**Abstract**

*Accessing cardiovascular disease risks across different population groups is crucial for improving healthcare treatments and tailoring medical interventions to the specific needs of diverse communities. This research seeks to create a predictive model to assess risk of cardiovascular disease based on demographic factors such as gender, ethnicity, socioeconomic status, education, etc. By processing through medical records and public health datasets, data processing techniques such as data cleaning, normalization, and feature selection will be used to ensure the quality of the input data within the model. Additionally advanced machine learning techniques such as multi-layer perceptron (MLPs), random forests, logistic regression, and support vector machines (SVM) will be trained to predict outcomes and identify key risk factors for cardiovascular diseases and hospitalization rates. After training and testing medical and public health datasets with the model, we achieved a 89% accuracy, demonstrating the model’s effectiveness in identifying cardiological disease risks and predicting potential pulmonary diseases within different population groups.​*

**Introduction**

In 2024, cardiovascular disease, or CVD, claimed 931,478 deaths [1], leading the United States with around 20% of all deaths in 2022 [2]. The most common type of CVD was coronary heart disease, claiming 40% of all CVD deaths [1]. Additionally, between 2004 and 2015 there was an increase of 605,000 new heart attacks [1]. Consequently, “Between 2019 and 2020, direct and indirect costs of total CVD were $422.3 billion ($254.3 billion in direct costs and $168.0 billion in lost productivity/mortality)” [1].

Between 2017 and 2020, 59% of Black Females and 58.9% of Black Males had some form of CVD, leading all racial groups within the United States [1]. Moreover, CVD also claims other racial groups such as Hispanics, who have 7.4% of all CVD diagnoses within the United States [3]. African Americans and Hispanics together claim about 17% of all CVD diagnoses in 2017. Due to various factors such as obesity, African American men lead within 47.5% and Hispanics right behind with 46.6% [3], and diabetes, 21.6% of all Hispanics and 19.6% of African Americans over 20 [3].

Following this, the healthcare coverage towards these groups have shown to be lacking over the last decade. In 2022, about 18% of all hispanics were uninsured, which made up 40% of all uninsured non-elderly people. Additionally, 10% of African Americans were uninsured, making up 12% of the uninsured population [4]. These numbers bring into question whether appropriate care can be placed into these minority groups. Without the appropriate care to diagnose and treat Cardiovascular Disease early, figures shown above occur, leading to more deaths than necessary.

**Related Work**

The related work done on this subject in the past mainly aimed towards the goal of the common usage of machine learning and predictive models in medical settings to allow for better health outcomes in patient populations. Ward et al. and Sun et al. utilize extreme gradient boosting machine learning to train their models in their research [5, 6]. On a basic level, this is a large decision tree-type model. While this method is useful in classification and regression problems, most studies have limited their models to this particular method.

Our project aims to fill in the gaps in knowledge and test MLP, random forest (which is somewhat similar to extreme gradient boosting due to the usage of a decision tree), logistic regression, and SVM models to see if they can train a more accurate predictive model on this subject.

**Proposed Methods**

**Support Vector Machine (SVM)**  
 Support Vector Machine (SVM) is a supervised learning model widely used for classification tasks. It works by identifying a hyperplane that best separates data points into their respective classes while maximizing the margin between them. For this project, an SVM model will be developed to classify individuals as either negative ASCVD or positive ASCVD. The model leverages kernel functions to handle both linear and non-linear separability in the dataset, ensuring robust performance across varying data distributions. Additionally, the SVM model serves as a benchmark to compare classification efficacy with other models explored in this research.

**Logistic Regression**

Logistic Regression is a technique used for data analysis that finds a linear combination between data factors in order to compute the odds of an event. This makes it especially useful for classification tasks. For this project, a logistic regression model will be created in order to determine whether individuals have ASCVD or not. This model classifies individuals in the dataset as either negative ASCVD or positive ASCVD. The goal of this model is classification as well as comparison to the other models in the scope of this research.

**Random Forest**

Random forests use a number of decision trees during training in an ensemble to perform classification tasks, where classes are determined by the choice of the majority of the decision trees. For the purposes of this project, a random forest model will be created for the classification of ASCVD diagnosis. The model classifies individuals in the dataset as either negative ASCVD or positive ASCVD. The random forest model will be used not only for the purpose of classification, but also for the comparison between the other models in the scope of this research to determine the ideal method of classification for ASCVD diagnosis.

**Multi-Layer Perceptron (MLP)**

We utilize a Multi-Layer Perceptron (MLP) to classify the target variable using a dataset consisting of categorical and numerical features. MLPs are well-suited for this task as they are capable of capturing complex non-linear relationships between features and the target. Our approach involves preprocessing the dataset by encoding categorical variables using one-hot encoding and imputing missing values with the median. Features are then standardized to ensure uniform input scales, critical for optimizing the performance of the neural network.

**Experiments**

**Dataset**

The dataset used contained 30,040 samples, 27,111 of which were negative for ASCVD, and the remaining 2,929 of which were positive for ASCVD. The features used to predict the target feature (ASCVD) included ‘edu’ (education), ‘PIR’ (level of income), poverty, and ‘eth’ (ethnicity). The categorical among these features were replaced with dummy variables or had their values mapped to equivalent numerical values. The preprocessing applied to all of the models included adding class weights in order to give more influence to the data samples with positive ASCVD in order to balance out the difference in the target feature’s class numbers.

**SVM** The SVM model was used for binary classification of ASCVD, with class weights applied to address the imbalance in the dataset. Data standardization was performed using StandardScaler, and several kernel functions, including linear, polynomial, and RBF, were tested with minimal performance differences. Despite its flexibility, the SVM achieved an accuracy of 68%, highlighting its difficulty in handling the positive class. This model serves as a baseline for comparison with other techniques explored in this research.

**Logistic Regression**

The logistic regression model was initially created without class weights for the binary classification of ASCVD, but after discovering the inability of the model to predict positive ASCVD, respective class weights of 5.2 and 0.55 for positive and negative ASCVD were ultimately added. The maximum iterations the model was allowed was 1000. The data was split into 20% for testing and 80% for training.

**Random Forest**

The random forest model was also initially created with no class weights, but seeing the inability of the model to predict positive ASCVD, respective class weights of 5.2 and 0.55 for positive and negative ASCVD were added to the final model. The number of estimators (decision trees in the forest) used was 150. The maximum depth these estimators were allowed to reach was 100. The data was split with 80% for training and 20% for testing. Bootstrapping could not be employed in the final model; although it raised the model’s accuracy as a whole, it greatly worsened the prediction accuracy when predicting positive ASCVD. The same problem was encountered when using gini index as the measure of feature importance, so entropy was used instead.

