Occupation-based Predictive Model for Cardiovascular Health

Team 4: Sabari Priya Jaini, Junkang Gu, Siddharth Shah

1. ABSTRACT

This study examined the association between employment attributes and Cardiovascular Disease (CVD) risk using the National Health Interview Survey (NHIS) data. Key predictors including occupation, age, weight, BMI, poverty ratio, and hypertension were identified. Due to dataset imbalance, synthetic data augmentation was employed. The data underwent rigorous preparation and pre-processing for machine learning models. Our models, including Random Forest, SVM, and Neural Networks, yielded effective CVD predictions when leveraging synthetic data and comprehensive attributes. This research provides critical insights into employment-related CVD risks, informing future risk management and preventive strategies.

2. INTRODUCTION

The aim of this project is 2-fold: (1) create a predictive model for cardiovascular health status that uses occupation-related factors to identify individuals at risk for cardiovascular diseases (2) analyze the trends between various occupations/industries and the CVD risk of people belonging to them across years (2015 - 2021). The project addresses the issue of insufficient risk assessment tools for individuals in various occupations, as well as a lack of understanding of how occupation-related factors influence CVD risk. By incorporating occupation-related factors, the proposed predictive model aims to improve the accuracy of CVD risk assessment and provide tailored recommendations for individuals (based on their occupation) to reduce their risk of developing CVD. Furthermore, the project attempts to answer questions like “Is the risk of CVD increasing or decreasing in a particular industry over the years?” and “In recent years, how has COVID-19 affected the fraction of the population that has (had) CVD among various occupations/industries?”

3. DOMAIN UNDERSTANDING: LITERATURE REVIEW

We conducted a literature review to better understand existing methods for assessing CVD risk and ascertain that the occupation of a person is indeed predictive of their CVD risk. The papers/studies we reviewed are as follows:

- Cardiovascular Health Status by Occupational Group: This is a telephonic survey, which asked people about their health attributes, lifestyle choices, and occupation. It ascertained that occupation and CVD risk are highly correlated.
- Clinical Implications of Revised PCE for Estimating Atherosclerotic CVD: Pooled Cohort Equations is a commonly used method for CVD risk assessment. It includes features like age, gender, race, blood pressure, cholesterol levels, and smoking status, but not occupation. Major limitation: overlooks minorities like hispanics, etc.
- Performance of the PCE to Estimate Atherosclerotic CVD Risk by BMI: Same as point 2.
- A case control study of occupation and CVD risk: This study investigated the risks of cardiovascular diseases associated with specific occupations using an inpatient clinico-occupational survey and multivariate logistic regression analysis. It demonstrated associations between specific occupations and the risk of cardiovascular disease incidence and suggested that the risk may vary by occupation.
- Risk prediction of CVD using ML classifiers: This study employed two machine learning techniques, multi-layer perceptron (MLP) and K-nearest neighbor (K-NN) for CVD detection using UCI repository data. Limitation - did not include occupational attributes.
- The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey - This paper analyzes the usage of GANs to synthesize new data to reduce class imbalance in the dataset. It is relevant to this project because the dataset used here is also unbalanced, and hence, one of the ways in which we have

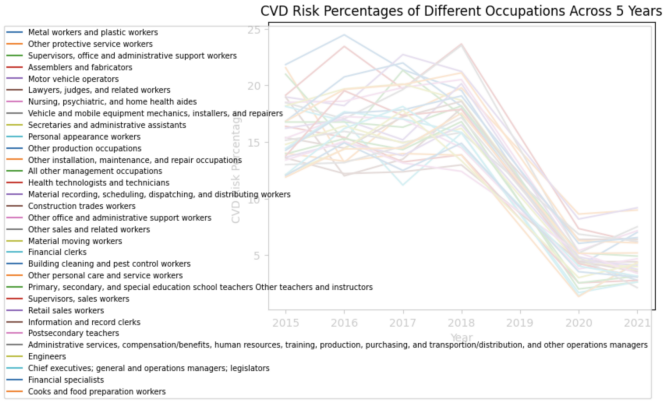
tried to alleviate this problem is by generating more samples (data synthesis) of the smaller sized class.

We discovered that the methodologies used in the above mentioned studies have some limitations, such as ignoring factors such as certain races, economic status, and occupation in CVD prediction. By incorporating all these factors in our predictive model, we hope to address these limitations.

