HEART DISEASE PREDICTION

HAN University of Applied Sciences

DECEMBER 19, 2023

PJOTR OTTEN

TABLE OF CONTENTS

ABSTRACT

This paper explores predictive modeling for heart disease utilizing the CRISP-DM framework and a diverse set of machine learning algorithms. Incorporating the Cleveland Heart Disease Database and four additional datasets, the study aims to enhance accuracy in cardiovascular disease prediction. Among the evaluated models, the Artificial Neural Network (ANN) emerges as the frontrunner, achieving an outstanding accuracy averaging around 93 to 94%. This high accuracy positions the ANN model for effective integration into real-world healthcare systems, contributing to proactive interventions and improved patient outcomes. Future work can focus on refining hyperparameters, enhancing model explainability, and leveraging longitudinal data for dynamic insights into cardiovascular risk changes over time.

For the exact code that has been constructed for this paper, please refer to Appendix 3.

1. INTRODUCTION

The systematic study of extensive datasets for hidden patterns and insights is known as data mining. It comprises obtaining and preprocessing data, analyzing its characteristics, and building models with machine learning algorithms. The goal is to extract relevant data so that decision-makers in a range of industries, such as finance and healthcare, can make well-informed choices. Gaining a competitive edge and optimizing processes both depend on data mining to apply insightful data. “Data mining has found extensive applicability in the healthcare industry such as classifying optimum treatment methods, predicting disease risk factors, and finding efficient cost structures of patient care.” (Thirugnanam, 2013)

Heart disease, which is a general term for a number of cardiovascular disorders, is a serious threat to global health. It is the world's leading cause of death, accounting for 17.9 million deaths per year. Its prevalence is influenced by risk factors such as poor diet, smoking, physical inactivity, and excessive alcohol consumption. Proactive public health measures are necessary to address heart disease in order to lessen its global impact. (2019, World Health Organization)

“Research in the field of cardiovascular diseases using data mining has been an ongoing effort involving prediction, treatment, and risk score analysis with high levels of accuracy.” (Thirugnanam, 2013) A lot of data has been gathered regarding cardiovascular diseases; the Cleveland Heart Clinic dataset is the most widely used one. This work will use five different heart disease datasets combined into one, including the Cleveland Heart Disease Database (CHDD). The source of this open-source dataset is Kaggle. The uniqueness of this dataset comes from the 11 shared features it incorporates from other datasets, providing a more thorough understanding of cardiovascular risk factors. Additionally, the dataset can be accessed via the "Heart Disease" index on the UCI Machine Learning Repository (Fedesoriano, 2021).

The methodology aims to develop a framework for methodically evaluating the likelihood of cardiovascular disease by utilizing several predictive models. The evaluation of these models aims to identify the most accurate predictor, contributing to the ongoing discourse on precision medicine in cardiovascular health.

The integration of this Kaggle-sourced dataset represents a pivotal step toward enhancing predictive capabilities and expanding the scope of cardiovascular research.

2. LITERATURE REVIEW

Predicting heart disease has seen many research papers with the adoption of various machine-learning models. Attempting for the highest accuracy possible, researchers have employed numerous models, each offering unique insights into cardiovascular risk prediction. These techniques include Support Vector Machine (SVM), Artificial Neural Networks (ANN), Random Forest (RF), Naïve Bayesian (NB), and Decision Trees (DT).

Research focused on predicting cardiovascular disease shares common ground, yet differs due to variations in datasets, models, and parameters. The use of different datasets introduces unique demographic and geographical factors, shaping the studies. Additionally, the adoption of diverse predictive models and parameter settings contributes to the richness of approaches, highlighting the dynamic nature of cardiovascular disease prediction research. In a study by Xing et al. (2007), it was found that Support Vector Machines (SVM) exhibited an accuracy of 92.1%, artificial neural networks achieved 91.0%, and decision trees yielded 89.6% accuracy in the results. In a similar study by Vardhan et al. (2022), algorithms like SVM, LR, ANN, RF, KNN, and more were used to predict heart disease. Out of all the models, Random Forest had the highest accuracy of 90.16 percent.

SVM, or support vector machines, is a potent algorithm used for regression and classification. By employing important support vectors to maximize class separation, it locates a hyperplane in feature space (Vora, 2022). By mapping data to a higher-dimensional space, SVM uses the kernel trick to handle both linear and non-linear relationships. Aryo Anggoro and Devi Kurnia's (2020) study demonstrated the efficacy of SVM in predicting heart diseases and their prognosis.

A statistical technique for binary classification that models the likelihood of an event based on independent variables is called logistic regression. It is useful for binary outcome prediction because it uses the logistic function to convert input features into probabilities between 0 and 1. Peng and colleagues, n.d. High accuracies were obtained in a study by Khanna et al. (n.d.), demonstrating the effectiveness of logistic regression as a method for accurately predicting heart disease.

