WILFRID LAURIER UNIVERSITY

LAURIER

Inspiring Lives.

Machine Learning: Heart Disease Risk Prediction

CP322 - Department of Physics and Computer Science

Group 9

December 4th, 2024

Dr. Yang Liu

Member 1 Name: Marc Niven Kumar	**ID:** 000006972
Member 2 Name: Charnel Dolon	**ID:** 212207670
Member 3 Name: Israt Erin Urbi	**ID:** 211836570
Member 4 Name: Brandon Pham	**ID:** 200973390
Member 5 Name: Simran Badwal	**ID:** 169033506

INTRODUCTION

Heart disease remains the leading cause of mortality globally, accounting for millions of deaths each year. Early detection and accurate prediction of heart disease risk are crucial to preventing fatal outcomes and improving patient care. However, the complexity of health-related data and the interaction between various risk factors make accurate prediction a challenging task.

In this project, we aim to leverage machine learning techniques to predict heart disease risk using a dataset of health indicators. By analyzing and interpreting patterns in the data, we strive to identify the most influential risk factors and create models that assist healthcare professionals in making data-driven decisions.

Context

Heart disease represents a significant public health challenge in Canada, with approximately 1 in 12 adults aged 20 and older—equivalent to 2.6 million people—living with a diagnosed condition. The impact of this disease is not only widespread but also severe; every hour, an estimated 14 Canadian adults with diagnosed heart disease succumb to the condition. These statistics highlight the urgent need for more effective strategies in prevention, diagnosis, and management.

Despite advancements in medical care, the burden of heart disease continues to grow, driven by factors such as aging populations, lifestyle changes in terms of food consumption, and the increasing prevalence of underlying risk factors like obesity, diabetes, and high blood pressure. This underscores the importance of early detection and intervention to reduce morbidity and mortality rates associated with heart disease.

Motivation

Our motivation lies in identifying the key health metrics associated with heart disease and determining the most accurate machine learning models for prediction, ultimately aiming to improve health outcomes and save lives. By analyzing patterns in the data, we aim to optimize prediction accuracy using four distinct machine learning models: K-Nearest Neighbors (KNN), Naive Bayes (NB), Ridge Regression (RIDGE), and Decision Trees (DT). Each of these models offers unique strengths—KNN excels at capturing local data relationships, NB provides probabilistic insights, RIDGE effectively handles multicollinearity, and DT offers interpretability. By evaluating and comparing these models, the project seeks to not only determine the most effective approach but also to support healthcare professionals in integrating predictive analytics into their workflow, ultimately advancing patient care and outcomes.

Related Works

For our project, related works were analyzed as a foundation to be used in order to determine which machine learning model is the best at predicting heart disease risk, and to identify the most important health indicators in regards to heart disease. In the first related work we analyzed, it discusses the use of k-nearest neighbors, decision trees, and random forests machine learning models to make heart disease predictions (Ali et al., 2021). It concluded that simple machine learning algorithms are able to make highly accurate predictions. In the second related work analyzed, the algorithms of support vector

machines, XGBoosts, bagging, decision trees and random forests were evaluated (El-Sofany et al., 2024). This related work identified key health indicators, allowing for an improvement in the model's performance. In the last related work we analyzed, logistic regression was found as the machine learning model that was able to identify and assess the risk factors for heart disease at a high efficiency (Zhang et al., 2024).

The baseline methods differ from our method of choice because our approach focuses on integrating and comparing a diverse range of machine learning models while emphasizing the optimization of their parameters to achieve the highest predictive accuracy. Unlike the works by Ali et al. (2021), which primarily demonstrated the effectiveness of simpler algorithms like k-nearest neighbors and decision trees, or El-Sofany et al. (2024), which concentrated on ensemble methods such as XGBoost and bagging, our project seeks to unify the strengths of these approaches by benchmarking multiple models under consistent evaluation metrics.