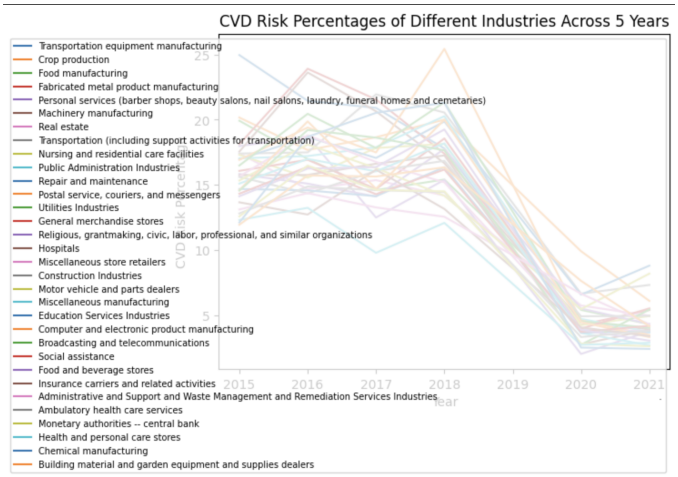
4. DATA SOURCES AND UNDERSTANDING

Data from the National Health Interview Survey (NHIS) by Centers for Disease Control and Prevention (CDC) is our primary data source. For building the predictive model, we concentrated on data from 2020 and 2021, which includes occupation-related attributes, cardiovascular health status, and other attributes which may affect the cardiovascular health of a person. On preliminary exploration, we found most of the columns in the dataset to be categorical, age and weight being the only exceptions. By careful examination, we found that some questions in the interview were asked based on the interviewees' responses to some other questions. This created complex dependencies among attributes and a lot of columns with missing values, making data preparation challenging. While determining the target variable, we observed that there were 4 columns (coronary heart disease, stroke, angina, myocardial infarction) which needed to be combined effectively to make the binary target variable (CVD). After that, we performed data visualization to better understand it. Among other plots, we made plots relating occupations/industries with CVD risk for several years and found that certain occupation/industry groups showed more susceptibility towards CVD risk than others. Also, for the second goal of this project (analyzing trends between occupations/industries and CVD risk across years) we made plots showing occupations/industries and the proportion of positive CVD cases in them across several years.

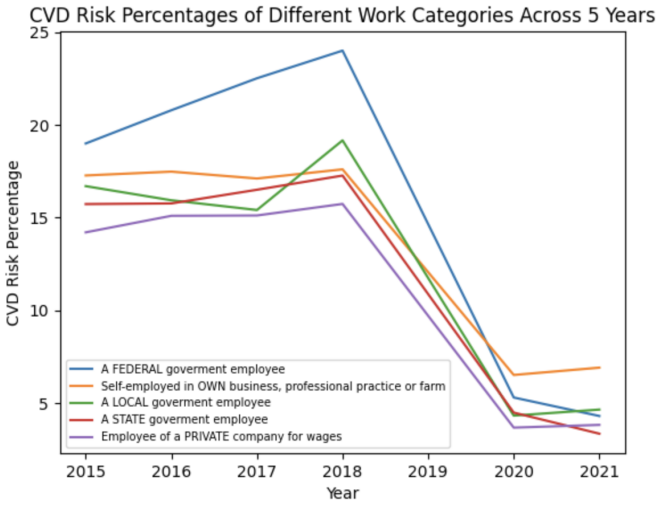
4.1 Detailed analysis of relationship between CVD risk and occupations/industries/work categories across years (2015, 2016, 2017, 2018, 2020, 2021)



The above graph shows years on the x-axis and percentage of people who have (had) CVD on the y-axis, for different occupations. It can be seen that throughout the years 2015-2018 the percentage of people with CVD fluctuated with some peaks and valleys, but in the years 2020-2021, the percentages reduced dramatically. We speculate that this is because our dataset is taken from a survey which contains adults who were alive at the time of the survey. And, in the first year of COVID-19, CVD deaths saw a steep rise in the US according to this article. Hence, a lot of people with medium to serious cases of CVD died during 2020-2021. The reason for this, as per the article, is that due to the pandemic, hospitals were under a lot of stress with a lack of beds, oxygen, and other facilities, and people did not want to seek medical help like regular checkups, etc due to the fear of the virus. Due to this, a lot of CVD deaths were reported during the pandemic. As a result, the overall proportion of CVD in the 2020 and 2021 datasets decreased dramatically, leading to a decrease in the proportions of CVD in each occupation/industry.