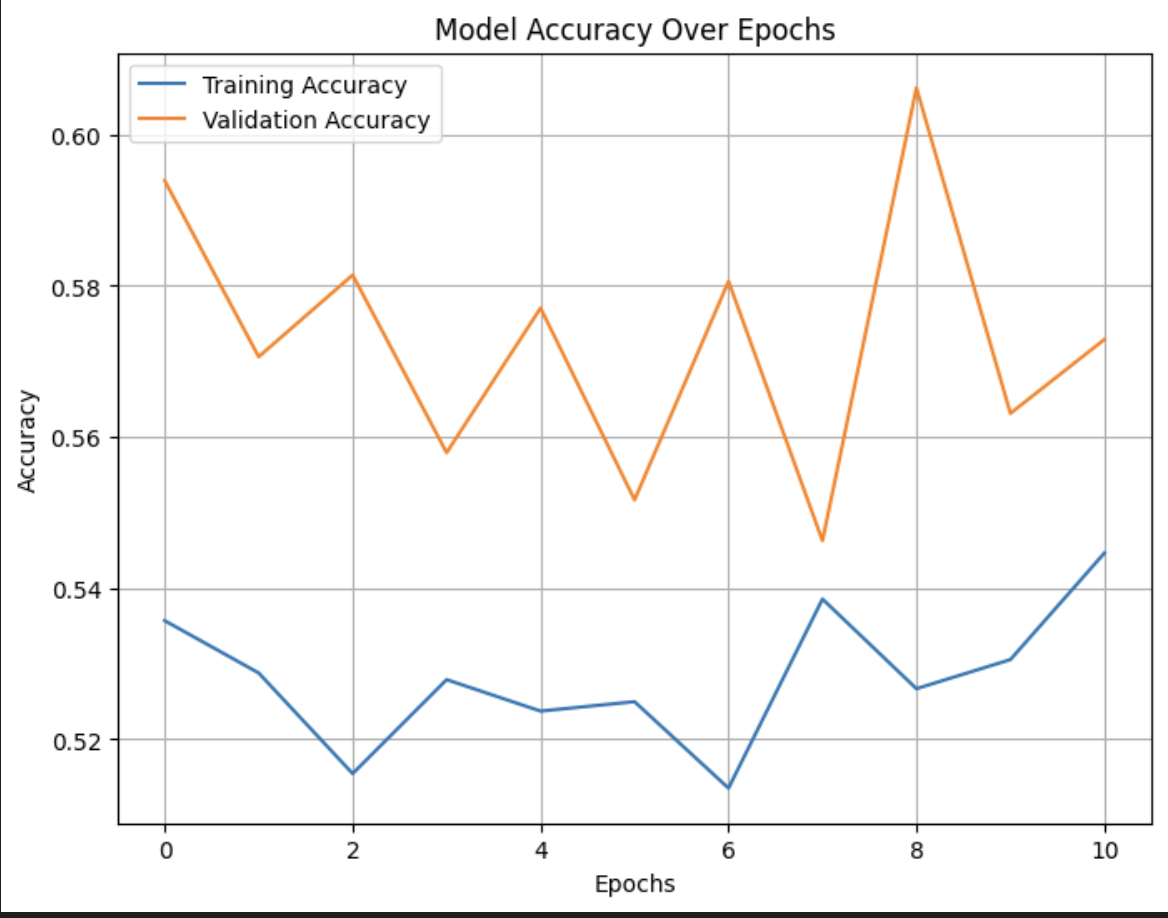
**MLP**

Initially, a simple Multilayer Perceptron (MLP) model with 4 hidden layers featuring decreasing neuron counts (128, 64, 32, 16) was used for binary classification. The structure of the model consisted of fully connected layers (Dense layers) with ReLU activation functions and minimal dropout for regularization. Adam optimizer was used with a learning rate of 0.01 to train the model. The reason for using Adam optimizer was the scale of the dataset. Adam keeps track of the moving average of the gradients to smoothly update. This prevents oscillations in the gradient descent path and accelerates convergence in relevant directions.

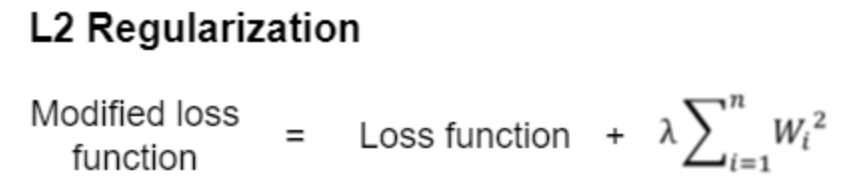
ReLU, a widely used activation function, was chosen for its computational efficiency and straightforward gradient behavior. However, this basic architecture faced challenges, particularly with overfitting and the potential for “dying neurons”—a problem where neurons permanently output zero due to negative inputs in ReLU.

The dropout rates in this initial model were relatively low, which limited the network’s ability to generalize by not sufficiently regularizing the connections. While the model showed some promise, it struggled with generalization to unseen data, especially in cases where the dataset exhibited high variance or noise. This initial model had an accuracy of 54%.

**MLP Model with LeakyRelu**

Instead of ReLU, LeakyReLU was introduced as the activation function. LeakyReLU solves the “dying neuron” problem by allowing a small, non-zero gradient for negative inputs. This change enabled the model to learn more robust representations,especially in deeper layers. The introduction of LeakyRelu increased the accuracy to 59%

**Adding Kernel Regularization to the MLP**

To further improve the performance and generalization of the MLP model, L2 kernel regularization was introduced. Regularization helps prevent overfitting by penalizing large weights, encouraging the model to learn simpler, more generalized patterns.

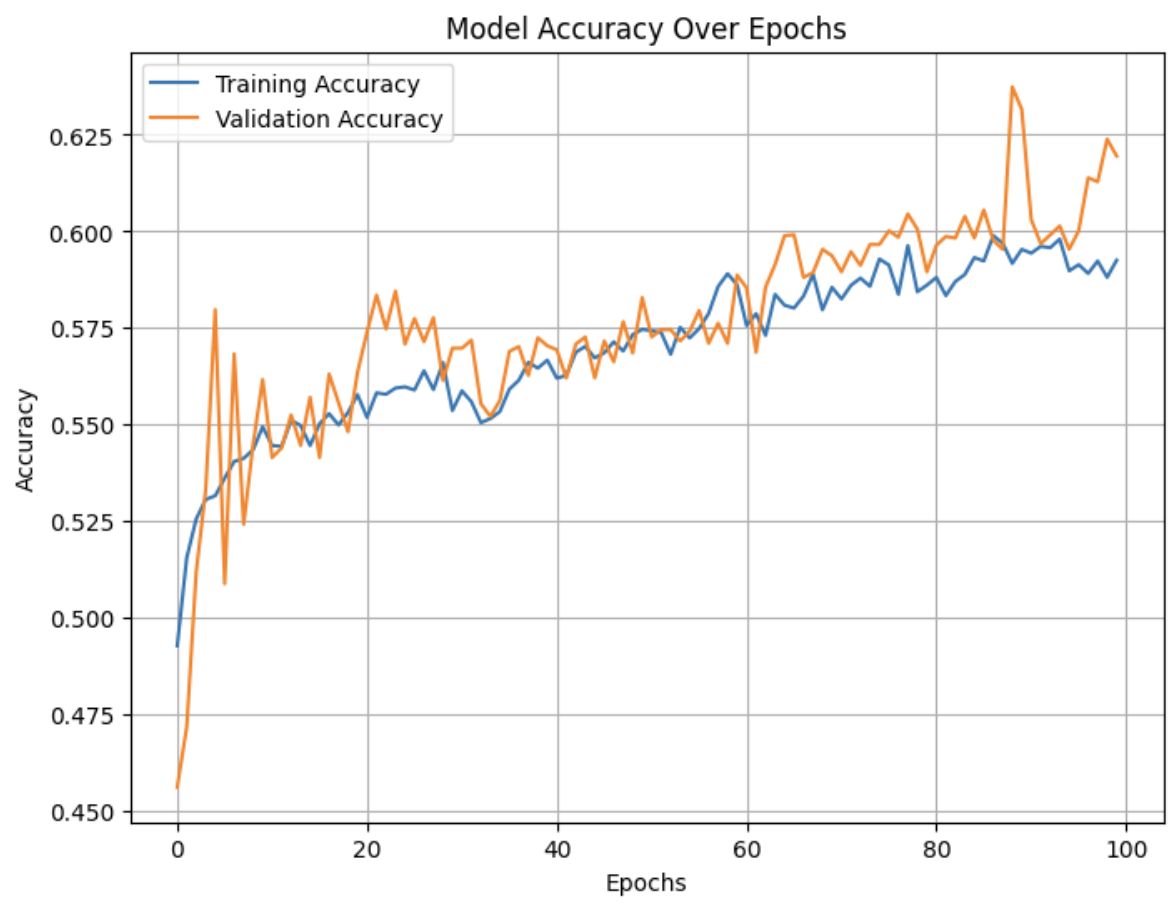
The introduction of L2 regularization increased the accuracy to 61%.

**Modifying the Adam Optimizer Learning Rate in the MLP Model**

Due to instability in the learning, the learning rate was lowered to 0.001 so that the model could take smaller, more precise steps. Lowering the learning rate also works well with regularization and dropout to ensure that the updates are gradual and consistent with the regularization goals. The learning rate of 0.001 was used after multiple experiments with the model.