Artificial Neural Networks (ANN) serve as computational models inspired by the neural structure of the human brain, featuring interconnected nodes or artificial neurons arranged in layers (Tick et al., n.d.). Their effectiveness in predicting heart disease lies in their capacity

to learn intricate patterns from diverse data inputs, enabling precise classification and early detection (Khan et al., 2022). According to Talukdar and Singh (2023), ANNs have proven highly efficient in navigating complex relationships within medical datasets, establishing them as a valuable tool for improving the accuracy of heart disease prediction.

A machine learning algorithm called K-Nearest Neighbors (KNN) uses the majority class of the closest neighbors of each data point in the feature space to classify the data points. KNN is useful in the prediction of heart disease because it takes advantage of patient profile similarity, which makes it efficient in finding patterns in medical datasets (Jabbar et al., 2013). Its suitability for heart disease prediction is attributed to its simplicity and capacity to manage nonlinear relationships in data, enabling precise risk assessment and early detection (Baskar, 2022).

Random Forest and XGBoost, two ensemble learning techniques, are powerful in predicting heart disease by leveraging the combination of multiple decision trees. Random Forest constructs an ensemble of decision trees and amalgamates their outputs, creating a robust model that mitigates overfitting and improves generalization (Asif et al., 2023). On the other hand, XGBoost, an optimized gradient boosting algorithm, sequentially builds weak learners to minimize errors, offering higher precision and efficiency in handling intricate relationships within complex medical datasets (Budholiya et al.,2022). These models stand out for their ability to provide reliable risk assessments and contribute to early detection of heart disease (Alqahtani et al., 2022).

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, assuming independence between features. In predicting heart disease, Naive Bayes analyzes relevant attributes such as age, sex, and clinical indicators, making it effective for risk assessment (Reddy et al., 2022)). Its simplicity and computational efficiency are advantageous for handling large datasets, contributing to accurate cardiovascular risk predictions ((Miranda et al., 2016). Naive Bayes has been applied successfully in the development of models detecting cardiovascular risk levels, showcasing its utility in healthcare data analysis (Patel et al., 2023)

A robust framework is essential for predicting heart disease as it provides a systematic approach, ensuring comprehensive data exploration and model refinement. For instance, a study by Felix et al. (2021), utilized the CRISP-DM framework to predict major bleeding in MCS patients, showcasing its effectiveness in guiding predictive modeling for cardiovascular outcomes.

3. CRISP-DM FRAMEWORK

Predictive modeling requires a well-defined data mining framework, especially when dealing with difficult problems like heart disease prediction. For several reasons, it is critical to implement an organized methodology, such as the Cross-Industry Standard Process for Data Mining (CRISP-DM). First of all, a systematic approach guarantees that researchers can

purposefully navigate the complexities of data regarding cardiovascular disease prediction, where datasets are diverse and complex. This is evident in the Data Understanding and Data Preparation phases of CRISP-DM, which empower scientists to thoroughly explore and preprocess data, laying the foundation for accurate predictive models (Wirth & Hipp, 2000).

Furthermore, CRISP-DM's Modeling stage fits in perfectly with the dynamic character of research on cardiovascular disease prediction. Through the selection and implementation of a variety of machine learning algorithms, researchers are able to customize their methodology to the particularities of the given problem. With the wide range of datasets, models, and parameters found in heart disease prediction studies, this flexibility is crucial.

The Evaluation phase, supported by CRISP-DM, introduces a quantitative dimension to model assessment, aiding in the identification of the most effective algorithms. In cardiovascular risk prediction, accuracy and reliability are paramount, making the systematic evaluation facilitated by CRISP-DM a crucial aspect of the research process (Chapman et al., 2000).

Because CRISP-DM is continuous, it can be repeatedly improved upon in light of evaluation and results, which helps to shape the constantly changing field of cardiovascular risk prediction research. Finally, the Deployment phase ensures the translation of successful models into real-world applications, impacting healthcare decision-making and underscoring the practical significance of a well-structured data mining framework (Wirth & Hipp, 2000). Through these attributes, CRISP-DM serves as a robust and adaptable guide, enhancing the effectiveness and reliability of predictive modeling in the context of heart disease.

4. BUSINESS UNDERSTANDING

The first phase of the CRISP-DM framework involves gaining a comprehensive understanding of the business problem. Understanding the business domain and its complexities is crucial for effectively executing the CRISP-DM model process. By establishing a solid foundation in business understanding, it is ensured that the data analysis and modeling activities directly contribute to solving the business problem.

4.1 BUSINESS PROBLEM

Cardiovascular disease poses a significant business problem due to its status as a leading cause of global mortality. The challenge is to predict and mitigate the risk of cardiac illnesses through advanced data mining and machine learning techniques. Researchers leverage large datasets, such as the Cleveland Clinic Heart Disease Dataset, to enhance the accuracy of predictions. The business goal is to employ various supervised machine learning algorithms, like Support Vector Machine (SVM), Naïve Bayes, Decision Trees, and Artificial Neural Networks, to analyze extensive datasets and forecast the likelihood of developing heart diseases (Uddin et al., 2019). This proactive approach aims to save lives by enabling early

intervention and personalized healthcare based on accurate risk predictions (Bhatt et al., 2023).

