Appendix

Mingju Kim, Ling Li, Lei Chen

December 12, 2023

1. Other classification models

1.1 Random Forest model

The Random Forest model is an ensemble learning method that constructs multiple decision trees during training and merges their predictions to improve accuracy and reduce overfitting. It operates by creating a "forest" of decision trees where each tree is trained on a random subset of the data and a random subset of features. The final prediction is made by aggregating the predictions of all individual trees, often through voting (for classification) or averaging (for regression).

Table 1: Random Forest assessment before parameter tuning

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Confusion Matrix
Model_Overall	0.6989	0.6877	0.6878	0.6877	0.6985	[[3811 1562] [1561 3439]]
Model_Female	0.7147	0.7053	0.6944	0.6998	0.7139	[[2572 935] [985 2238]]
Model_Male	0.7044	0.7098	0.6851	0.6972	0.7042	[[1326 507] [570 1240]]

As shown in Table 3, the Random Forest model achieves accuracy levels ranging from around 70% for the entire dataset and subsets. The balance between precision and recall across subsets indicates a fair balance in correctly identifying positive instances and capturing most of the actual positive instances. However, this model performs slightly better on the women's data in terms of accuracy and F1-Score compared to the entire dataset and men's data, indicating potential differences in predictive patterns among subsets. Additionally, The

ROC-AUC scores, which measure the model's ability to distinguish between classes, are around 70% for all subsets. This suggests a reasonably good ability of the model to separate classes.

Table 2: Random Forest assessment after parameter tuning

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Confusion Matrix
Model_Overall	0.7226	0.7404	0.6538	0.6944	0.7203	[[4227 1146] [1731 3269]]
Model_Female	0.7400	0.7492	0.6869	0.7167	0.7378	[[2766 741] [1009 2214]]
Model_Male	0.7376	0.7773	0.6613	0.7146	0.7371	[[1490 343] [613 1197]]

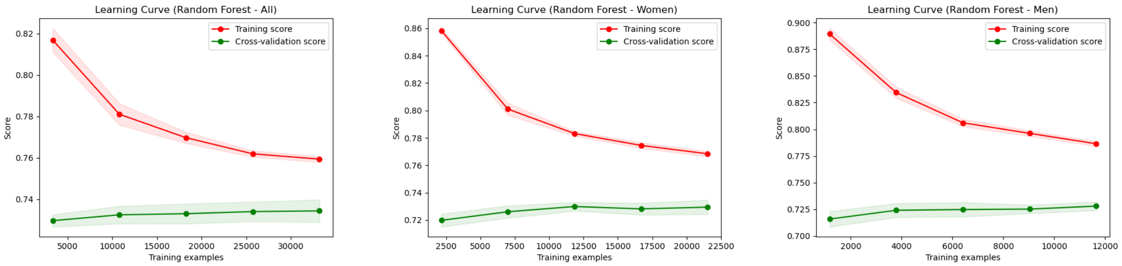

Figure 1: Learning Curve of Random Forest models

The parameter tuning led to substantial enhancements across all subsets in terms of accuracy, precision, recall, F1-Score, and ROC-AUC. This signifies a more robust and accurate Random Forest model. Like previous versions, the Random Forest model maintains consistent and improved performance metrics across all subsets without significant variations. What's more, the precision and recall values continue to be balanced, indicating a good trade-off between correctly identifying positive instances (precision) and capturing all positive instances (recall). In addition, The ROC-AUC values have increased to around 72% for all subsets, suggesting an enhanced ability of the model to differentiate between classes compared to the previous version. The learning curve also showed that trends in data captured well for both training and validation datasets with high train set scores and similar validation set scores.

In summary, these results suggest that the parameter tuning has successfully improved the Random Forest model's performance, making it more accurate, reliable, and better at distinguishing between classes across different subsets of the data.

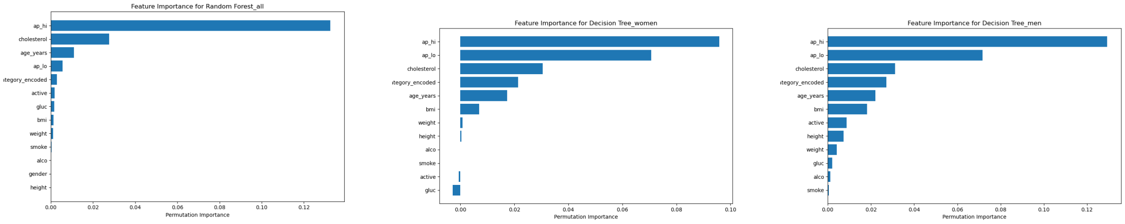

Figure 2: Feature Importance of Random Forest models

However, given the feature importance analysis results, most of the features show different contributions to heart attack prediction in the Random Forest models between different gender groups, except for the most important feature ”Systolic blood pressure”.