**Using Early Stopping with the MLP**

To avoid wasting computational resources, a method called early stopping was used. In this case, the early stopping method tracked the validation loss during training with a patience of 10. Patience essentially determines the number of epochs that the early stopper tolerates before it stops the training procedure.

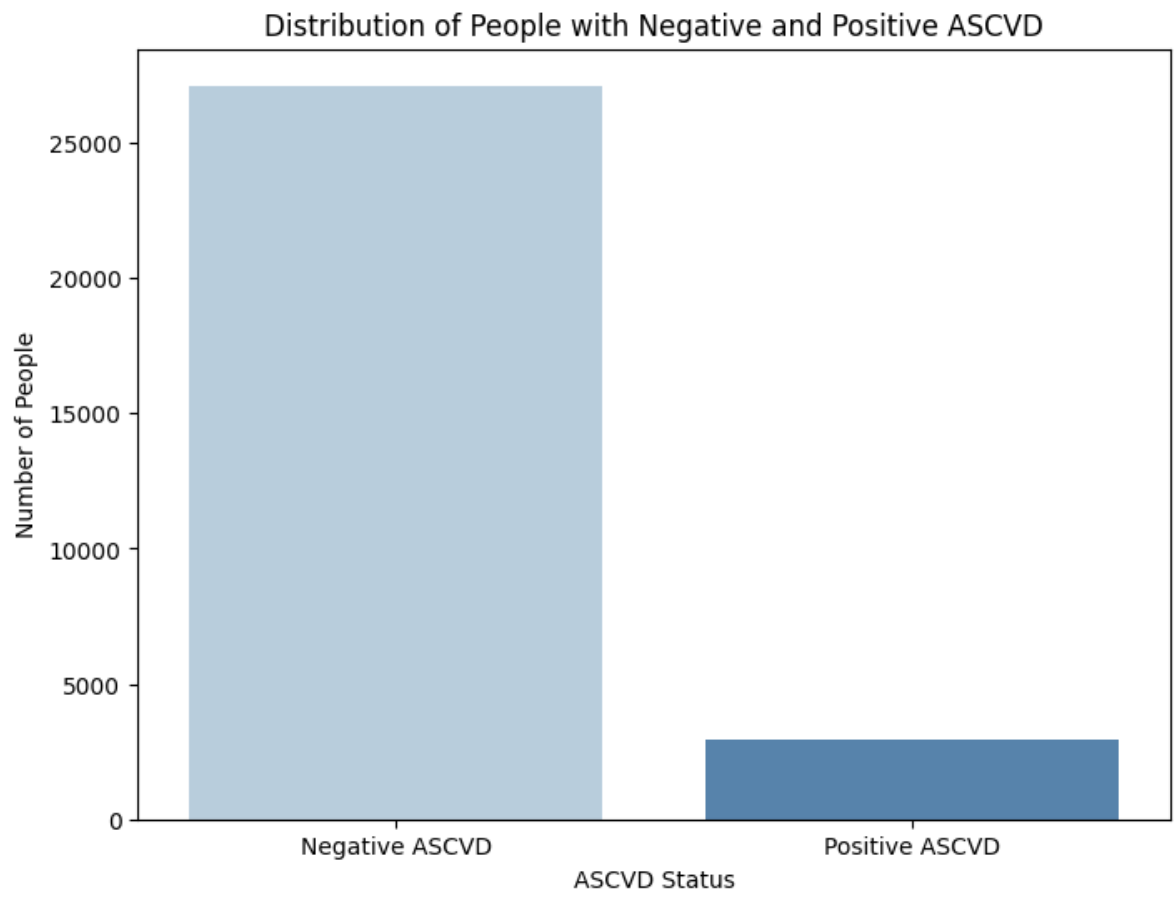
****

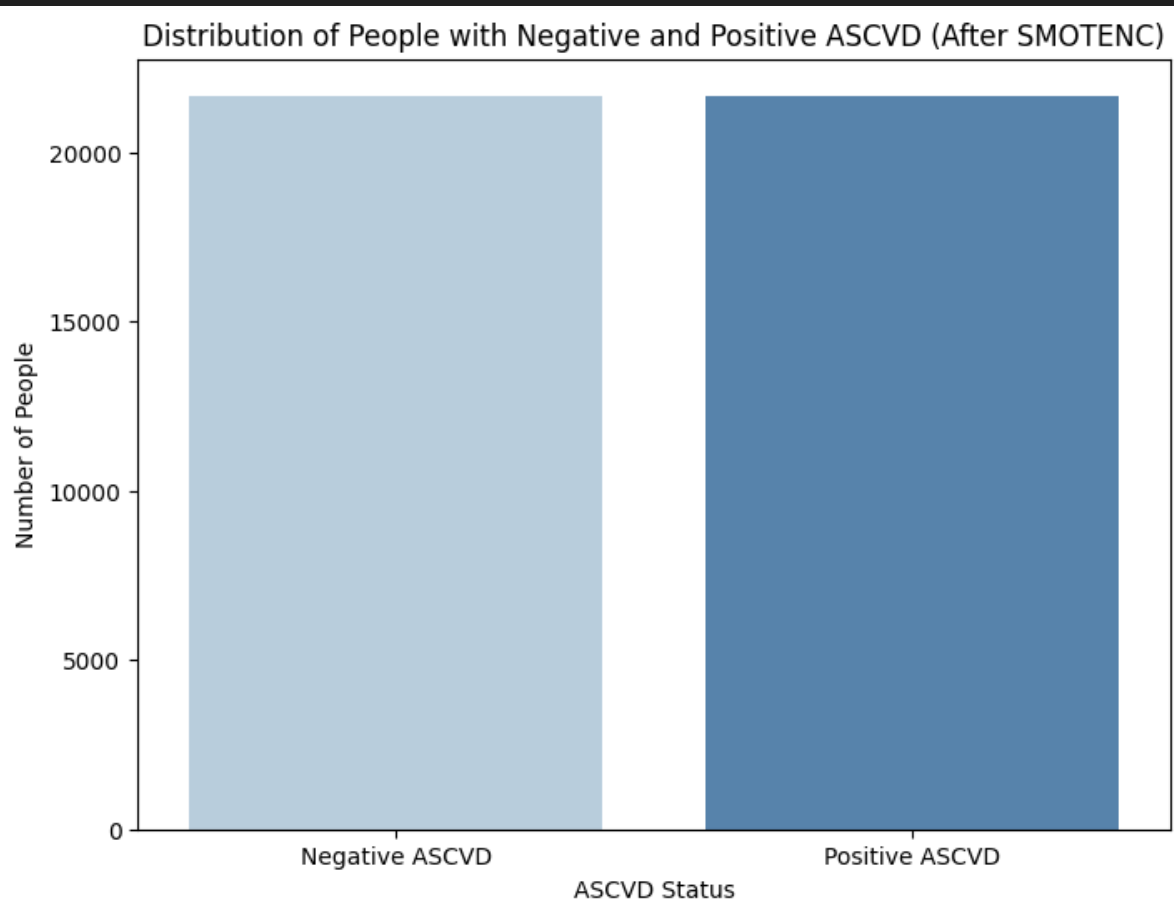
The accuracy for this model reached around 60%.

**Training the Model with Synthetic Data**

Due to the huge imbalance in the data, a method called SMOTENC was used to generate synthetic data.