OBJECTIVE

The objective of enhancing predictive capabilities in the context of cardiovascular disease research is to utilize advanced technologies, particularly machine learning (ML), to improve the accuracy and efficiency of predicting cardiovascular events. By using predictive analytics, this paper aims to develop highly accurate models that can reliably forecast the likelihood of cardiovascular diseases, allowing for timely interventions, personalized healthcare strategies, and ultimately contributing to better patient outcomes. (Weng et al., 2017)

TOOLS & TECHNIQUES

To tackle this business problem, good data science techniques and tools are necessary. For this paper Python will be used as the programming language for cardiovascular disease prediction as this aligns with research emphasizing the language's efficacy in data science and machine learning applications (Bowles, 2015). This decision is reinforced by the adoption of essential libraries, including pandas and NumPy for data manipulation.

In predictive modeling, scikit-learn is a preferred tool, known for its diverse machine-learning algorithms and ease of use (Raval, 2023). TensorFlow, a popular deep learning framework, further enriches the toolkit, enabling the development of neural network models for more intricate patterns in cardiovascular data.

The introduction of SMOTE (Synthetic Minority Over-sampling Technique) demonstrates an awareness of addressing class imbalance, and enhancing the model's performance, as supported by literature on machine learning-based heart disease prediction systems (Bowles, 2015).

Matplotlib and Seaborn, used for data visualization, align with best practices in conveying insights effectively, as recognized in projects implementing hands-on tutorials for heart disease detection (Raval, 2023).

While alternative tools and languages like R, Julia, Jupyter Notebooks, MATLAB, PyTorch, Plotly, and Dask offer unique features, the decision to exclude them is justified by Python's versatile ecosystem, extensive community support, and pragmatic advantages, as observed in the broader data science and machine learning community (Raval, 2023).

5. DATA UNDERSTANDING

The Cleveland Heart Disease Database (CHDD), Hungarian Institute of Cardiology, University Hospital of Zurich, Long Beach VA, and Statlog (Heart) are the five independent heart disease datasets that make up the Kaggle dataset used to predict heart disease (Fedesoriano, 2021). After the removal of duplicates, this combined dataset provides the

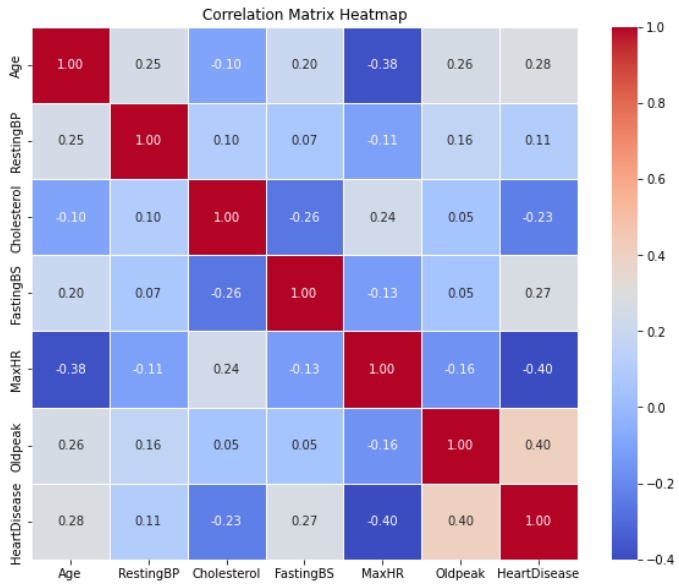
largest available dataset for research on heart disease, with 918 unique observations. The dataset, which has 11 common features, is formatted as a CSV file called "heart.csv." The UCI Machine Learning Repository's Index of "Heart Disease" datasets contains all of the used datasets (Fedesoriano, 2021).

The attribute information provides a detailed overview of each feature, including age, sex, chest pain type, resting blood pressure (RestingBP), serum cholesterol (in mm/dl), fasting blood sugar (FastingBS > 120 mg/dl), resting electrocardiogram results (RestingECG), maximum heart rate achieved (MaxHR), exercise-induced angina (exercise angina), oldpeak (St Depression induced by exercise relative to rest), Slope of the peak exercise ST segment ST Slope, and the output class (HeartDisease).

