Heart Disease Classification on Survey Data Utilising Component Analysis and Machine Learning Methods

Cheong Beng Chuan
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
bengchuan900@outlook.com

Khor Yong Teng
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
yongtengkhor338@gmail.com

***Abstract*— Heart disease is caused by several factors and it is vital to determine which factor causes it. Gathering information of the general public is usually in the form of a survey and is dependent on the groups of respondents as this will affect the balance of the class. Real world datasets are generally imbalanced with mixed data types, therefore requiring different approaches from visualization to data transformation and modelling. This paper evaluates different component analysis methods and machine learning models in classifying heart disease and determining factors related to it. Comparison between work done by others on this dataset and the chosen methods in this paper describe relatively similar performance. However, in the medical world, a single metric such as accuracy is not sufficient to determine methods' performance. Hence, this paper will also delve into additional metrics for a more accurate and perhaps situational evaluation of methods.**

Keywords—Classification, PCA, FAMD, Machine Learning, Imbalance Data

I. Introduction

The dataset chosen to be used in this paper is the Personal Key Indicators of Heart Disease dataset [1], which has been cleaned by Kamil Pytlak and published on Kaggle.com. The original uncleaned dataset was collected by CDC in the year 2020 through data collection involving telephone surveys, gathering information regarding the status of health. The resulting datasets include many questionnaire type columns such as "Do you have serious difficulty walking or climbing stairs?" or "Have you smoked at least 100 cigarettes in your entire life?". The methodology of cleaning done on the original datasets involves selecting variables that have direct or indirect effect on heart disease and reducing it down to the most relevant variables. Due to the nature of surveys, there are bound to be missing values and therefore, such rows are removed. The cleaned dataset dubbed heart_2020_cleaned consists of 18 variables of types Booleans, characters and decimals and totalling at 319795 observations.

Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare, and is one of the main goals of the survey. As stated from the description of the dataset, computational developments will allow application of machine learning to predict a patient's condition. As such the main focus of the paper will be to determine the factors that may cause heart disease and attempt to develop machine learning models that effectively predict and classify heart disease using the survey based data collected.

II. Aims and Objectives

The aim of this work is to determine the factors that may cause heart disease and develop machine learning models that are able to effectively classify and predict individuals that have it, using survey-based data.

To achieve these aims, the following key objectives will need to be met:

- Determination of the relationship between variables within the dataset and heart disease.
- Production of machine learning models that are able to effectively classify heart disease.
- Evaluation of data transformation and machine learning methods in its ability to effectively classify heart disease.

III. Related Works

Heart disease prediction using machine learning techniques is not a newly explored task and has been explored in multiple works such as [2] and [3]. However, within these works, only medically collected data is utilised with relatively low amounts of data. Supervised machine learning techniques such as Naive Bayes, Linear Regression, Deep Learning, Decision Tree, RF, GBT and Support Vector Machine K-nearest Neighbour were used within these works.

The dataset [1] being used within this work is different from conventional datasets used for the task of heart disease prediction as it consists of only survey data. A process of exploratory data analysis and prediction was done by Mushfirat Mohaimin can be found at the code section of the dataset [1] Kaggle page, detailing his approach. Utilising a random forest classifier, he was able to achieve an f1-score of 0.35 of predicting positive cases of heart disease.

Aside from typical supervised machine learning techniques, unsupervised machine learning techniques namely PCA and FAMD are used within this work for data transformation and dimensionality reduction.

Principal Component Analysis (PCA) [4] is one of oldest and most widely used dimensionality reduction methods, producing principal components that preserve as much variability as possible in the dataset with the least number of features. While PCA is designed to be used on numerical features, it seems to still be effective in analysing mixed data, provided with the prerequisite step of performing one hot encoding on categorical variables. [5]

An issue present in PCA is that the idea of variability collapses when we have binary data obtained by encoding, thus Factor analysis of mixed data (FAMD) is suggested to be utilised instead for mixed data, although being not as popular. [6] FAMD is a principal component analysis method that is used to analyse a dataset containing both quantitative (numerical variables) and qualitative (categorical) data [7], able to accept mixed data without one-hot-encoding categorical variables, unlike PCA.

IV. Methodology

For the purposes of our investigation, we obtained the "Personal Key Indicators of Heart Disease" dataset from Kaggle to be used as our sample dataset due to its suitability as it contains a decent mix of numerical and categorical variables. The analysis and modelling is then approached in a series of stages, alternating between data analysis, pre-processing, modelling and evaluation. RStudio was utilised throughout all steps stated.

A. *Exploratory Data Analysis*

An initial round of analysis was done on the dataset to gain a basic understanding of the dataset and determine if data cleaning was necessary for the dataset. The idea of what each variable represented was retrieved from Kaggle. The datatype of each column was then displayed to understand the variables involved, followed by checking the number of missing values for each column. A boxplot of numerical variables is also done to understand the distribution and check for the presence of outliers needed to be removed. It was then determined that no cleaning or imputation was required, which matches the description on kaggle stating that the dataset being a pre-cleaned dataset.

Being a classification task, the class ratios were then visualised using 2 different methods, namely barplot and pie chart, as to ascertain if data balancing methods are required before modelling.

Relationships between the variables were then analysed by using different types of visualisations. Firstly, the relationship between the "HeartDisease" variable (target variable), and the other variables are investigated to understand which variables are most relevant to the classification task. Numerical variables and categorical variables are visualised separately as they require different techniques. Numerical variables were visualised using Kernel Density plots and Histograms separated by the target variable to show the differences between their distributions. Categorical variables were visualised using barplots. Due to the unbalanced classes, percentage of categories within classes were shown on the plots. The relationship between each pair of variables was then investigated through the use of correlation plots and pair plots. A few types of correlation plots were done in this case. Pearson's correlation coefficient was used on numerical columns and as for categorical columns, Pearson's Chi Square was used. Pearson's correlation coefficient allows the measuring of direction and magnitude of relationship between two continuous variables while Pearson's Chi Square statistics can be used to show the relationship between two categorical variables. To visualise the results, pairwise comparison in the form of heatmap is done. Binary columns are visualised separately from categorical columns. Due to the dataset being of mixed data types, estimated latent correlations for mixed type data was visualised within a single plot as an alternative. Paired scatterplots of numerical values coloured based on the target were done as well to better understand class separability.

B. *Pre-processing*

After the initial exploration, some preprocessing steps were taken. Ordinal variables of character data types such as AgeCategory and GenHealth were converted to numerical variables. For AgeCategory, the data are encoded in the form of ranges whereas for GenHealth, the data are encoded in the satisfaction scale. Both were converted into integers following the appropriate order.