1.2 Decision Tree model

The model used here is a Decision Tree classifier, a popular algorithm for both classification and regression tasks. It works by recursively partitioning the data based on features to create a tree-like structure where each internal node represents a feature, each branch represents a decision based on that feature, and each leaf node represents the outcome or class label.

Table 3: Decision Tree assessment before parameter tuning

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Confusion Matrix
Model_Overall	0.6290	0.6157	0.6132	0.6144	0.6285	[[3459 1914] [1934 3066]]
Model_Female	0.6429	0.6276	0.6258	0.6267	0.6422	[[2310 1197] [1206 2017]]
Model_Male	0.6237	0.6224	0.6166	0.6195	0.6236	[[1156 677] [694 1116]]

Based on the results in Table 1, we found that this model performs reasonably well across the entire dataset and subsets (women’s and men’s data) with accuracy, precision, recall, and F1-Score hovering around 62-64%. Women’s data shows slightly better performance metrics compared to the entire dataset and men’s data, suggesting potential differences in predictive patterns between genders in the dataset. However, there’s no substantial difference in performance between the subsets. What’s more, the confusion matrices provide insight into the model’s performance, showing the distribution of correctly and incorrectly classified instances for each class within the subsets.

Table 4: Decision Tree assessment after parameter tuning

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Confusion Matrix
Model_Overall	0.7134	0.7362	0.6318	0.6800	0.7106	[[4241 1132] [1841 3159]]
Model_Female	0.7245	0.7638	0.6150	0.6813	0.7201	[[2894 613] [1241 1982]]
Model_Male	0.7115	0.7505	0.6282	0.6839	0.7110	[[1455 378] [673 1137]]

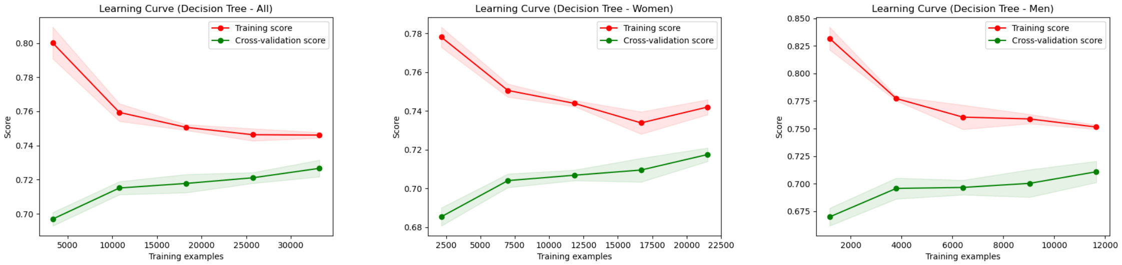

Figure 3: Learning curves of Decision Tree models

The parameter tuning resulted in significant enhancements across all subsets in terms of accuracy, precision, recall, F1-Score, and ROC-AUC. This signifies a more robust and accurate model. The improvements in performance metrics are consistent across the entire dataset, women's data, and men's data, indicating a more balanced and robust model that performs consistently across different subsets. Notably, precision has increased significantly across all subsets, indicating a reduction in false positive predictions and improved precision in correctly identifying positive instances. The confusion matrices show a reduction in misclassifications (both false positives and false negatives), indicating the model's improved ability to correctly classify instances in all subsets. While there are improvements in multiple metrics, it's important to consider trade-offs between precision and recall. Adjusting the model parameters may further balance these metrics based on specific needs. From the learning curve of the Decision Tree model in different sub-datasets, we can also find that both the training and validation scores increase and stabilize as more data is provided. According to these results, the parameter tuning has successfully improved the Decision Tree model's performance, increasing its accuracy and reliability across different data subsets

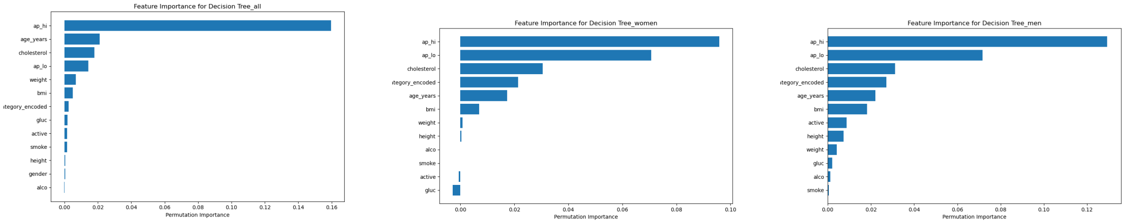

Figure 4: Feature Importance of Decision Tree models

The feature importance analysis of Decision Tree model demonstrated the same five top important features involved in the prediction of heart attack, which showed no distinct gender-specific influence between the subdatasets.