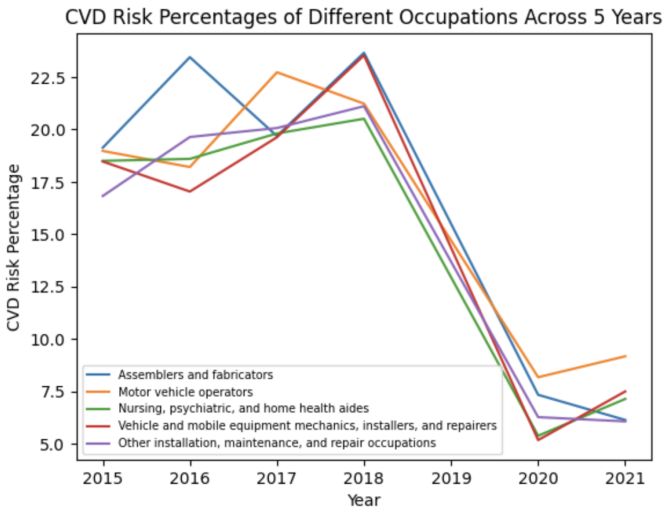


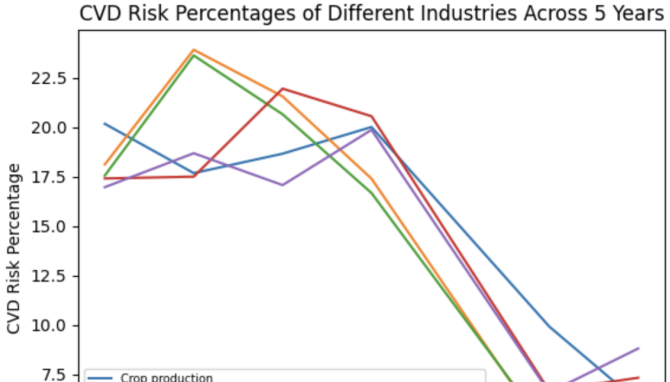
A similar trend can be observed in the above graph, which shows years on the x-axis and percentage of people who have (had) CVD on the y-axis, for different occupations.



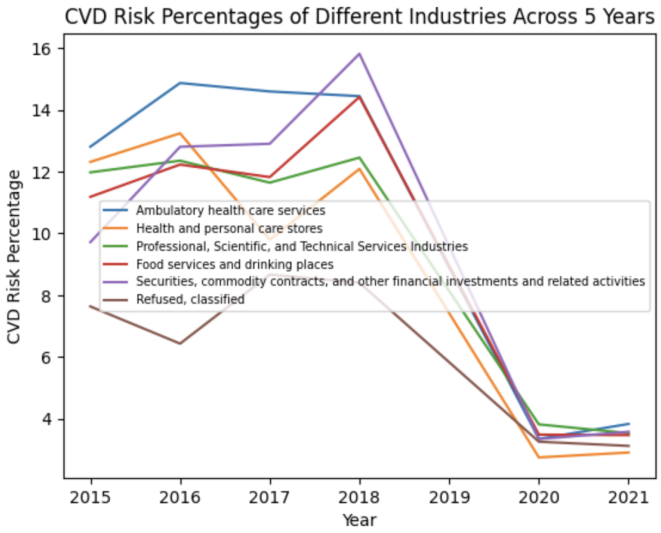
The above graph shows years on the x-axis and percentage of people with CVD on the y-axis for different work categories and a similar trend can be seen in it as well.

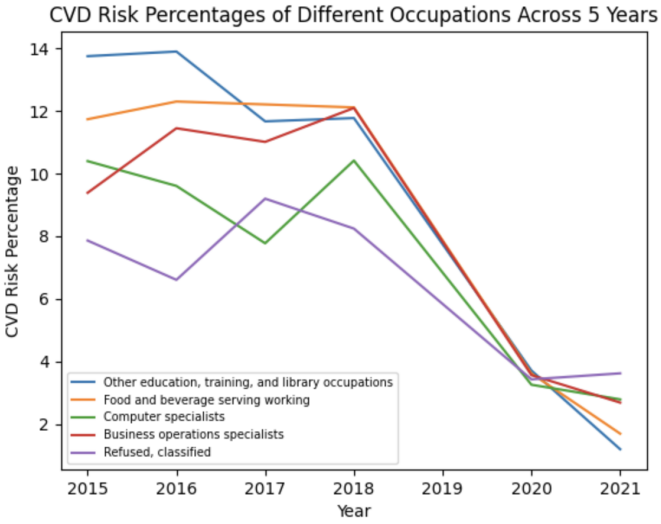
4.2 Industries/Occupations that are always in top 25 in terms of risk of CVD across years