To prepare the dataset for further transformation (PCA) and modelling, the remaining nominal categorical variables are required to be converted into numeric columns. Binary variables are binarized into 1 (yes) and 0 (No), with the exception of "Sex" which is one-hot-encoded into 2 features (Male, Female) as a means to prevent inducing bias. The "Race" being a variable with multiple levels was one-hot-encoded into variables named "Race_White", "Race_Black", "Race_Asian", "Race_Indian_Native", "Race_Hispanic", "Race_Other" according to its levels. "Diabetic" being a special case was separated into 3 binary variables namely, "Diabetic", "Diabetic_Pregnancy" and "Diabetic_Borderline" for better representation. The processed dataset was saved as a separate dataset from the non-one-hot-encoded dataset as both versions are necessary.

C. *Data Transformation and Dimensionality Reduction*

Standardization was done to numerical variables of the dataset in preparation of Component Analysis Methods (CA). CA methods were then used for dimensionality reduction, using PCA and FAMD separately. As a means of visualizing the captured variance of the CA methods, a scree plot was used. Based on the Scree plots, 6 principal components were selected for both methods.

D. *Post Transformation Analysis*

The resulting principal components were then plotted to better understand the effects of the CA methods. For PCA, a biplot of the first 2 principal components were plotted where the points are colored according to their class, providing insight into class separability and the influence of the original variables on the principal components and by extension the classes themselves. For FAMD, a squared loading plot of the first 2 cos squared components were plotted in a new feature space that displayed the captured variation of individual variables by the dimension. For both FAMD and PCA, paired scatter plots were plotted to show the separability across different principal components. The first 3 components were plotted on a 3D scatter plot as well to provide further insight on data separation, in a higher dimensional space for better visualization.

E. *Data Balancing*

Being an imbalanced classification problem as determined from previous analysis, majority class down sampling was chosen as the data sampling method to solve this issue. Instead of down sampling the dataset before k-fold cross validation, down sampling was implemented within each iteration of the cross-validation, allowing for different data points to be selected for training, thus enabling a more representative evaluation. The majority class will be down sampled by approximately 90.64% within each fold.

F. *Classification*

K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes classifiers were selected to be used in the classification tasks to determine which was most suitable.

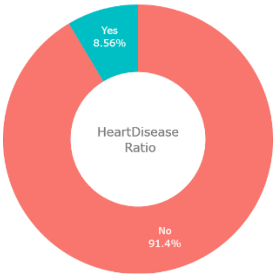
The performance of each classifier is evaluated within cross-validation, using down sampled datasets of the original, post-PCA, and post-FAMD data.

G. *Evaluation*

K-Fold cross-validation was used for the evaluation of the different classification methods paired with different transformations of the data (Original, PCA, FAMD) for better usage of the data and a more holistic evaluation. Within each fold, the data is split into 90% train set and 10% test set. The training data is then processed by down sampling followed by the designated transformation method (no transformation, PCA, FAMD) before feeding it into the 3 classifiers. After cross-validation, a prediction will be produced for each combination of transformation method and classifier (KNN, SVM, Naive Bayes). The predictions are then evaluated using confusion matrices and evaluation metrics such as accuracy, precision, recall and F1-measure.

V. Results

A. *Exploratory Data Analysis*



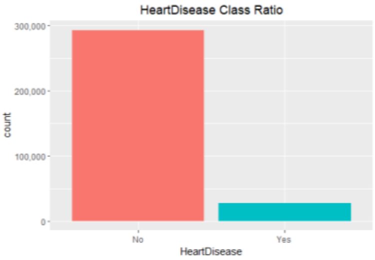


Fig. 1. Pie Chart of HeartDisease Fig. 2. Bar Chart of HeartDisease

Figure 1 and 2 shows that the dependent variable, HeartDisease, in the dataset, consists of 91.4% of No class and 8.6% of Yes class.

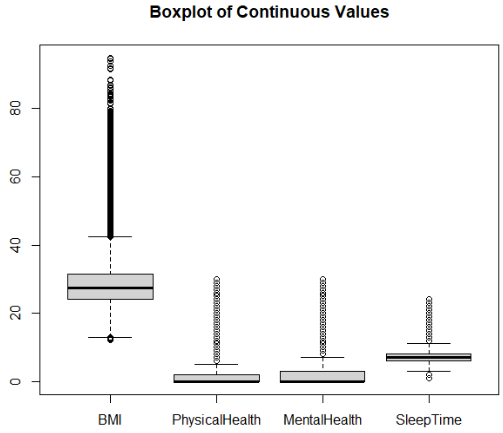
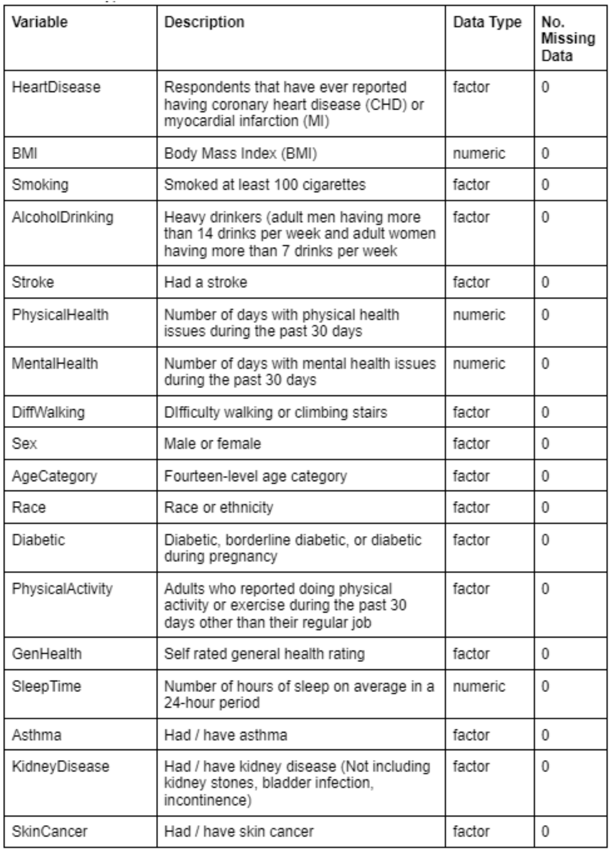


Fig. 3. Boxplot of Continuous Values

Figure 3 shows that the interquartile range of BMI is within normal and overweight ranges while there are more outliers that are obsese than underweight. PhysicalHealth and MentalHealth are generally the same with MentalHealth having slightly larger interquartile range.