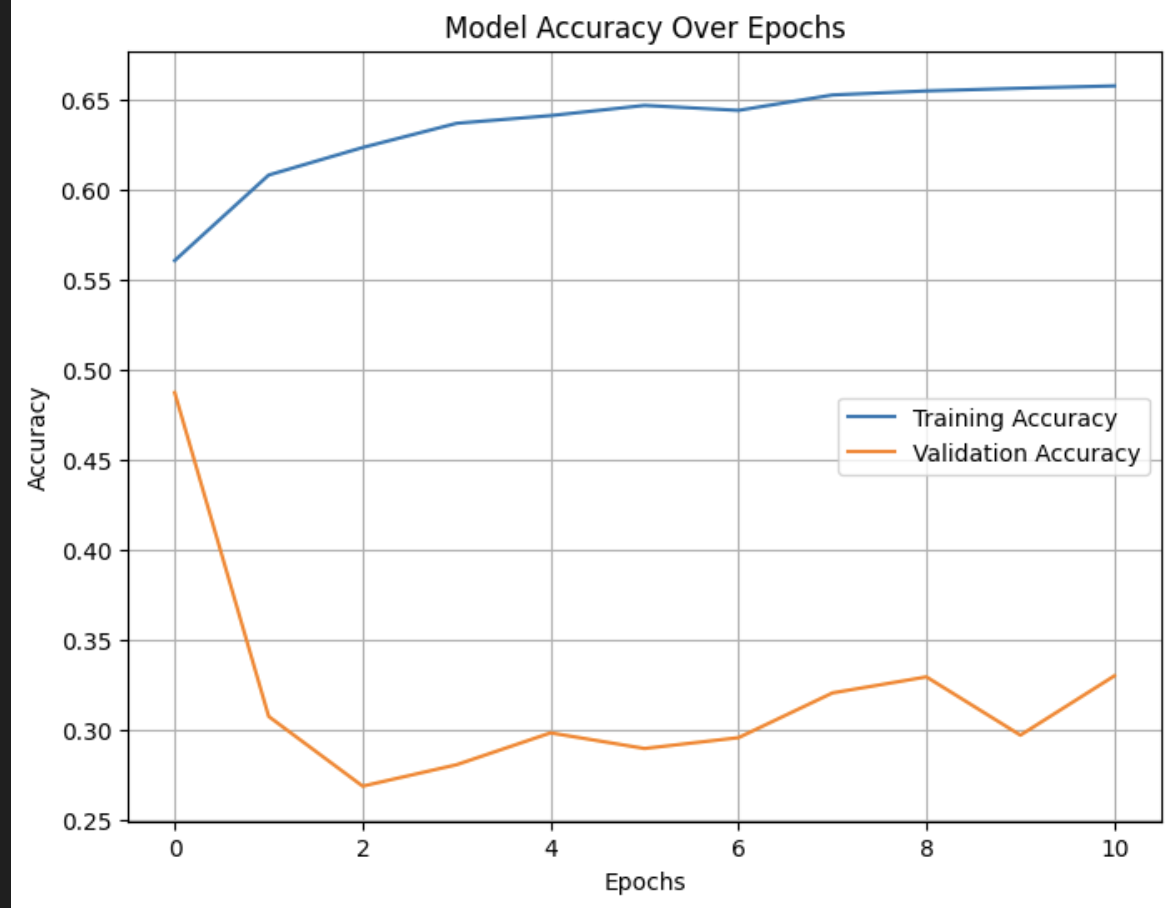
The image below shows the distribution of data before adding the synthetic data.