The exploration of the data involves various techniques to understand its characteristics and relationships. Initial exploration steps include loading the dataset using pandas and employing methods like head(), tail(), info(), and describe() to gain insights into the dataset's structure, data types, and summary statistics. Further exploration delves into the distribution of key features, including age, RestingBP, cholesterol levels, and MaxHR, through visualizations such as histograms. Bar graphs provide insights into the distribution of people based on sex, fasting blood sugar levels, exercise-induced angina, and resting electrocardiogram results. These visualizations aid in identifying patterns and potential relationships within the data, contributing to a comprehensive understanding of the dataset's characteristics and informing subsequent steps in the predictive modeling process.

The correlation matrix heatmap offers a detailed insight into the relationships between selected features. Notably, there is no strong correlation between variables, indicating that multicollinearity is not a significant concern. This observation is crucial for model development, as it suggests that the selected features provide unique information without redundancy.

FIGURE 1 (CORRELATION HEATMAP



For a detailed view of all visualizations, please refer to Appendix 1.1-11. These visual representations enhance the interpretability of the dataset and serve as a reference for the insights gained during the exploratory data analysis phase.

6. DATA PREPARATION

The data preparation phase, often referred to as "data munging," plays a crucial role in shaping the dataset for modeling. This phase involves five key tasks:

Data cleaning is essential for ensuring the quality and accuracy of the dataset. As this dataset was already cleaned no missing values, NaNs (Not a Number), or outliers had to be handled. The data cleaning process ensures that the dataset is free from inconsistencies and ready for further analysis.

Feature engineering is performed to derive new attributes that may enhance the predictive power of the model. Three new features are constructed:

- 'BP_Category': Categorization of resting blood pressure into stages.
- 'Chol_Category': Categorization of serum cholesterol levels.
- 'Age_Group': Grouping ages into predefined categories.

Additionally, the interaction between resting blood pressure and cholesterol, named 'BP_Chol_Interaction,' is calculated to capture potential combined effects.

The dataset, already a compilation of five independent heart datasets, does not require further integration from external sources. The consolidation of data during its creation ensures a unified and comprehensive dataset for analysis.

Data formatting involves converting categorical variables into a suitable format for modeling. LabelEncoder is applied to transform categorical variables such as 'Sex,' 'ChestPainType,' 'FastingBS,' 'RestingECG,' 'ExerciseAngina,' 'ST_Slope,' and 'HeartDisease' into numeric values. Additionally, one-hot encoding is employed for the 'BP_Category' and 'Chol_Category' features to ensure compatibility with machine learning algorithms.

The prepared dataset, represented by features in 'X' and the target variable 'y,' is now ready for the subsequent modeling phase. These transformations and enhancements contribute to the dataset's readiness for accurate and effective predictive modeling of heart disease.

7. MODELING

The selection of Logistic Regression, Artificial Neural Networks (ANN), Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, Decision Tree, and XGBoost for heart disease prediction reflects a strategic approach to leverage diverse

machine learning algorithms, each offering unique strengths. Logistic Regression, known for its simplicity and effectiveness in binary classification tasks, is cited as a reliable technique for accurate heart disease prediction in the research literature (Rahman et al., 2023). ANN, with its capacity to learn intricate patterns, has been proven efficient in navigating complex relationships within medical datasets, establishing it as a valuable tool for improving the accuracy of heart disease prediction (Rahman et al., 2023). Random Forest, SVM, KNN, Naïve Bayes, Decision Tree, and XGBoost are acknowledged for their predictive capabilities in cardiovascular disease research, with Random Forest specifically highlighted for its ability to provide reliable risk assessments and contribute to early detection of heart disease (Jain et al., 2022).

Numerous machine-learning algorithms were used to create a very accurate predictive model for heart disease. Training and testing sets of the dataset were separated, starting with the Logistic Regression (LR) model. Prior to training the LR model, a standardization procedure was put in place to guarantee consistency across features. Predictions were then performed on the test set, and the accuracy was found to be 91.85%. The model's complex performance was explained by the confusion matrix and classification report that went along with it. Precision indicates the accuracy of positive predictions, recall assesses the model's capacity to correctly identify instances of a given class, and the F1-score is a metric that finds a balance between the two. The LR model obtains an F1-score of 90% for class 0 (no heart disease), which shows a well-balanced combination of recall and precision for cases without heart disease. In a similar vein, class 1's F1-score of 93% denotes a balanced performance in spotting heart disease cases.

Artificial Neural Network (ANN)

Including all the important features, the Artificial Neural Network (ANN) model demonstrated the highest accuracy, averaging around 93 and 94%. By initializing random seeds, the code prioritizes the reproducibility of results, guaranteeing consistent results across runs for trustworthy validation and comparison.