1.3 Support Vector Machines (SVM)

Support Vector Machines (SVM) are a class of supervised learning models used for classification and regression analysis. They excel by finding the optimal hyperplane that best separates different classes in the feature space. SVMs are especially effective in high-dimensional spaces and can handle both linear and non-linear relationships in data.

Table 5: SVM Performance Metrics (Before Tuning)

Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Entire Data	0.7194	0.7408	0.6426	0.6882	0.7167
Women's Data	0.7397	0.7578	0.6708	0.7117	0.7369
Men's Data	0.7343	0.7702	0.6630	0.7126	0.7338

Table 6: SVM Performance Metrics (After Tuning)

Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Entire Data	0.7198	0.7416	0.6428	0.6887	0.7172
Women's Data	0.7394	0.7567	0.6717	0.7117	0.7366
Men's Data	0.7304	0.7614	0.6663	0.7107	0.7300

In this machine learning project, the Support Vector Machine (SVM) model, before and after hyperparameter tuning, exhibited noteworthy performance. Initially, the SVM's accuracy and precision slightly surpassed the baseline logistic regression model, demonstrating its inherent effectiveness. Following hyperparameter tuning, the SVM showed incremental improvements across key metrics, highlighting the value of fine-tuning in enhancing model efficacy.

Crucially, the model's performance varied when analyzing data based on gender, with women's data consistently yielding higher accuracy and F1-scores. This pattern underscores the importance of gender as a significant variable in predicting outcomes, suggesting that gender-specific approaches may offer more precise insights in health-related predictions.

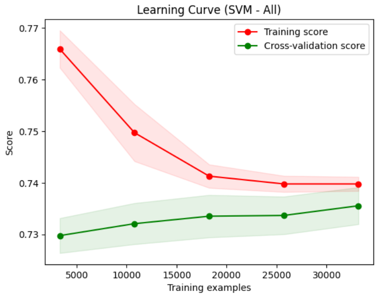

Figure 5: SVM - All

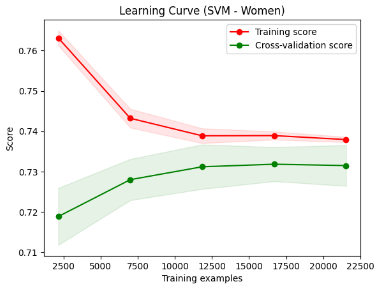

Figure 6: SVM - Women

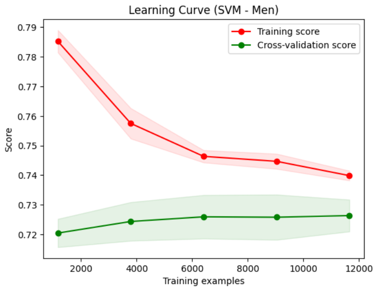

Figure 7: SVM - Men

- Learning Curve (SVM - All): The training and cross-validation scores converge with more training examples, suggesting good generalization but limited improvement with additional data. A gap persists, indicating some overfitting.

- Learning Curve (SVM - Women): Shows less variability and stable cross-validation scores, indicating robustness against variance and potentially better generalization for women's data.

- Learning Curve (SVM - Men): Begins with high training scores, which decrease sharply with more data, correcting initial overfitting. The narrowing gap between scores demonstrates effective learning but hints at performance limits.

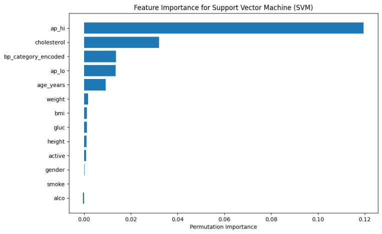

Figure 8: SVM - All

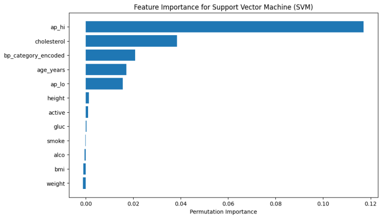

Figure 9: SVM - Women

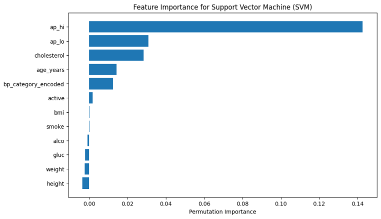

Figure 10: SVM - Men

From the feature importances of the Support Vector Machines (SVM), it is evident that 'ap_hi' (systolic blood pressure) consistently holds the highest importance across all data subsets, suggesting its significant role in predicting cardiovascular outcomes. 'cholesterol' also exhibits notable importance, emphasizing its relevance in risk assessment. Intriguingly, the primary three features impacting predictions in the entire dataset and the women dataset are identical, comprising ap_hi, cholesterol, and bp_category_encoded. Contrastingly, the top three features for the male subset include ap_hi, ap_lo (diastolic blood pressure), and cholesterol. This variation suggests gender-specific differences in the determinants of heart attack risk.