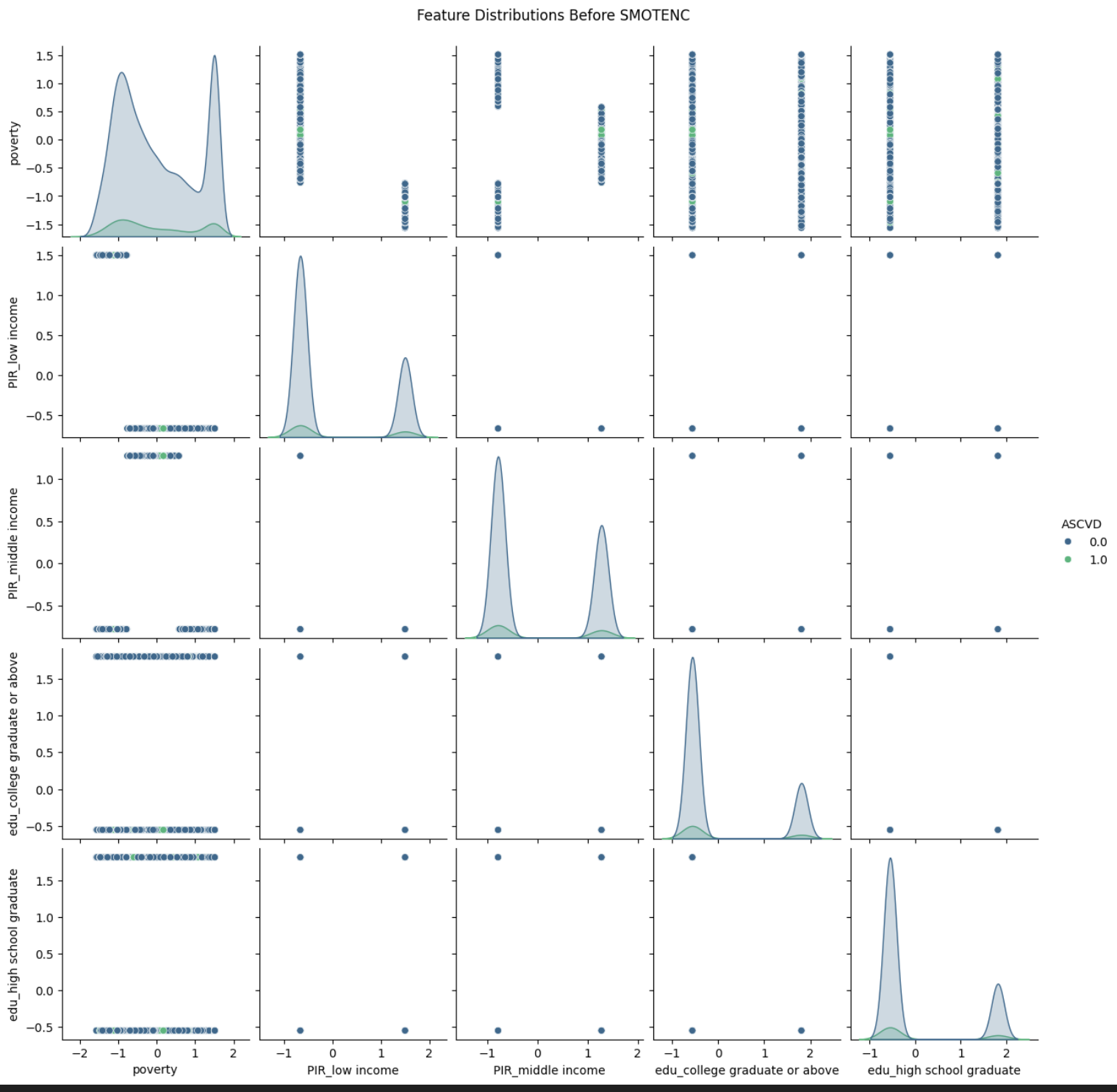


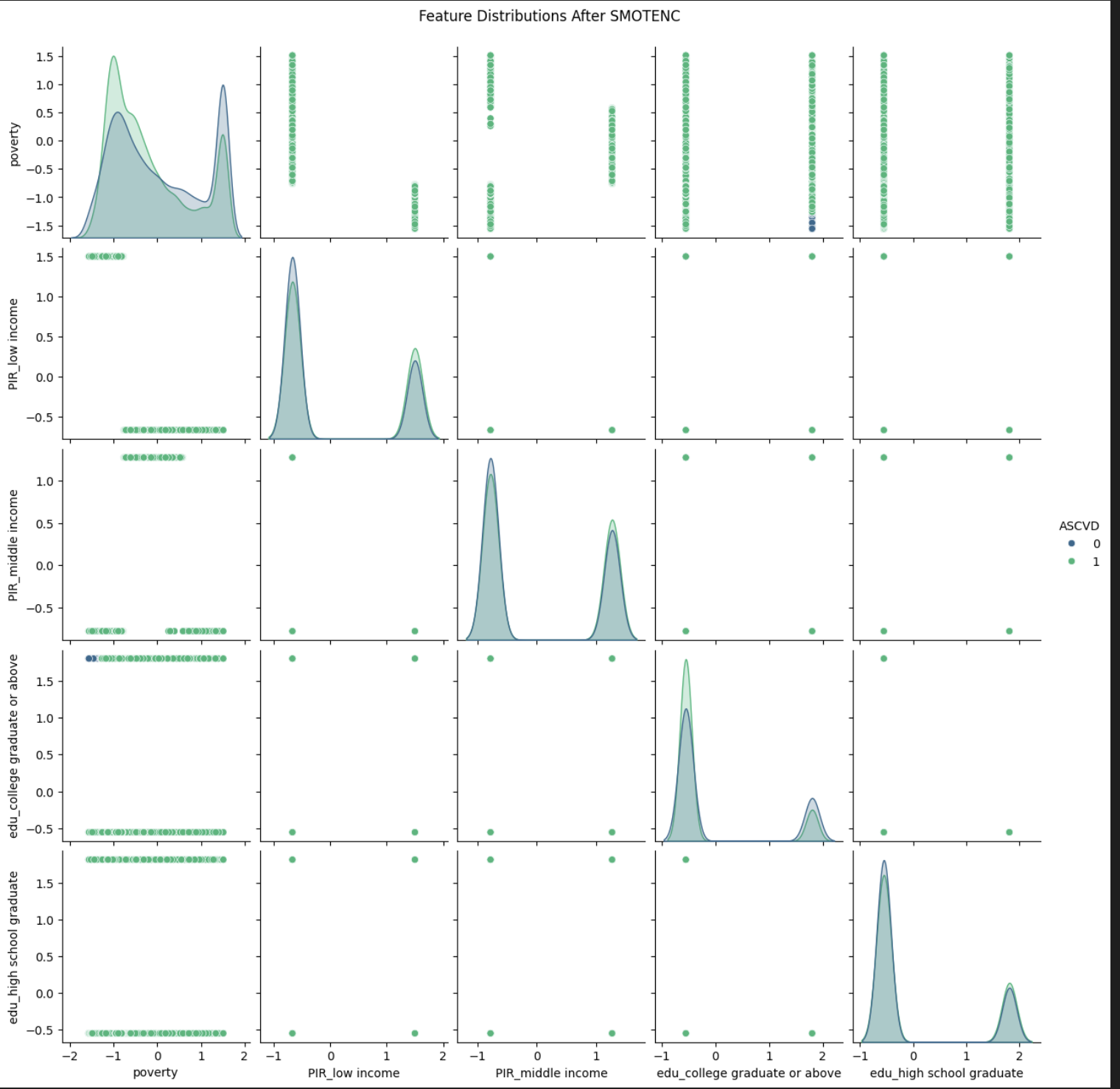


The image above shows the distribution of data after adding the synthetic data.

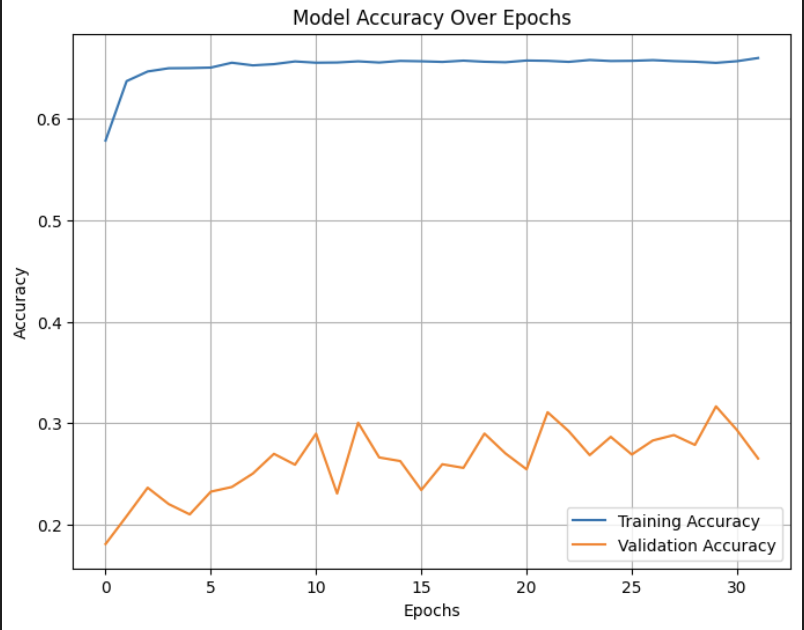
The resulting model gets to a training accuracy of 67%. However, it does not perform well for the validation. This indicates that the model is struggling to generalize.