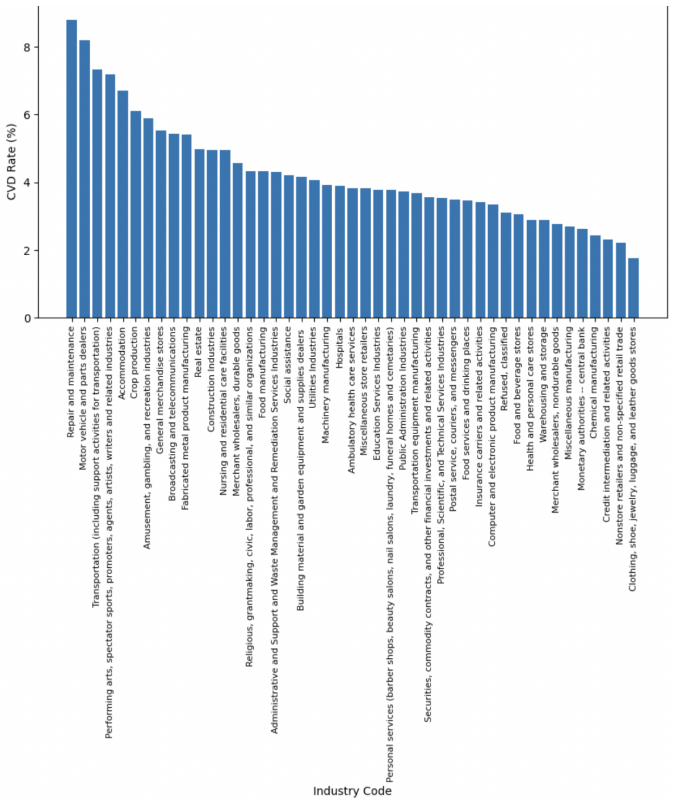
4.3 Industries/Occupations that are always in the least 25 in terms of risk of CVD across years