MODELS IMPLEMENTED

Decision Tree

A decision tree is easy to implement and interpret machine learning models used for both classification and regression taks. It splits data into subsets based on feature values, making decisions based on a series of rules that resemble a flow chart or tree structure. Here are some reasons why the decision tree is beneficial for our project.

- Interpretability: The model is easy to interpret and visualize, making it a good choice when understanding the logic behind predictions is important.
- Non-linear Relationships: Unlike linear models, decision trees can model complex, non-linear relationships between features and target variables.
- Handling Mixed Data Types: The dataset includes both numerical (e.g BMI, Age) and categorical (e.g. HighBP, GenHlth) variables. Decision trees handle mixed data types seamlessly, making it ideal for this project
- Feature Importance: Decision trees inherently rank features by their importance during the model training process. This ranking helps identify the most critical health indicators, which aligns with our goal of identifying key contributors to heart disease

For this project, the decision tree classifier was implemented using the ID3 algorithm with the entropy criterion. This ensures that splits in the tree are made based on the maximum reduction in entropy, leading to more informative splits.

As a result, the tree structure helped us understand how combinations of features influence the likelihood of heart disease, and by focusing on features with the highest information gain, the decision tree improves accuracy of predictions compared to random guessing or models trained on irrelevant features. These insights further support the robustness of our feature engineering process.

The implementation of a decision tree not only provides a solid foundation for heart disease risk prediction but also bridges the gap between complex machine learning algorithms and human interpretability. Its ability to highlight key health indicators and deliver accurate predictions ensure that it is both an important and helpful component to this project.

Naïve Bayes

Naïve Bayes is a probabilistic machine learning algorithm that assumes no dependency between features, based on Bayes' theorem, particularly effective for classification tasks**.** The algorithm computes the posterior probability of each class given the input features. In the context of heart disease prediction, where the target variable is binary, the algorithm computes probabilities for both classes and assigns the label with the highest probability.

Naïve Bayes was selected for this project due to its computational efficiency and simplicity, which allowed it to be scaled to handle the large dataset used in this project with speed. Its reliance on probabilistic principles also makes it well suited for categorical features, such as physical activity levels.

The Naïve Bayes algorithm has also been applied in several studies focusing on heart disease risk prediction with reasonable success, where cardiologists claim that “the result can support medical analysis related to cardiovascular disease”(Miranda et al., 2016), which gave us the confidence to apply this model in our project.

Logistic Regression

Logistic Regression is the third model we have implemented. This model was chosen mainly because

- Logistic regression is suited for binary classification tasks like this one, where the goal is to predict the presence or absence of heart disease.
- Its probabilistic framework allows us to interpret results very easily and is efficient for large datasets such as this one.
- RidgeClassifier extends logistic regression with L2 regularization, making it a great choice for this dataset with multiple potentially correlated features.
- L2 regularization minimizes the sum of squared coefficients. It prevents overfitting by shrinking coefficients closer to zero without completely eliminating any. This is especially useful in datasets with many features, as it retains all features but reduces their impact proportionally.
- Compared to L1 regularization (Lasso), which enforces sparsity by setting some coefficients to zero, Ridge was preferred here to avoid losing any potentially relevant features in predicting heart disease.

K-Nearest Neighbor (KNN)

KNN is a simple and effective machine learning algorithm suitable for datasets with both numerical and categorical features. It relies on calculating distances between data points, which makes it intuitive and interpretable. Specifically for this dataset:

- **Non-parametric Nature**: KNN does not make assumptions about the data distribution, making it suitable for datasets with diverse patterns and relationships.
- **Ease of Implementation**: It is straightforward to implement and requires minimal tuning compared to other algorithms.
- **Mixed Features**: After preprocessing (scaling), KNN can effectively handle datasets with mixed feature types (e.g., BMI, blood pressure).
- **Balanced Classes**: After downsampling, the dataset is balanced, making KNN a great choice for binary classification.
- **Scaling Importance**: KNN works well when the data is scaled, as it reduces the impact of features with larger ranges (e.g., BMI vs. binary indicators).