Variable	Description	Data Type	No. Missing Data
HeartDisease	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)	factor	0
BMI	Body Mass Index (BMI)	numeric	0
Smoking	Smoked at least 100 cigarettes	factor	0
AlcoholDrinking	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week	factor	0
Stroke	Had a stroke	factor	0
PhysicalHealth	Number of days with physical health issues during the past 30 days	numeric	0
MentalHealth	Number of days with mental health issues during the past 30 days	numeric	0
DiffWalking	DIfficulty walking or climbing stairs	factor	0
Sex	Male or female	factor	0
AgeCategory	Fourteen-level age category	factor	0
Race	Race or ethnicity	factor	0
Diabetic	Diabetic, borderline diabetic, or diabetic during pregnancy	factor	0
PhysicalActivity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job	factor	0
GenHealth	Self rated general health rating	factor	0
SleepTime	Number of hours of sleep on average in a 24-hour period	numeric	0
Asthma	Had / have asthma	factor	0
KidneyDisease	Had / have kidney disease (Not including kidney stones, bladder infection, incontinence)	factor	0
SkinCancer	Had / have skin cancer	factor	0

Fig. 4. Variable Property Table

There are 4 numerical variables in the dataset while the rest are of factor data types with the variables consisting of binary and categorical classes. There is no missing data which is consistent with the methodology of cleaning.

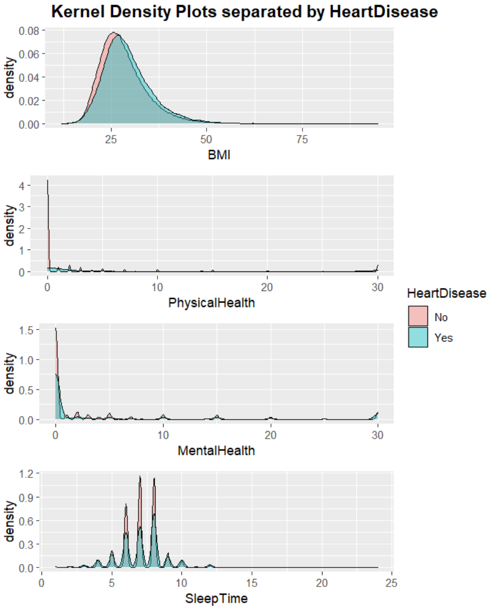


Fig. 5. Kernel Density Plots separated by HeartDisease

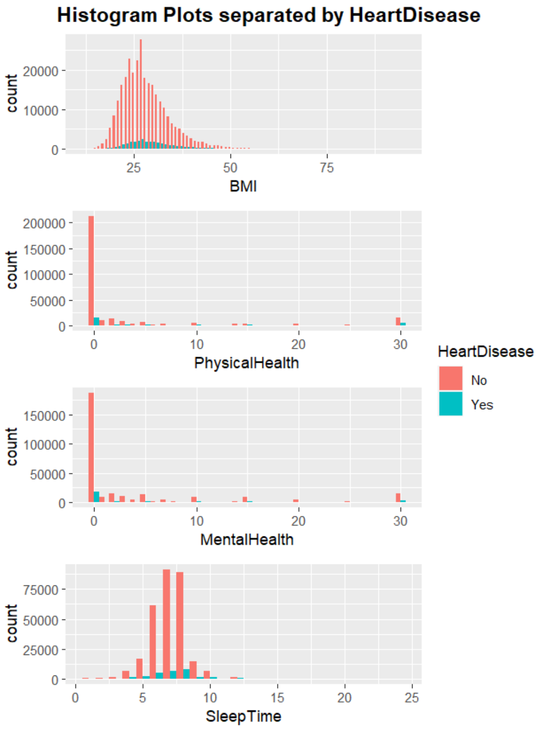


Fig. 6. Histogram Plots separated by HeartDisease

Figure 5 and 6 show no apparent separation in distribution under HeartDisease. Figure 6 displays the overwhelming amount of No compared to Yes while the distribution shape remains mostly the same under HeartDisease. Regardless, it is shown that there's a slight shift in having heart disease towards higher BMI and SleepTime, however it is not very apparent due to the class imbalance present.

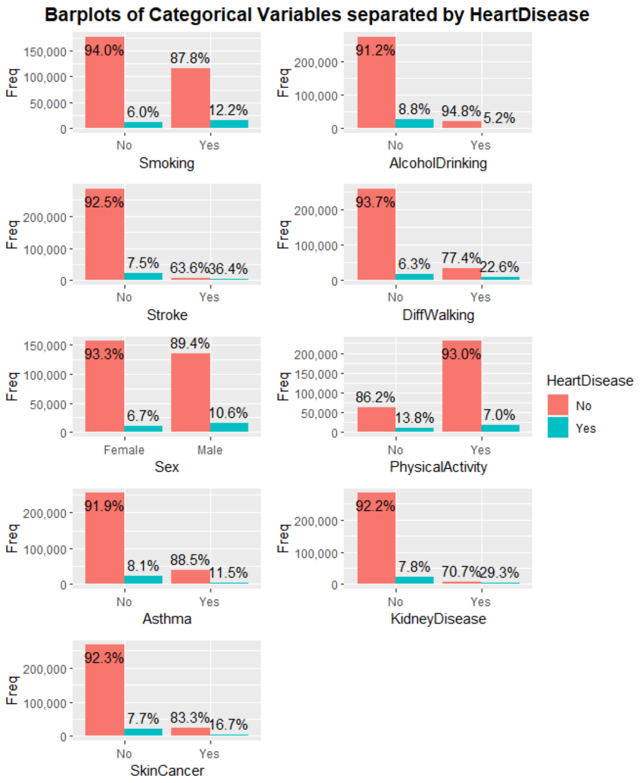


Fig. 7. Barplots of Binary Variables separated by HeartDisease

Figure 7 shows that Smoking and Sex variables has rather balanced classes while other categorical variables are unbalanced. As expected, stroke shows a higher percentage of respondents having heart disease. Respondents having smoking, skin cancer, asthma and kidney disease also show a higher percentage of heart disease. However, alcohol drinking suggests there are more respondents that do not drink having heart disease than those that drink. This could be due to the unbalance of the class. The physical activity shows that the majority of the respondents does physical activity which results in a lower percentage of respondents having heart disease.