****



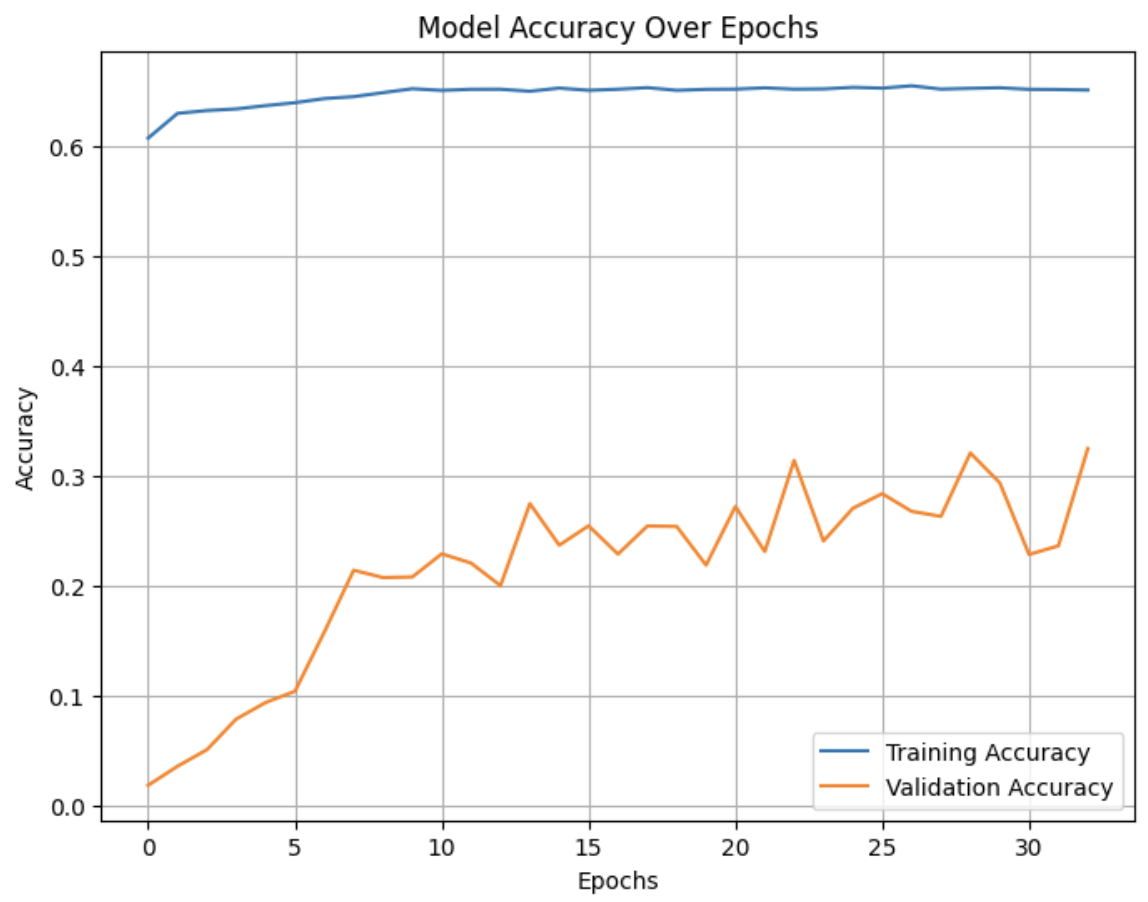


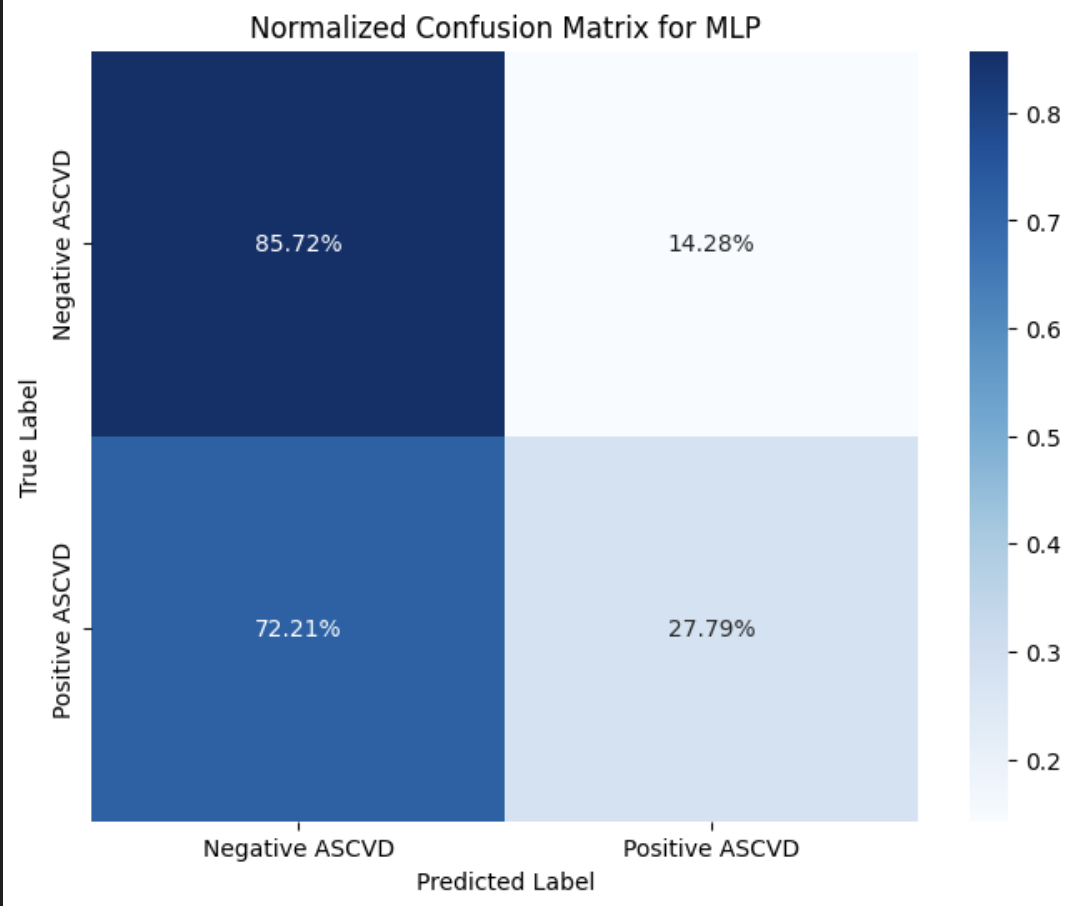
After analyzing the distribution of post-addition of the synthetic data, it seems like the synthetic data did not generate a lot of noise. Therefore, an experiment was made using a much simpler MLP with 2 hidden layers and decreasing neuron counts (32, 16).



The model performed well on the training validation with an accuracy of 73%, but there were still signs of the model being overfitted.

Next, the neural network’s architecture was further modified, and a model with 3 hidden layers with decreasing neuron counts (64, 32, 16) was made.

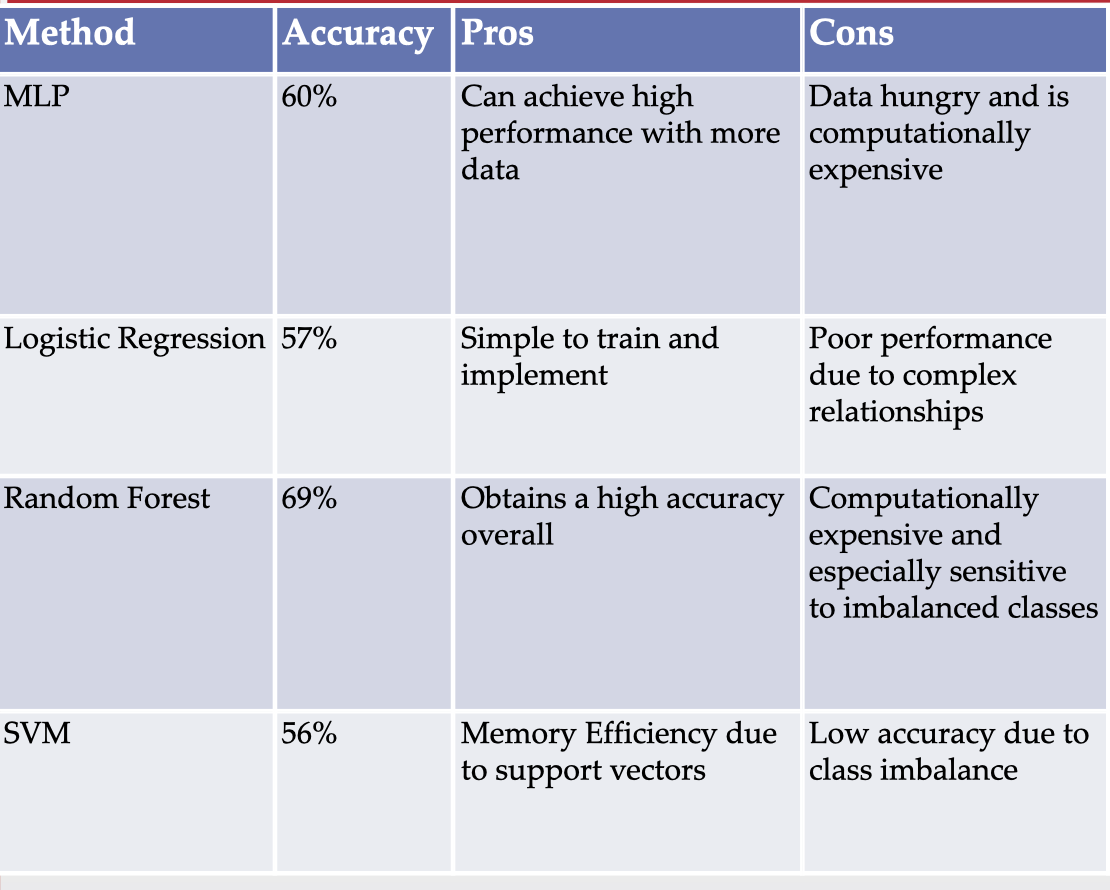


However there were still signs of the model being overfitted, and the model was predicting a great number of false negatives.

**Evaluation of Models**

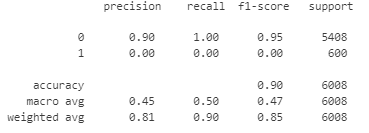
The primary goal of this project was to further research the impact of a person’s socioeconomic status on their health, and especially on their propensity for having ASCVD. In doing this, we compare the performance of logistic regression, random forest, SVM and multilayer perceptron models on their ability to predict ASCVD in the scope of this research.

**Results**



**SVM**

Prior to applying class weights, our accuracy achieved was 90% for the binary classification of ASCVD.



After applying class weights, the SVM model achieved an accuracy of 68%, with a high precision for negative ASCVD (0.97). However, the model struggled with positive ASCVD, showing low precision (0.21). This indicates the model's bias towards predicting negative ASCVD, likely due to the class imbalance in the dataset. The F1-score for negative ASCVD was strong (0.80), but the low F1-score for positive ASCVD (0.33) highlights the model's challenges with the minority class. Further adjustments, such as tuning class weights or using resampling techniques, could help improve predictions for positive ASCVD.