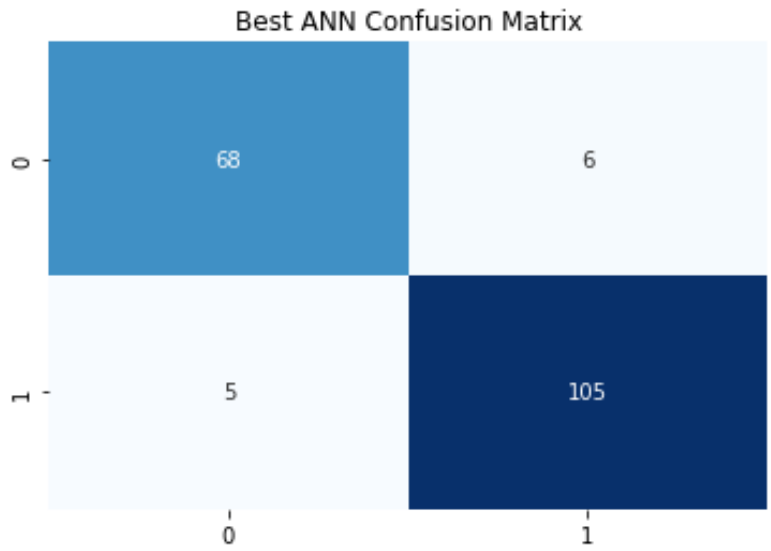
The model's architecture is built for robustness over the course of ten runs. It includes a well-thought-out three-layer structure, standardization with StandardScaler, and a rigorous division of the dataset into training and testing sets. Using the sigmoid activation function, the architecture consists of an input layer with 64 neurons, a hidden layer with 32 neurons, and an output layer with a single neuron. Issues like the vanishing gradient problem and overfitting are mitigated by the use of ReLU activation, He normal initializer, and dropout layers.

The RMSprop optimizer with a learning rate of 0.001 and binary cross-entropy loss function contribute to effective training. In simpler terms, the optimizer guides the learning process, adjusting the model's parameters to minimize errors. The binary cross-entropy loss function measures the difference between predicted and actual values for binary classification, guiding the model towards accurate predictions.

The model undergoes 100 epochs of training with a batch size of 32, reserving 20% for validation to ensure adaptability without overfitting. The iterative loop allows for multiple training runs, with the best-performing model based on test set accuracy being retained.

The results of the best ANN model showcase its superior accuracy, a confusion matrix detailing performance, and a comprehensive classification report providing insights into precision, recall, and F1-score for each class. The confusion matrix is visually depicted using a heatmap, offering a clear overview of the model's accuracy in classifying instances.

FIGURE 2: ANN CONFUSION MATRIX



This ANN model excels with a best test set accuracy of 94.02%, underscoring its proficiency in predicting heart disease presence or absence. The confusion matrix highlights accurate predictions for both classes, particularly for instances of heart disease. Precision, recall, and F1-score metrics in the classification report affirm the model's effectiveness, showcasing a balanced performance across different classes.

The ANN model not only achieves outstanding overall accuracy but also demonstrates robust and balanced performance, emphasizing its potential in clinical applications for accurate heart disease prediction.

TABLE 1: MODEL ACCURACIES

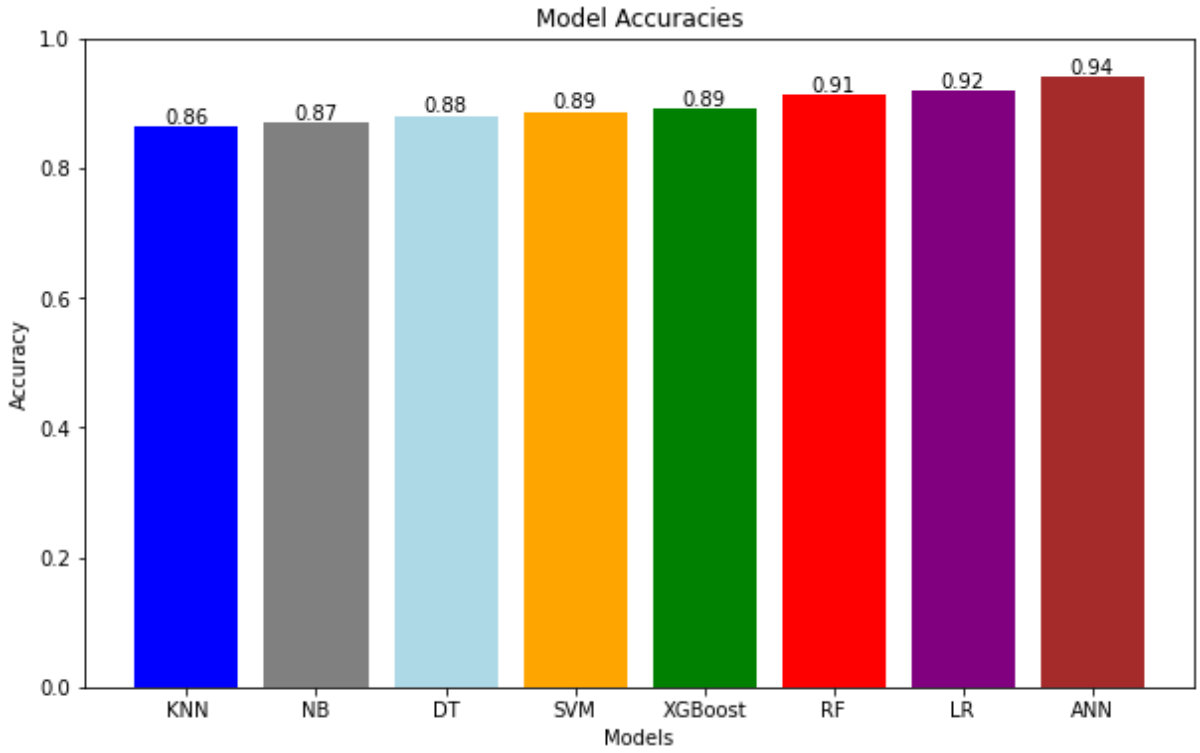
	Model Accuracy Score %				
	Model	This study	(Parmar, 2020)	(Emil, 2023)	(Bashir et al., 2019)
1	ANN	94.02	-	-	-
0	Logistic Regression	91.85	85.25	92.39	82.56
2	Random Forest	91.30	85.15	89.13	84.17
7	XGBoost	89.13	-	-	-
3	SVM	88.59	81.97	90.76	84.85
6	Decision Tree	88.04	-	74.46	82.22
5	Naïve Bayes	86.96	85.25	90.76	84.24
4	KNN	86.41	90.16	90.22	-