1.4 KNN (K-Nearest Neighbors)

K-Nearest Neighbors (KNN) is a simple, non-parametric algorithm used in machine learning for both classification and regression tasks. It identifies the K nearest data points in the feature space and makes predictions based on the majority vote for classification or the average for regression. While KNN is straightforward and easy to implement, it can become computationally intensive as dataset size grows.

Table 7: KNN Performance Metrics (Before Tuning)

Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Entire Data	0.6886	0.6792	0.6708	0.6750	0.6880
Women's Data	0.7031	0.6923	0.6841	0.6882	0.7024
Men's Data	0.6934	0.6963	0.6790	0.6876	0.6933

Table 8: KNN Performance Metrics (After Tuning)

Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Entire Data	0.6951	0.6886	0.6706	0.6795	0.6942
Women's Data	0.7153	0.7085	0.6891	0.6986	0.7142
Men's Data	0.7027	0.7098	0.6796	0.6943	0.7026

In this machine learning project, the K-Nearest Neighbors (KNN) model demonstrated significant changes in performance before and after hyperparameter tuning, particularly when assessed across different data subsets. While the baseline logistic regression model slightly outperformed KNN in overall metrics, the KNN model showed a notable improvement post-tuning, especially in accuracy and precision, underscoring the impact of fine-tuning on its performance.

When analyzing the KNN model's results based on gender, it was observed that the model performed better with women's data, achieving higher accuracy and F1-scores compared to men's data. This variation highlights the influence of gender on the model's predictive accuracy and suggests that a gender-specific approach could be more effective in certain applications.

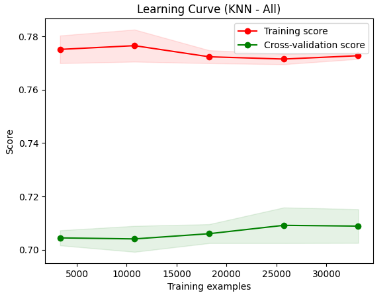

Figure 11: KNN - All

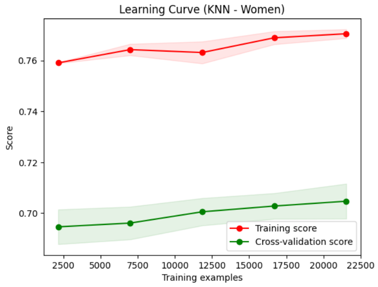

Figure 12: KNN - Women

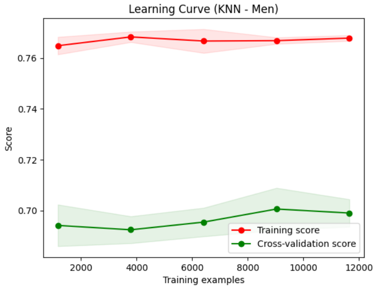

Figure 13: KNN - Men

- Learning Curve (KNN - All): Convergence of training and cross-validation scores indicates good generalization, with a persistent gap suggesting slight overfitting.

- Learning Curve (KNN - Women): Steady increase in cross-validation score with added data points to a plateau suggests robust generalization with a potential limit on performance gains.

- Learning Curve (KNN - Men): Stable scores with minimal gaps imply less initial overfitting and good generalization, yet the flat trend suggests a ceiling to learning improvement.

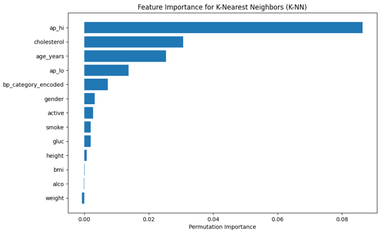

Figure 14: KNN - All

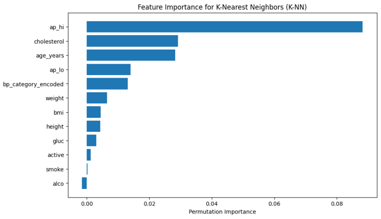

Figure 15: KNN - Women

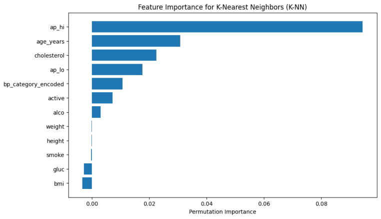

Figure 16: KNN - Men

Analysis of feature importances derived from the K-Nearest Neighbors (KNN) algorithm indicates that ap_hi (systolic blood pressure), cholesterol, age_years, ap_lo (diastolic blood pressure), and bp_category_encoded consistently rank as the top five key features across all three datasets. A gender-specific analysis reveals notable differences: for women, weight and bmi emerge as significant predictors of heart attack risk, whereas for men, 'active' and 'alco' (alcohol consumption) are more influential. This suggests the existence of gender-specific risk factors in predicting cardiovascular health.