**A screenshot of a computer code

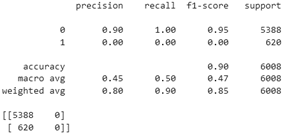
AI-generated content may be incorrect.**

**A blue squares with white text

AI-generated content may be incorrect.**

**Logistic Regression**

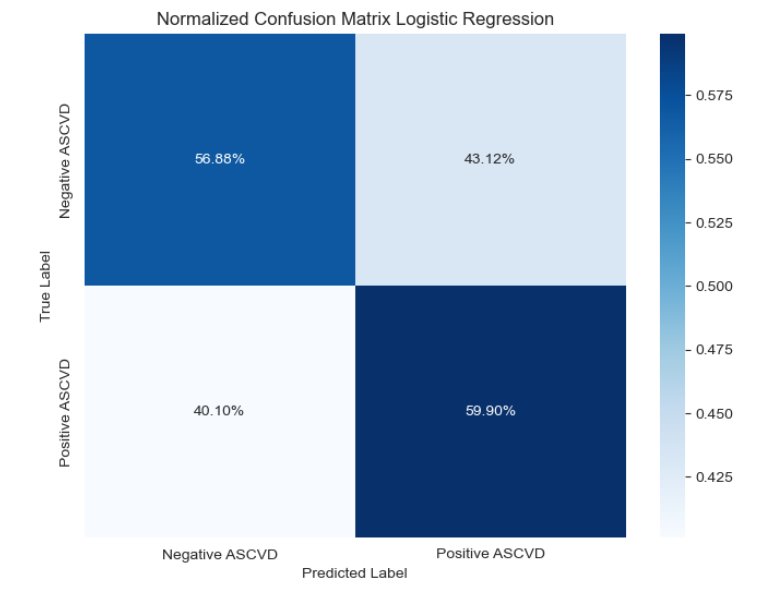
The logistic regression model that was initially created without class weights achieved an accuracy of 89.7% for the binary classification of ASCVD.



After implementing class weights to the logistic regression model, its accuracy was greatly reduced to 58.4%. This drop in accuracy was a direct result of the class weights, but so was the increase observed when the model predicted positive ASCVD. The decrease in the model’s accuracy was also seen by the f1 score drop when the model predicts negative ASCVD, though the final model was still worse at predicting positive ASCVD.

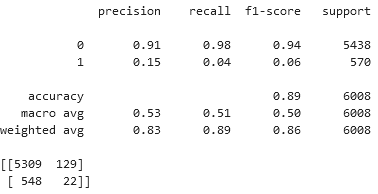
A screenshot of a graph

Description automatically generated



**Random Forest**

The initial model created for random forest had an 89% overall accuracy for the binary classification of ASCVD, but was only able to predict when someone did not have ASCVD.



After implementing the same class weights as the rest of the models, the random forest overall accuracy for the binary classification of ASCVD came out to 69%. This was much lower, but the model was then able to predict when someone was positive for ASCVD much more often.