The Random Forest model, with a preliminary split and scaling of the data. Addressing class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied during the training of the Random Forest classifier. The model demonstrated robust performance with an accuracy of 91.30%. The associated confusion matrix and classification report delineated the model's efficacy in classifying instances for both classes.

For the Support Vector Machine (SVM) model, a meticulous hyperparameter tuning process unfolded using GridSearchCV. The optimal SVM model, characterized by a set of hyperparameters, was then evaluated on the test set, yielding an accuracy of 88.59%. The subsequent confusion matrix and classification report detailed the model's precision, recall, and F1-score metrics for each class.

Moving to the K Nearest Neighbors (KNN) model, hyperparameter tuning via GridSearchCV led to an accuracy of 86.41%. Evaluation through the confusion matrix and classification report highlighted the model's proficiency. The Naïve Bayes model, post-standardization, achieved 86.96% accuracy, exhibiting balanced performance. The Decision Tree, with scaled data and a specified depth, attained an 88.04% accuracy, as detailed in the accompanying confusion matrix and classification report. Finally, the XGBoost model, post-split and scaling, delivered an 89.13% accuracy.

FIGURE 3: MODEL ACCURACIES BAR CHART



To provide a visual summary of the model performances, a bar graph depicted the accuracies of each model, emphasizing the ANN's remarkable accuracy, making it the standout model for this heart disease prediction paper.

8. EVALUATION

The in-depth evaluation of different machine-learning models for heart disease prediction entails looking at key elements associated with business success metrics, going over the process in detail, and describing the next steps.

It is clear from evaluating the models in relation to the goal of developing a highly accurate model that different algorithms provide different levels of accuracy, emphasizing the necessity of selecting the best deployment model. With an accuracy of 91.85%, the Logistic Regression model demonstrates excellent predictive abilities. But among the algorithms taken into consideration, the Artificial Neural Network (ANN) model performs better than the others, achieving an astounding accuracy of 94.02%, making it the most accurate predictor. This high accuracy supports the ANN model's ability to accurately predict the presence or absence of heart disease, which is in line with the main objective of precise risk assessment.

A review of the process shows a good execution, covering data understanding, preparation, and modeling phases. The adherence to the CRISP-DM framework provided a structured and iterative approach, ensuring that all aspects, from dataset exploration to model evaluation, are addressed. Visualizations, correlation analyses, and detailed model assessments contribute to a robust understanding of the dataset and model performances. The chosen models, representing a variety of machine-learning techniques, offer a diverse yet strategic approach to tackling the business problem of heart disease prediction. No significant oversights or shortcomings are apparent, affirming the integrity of the conducted analysis.

Considering the evaluation results and the review of the process, the next steps involve making informed decisions on deployment, further iteration, or potential initiation of new projects. The ANN model's exceptional accuracy positions it as the primary candidate for deployment in predicting heart disease. Its highly accurate predictions, emphasize its potential utility in clinical applications. Further iterations could involve fine-tuning hyperparameters or exploring additional data sources to enhance model performance. If resources allow, initiating new projects could involve expanding the scope of heart disease prediction research, and considering emerging technologies or methodologies for even more accurate predictions.