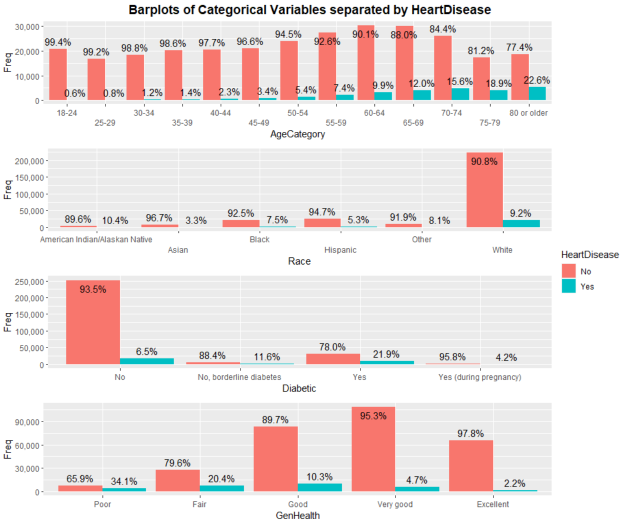


Fig. 8. Barplots of Nominal Variables separated by HeartDisease

Figure 8 shows that the tendency of having HeartDisease increases as age increases. For Race, the figure shows that the majority of the respondents are White, with the American Indian/Alaskan Native having the highest percentage of HeartDisease while Asians have the lowest percentage. For diabetic, the percentage of heart disease increases along with the tendency of having diabetic. However, it is shown that pregnant respondents have the lowest percentage. For GenHealth, the percentage of having HeartDisease decreases from Poor to Excellent.

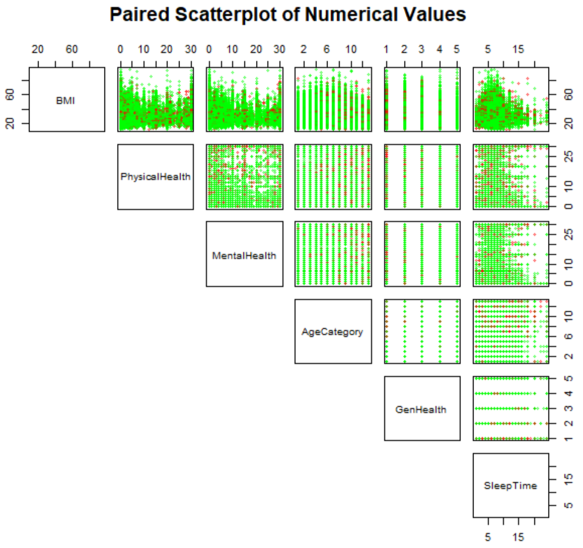


Fig. 9. Paired Scatterplot of Numerical Values

Figure 9 (after converting ordinal categories to numerical variables) visualizes the distribution of HeartDisease classes, where red points represent positive cases and green points represent negative. Only AgeCategory and GenHealth can be clearly observed to affect the distribution. More positive cases of HeartDisease can be seen at a higher AgeCategory and at a lower GenHealth.

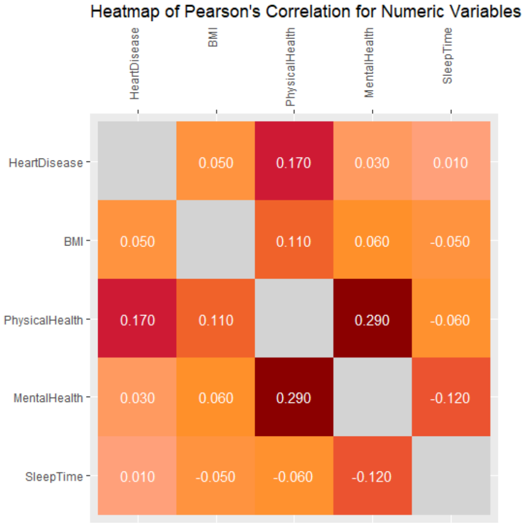


Fig. 10. Heatmap of Pearson's Correlation for Numeric Variables

Figure 10 shows that PhysicalHealth has the highest positive correlation with HeartDisease. While MentalHealth and PhysicalHealth have higher correlation, MentalHealth has much lower correlation with HeartDisease. SleepTime has the highest negative correlation with MentalHealth which suggests that higher sleep time leads to lower MentalHealth. Both MentalHealth and SleepTime have relatively no correlation with HeartDisease.

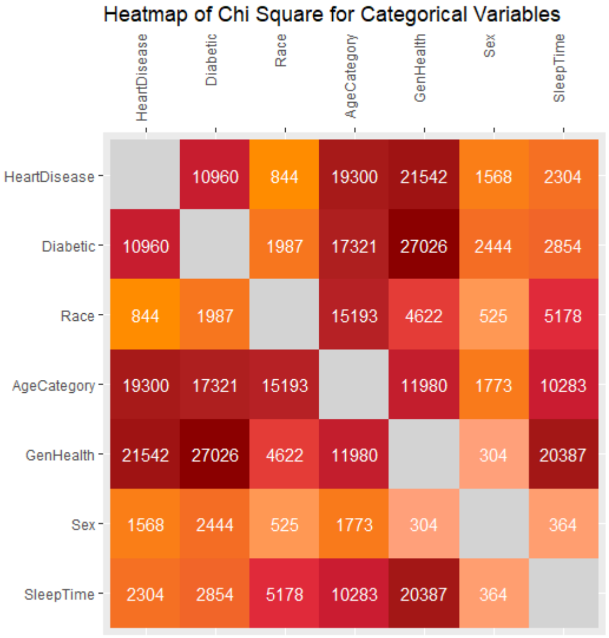


Fig. 11. Heatmap of Chi Square for Categorical Variables

Figure 11 shows that GenHealth variable has the highest association with HeartDisease followed by AgeCategory and Diabetic variables. The lowest association variable with HeartDisease is Race followed by Sex and SleepTime variables. GenHealth displayed the highest association with Diabetic while having the lowest association with Race variable.