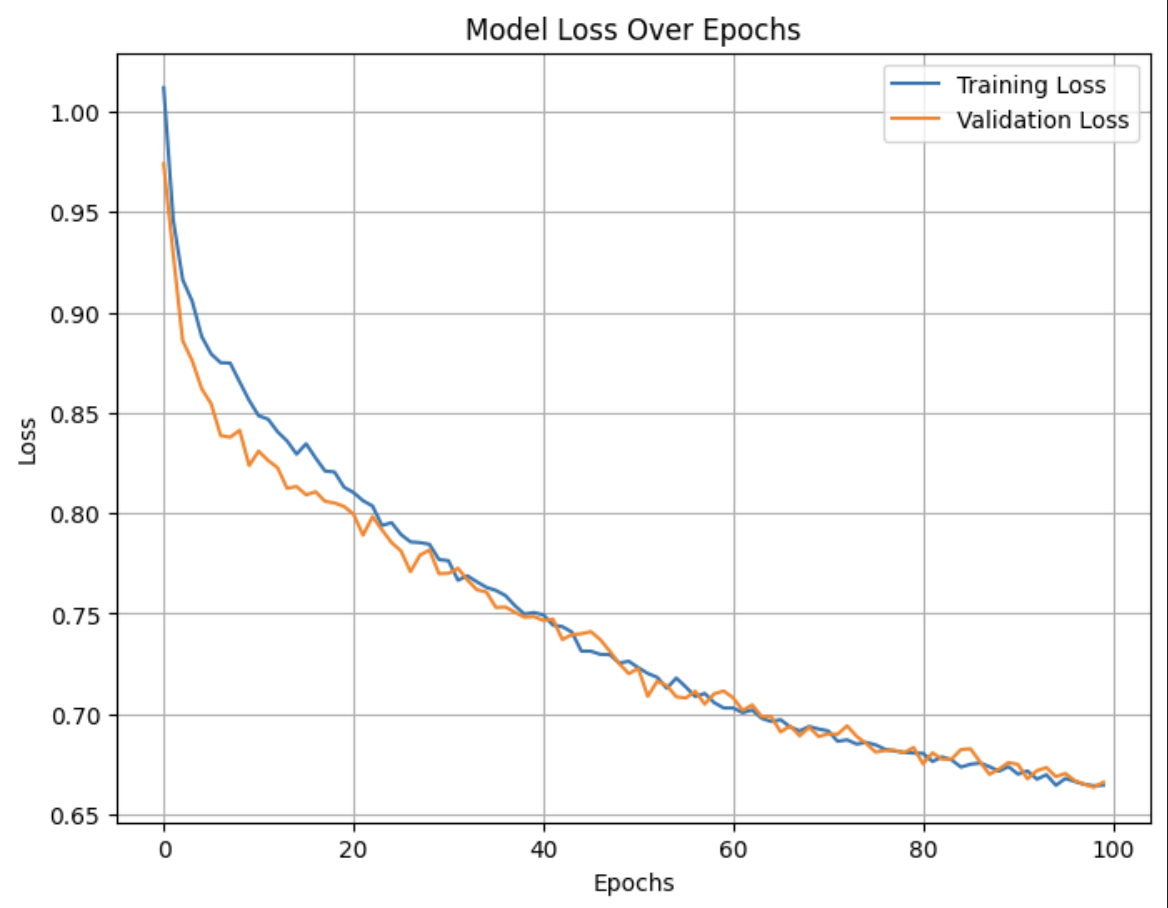
A screenshot of a computer

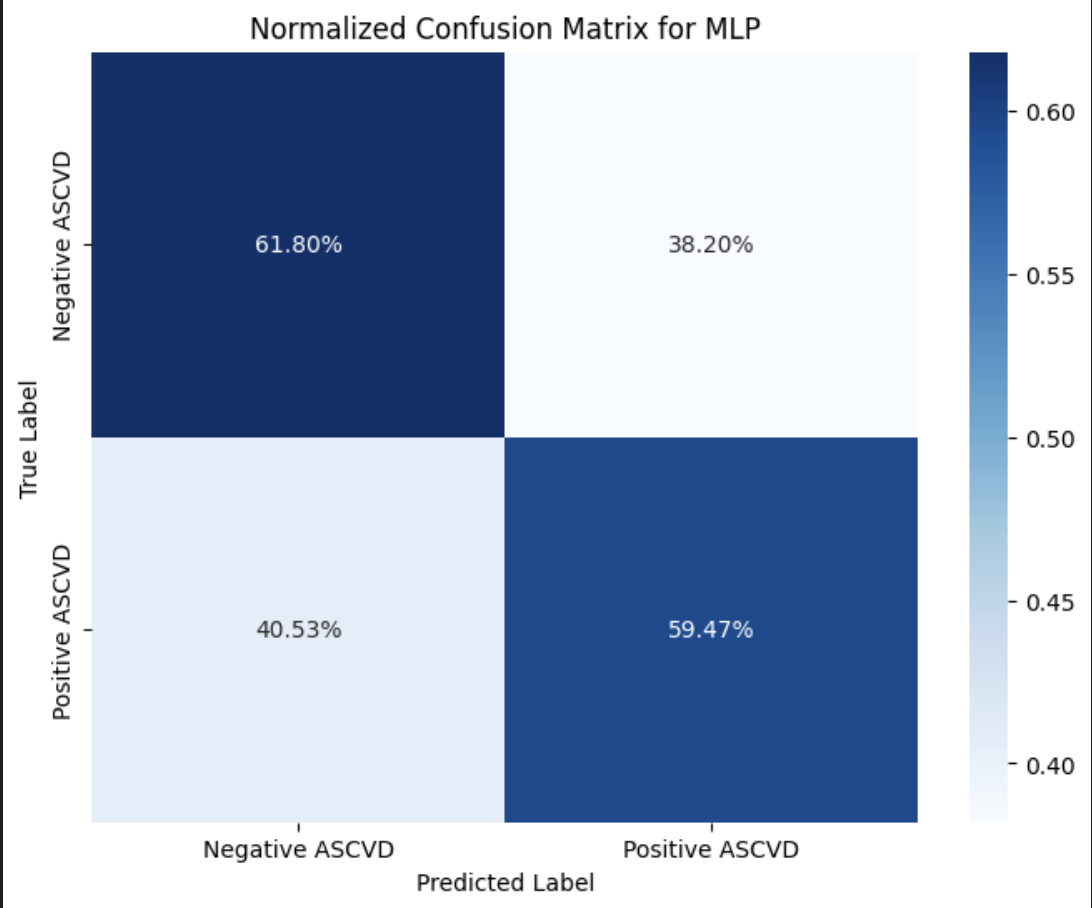
Description automatically generated

A blue squares with white text

Description automatically generated

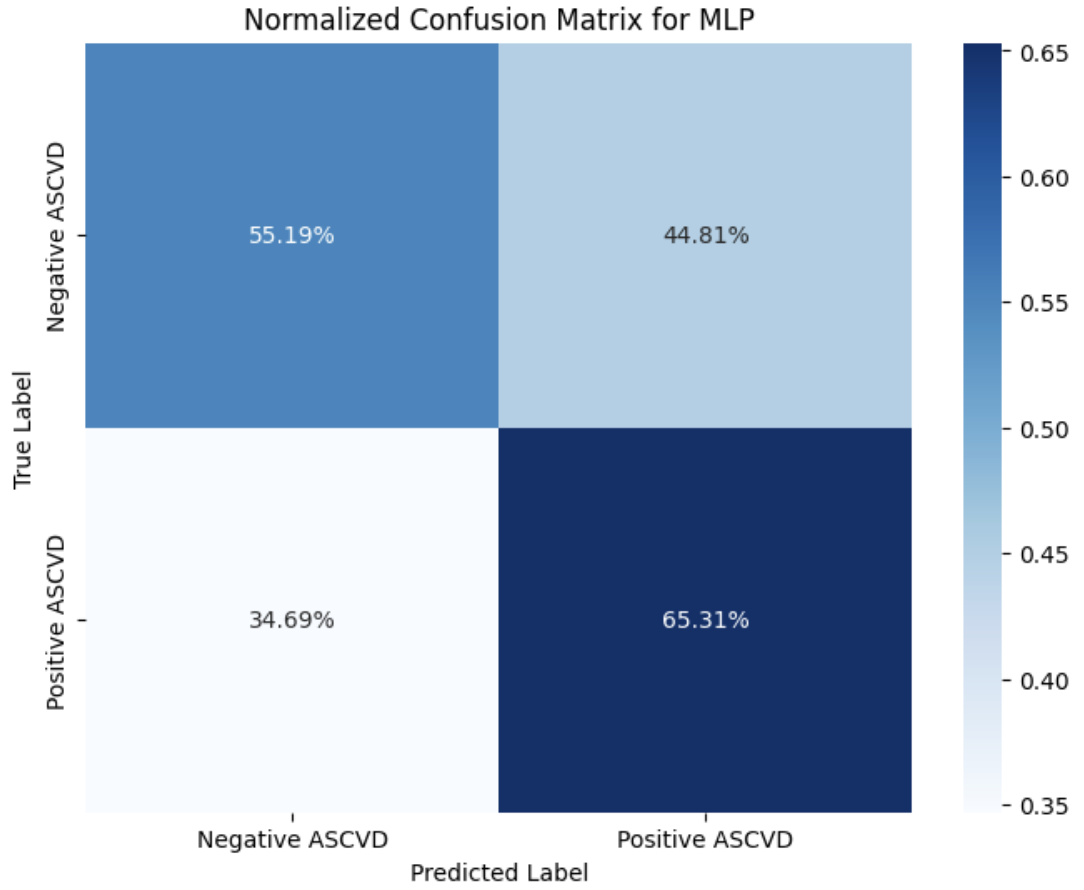
**MLP**

The results for the Multilayer Perceptron (MLP) model with 4 hidden layers featuring decreasing neuron counts (128, 64, 32, 16) using LeakyRelu, L2 regularization and batch normalization using Adam optimizer with the learning rate of 0.001 are as follows:



Due to the data not being diverse enough, the model is struggling and producing many false negatives and false positives.

The results for the model with the simpler architecture are as follows:



Similar to the previous model, there are many false predictions.

**Discussion**

The best model we tested was the random forest. It had the highest accuracy, beating out the next highest, MLP, by 9 points. However, due to the size of the data, it became computationally expensive compared to the other models. Furthermore, its success can be attributed to random forest’s ability to incorporate class weights, balancing out the imbalance between negative and positive diagnoses. Other models such as MLP and Logistic Regression would need to oversample/undersample or train with synthetic data to achieve the same step.

The random forest model without any class weights could certainly be promising as a way to predict who would not be a likely candidate for having ASCVD, which could also be helpful in medicine.

**Conclusion**

Overall, the models we trained for this project did not reach the goal of efficiently predicting when someone would be at high risk of having ASCVD based on their socioeconomic status. However, it could be promising to use this dataset in order to effectively predict who is not at risk for ASCVD based on their socioeconomic status.

Going forward, we could find better data with more balanced sampling between positive and negative diagnoses, leading to more robust models and higher accuracies across the board. Additionally, we could use better neural network models that could help better fit for these scenarios such as adjusting the loss function, modifying the output threshold, or increasing data preprocessing to add more emphasis to the minority class.

**References**

**[1]** Martin, S., Aday, A., Almarzooq, Z., & Anderson, C. (2024, January 24). *2024 heart disease and stroke statistics: A report of us ...* 2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. https://www.ahajournals.org/doi/10.1161/CIR.0000000000001209

**[2]** Dey, M. (2024, November 5). *US death statistics by death rate, age and gender, number of people shot to death, states and causes*. US Death Statistics By Death Rate, Age And Gender, Number Of People Shot To Death, States And Causes. https://www.sci-tech-today.com/stats/us-death-statistics/#Causes\_Of\_Deaths\_In\_US

**[3]** CDC, NVSS, & NCHS. (2019, April). *Racial and ethnic disparities in heart disease*. Health, United States Spotlight Racial and Ethnic Disparities in Heart Disease . https://www.cdc.gov/nchs/hus/spotlight/HeartDiseaseSpotlight\_2019\_0404.pdf

**[4]** Vankar, P. (2023, December). *Uninsured percentage and rate by ethnicity U.S. 2022*. Statista. https://www.statista.com/statistics/498567/share-and-rate-of-us-non-elderly-without-health-insurance-by-ethnicity/

**[5]** Ward, Andrew, et al. “Machine Learning and Atherosclerotic Cardiovascular Disease Risk Prediction in a Multi-Ethnic Population.” *Nature News*, Nature Publishing Group, 23 Sept. 2020, www.nature.com/articles/s41746-020-00331-1.

**[6]** Sun, Feinuo, et al. “Social Determinants, Cardiovascular Disease, and Health ...” *Social Determinants, Cardiovascular Disease, and Health Care Cost: A Nationwide Study in the United States Using Machine Learning*, Journal of the American Heart Association, 7 Mar. 2023, www.ahajournals.org/doi/full/10.1161/JAHA.122.027919.