1.5 XGBoost (eXtreme Gradient Boosting)

XGBoost (eXtreme Gradient Boosting) stands out as an advanced implementation of gradient boosting algorithms, celebrated for its speed and efficiency. It is a scalable and precise approach, frequently used for structured data challenges. XGBoost offers efficient tree boosting, capable of tackling complex problems, and is versatile in accommodating custom optimization objectives and evaluation criteria.

Table 9: XGBoost Performance Metrics (Before Tuning)

Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Entire Data	0.7196	0.7265	0.6706	0.6975	0.7179
Women's Data	0.7328	0.7368	0.6879	0.7115	0.7310
Men's Data	0.7203	0.7390	0.6757	0.7059	0.7200

Table 10: XGBoost Performance Metrics (After Tuning)

Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Entire Data	0.7240	0.7397	0.6594	0.6973	0.7218
Women's Data	0.7403	0.7504	0.6857	0.7166	0.7381
Men's Data	0.7395	0.7768	0.6674	0.7180	0.7390

For the XGBoost model, the impact of hyperparameter tuning was more modest, with only slight improvements observed in accuracy and F1-score across all data sets. Despite these marginal gains, the consistent enhancement across different data subsets suggests a uniform boost in the model's robustness. The relatively limited improvement indicates that the initial parameters of the XGBoost model were already quite effective, or that the model has approached its performance limit for the given dataset. This outcome also suggests a potential for further optimization, either through more in-depth tuning processes, exploration of advanced feature engineering techniques, or experimentation with different model configurations.

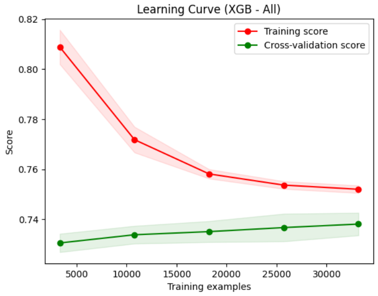

Figure 17: XGB - All

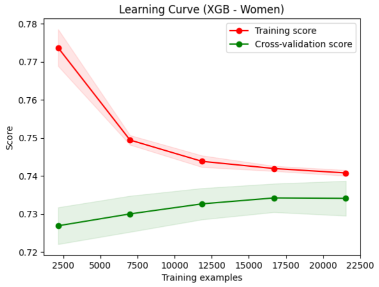

Figure 18: XGB - Women

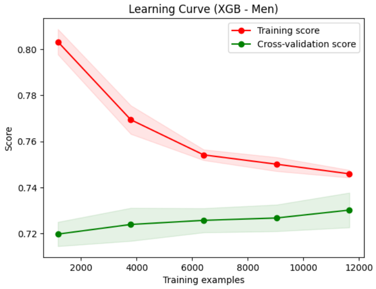

Figure 19: XGB - Men

- Learning Curve (XGB - All): The initial high performance decreases, while the cross-validation score increases with more data, suggesting an improvement in generalization and a reduction in overfitting.

- Learning Curve (XGB - Women): The training score starts higher and declines more gradually, with a steady improvement in cross-validation scores, indicating effective generalization and a smaller initial gap, implying less overfitting.

- Learning Curve (XGB - Men): The training score decreases while the cross-validation score increases, both converging at a lower score than the entire dataset, suggesting a good fit with less initial overfitting but a similar limit to performance gains.

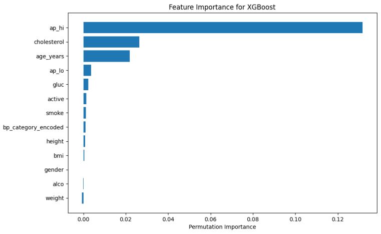

Figure 20: XGB - All

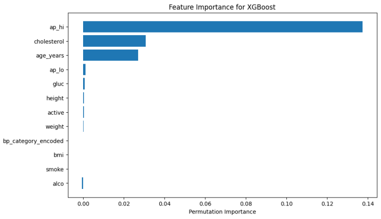

Figure 21: XGB - Women

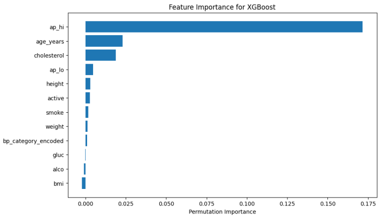

Figure 22: XGB - Men

From the feature importances of the XGBoost (eXtreme Gradient Boosting), we observe that 'ap_hi' (systolic blood pressure) consistently holds the highest importance across all three datasets, suggesting its significant role in predicting cardiovascular outcomes. 'cholesterol', 'age_years', and 'ap_lo' also exhibit notable importance in all datasets, emphasizing their relevance in risk assessment. Additionally, it's interesting to note that for women dataset, 'gluc' (glucose level) is the next important feature, which are comparatively lower feature importance in men dataset, indicating that there are certain differences in features affecting prediction of heart attack between men and women.