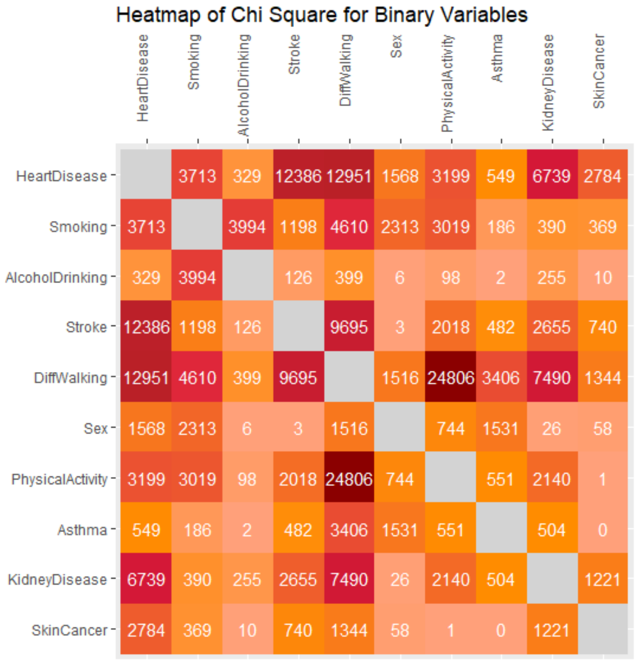


Fig. 12. Heatmap of Chi Square for Binary Variables

Figure 12 shows that DiffWalking has the highest association with HeartDisease, followed by Stroke and Kidney Disease. AlcoholDrinking has the lowest association with HeartDisease followed by Asthma and Sex variables. DiffWalking has much higher association with PhysicalActivity while PhysicalActivity has a rather low to moderate association with HeartDisease. The lowest association pair of variables are SkinCancer and Asthma with 0 Chi Square followed by SkinCancer and PhysicalActivity with 1 Chi Square.

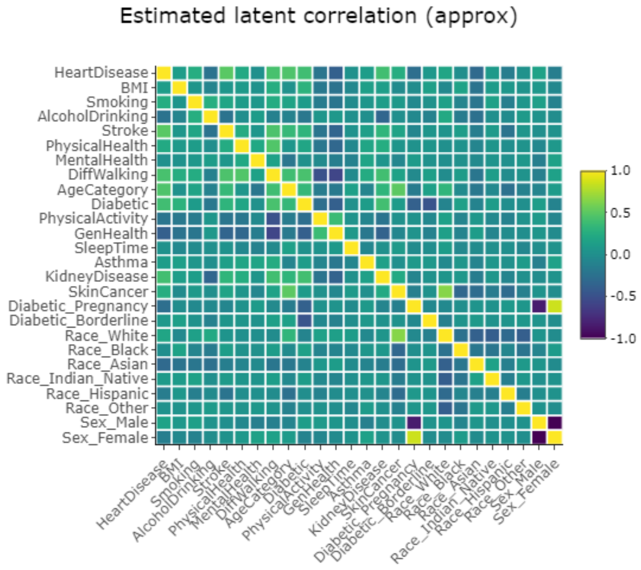


Fig. 13. Estimated latent correlation (approx)

The one-hot-encoded version of the dataset was used to display Figure 13. As observed with Barplots, Pearson's Correlation and Chi Square as well, the variables that are more correlated with HeartDisease are Stroke, DiffWalking, AgeCategory, KidneyDisease, GenHealth, Diabetic, Physical Health, PhysicalActivity. With estimated latent correlation, excluding pairs such as Diabetic_Pregnancy and Sex_Male/Sex_Female which have a high correlation due to males not being able to be pregnant, the other variables do not show a strong correlation with each other.

B. *Post Transformation Analysis*

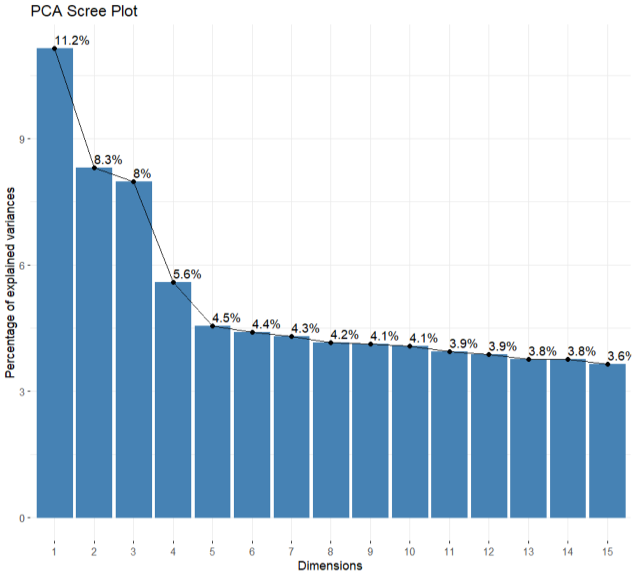


Fig. 14. PCA Scree Plot

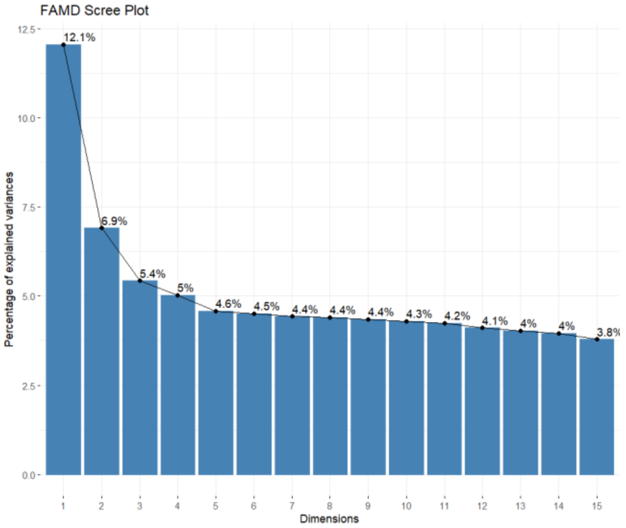


Fig. 15. FAMD Scree Plot

Figure 14 and 15 show that the percentage of explained variance per dimension plateaus starting from the 5th to 6th dimension. Based on the plots, 6 principal components were selected for both methods to be used in later classification tasks for consistency, resulting in a total captured explained variance of 42% and 38.5% for PCA and FAMD respectively.

