0.774 | 0.774 | 0.774 |
| SVM ensemble | 0.753 | 0.753 | 0.753 |

# DISCUSSION

**EDA.** We were able to draw some conclusions from exploratory analysis by visualizing the data with different plots such as pair plot (Figure 1), box plot, correlation plot (Figure 2), and. kernel density plot. PhysicalHealth isn't the only good indicator to conclude the presence of heart disease. MentalHealth follows the same distribution of PhysicalHealth and isn't the only good indicator. We can also see that healthy sleep hours can also cause heart disease like the odd cases in BMI. There are about 13 categorical columns where most of them are binary. After observing the Kernel Density plot (not shown) there is no strong correlation in categorical variables also.

**Decision Tree.** The characteristics of this model were described in detail in the previous section.

**LDA.** Since the accuracy of LDA did not improve between the 10000-sample model and the model training on the full dataset (75.6% and 76.02%), we can clearly state that even if we increase the dataset size the accuracy is not getting increased much, so our model performs well even if the dataset size is moderate or huge.

LDA is used to determine group means and also for each individual. The histogram (Figure 4) tells that some part of group 0 is colliding with group 1, it clearly indicates that we did not get the best accuracy but we achieved an accuracy of which is a good accuracy and we clearly see that there are no outliers, indicates we removed the outliers to increase the accuracy and fit the model properly.

Chart, histogram

Description automatically generated

**Figure 4:** Group distribution for ‘HeartDisease’

**SVM.** The SVM model achieved accuracy ranging from 0.49 for the sigmoid kernel exploratory model to 0.75 for the quadratic kernel. The sigmoid model’s low accuracy was due to a bias in prediction toward a negative response for heart disease, which is illustrated by a low TPR (0.0804124) and high TNR (0.8757282). All other models had similar accuracy results, so in selecting which kernel was best, it was important to consider context. Assuming these predictions would most likely be used for heart disease intervention and prevention, the TPR or sensitivity is the most important factor.

Since the SVM model is too complex to fit 467702 observations, an ensemble of 104 SVM models with linear kernels and cost parameter 0.001 were trained on random samples sized n = 15000 from the full training dataset. The final model accuracy was 0.746, which better than the decision tree, but not as good as LDA. Options to improve this analysis would be to use a high-powered computer processor to fit a larger dataset to the SVM model, reduce the number of covariates, or apply stochastic gradient descent (SGD).

Since the complexity of SVM is exponential, it’s infeasible to fit the model to very large datasets. Menon found that SVM can be extended using stochastic gradient descent (SGD), which takes advantage of the gradient of a function as an indication of the direction of greatest increase. The SGD, which is a gradient calculated iteratively using a sample of the training set, can be applied to the hinge-loss of the SVM to reduce the computational burden (Menon, 2009).

LDA likely outperformed SVM because of it’s feature reduction capabilities. SVM is a more complex model, and was expected to achieve the highest accuracy. If the feature importance discoveries had been included in the SVM model to reduce covariates, accuracy may have been improved. The ensemble SVM was the most sensitive of all three models, which is notable because if the goal for this research is intervention for individuals at risk for heart disease, it is more important to achieve a high true positive rate than it is to avoid false negatives.

**Conclusion.** In our work, we applied LDA, Decision trees and SVM techniques on our dataset to check if the person has a heart disease and how all the covariates are in relation with the target variable. As the data is highly imbalanced, and we have performed oversampling techniques in python to overcome the underfitting/overfitting of the data. We observed that LDA is the most effective technique with respect to our dataset as LDA performs well on binary classification data.

Our analysis could be improved with several approaches, such as applying discoveries concerning feature importance from LDA and decision tree to the SVM model (feature reduction), stratifying subsamples for the SVM exploratory and ensemble models, and applying a stochastic gradient descent (SGD) to the SVM, which would lower the complexity of the loss function and allow the model to process the full dataset.

# REFERENCES:

Ariyo, A. A., Haan, M., Tangen, C. M., Rutledge, J. C., Cushman, M., Dobs, A., & Furberg, C. D. (2000). Depressive symptoms and risks of coronary heart disease and mortality in elderly americans. *Circulation*, *102*(15), 1773–1779.<https://doi.org/10.1161/01.CIR.102.15.1773>

CDC. (2019). *Know your risk for heart disease*. Retrieved from <https://www.cdc.gov/heartdisease/risk_factors.htm>

CDC. (2020). *Personal key indicators of heart disease* (K. Pytlak, Ed.). Kaggle. Retrieved from<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>

Dahlöf, B. (2010). Cardiovascular disease risk factors: Epidemiology and risk assessment. *The American Journal of Cardiology*, *105*(1), 3A–9A.

Fryar, C. D., Chen, T.-C., & Li, X. (2012). *Prevalence of uncontrolled risk factors for cardiovascular disease: United states, 1999-2010*. US Department of Health; Human Services, Centers for Disease Control and ….