2. Ensemble model

Table 11: Random Forest assessment after parameter tuning

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Confusion Matrix
Model_Overall	0.7320	0.7556	0.6602	0.7047	0.7298	[[5345 1341] [2134 4146]]
Model_Female	0.7342	0.7467	0.6697	0.7061	0.7314	[[3491 911] [1325 2686]]
Model_Male	0.7332	0.7510	0.6701	0.7083	0.7312	[[[1864 489] [726 1475]]

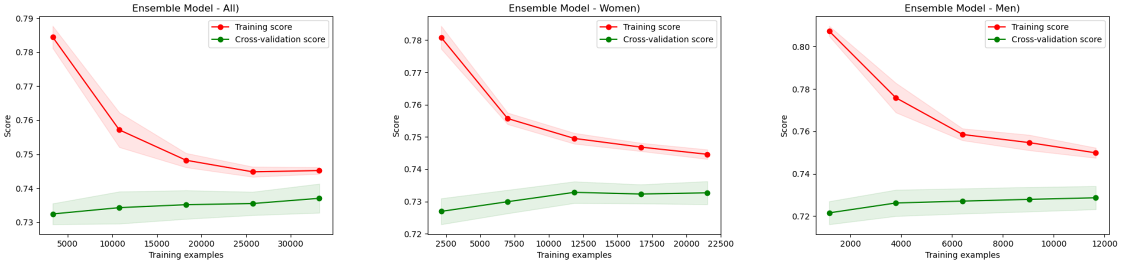

Figure 23: Learning Curve of Ensemble Model

As a result of combining Random Forest, SVM, XGBoost, and Logistic models in the Ensemble model, the overall accuracy across all datasets, as well as both gender-based subsets, is relatively consistent, hovering around 73%, although it is slightly higher than using all four models separately. The recall scores, measuring the model's ability to correctly identify positives, are also consistent and reasonable, around 66% to 67%. The model captures a significant portion of actual positives for both genders. As shown by the learning curve, this is a proper model that is neither overfitting nor underfitting.

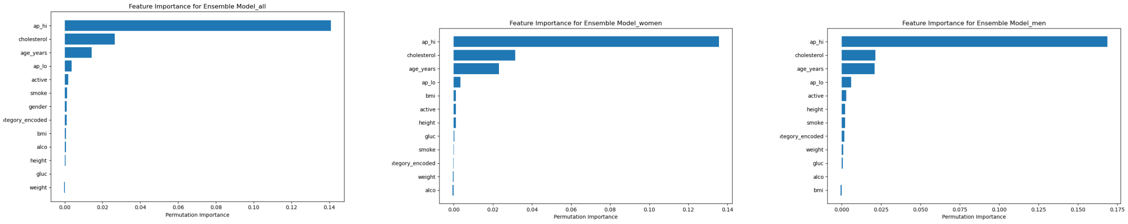

Figure 24: Feature Importance of Ensemble Model

The feature importance analysis showed that the top three important features contributing to heart attack prediction in the Ensemble Model are consistent between entire and both gender subgroups.

Overall, the ensemble model demonstrates a balanced and consistent performance across both genders, indicating its potential usefulness for prediction tasks on this dataset.

3. Deep learning models

The Perceptron is a simple neural network for binary classification, adjusting weights to learn patterns from data but limited to linearly separable problems. Meanwhile, the Multi-Layer Perceptron (MLP) is a more complex neural network with multiple layers, capable of handling nonlinearity and intricate patterns using activation functions and backpropagation. MLPs excel at learning complex data relationships, making them suitable for diverse tasks like classification and pattern recognition.

Table 12: Perceptron assessment

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Confusion Matrix
Model_Overall	0.6950	0.6866	0.6756	0.6810	0.6943	[[3831 1542] [1622 3378]]
Model_Female	0.5808	0.5354	0.9442	0.6833	0.5955	[[866 2641] [180 3043]]
Model_Male	0.6830	0.6540	0.7685	0.7066	0.6835	[[1097 736] [419 1391]]

Table 13: MLP (multi-layered perceptrons) assessment

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Confusion Matrix
Model_Overall	0.6951	0.6916	0.6630	0.6770	0.6940	[[3895 1478] [1685 3315]]
Model_Female	0.7083	0.7106	0.6596	0.6842	0.7063	[[2641 866] [1097 2126]]
Model_Male	0.6783	0.6700	0.6945	0.6820	0.6784	[[1214 619] [553 1257]]

Comparing the performance of the Perceptron and Multi-Layer Perceptron (MLP) models reveals nuanced differences across the entire dataset and gender-specific subsets. The Perceptron achieved an accuracy of 69.50% overall, demonstrating varied precision, recall, and F1-Score between women's (58.08%) and men's (68.30%) data. In contrast, the MLP showcased slightly improved performance with an accuracy of 69.51% across the entire dataset, showing enhanced results particularly in women's data (70.83%) compared to men's (67.83%). These models' assessments highlight distinctions in their predictive capabilities concerning gender subsets, with the MLP generally displaying more consistent and balanced performance across metrics.