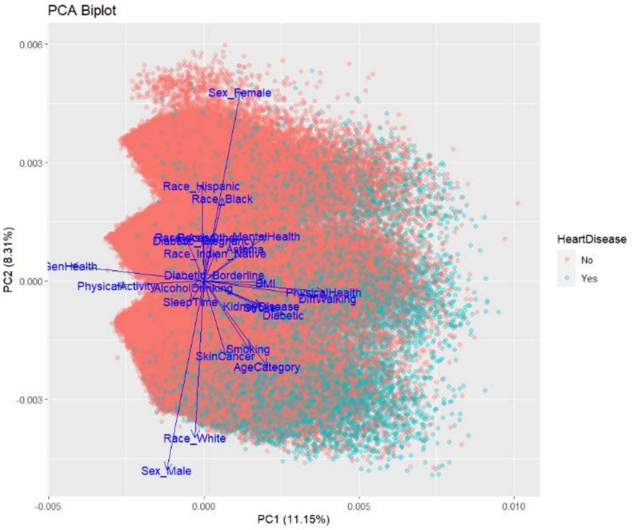


Fig. 16. PCA Biplot of the first 2 principal components

Figure 16 was able to represent the relationships between the independent variables and the tendency for classifying HeartDisease based on those variables. It can be observed that following the direction of PhysicalHealth, DiffWalking, KidneyDisease, Diabetic, Smoking, SkinCancer, AgeCategory, BMI, MentalHealth and Asthma vectors there is an increase in tendency for positive cases of HeartDisease while following the direction of the GenHealth and PhysicalActivity vectors shows the opposite. It is also shown that the distributions for positive and negative cases of HeartDisease are not linearly separable.

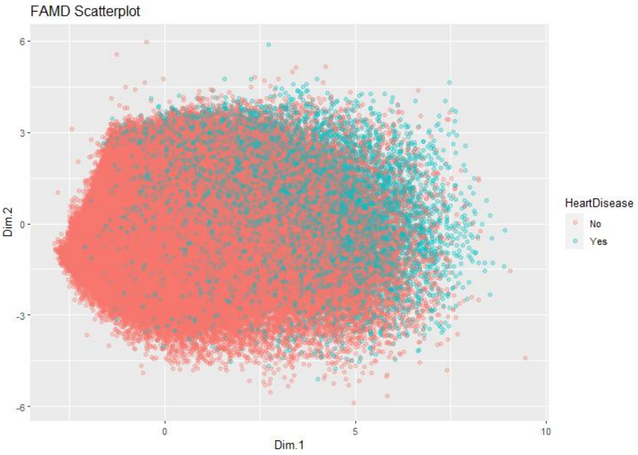


Fig. 17. FAMD Biplot of the first 2 principal components

Figure 17 shows that there is an increase in tendency for positive cases of HeartDisease as both the values of both principal components increase. It is also shown that the distributions for positive and negative cases of HeartDisease are not linearly separable, similar to PCA.

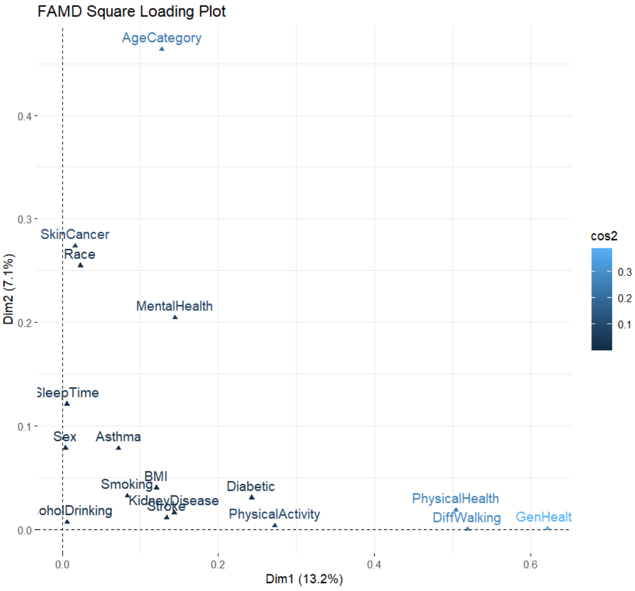


Fig. 18. FAMD Square Loading Plot

Figure 18 visualizes the importance of variables in explaining the variance captured by the principal components. It can be seen that PhysicalHealth, GenHealth and DiffWalking have high importance in explaining the variance captured by the first principal component while Diabetic and PhysicalActivity have medium importance in doing so. At the same time, AgeCategory has high importance in explaining the variance captured by principal component 2 while SkinCancer, MentalHealth, Race have medium importance. Combined with the above scatterplot, we can observe which variables contributed the most to the different distributions between positive and negative cases of HeartDisease.

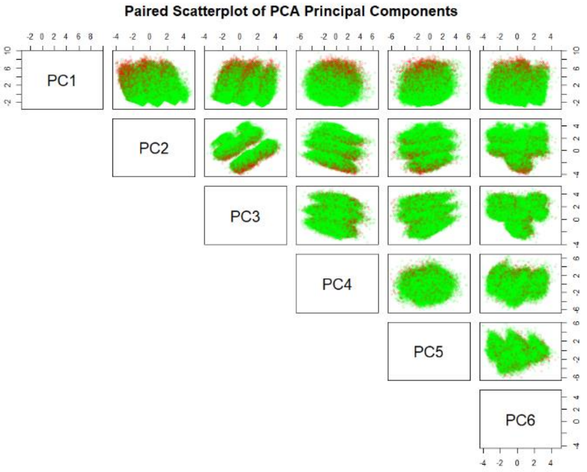


Fig. 19. Paired Scatterplot of PCA Principal Components

Figure 19 shows that principal components 1 and 2 have a more obvious difference in distribution for positive and negative classes of HeartDisease, while proceeding towards principal components 3, 4, 5, and 6, no obvious tendencies can be observed.

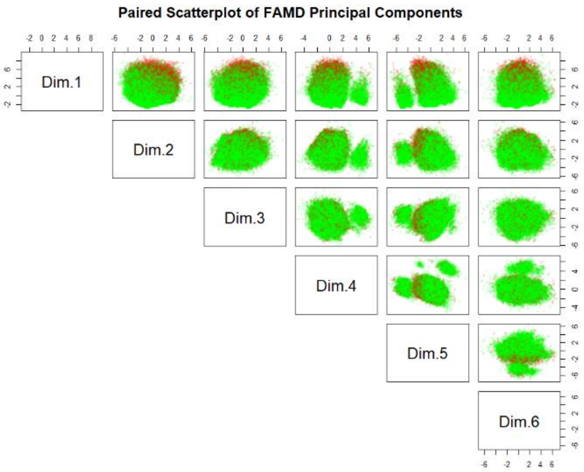


Fig. 20. Paired Scatterplot of FAMD Principal Components

Figure 20 shows that principal components 1 and 2 have a more obvious difference in distribution for positive and negative classes of HeartDisease, while proceeding towards principal components 3, 4, and 6, no obvious tendencies can be observed. Interestingly, a difference in the distribution can be observed for principal component 5.