Feature Engineering

Why Feature Engineering?

Heart disease prediction datasets often include numerous variables, some of which may not significantly contribute to prediction. Incorporating irrelevant or redundant features can lead to overfitting and reduced model performance. By systematically selecting the most informative features, we can:

- Improve the generalizability of the model.
- Reduce training and testing times.
- Provide insights into the most critical health indicators for heart disease risk.

The ID3 Algorithm

To enhance the predictive power of our model, we employ the ID3 (Iterative Dichotomiser 3) algorithm for optimal feature selection. The ID3 algorithm utilizes an entropy-based approach to determine the most informative features, ensuring that the machine learning models are trained on the most relevant subset of the data. By reducing redundancy and focusing on key indicators, the ID3 algorithm improves both model interpretability and efficiency.

Advantages of Using ID3 for Feature Selection

1. **Focus on Key Indicators:** The entropy-based approach ensures that only features with the highest impact on reducing uncertainty are chosen.
2. **Improved Model Performance:** By removing irrelevant features, the risk of overfitting is minimized, leading to a more robust model.
3. **Interpretable Results:** The selected features provide insights into the most important health indicators, aiding in understanding the underlying factors contributing to heart disease risk.

Feature Selection Results

Using the ID3 algorithm, we identified the following features as the most critical indicators for predicting heart disease:

1. **General Health Status (GenHlth):** A self-reported measure of overall health, indicative of lifestyle and underlying health conditions.
2. **High Blood Pressure (HighBP):** A well-established contributor to heart disease.
3. **Age:** A primary risk factor for heart disease, with older individuals at higher risk.

By focusing on these features, our machine learning models achieved higher interpretability and efficiency, enabling more accurate predictions.

Implementation Tools

- **Pandas**: For data manipulation and preprocessing.
- **Scikit-learn**: For machine learning model implementations, scaling, and evaluation metrics.
- **Matplotlib & Seaborn**: For data visualization, including feature importance and performance metrics like ROC curves.
- **NumPy**: For numerical operations and array manipulation.

Dataset Description

The dataset consisted of 22 features, including:

- **Health Indicators**: Blood pressure (HighBP), cholesterol (HighChol), BMI, smoking status (Smoker), physical activity (PhysActivity), and general health (GenHlth).
- **Target Variable**: HeartDiseaseorAttack (binary: 1 for presence, 0 for absence of heart disease).

Data Preprocessing

Handling Missing Values:

- Checked for missing values using NaN and found none.

Resampling to Address Imbalance:

- The dataset shows an imbalance between the target classes, with significantly more samples in the non-heart disease category. Downsampling the majority class to match the minority class resulted in a balanced dataset, ensuring unbiased model performance.

- **Impact of Resampling on Accuracy**: Resampling helped the models learn patterns in both classes, improving their ability to correctly classify heart disease cases, as reflected in balanced evaluation metrics like precision and recall.

Train-Test Split:

- The dataset was split into training (50%) and testing (50%) sets, ensuring the model was evaluated on 1:1 data to simulate real-world performance.

MODEL EVALUATION

Evaluation Metrics

For all the models these evaluation techniques were used: Accuracy, Precision, Recall, F1-Score, ROC-AUC.

Decision Tree

The Decision Tree model yielded the following performance metrics on the test dataset:

Accuracy: 0.79
Precision: 0.76
Recall: 0.83
F1-Score: 0.80
ROC-AUC: 0.79

The Decision Tree model performed well, with an accuracy of 79%, meaning 79% of its predictions were correct. The precision of 76% indicates that, when the model predicted the presence of heart disease, it was correct 76% of the time. This shows that the model is fairly reliable when diagnosing heart disease.

The recall of 83% is particularly noteworthy, as it indicates the model successfully identified 83% of actual heart disease cases. This high recall suggests that the model is effective at flagging those who may have heart disease, which is crucial for medical interventions.