9. DEPLOYMENT

To implement the heart disease prediction model in the real world, a systematic deployment plan is crucial. The following steps outline an effective deployment strategy:

1. Integration with Healthcare Systems: Collaborate with healthcare institutions to seamlessly integrate the predictive model into their existing systems, incorporating it

into Electronic Health Record (EHR) systems or developing user-friendly applications for healthcare professionals (Karthick et al., 2022).
2. User Training: Conduct comprehensive training sessions for healthcare practitioners who will utilize the model. Ensure that medical professionals understand the model's predictions, limitations, and how to interpret the results for informed decision-making.
3. Scalability Considerations: Ensure the model is scalable to handle a diverse and extensive patient database. Scalability is crucial for accommodating the growing volume of health data and ensuring the model's efficiency in different healthcare settings (Nashif et al., 2018).
4. Privacy and Ethical Compliance: Prioritize patient privacy and comply with ethical standards. Develop protocols and security measures to safeguard patient information, ensuring adherence to healthcare regulations and standards like HIPAA (Nass et al., 2009).

To ensure sustained performance and reliability, a thorough monitoring and maintenance plan is essential:

1. Continuous Monitoring: Implement real-time monitoring to track the model's performance in different healthcare environments. Monitor accuracy, false positives, and false negatives to identify any deviations from expected outcomes (Mohapatra et al., 2023).
2. Regular Model Updates: Periodically update the model using new data to enhance its predictive capabilities. This involves retraining the model with the latest information to account for evolving patterns and risk factors associated with heart disease (Bebortta et al., 2023).
3. Feedback Mechanism: To get feedback on the model's predictions, set up a loop that includes medical professionals. To continuously hone and enhance the model, gather input on both accurate and inaccurate forecasts(Mohapatra et al., 2023).
4. Security Measures: To safeguard patient data, audit and update security procedures on a regular basis. Keep up with the most recent cybersecurity threats and put precautions in place to guarantee the security of the model and related data (Nashif et al., 2018).

By systematically addressing these deployment tasks, the heart disease prediction model can be effectively integrated into real-world healthcare systems, contributing to proactive interventions and improved patient outcomes. Continuous monitoring and refinement ensure the model's ongoing success in predicting heart disease with accuracy and reliability.

10. CONCLUSION AND FUTURE WORK

In this paper, a comprehensive exploration of predictive modeling for heart disease was undertaken, leveraging the CRISP-DM framework and an ensemble of machine learning algorithms. The goal was to enhance accuracy in cardiovascular disease prediction, ultimately contributing to proactive healthcare interventions. The study incorporated the unique Cleveland Heart Disease Database and four additional datasets, providing a rich foundation for a robust predictive framework.

The models, including Logistic Regression, Artificial Neural Networks (ANN), Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, Decision Tree, and XGBoost, were systematically evaluated. Notably, the ANN model emerged as the frontrunner, achieving an outstanding accuracy of 94.02%. As this model has such high accuracy, it can be effectively integrated into real-world healthcare systems, contributing to proactive interventions and improved patient outcomes.

While achieving remarkable accuracy, future work can focus on several avenues for refinement and expansion:

While achieving high accuracy, future work for the Artificial Neural Network (ANN) model involves refining hyperparameters, exploring additional features for enhanced predictability, and validating on diverse datasets. Improving model explainability is crucial for clinical trust, and seamless real-world deployment, with continuous collaboration, is a priority. Additionally, leveraging longitudinal data can offer dynamic insights into cardiovascular risk changes over time.

By addressing these areas, future endeavors can contribute to the refinement, robustness, and practical applicability of predictive models for heart disease, further advancing precision medicine in cardiovascular health.

REFERENCES

Alqahtani, A., Alsubai, S., Sha, M., Vilčeková, L., & Javed, T. (2022). Cardiovascular Disease Detection using Ensemble Learning. *Computational Intelligence and Neuroscience*, *2022*, 1–9. https://doi.org/10.1155/2022/5267498

Aryo Anggoro, D., & Devi Kurnia, N. (2020). Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease. *International Journal of Emerging Trends in Engineering Research*, *8*(5). https://doi.org/10.30534/ijeter/2020/32852020

Asif, D., Bibi, M., Arif, M., & Mukheimer, A. (2023). Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization. *Algorithms*, *16*(6), 308. https://doi.org/10.3390/a16060308

Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019). Improving Heart Disease Prediction Using Feature Selection Approaches. *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan*. https://doi.org/10.1109/ibcast.2019.8667106

Baskar, C. (2022, October 18). Cardiovascular disease prediction using KNN algorithm. *Medium*. https://medium.com/analytics-vidhya/heart-disease-prediction-using-knn-algorithm-be78f800e2a9

Bebortta, S., Tripathy, S. S., Basheer, S., & Chowdhary, C. L. (2023). FedEHR: A Federated Learning Approach towards the Prediction of Heart Diseases in IoT-Based Electronic Health Records. *Diagnostics*, *13*(20), 3166. https://doi.org/10.3390/diagnostics13203166

Bhatt, C., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), 88. https://doi.org/10.3390/a16020088

Bowles, M. (2015). *Machine learning in Python: Essential Techniques for Predictive Analysis*. John Wiley & Sons.