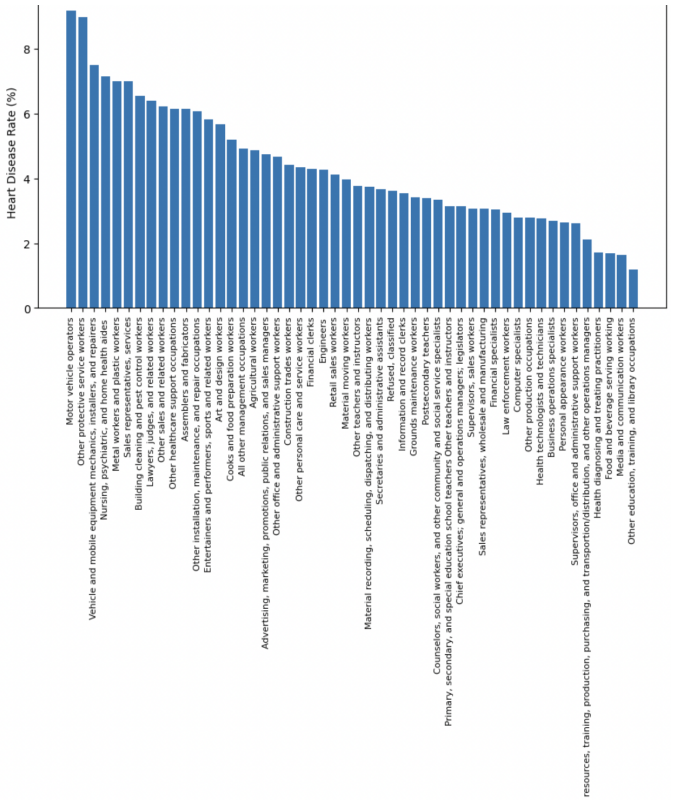


4.4 Employment attributes vs CVD in 2021

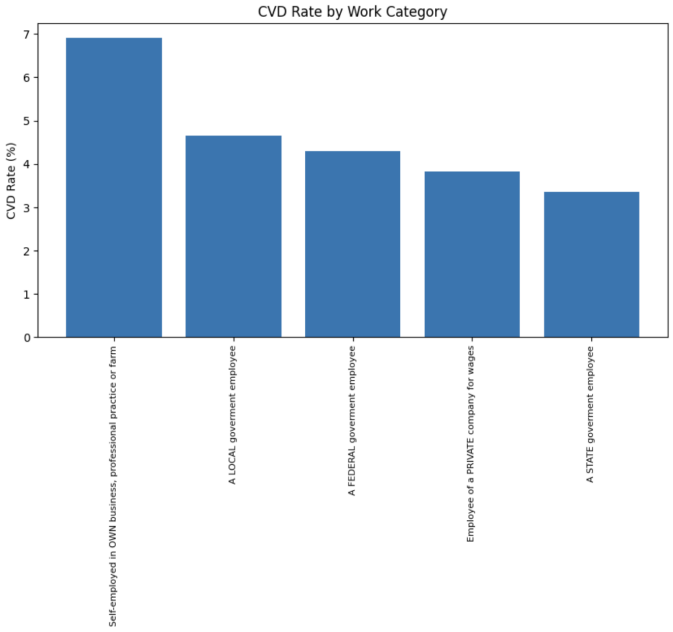
4.4.1 Industry vs CVD



4.4.2 Occupation vs CVD

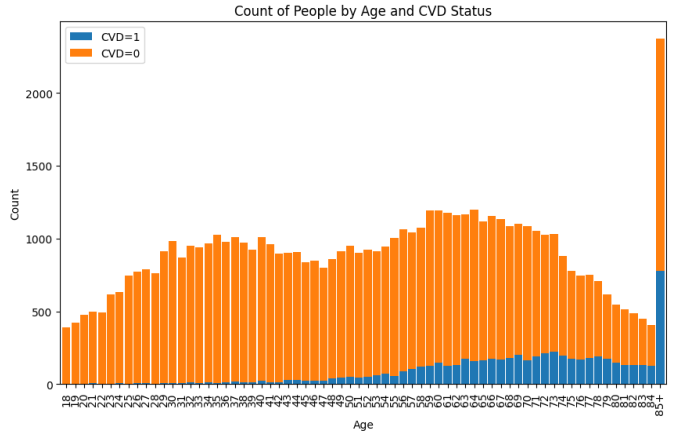


4.4.3 Work category vs CVD risk

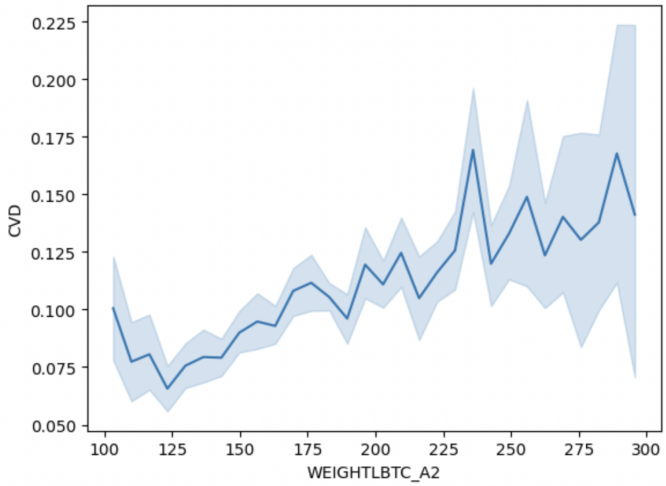


4.4.4 Other Attributes vs CVD

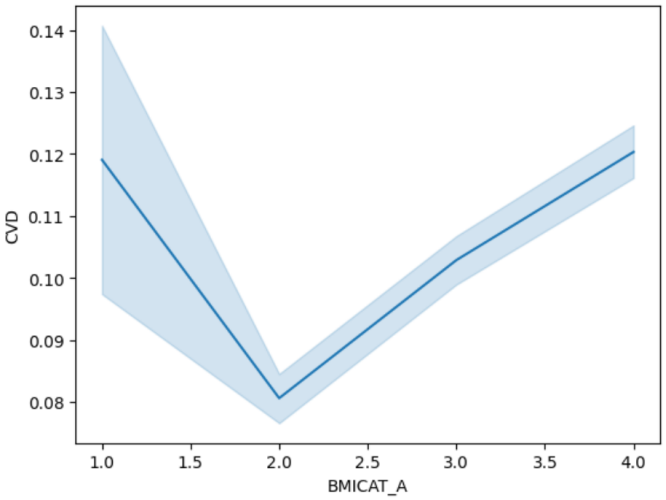
4.4.4.1 Age vs CVD



4.4.4.2 Weight vs CVD

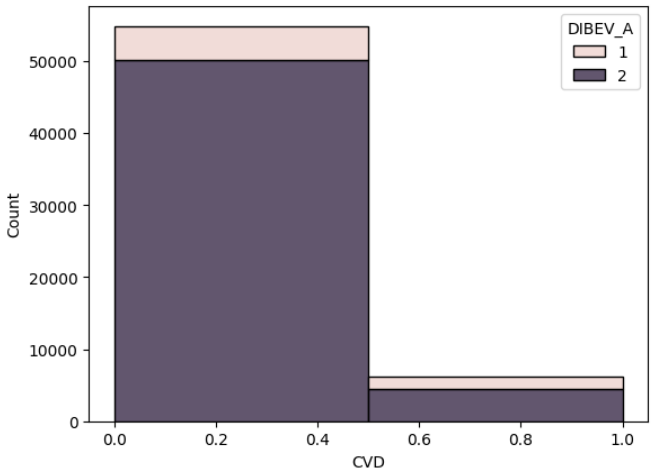


4.4.4.3 BMI Category vs CVD

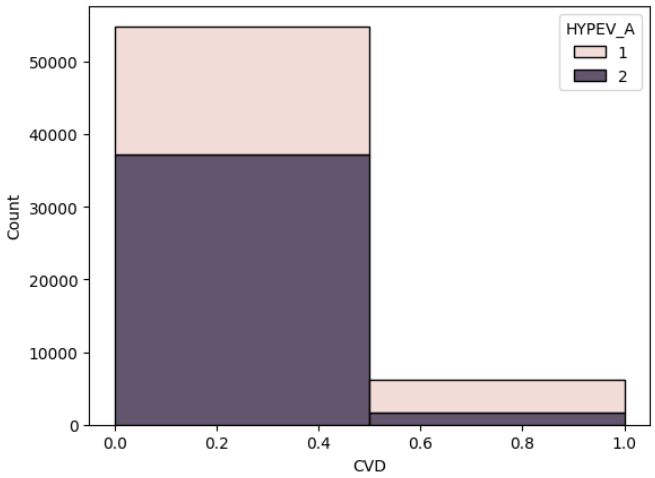


We examined the relationship between categorical Body Mass Index (BMI) and the rate of Cardiovascular Disease (CVD), as shown in the above graph. The BMI categories were classified as Underweight (1), Healthy weight (2), Overweight (3), and Obese (4). The graph reveals a V-shaped relationship between BMI categories and CVD rate. Notably, we found that both Underweight and Overweight categories had high CVD rates. It is important to note, however, that the Underweight category contained very few samples (463), which may have introduced a considerable amount of noise in the data, and therefore, these results should be interpreted with caution. On the other hand, the Overweight category, having a higher sample size, provides a more reliable indication of increased CVD risk. This analysis underscores the complex relationship between BMI and CVD, suggesting that both extremes of the weight spectrum could be associated with higher CVD rates. Further investigation is needed to understand the underlying mechanisms and to control for potential confounding variables.

4.4.4.5 Diabetes ever vs CVD



4.4.4.6 Hypertension ever vs CVD



5. DATA UNDERSTANDING CONCLUSION

Summarizing our dataset exploration and understanding phase, we identified potential key predictors of cardiovascular disease (CVD) risk, which include occupation information, age, weight, BMI, poverty ratio, and hypertension. We noticed a significant decline in the percentage of people with CVD across all occupations/industries/work categories during the years 2020-2021, indicating a shift in the data pattern starting from 2020. As a result, we decided to train our models

exclusively on data from this period to capture the most recent trends. We also recognised that the dataset is imbalanced, with only 10% of the samples having (had) CVD, a factor we addressed by employing synthetic data augmentation in our models. Additionally, most of the attributes in our dataset are ranked or categorical, with only a few being numerical. This discovery informed our choice of modelling techniques and necessitated appropriate preprocessing steps to ensure our models could effectively learn from the data.