The F1-Score of 0.80 strikes a good balance between precision and recall, making the model reliable in terms of both identifying true cases and avoiding false positives. The ROC-AUC score of 0.79 further supports the model's ability to distinguish between heart disease and non-heart disease cases, indicating solid performance overall.

In conclusion, the Decision Tree model is a strong candidate for predicting heart disease, offering a good mix of accuracy, precision, recall, and AUC. Its ability to identify most true positives (heart disease cases) with minimal false positives makes it suitable for use in medical settings where early detection is crucial.

Naïve Bayes

The Naïve Bayes model yielded the following performance metrics on the test dataset:

Accuracy: 0.73
Precision: 0.75
Recall: 0.70
F1-Score: 0.72
ROC-AUC: 0.81

In terms of accuracy and precision, we observed decent results, where 73% of the predictions were correct and 75% of the cases it predicted as heart disease were correctly identified. This shows that the model does reasonably well in distinguishing between heart disease and non-heart disease cases, and when the model predicts the presence of heart disease, it is likely to be correct.

The ROC-AUC score of 0.81 indicates that the Naïve Bayes model has a good ability to distinguish between heart disease and non-heart disease cases, and is capable of making reliable predictions. The F1-Score of 0.72 indicates a decent ability to maintain both precision and recall. However, while both the F1-Score and ROC-AUC are solid, the priority in this scenario is to improve recall.

When it comes to recall, it correctly identified 70% of the true heart disease cases. This result is not ideal, especially in the context of heart disease, as the model misses around 30% of the actual heart disease cases. This means that some patients who are at risk of heart disease might not be flagged for further medical intervention, potentially leading to adverse outcomes.

Thus, while the model generally produced decent results, it is not yet suitable for critical medical applications like heart disease prediction where high recall is important. This could be because of the model's assumption of feature independence, which may not hold true in this dataset. At first glance, the features in the dataset may appear independent. However, in the medical field, health indicators rarely function in isolation. For example, physical activity levels can influence body mass index, which in turn can affect blood pressure. These interdependencies affect the Naïve Bayes model performance.

Logistic regression

Evaluation Metrics

- **Accuracy**: 77.24%
    - This indicates that approximately 77% of the predictions made by the model are correct. While not perfect, it demonstrates a decent ability to differentiate between heart disease and non-heart disease cases.
- **Precision**: 76.10%
    - Precision measures the proportion of predicted positive cases (heart disease) that are actual positives. A precision of 76.10% suggests that the model is relatively reliable in its positive predictions, but some false positives exist.
- **Recall**: 79.58%
    - Recall indicates the proportion of actual positive cases (heart disease) that the model correctly identifies. A recall of 79.58% suggests that the model captures most of the positive cases with a 20% chance of a false positive.

- **F1-Score**: 77.80%
    - The F1-Score balances precision and recall, providing an overall measure of the model's classification performance. A value of 77.80% reflects a balance between false positives and false negatives.
- **ROC-AUC**: 81.41%
    - The ROC-AUC score measures the model's ability to distinguish between classes. A score of 81.41% indicates good overall performance, as the model effectively ranks positive cases higher than negative ones most of the time.

ROC Curve Analysis

The ROC curve demonstrates the trade-off between true positive and false positive rates at different threshold levels. The Area Under the Curve (AUC) of 0.8141 highlights the model's ability to distinguish between heart disease and non-heart disease cases effectively.

Confusion Matrix

- **True Positives (TP)**: 11431 cases where the model correctly predicted heart disease.
- **True Negatives (TN)**: 10711 cases where the model correctly predicted no heart disease.
- **False Positives (FP)**: 3590 cases where the model predicted heart disease incorrectly.
- **False Negatives (FN)**: 2934 cases where the model missed actual heart disease cases.

Performance

- **Strengths**:
    - The ROC-AUC score demonstrates strong discrimination ability.
    - Balanced precision and recall indicate reliable predictions without overly favoring either positives or negatives.
- **Weaknesses**:
    - The false negative rate (2934 missed cases) is notable and could lead to missed diagnoses of heart disease, which is critical in healthcare scenarios.
    - The accuracy of 77.24% leaves room for improvement, particularly in minimizing false negatives and false positives.