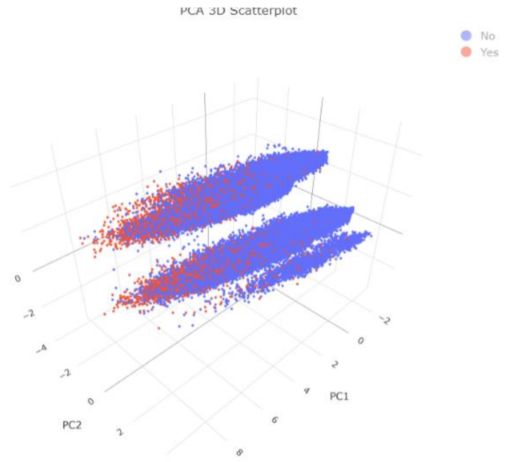


Fig. 21. PCA 3D Scatterplot

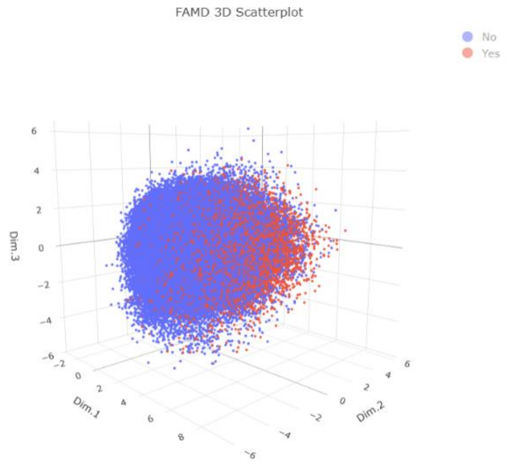


Fig. 22. FAMD 3D Scatterplot

Figure 21 and 22 show the distribution of HeartDisease classes using 3 principal components each. The distribution of positive and negative classes can be seen to show a tendency similar to previous plots. However, the shape of distribution of points between PCA and FAMD in the 3D space shows an obvious difference, where all points in FAMD are seemingly part of a single cluster while PCA has separate clusters.

C. Classification Results

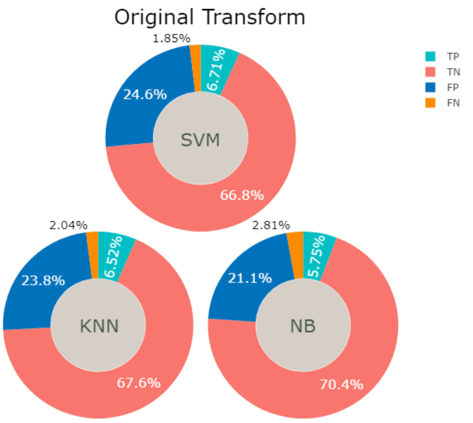


Fig. 23. Tri-Donut Chart for Original Transform

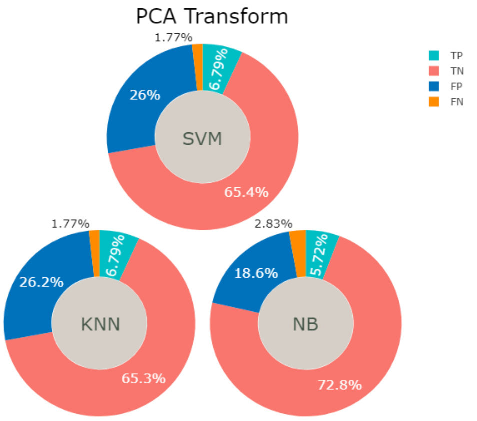


Fig. 24. Tri-Donut Chart for PCA Transform

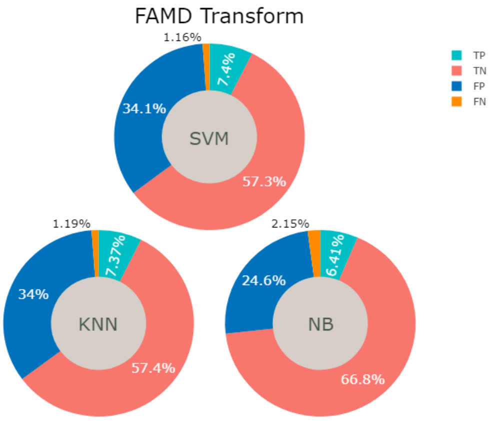
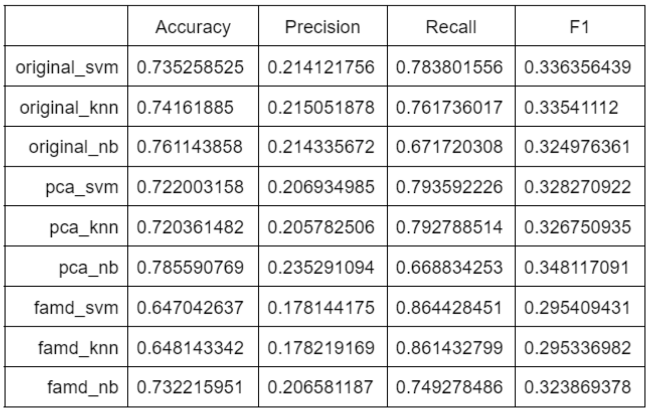


Fig. 25. Tri-Donut Chart for FAMD Transform

Figure 23-25 representing the classification results of classifying HeartDisease using different transformation methods paired with different machine learning classifiers visualises the resulting percentage of true positives (TP), true negatives (TN), false positives (FP), false negative (FN) predicted.



	Accuracy	Precision	Recall	F1
original_svm	0.735258525	0.214121756	0.783801556	0.336356439
original_knn	0.74161885	0.215051878	0.761736017	0.33541112
original_nb	0.761143858	0.214335672	0.671720308	0.324976361
pca_svm	0.722003158	0.206934985	0.793592226	0.328270922
pca_knn	0.720361482	0.205782506	0.792788514	0.326750935
pca_nb	0.785590769	0.235291094	0.668834253	0.348117091
famd_svm	0.647042637	0.178144175	0.864428451	0.295409431
famd_knn	0.648143342	0.178219169	0.861432799	0.295336982
famd_nb	0.732215951	0.206581187	0.749278486	0.323869378

Fig. 26. Metrics Table for Results

Using the confusion matrix, evaluation metrics such as accuracy, precision, recall and F1-score in correctly classifying positive cases of heart disease are obtained. Naive Bayes classifier paired with PCA achieved the highest accuracy, precision and F1-score but with the lowest recall among all methods. The highest recall was however achieved by SVM classifier paired with FAMD. It can also be observed that classifiers utilizing FAMD produce higher recall but lower precision compared to the other methods. Naive bayes classifiers can also be seen to provide better precision while decreasing recall compared to other classifiers across all transformations.