6. DATA PREPARATION AND PRE-PROCESSING

- **Data Integration** - We combined the 2020 and 2021 NHIS data by matching the corresponding columns and removing incompatible columns. We concentrated on occupation, cardiovascular health, and other health-related variables such as age, gender, race, blood pressure, cholesterol levels, smoking status, etc. Our plan was to integrate the 2019 data as well. But, this integration was infeasible because employment industry and occupational attributes, which are very important for our hypothesis, were not asked in the interview in 2019.
- **Dependencies and Missing Values** - We analyzed the dependencies discovered in the previous phase and dealt with all these columns by understanding what the missing values meant in each column and filled them accordingly. For example, HYPMED_A (hypertension medication) is asked only if HYPEV_A (ever had hypertension) = 1. So, we filled the missing values in HYPMED_A with a value 2 which indicates no hypertension medication. There were several interdependent variables of this kind, with some having very complex dependencies which related multiple variables. Also, for employment related columns, we used KNN imputers from scikit-learn library to fill the missing values. We did this imputation for occupation, industry and work category in a one-after-the-other fashion so none of the values of any of them affected the imputations of others.
- **Data Transformation** - After analyzing all the columns, we made adjustments to some categories of certain columns based on our requirements. This involved merging or grouping together specific categories of the data in order to better suit our needs. For example, ANXFREQ_A had categories 7, 8 and 9, which indicated Refused, Not Ascertained, and Don't know, which all basically meant unknown data. Hence, we treated them as missing values and replaced them with the mode of that column. We also combined the 4 CVD related attributes (coronary heart disease ever, angina ever, myocardial infarction ever, stroke ever) in the dataset into one column called CVD indicating if the person ever had a cardiovascular disease or not, to suit our project hypothesis. Then, as the smoking status of a person is crucial for determining their cardiovascular health status, we gave it dedicated attention for preprocessing.

```
Total CVD=1                                  6207
Former smokers who have had CVD              2837
Current smokers who have had CVD             941
Never tobacco in any form who have had CVD 2137
```