Now analyzing the findings, it becomes evident that the MLP generally outperformed the Perceptron in most evaluation metrics across the entire dataset and gender-segregated subsets. The MLP consistently exhibited slightly higher accuracy, precision, and F1-Score compared to the Perceptron. Notably, the Perceptron demonstrated notably higher recall in predicting women's outcomes but suffered from substantially lower precision, leading to a lower F1-Score. On the other hand, the MLP maintained a better balance between precision and recall for women's data. Regarding men's data, both models showcased relatively similar performance in accuracy and F1-Score, but the Perceptron demonstrated notably higher recall compared to the MLP. These differences suggest that while the Perceptron excelled in certain recall-oriented scenarios, the MLP demonstrated more balanced and robust performance across various metrics, indicating its superiority in most classification scenarios.

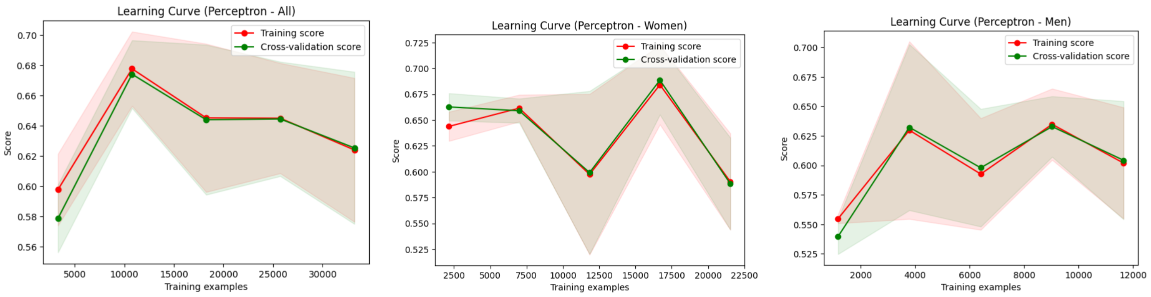

Figure 25: Learning curves for the Perceptron model

For the learning curve in the Perceptron model for all genders, the learning curve initially shows improvement in both training and cross-validation scores, suggesting effective learning. However, a subsequent decline in both scores indicates overfitting as the model's complexity increases. The overlapping curves highlight a phase of balanced performance before overfitting occurs, emphasizing the need for strategies like regularization or early stopping to maintain optimal model performance while preventing overfitting. However, in the separate models for man and women, the overlapping training and cross-validation score curves repeatedly fluctuate, indicating instability. These oscillations suggest challenges in finding a stable model performance due to sensitivity to dataset changes or noise. Adjusting model architecture or regularization could help achieve a more consistent and robust performance.

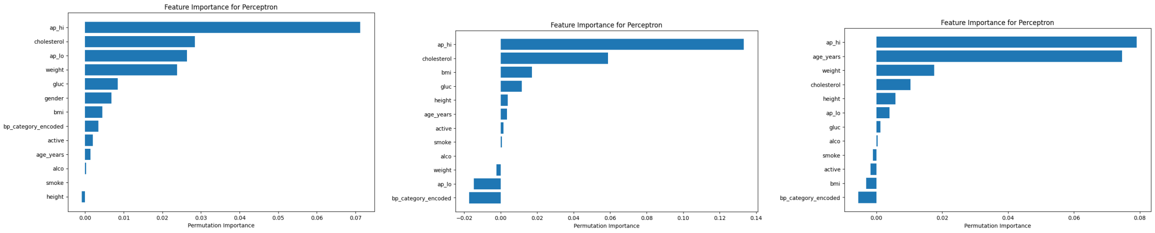

Figure 26: Feature importance in the Perceptron model

The perceptron model analysis reveals distinct key predictors of cardiovascular disease across genders. Systolic blood pressure and cholesterol level (cholesterol) emerge consistently significant for all genders, signifying their universal importance. For women, body mass index and surprisingly, height, join the core predictors, suggesting unique gender-specific influences. In contrast, among men, age becomes a notable factor alongside weight (weight) and height, highlighting the potential impact of age on cardiovascular risk for men. While certain factors like blood pressure and cholesterol maintain their relevance, variations in the significance of BMI, age, and height in gender-specific models underscore potential nuanced associations influencing cardiovascular disease predictions based on gender-specific data subsets.