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *An introduction to statistical learning*.

Kotsiantis, S.B. Decision trees: a recent overview. *Artif Intell Rev* **39,** 261–283 (2013).<https://doi.org/10.1007/s10462-011-9272-4>

Menon, A. K. (2009). Large-scale support vector machines: Algorithms and theory. *Research Exam, University of California, San Diego*, *117*.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2021). *e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), TU wien*. Retrieved from [https://CRAN.R-project.org/package=e1071](https://cran.r-project.org/package=e1071)

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from[https://www.R-project.org](https://www.r-project.org)

Rüping, S. (2001). *SVM kernels for time series analysis*. Technical report.

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130–135.<https://doi.org/10.11919/j.issn.1002-0829.215044>

<https://dzone.com/articles/decision-trees-and-pruning-in-r>

Terry Therneau and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15.[https://CRAN.R-project.org/package=rpart](https://cran.r-project.org/package=rpart)

Stephen Milborrow (2021). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.1.0. [https://CRAN.R-project.org/package=rpart.plot](https://cran.r-project.org/package=rpart.plot)

Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-90. https://CRAN.R-project.org/package=caret

Brian Ripley (2021). tree: Classification and Regression Trees. R package version 1.0-41. [https://CRAN.R-project.org/package=tree](https://cran.r-project.org/package=tree)

<http://www.sthda.com/english/articles/36-classification-methods-essentials/146-discriminant-analysis-essentials-in-r/>

<https://www.geeksforgeeks.org/linear-discriminant-analysis-in-r-programming/>

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

<https://cran.r-project.org/web/packages/MASS/citation.html>

<https://www.analyticsvidhya.com/blog/2021/08/a-brief-introduction-to-linear-discriminant-analysis/>

Python Packages:

pandas : <https://pandas.pydata.org/docs/reference/index.html>

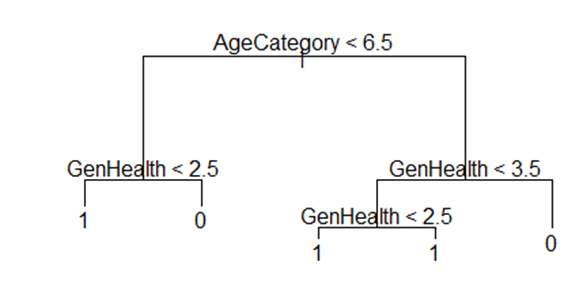
Numpy : <https://numpy.org/doc/stable/reference/>

Matplotlib:<https://matplotlib.org/stable/index.html>

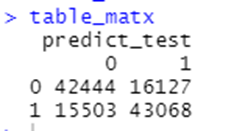
Seaborn: <https://seaborn.pydata.org/api.html>

Scikit-learn:<https://scikit-learn.org/stable/modules/classes.html>

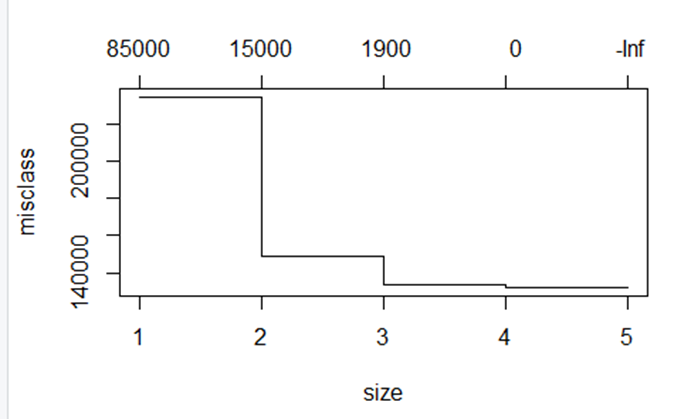
# SUPPLEMENTAL MATERIAL



**Supplemental Figure 1:** Pruned decision tree



**Supplemental Figure 2:** Confusion matrix for decision tree on test data



**Supplemental Figure 3:** Cross validation plot used to choose decision tree size.

**Supplemental Table 1:** Summary of results for the LDA model by group, where 0 indicates no heart disease, and 1 indicates heart disease.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **f1\_score** |
| 0 | 0.7460177 | 0.7678182 | 0.7567610 |
| 1 | 0.7744105 | 0.7530299 | 0.7635706 |

**Supplemental Table 2**: Summary of results for the exploratory SVM models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Kernel** | **Cost** | **Accuracy** | **Sensitivity** | **Precision** |
| Linear | 0.001 | 0.745 | 0.814433 | 0.7053571 |
| Radial | 0.1 | 0.747 | 0.8061856 | 0.7109091 |
| Quadratic | 0.1 | 0.75 | 0.7690722 | 0.7299413 |
| Cubic | 0.1 | 0.745 | 0.7587629 | 0.7272727 |
| Sigmoid | 0.01 | 0.49 | 0.0804124 | 0.3786408 |