The above figure shows the total number of data points in class, CVD = 1 and the corresponding numbers for different kinds of smoking statuses. Based on the numbers, it can be observed that among people who have (had) CVD, former smokers were manyfold more than current smokers. This is contrary to the general assumption that current smokers are more susceptible to have (had) CVD than former smokers. The reason, then, for such numbers is that once a person gets CVD, they become more responsible towards their health and quite smoking, thus, becoming former smokers. Hence, contrary to general belief, for our dataset, former smoker status may make a data point more susceptible to belonging to class, CVD = 1, than current smoker status. Hence, we decided to make the smoking-related variable (with categories, current heavy smoker, current light smoker, former smoker, and never smoked) one-hot encoded, as opposed to ranked. Next, we normalize the data using the min-max normalization method to ensure that different features or variables were on the same scale and had equal weight in the analysis, thereby improving the overall quality and accuracy of the results.

- **One-hot encoding** - We applied the technique of one hot encoding to transform our categorical data into numerical data that could be used as input for our machine learning models. This involved creating a binary variable for each category within our categorical variables, where the binary variable would take on the value of 1 if the category was present and 0 otherwise. To perform one hot encoding, we used Python libraries such as scikit-learn or pandas.
- **Correlation analysis** - The correlation matrix provided us with a visual representation of the relationships between the variables. We analyzed the correlation matrix to identify any strong correlations between variables, which could indicate that the variables are related and may have a similar impact on our model's output. We also checked for any correlations that were close to 1 or -1, which could indicate the presence of multicollinearity in our dataset. If we detected any strong correlations or multicollinearity, we took appropriate measures. For instance, if there was a strong correlation between 2 input variables (eg. Diabetes ever and diabetes type 4, which effectively means diabetes never), then we removed one of them.
- **Feature ranking and selection** - First, we created the correlation matrix for the entire dataset and manually inspected highly correlated input features. Based on that just one feature got removed: diabetes type 4. Then, we looked at the correlation of all remaining input features with CVD and ranked the features. Next, we ranked the features based on the chi-square test. Then, we ranked the features based on information gain. Finally, we added all the 3 ranks for each input feature and then selected the top 25 features with the lowest ranks.
- **Synthetic data generation -** After our data analysis and experimenting with some basic models on the data, we realized that the class imbalance problem in the dataset was hindering the models' ability to learn the underlying patterns. As a result, the models were not able to have a high recall for class, CVD = 1. After some literature review and research on how to alleviate class imbalance, we decided on using synthetic data generation using GANs and Gaussian Copula methods to improve the model's performances. Thus, we have used synthetic data for class, CVD = 1, in addition to real data with class, CVD = 1.

7. MODELING AND TESTING

We experimented with detailed employment attributes and simple employment attributes. Detailed employment attributes gave us better performance, and hence, we trained the models with it.

Also, in all the models that were trained and tested on an imbalanced dataset (i.e. without the use of synthetic data), we have included class weights in the loss function, giving more weight to class, CVD = 1. This is done so that the learning algorithms are penalised more when they wrongly classify data points of class, CVD = 1. As a result, using weights improved the recall of class, CVD = 1 (even though it may have potentially decreased the overall accuracy). Note, that we used weight = 1 for class, CVD = 0, and experimented with a lot of values for the weight of class, CVD = 1, ranging from 1 to 11. From the results of these experiments, we found that the best value for class, CVD = 1's weight is 9. Hence, in all the results in the following sections, where the dataset is imbalanced, the same weight is used.

7.1 Random Forest

Dataset used	Accuracy	Recall (Class 1)
Original dataset with PCE attributes	80%	0.57
Original dataset with PCE+Occupation attributes	90%	0.32
Synthetic data augmented with PCE attributes	80.41%	0.79
Synthetic data augmented with PCE+Occupation attributes	82.11%	0.83
Original dataset with top 25 features	90%	0.55

In our study of Cardiovascular Disease (CVD) prediction using the Random Forest algorithm, we experimented with different dataset configurations. When employing the original dataset with PCE attributes, the model achieved an accuracy of 80% but had a lower recall for class 1 of 0.57. The incorporation of occupation attributes into the original dataset substantially increased accuracy to 90%, but the recall for class 1 dropped to 0.32, indicating a challenge in correctly identifying true positive cases. To address this, we augmented the original dataset with synthetic data. For the synthetic dataset using PCE attributes, we achieved a balanced improvement with an accuracy of approximately 80.41% and a recall for class 1 of 0.79. This improvement was more pronounced when we added occupation attributes to the synthetic data, resulting in an accuracy of 82.11% and an improved recall of 0.83. Interestingly, the use of the original dataset with only the top 25 features led to an accuracy of 90%, comparable to when we included the occupation attributes, but with a higher recall for class 1 of 0.55. This shows the effectiveness of feature selection in improving the model performance.