The model demonstrates good overall performance, as indicated by its ROC-AUC of 81.41%. However, further tuning, additional data, or more advanced models could be explored to improve its recall and reduce false negatives, ensuring that fewer cases of heart disease are missed.

K-Nearest Neighbor (KNN)

Evaluation Metrics

- Accuracy: 77.15%
    - The accuracy indicates that approximately 77% of predictions made by the model are correct.
- Precision: 74.50%

    - Precision measures the proportion of positive predictions (heart disease cases) that are actual positives. A value of 74.50% suggests good reliability, although some false positives still exist.
- Recall: 82.71%
    - Recall (sensitivity) indicates the proportion of actual positive cases that the model successfully identifies. A recall of 82.71% reflects the model's ability to capture the majority of heart disease cases.
- F1-Score: 78.39%
    - A value of 78.39% highlights that the model maintains a good balance between false positives and false negatives.
- ROC-AUC: 83.73%
    - The ROC-AUC measures the model's ability to distinguish between the two classes (heart disease and non-heart disease). With a value of 83.73%, the model demonstrates good discriminatory power.

ROC Curve Analysis

The ROC curve shows the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) at various thresholds. An AUC of 0.8373 indicates good model performance, with the curve staying significantly above the diagonal baseline.

Confusion Matrix

- **True Positives (TP)**: 11881 cases correctly predicted as heart disease.
- **True Negatives (TN)**: 1e+04 cases correctly predicted as no heart disease.
- **False Positives (FP)**: 4067 cases incorrectly predicted as heart disease.
- **False Negatives (FN)**: 2484 cases missed as heart disease.

The confusion matrix highlights a notable reduction in false negatives compared to the earlier model, which is critical for healthcare applications to minimize missed diagnoses.

Performance

- **Strengths**:
    - **ROC-AUC** confirms the model's ability to distinguish between the two classes effectively.
- **Weaknesses**:
    - While precision (74.50%) is relatively good, the number of false positives (4067) remains an area for improvement. Reducing false positives would help avoid unnecessary healthcare interventions.

The high ROC-AUC and F1-score increases the model's reliability. However, future efforts should aim to reduce false positives while maintaining or improving recall for better diagnostic performance.

CONCLUSION

This project used a variety of machine learning models in order to predict the risk of heart disease through the use of various health indicators. Through feature engineering, the key health indicators in predicting heart disease were identified to be general health, high blood pressure, and age. The models of decision trees, naive bayes, regression and k-nearest neighbors were implemented and evaluated, with the various similarities and differences being compared with each other.

Our technique effectively tackles the problem of predicting heart disease risk through a variety of health indicators through use of data preprocessing, feature engineering, and the use of machine learning models. Through the use of resampling, we were able to handle class imbalances that reduced the bias towards the majority class of people without heart disease, which improved metrics such as precision and recall.

In order to improve upon this project, future research should be focused on exploring different machine learning models, such as deep learning models, and to use larger, more extensive datasets to increase the accuracy of the model.

References

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in biology and medicine*, 136, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

Canada, P. H. A. of. (2022). Government of Canada. Retrieved from https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html

El-Sofany, H., Bouallegue, B. & El-Latif, Y.M.A. A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. *Sci Rep* **14**, 23277 (2024). https://doi.org/10.1038/s41598-024-74656-2

Miranda, E., Irwansyah, E., Amelga, A. Y., Maribondang, M. M., & Salim, M. (2016, July). *Detection of cardiovascular disease risk's level for adults using naive Bayes classifier*. Healthcare informatics research. https://pmc.ncbi.nlm.nih.gov/articles/PMC4981580/#B20

Zhang, M., Wang, H., & Zhao, J. (2024). Use machine learning models to identify and assess risk factors for coronary artery disease. *PloS one, 19*(9), e0307952. https://doi.org/10.1371/journal.pone.0307952