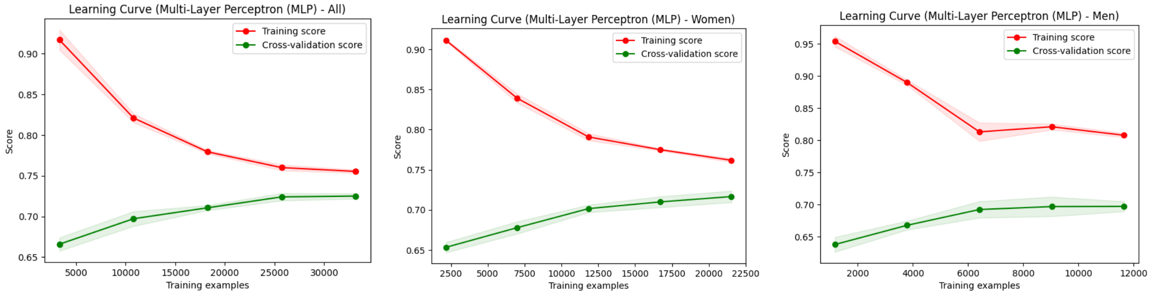

Figure 27: Explained variance by principal component

The scenario where the training score decreases while the cross-validation score curve increases but fails to converge signifies a unique situation in the learning curve analysis. The decreasing training score suggests the model's struggle with the training data, possibly due to overfitting or inadequate learning. Simultaneously, the increasing cross-validation score indicates the model's better generalization to unseen data despite the training score decline. The lack of convergence between the two curves indicates ongoing challenges in finding a balance between model complexity and generalization. This scenario suggests that the model could benefit from adjustments to mitigate overfitting tendencies, such as regularization techniques or reducing model complexity, to achieve improved generalization without compromising its learning capability.

The analysis from the multiple layer perceptron (MLP) model unveils distinctive insights into the influential factors determining the likelihood of cardiovascular disease across different genders. For the combined gender model, systolic blood pressure, cholesterol level (cholesterol), age (age), diastolic blood pressure, and physical activity (active) emerge as the top five predictors. This underscores the critical role of blood pressure, cholesterol, and age in predicting cardiovascular risk among diverse populations. In the MLP model exclusive to women, the factors of systolic and diastolic blood pressure, cholesterol level, age, and physical activity maintain consistent importance, reinforcing the impact of these factors on cardiovascular health in female-specific datasets. In contrast, the model tailored specifically to men highlights systolic and diastolic blood pressure, cholesterol level, age, and physical activity as the primary predictive factors, underlining the commonalities between men and the overall model while emphasizing the distinct emphasis on blood pressure metrics in predicting cardiovascular risk in male-centric datasets. This analysis signifies the shared significance of blood pressure, cholesterol, age, and physical activity in predicting cardiovascular disease across genders, while also showcasing nuanced differences that might exist in gender-specific models, offering valuable insights for targeted healthcare interventions and risk assessments based on gender-based data subsets.

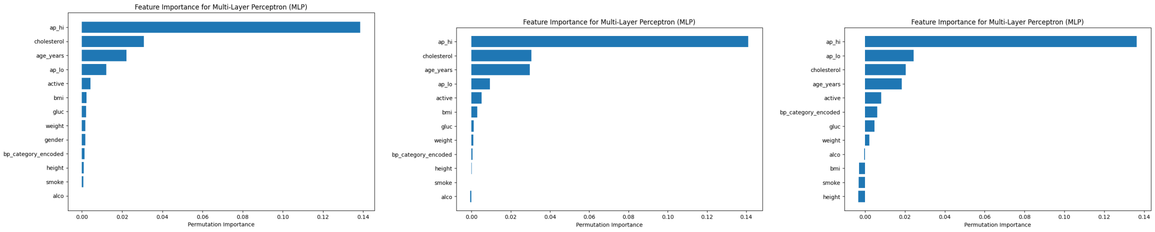

Figure 28: Explained variance by principal component

The distinction between the perceptron model and the multiple layer perceptron (MLP) model in predicting cardiovascular disease lies in the nuanced prioritization of influential factors across gender-specific datasets. The perceptron model highlighted consistent factors like systolic blood pressure, cholesterol level (cholesterol), and other health indicators such as weight, BMI, and glucose levels as significant predictors across genders, albeit with varying importance. In contrast, the MLP model emphasizes blood pressure metrics, cholesterol level, age (age), and physical activity (active) as the primary predictors for both combined genders and gender-specific datasets. Notably, the MLP model tends to place more emphasis on blood pressure variables for men, showcasing a clearer focus on these metrics as crucial predictors in male-centric datasets while maintaining a consistent emphasis on cholesterol, age, and physical activity across all datasets. This nuanced distinction suggests that while some core factors remain consistent between the models, the MLP model offers a refined understanding by highlighting blood pressure metrics as pivotal predictors, especially in gender-specific subsets, potentially enabling more precise risk assessment and targeted interventions for cardiovascular disease based on gender-specific health indicators.