VI. Discussion

A. Exploratory Data Analysis

The initial pie chart and bar chart visualization of the HeartDisease shows that the class is imbalanced with the majority having no HeartDisease, where the pie chart is able to visualize the difference in ratios better. This results in a higher percentage of No HeartDisease throughout all the variables. Regardless, with the help of percentage labels on the barplots, trends can be seen on variables such as AgeCategory, Diabetic and GenHealth. For further insight, a Chi Square heatmap is plotted to confirm the trend and association. Overall, sickness type variables such as Stroke, SkinCancer, Asthma and KidneyDisease show higher percentage of having HeartDisease.

By observing the boxplot, BMI does show more outliers than other continuous variables but were not removed as it is still within the realm of possibility. However, the outliers for SleepTime span till 24 in hours but without further information (the possibility of comatose), the outliers are not removed as it is still within a valid range.

Kernel Density and histogram plots both summaries the lack of differences in distribution of separation by HeartDisease. Kernel density plots can be seen to be more suitable for decimal variables whereas histogram plots were more suitable for integer variables due to integer variables being visualized as multimodal density plots. A Pearson's

Correlation heatmap further shows that the numerical variables have weak correlation with HeartDisease. The estimated latent correlation and pairwise scatterplot generally summaries the lack of differences in distribution and weak association between variables.

In summary, the classes in the dataset are imbalanced and the weak correlation leads to the decision of down sampling and the use of CA methods to reduce the dimension using all the variables.

B. Post Transformation Analysis

PCA allowed for a better understanding of the variables of the dataset in relation to its correlation to HeartDisease, more so than exploratory data analysis or FAMD. The PCA Biplot was able to provide a great picture of the variables that are most likely to correlate with HeartDisease or even cause it, providing information on what variables increase or decrease the tendency for HeartDisease or even the ones that show weak correlation to it in a clear and distinct way.

A biplot was attempted to be plotted for FAMD, but was unsuccessful due to the lack of implementations. A replacement where a scatterplot and square loading plot was used instead. Although information regarding which variables are affecting the variation within the principal components, resulting in the different distributions between HeartDisease classes are able to be obtained, FAMD lacks descriptive ability as to which way the variables do so. In terms of data exploration, PCA conveys more information than FAMD.

The paired scatter plots and 3D scatter plots of PCA and FAMD shows that the HeartDisease classes are most likely unable to be linearly separable even in a higher dimensional space. The both classes appear seemingly together within the same populations across different principal components, with only tendencies towards distributions visible across the principal components.

C. Classification Results

As ascertained from the results Naive Bayes classifier using PCA data provides the best accuracy, precision and F1-score while SVM classifier using FAMD data produces the best recall. Initially, it would appear that Naive Bayes has the best performance overall as it is able to correctly classify a larger portion of the data. However, it can be seen that it is actually the worst at classifying positive cases compared to other methods, resulting in the poor recall. Given that the data obtained is from a survey, one may argue that it may be more beneficial to capture and predict as many positive cases as possible, such that when the model is deployed, more individuals would be alerted as to allow them to be aware of their condition and the risks of developing heart disease. In such a case, recall is a much more suitable metric to be used for evaluation, and SVM with FAMD would be much more suitable for the task.

Due to the nature of the dataset whereby the population of both positive and negative cases for heart disease are not visibly linearly separable, a large portion of the negative class is classified as false positive. This may not be due to bad implementation of the model. According to [8], “non-fracture case might be a fracture case waiting to happen”, and provided this dataset is based on survey data, some negative heart disease cases might be positive cases waiting to happen. Solidifying further that recall may be a better evaluation metric to be used and that SVM with FAMD might be a better classifier for the task. Further investigations must be done to conclude these speculations.

Compared to the code on Kaggle mentioned [1], where a random forest classifier is used on original up sampled data, the performance we achieved was slightly worse in terms of F1-score achieved but better in terms of accuracy, where an accuracy of 0.59, precision of 0.22, recall of 0.83 and F1-score of 0.35 was achieved. Investigating data sampling methods and random forest classifiers could be done in an attempt to improve performance.

VII. Conclusion

Overall, the relationship between variables within the dataset was able to be determined from data analysis. PhysicalHealth, DiffWalking, KidneyDisease, Diabetic, Smoking, SkinCancer, AgeCategory, BMI, MentalHealth and Asthma variables increase the chance of heart disease while GenHealth and PhysicalActivity does the opposite within the dataset. Therefore, these variables are recommended to be focused on and further investigated in future works. The machine learning models produced were not able to predict heart disease very effectively, however have similar performance compared to other works mentioned. The lackluster performance could potentially be due to the nature of the dataset and its limitations rather than poor modelling approaches. Further investigations into the topic using different datasets could be done as well for validation. Further experimentation with different up sampling methods could be investigated to produce better results through solving the dataset imbalance issue without losing dataset variability caused by down sampling.

VIII. Author Contribution

Cheong, B. C. and Khor, Y. T. shared the task of exploratory data analysis using similar and separate approaches which was then resolved into the report to remove redundancy. Different data transformation methods and machine learning methods were then researched and implemented by Cheong, B. C. and Khor, Y. T. producing separate approaches that were evaluated within this paper. The experimentation and evaluation method were devised by Khor, Y. T. and implementation was aided by Cheong, B. C. Experimentation process was carried out by Cheong, B. C. and all authors evaluated and discussed the results and contributed to the final paper.

References

[1] Kamil, P., (2022). Personal Key Indicators of Heart Disease. Retrieved [1st April 2022] from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

[2] Mohan, S., Thirumalai, C., Srivastava, G. (2019) Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[3] Singh, A., Kumar, R. (2020). Heart Disease Prediction Using Machine Learning Algorithms. International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

[4] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

[5] Kalantan, Zakiah & Alqahtani, Nada. (2019). A study of principal components analysis for mixed data. International Journal of ADVANCED AND APPLIED SCIENCES. 6. 99-104. 10.21833/ijaas.2019.12.012.

[6] Mahmood, M. S. (2021, July 12). Factor Analysis of Mixed Data. Medium. https://towardsdatascience.com/factor-analysis-of-mixed-data-5ad5ce98663c

[7] Jérôme Pagès. (2015). Multiple factor analysis by example using R. Crc Press, Taylor & Francis Group.

[8] Sharpe, P. K., Caleb, P. (1998). Self Organising Maps for the Investigation of Clinical Data: A Case Study. Neural Computing & Applications(1998) 7:65-70