7.2 Support Vector Machine

Our exploration of cardiovascular disease (CVD) prediction also included experimentation with the Support Vector Machine (SVM) algorithm. With the original dataset using PCE attributes, we observed an accuracy of 75% and a class 1 recall of 0.74. When we included additional occupation attributes, the accuracy increased slightly to 78%, but the recall for class 1 decreased marginally to 0.7. Utilizing synthetic data augmented with PCE attributes significantly improved the model's performance, with accuracy reaching approximately 82% and class 1 recall increasing to 0.85. Further adding occupation attributes to the synthetic data resulted in a modest accuracy improvement to about 84% and a recall of 0.83. Using the original dataset with only the top 25 features resulted in an accuracy of 77.8% and a recall of 0.72, which is comparable to the performance on the original dataset with PCE attributes. Overall, these results underscore the effectiveness of SVM in CVD prediction, especially when synthetic data augmentation and a more comprehensive set of attributes are employed.

Dataset used	Accuracy	Recall (Class 1)
Original dataset with PCE attributes	75%	0.74
Original dataset with PCE+Occupation attributes	78%	0.7
Synthetic data augmented with PCE attributes	81.99%	0.85
Synthetic data augmented with PCE +Occupation attributes	83.85%	0.83
Original dataset with top 25 features	77.8%	0.72

7.3 Neural Network

In our study of predicting Cardiovascular Disease (CVD) using a neural network, we tested various models with different dataset configurations and architectures. The architectures were optimized using grid search and L2 regularization to prevent overfitting and improve model performance. With the original dataset using PCE attributes, our model achieved an accuracy of 73% and a recall for class 1 of 0.82, using a neural network architecture of (40, 104, 8, 1). When we included occupation attributes to the original dataset, the accuracy slightly improved to 76% but the recall for class 1 dropped to 0.72, with the same architecture. To tackle the issue of imbalance in the dataset, we generated synthetic data and augmented it with PCE attributes. This resulted in a substantial improvement in both accuracy and recall (87% and 0.87, respectively), utilizing a more complex architecture of (40, 104, 8, 8, 1). A further enhancement was observed when we added occupation attributes to the synthetic data, which led to an impressive accuracy of 93% and a recall of 0.93 with a simplified architecture of (8, 104, 104, 1). Interestingly, using the original dataset with only the top 25 features resulted in an accuracy and recall similar to the original dataset with PCE attributes.

Dataset used	Accuracy	Recall (Class 1)	Best Architecture
Original dataset with PCE attributes	73%	0.82	(40,104,8,1)
Original dataset with PCE+Occupation attributes	76%	0.72	(40,104,8,1)
Synthetic data augmented with PCE attributes	87%	0.87	(40,104,8,8,1)
Synthetic data augmented with PCE +Occupation attributes	93%	0.93	(8,104,104,1)
Original dataset with top 25 features	72%	0.82	(40,104,8,1)

8. CONCLUSION

Our analysis of National Health Interview Survey data from 2015 to 2021 yielded important findings about employment attributes and their relation to Cardiovascular Disease (CVD) risk. We observed consistent patterns where private company employees demonstrated lower CVD risk compared to federal and self-employed individuals. Certain occupations and industries were consistently ranked at the top and bottom in terms of CVD risk. Furthermore, there was a notable decrease in CVD percentages from 2018 to 2020-2021, which we attribute to the effects of the COVID-19 outbreak.

Predictive modeling further reinforced the significance of employment attributes in estimating CVD risk. The utilization of a synthetic dataset indicated that incorporating occupation attributes boosted the accuracy across all the models. However, when evaluating the original dataset based on both accuracy and recall, the impact of including occupation attributes yielded inconsistent results. This underscores the complexity and multidimensionality of factors influencing CVD risk. The insights derived from this research offer critical knowledge to inform more targeted risk management and preventive efforts against CVD, especially among different employment categories. Further research is recommended to expand our understanding of these relationships and to refine predictive models for increased accuracy and consistency.

9. REFERENCES

- Shockey, T et. al., Cardiovascular Health Status by Occupational Group
- Yadlowsky, S et. al., Clinical Implications of Revised PCE for Estimating Atherosclerotic CVD
- Khera, R et. al., Performance of the PCE to Estimate Atherosclerotic CVD Risk by BMI
- Fukai, K et. al., A case control study of occupation and CVD risk
- Pal, M et. al., Risk prediction of cardiovascular disease using machine learning classifiers
- Cardiovascular deaths saw steep rise in U.S. during first year of the COVID-19 pandemic