Budholiya, K., Shrivastava, S. K., & Sharma, V. (2022). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University - Computer and Information Sciences*, *34*(7), 4514–4523. https://doi.org/10.1016/j.jksuci.2020.10.013

Emil, M. (2023, November 15). *HeartDisease Analysis, Visualization and Classify*. Kaggle. https://www.kaggle.com/code/minaemil329/heartdisease-analysis-visualization-and-classify/notebook

Fedesoriano. (2021). *Heart Failure Prediction Dataset* [Dataset]. https://www.kaggle.com/fedesoriano/heart-failure-prediction

Felix, S., Bagheri, A., Ramjankhan, F., Spruit, M., Oberski, D. L., De Jonge, N., Van Laake, L. W., Suyker, W. J., & Asselbergs, F. (2021). A data mining-based cross-industry process for predicting major bleeding in mechanical circulatory support. *European Heart Journal*, *2*(4), 635–642. https://doi.org/10.1093/ehjdh/ztab082

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using K- nearest neighbor and genetic algorithm. *Procedia Technology*, *10*, 85–94. https://doi.org/10.1016/j.protcy.2013.12.340

Jain, D., Yadav, V. K., Kumari, M. S., Chandravansi, K., Reddy, G. V., & Cheltha, J. (2022). Cardiovascular Disease Predictor. *Lovely Professional University*. https://doi.org/10.4108/eai.16-4-2022.2318172

Karthick, K., Aruna, S., Samikannu, R., Kuppusamy, R., Teekaraman, Y., & Thelkar, A. R. (2022). Implementation of a heart disease risk prediction model using machine learning. *Computational and Mathematical Methods in Medicine*, *2022*, 1–14. https://doi.org/10.1155/2022/6517716

Khan, M. U., Samer, S., Alshehri, M. D., Baloch, N. K., Khan, H., Hussain, F., Kim, S. W., & Zikria, Y. B. (2022). Artificial neural network-based cardiovascular disease prediction using spectral features. *Computers & Electrical Engineering*, *101*, 108094. https://doi.org/10.1016/j.compeleceng.2022.108094

Khanna, D., Sahu, R., Baths, V., & Deshpande, B. M. (n.d.). Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict

the Prevalence of Heart Disease. *International Journal of Machine Learning and Computing*, *5*(5), 414–419. https://doi.org/10.7763/ijmlc.2015.v5.544

Miranda, E., Irwansyah, E., Amelga, A. Y., Maribondang, M. M., & Salim, M. (2016). Detection of cardiovascular disease risk's level for adults using Naive Bayes Classifier. *Healthcare Informatics Research*, *22*(3), 196. https://doi.org/10.4258/hir.2016.22.3.196

Mohapatra, S., Maneesha, S., Mohanty, S., Patra, P. K., Bhoi, S. K., Sahoo, K. S., & Gandomi, A. H. (2023). A stacking classifiers model for detecting heart irregularities and predicting Cardiovascular Disease. *Healthcare Analytics*, *3*, 100133. https://doi.org/10.1016/j.health.2022.100133

Nashif, S., Raihan, M. R., Islam, M. R., & Imam, H. (2018). Heart disease detection by using machine learning algorithms and a Real-Time cardiovascular Health monitoring system. *World Journal of Engineering and Technology*, *06*(04), 854–873. https://doi.org/10.4236/wjet.2018.64057

Nass, S. J., Levit, L., & Gostin, L. O. (2009). *Beyond the HIPAA privacy rule: enhancing privacy, improving health through research*. https://pubmed.ncbi.nlm.nih.gov/20662116/

Patel, T. P., Patel, D. P., Sanyal, M., & Shrivastav, P. S. (2023). Prediction of Heart Disease and Survivability using Support Vector Machine and Naive Bayes Algorithm. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2023.06.09.543776

Peng, C. J., Lee, K. L., & Ingersoll, G. M. (n.d.). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, *96*(1), 3–14. https://doi.org/10.1080/00220670209598786

Rahman, P., Rifat, A., Chy, M. I., Khan, M. M., Masud, M., & Aljahdali, S. (2023). Machine learning and artificial neural network for predicting heart failure risk. *Computer Systems Science and Engineering*, *44*(1), 757–775. https://doi.org/10.32604/csse.2023.021469

Raval, P. (Ed.). (2023, October 12). Project on heart Disease Prediction using Machine Learning. *ProjectPro*. https://www.projectpro.io/article/heart-disease-prediction-using-machine-learning-project/615

Reddy, V. S. K., Meghana, P., Reddy, N. G., & Rao, B. A. (2022). Prediction on Cardiovascular disease using Decision tree and Naïve Bayes classifiers. *Journal of Physics*, *2161*(1), 012015. https://doi.org/10.1088/1742-6596/2161/1/012015

Sharma, S., & Parmar, M. (2020). Heart Diseases Prediction using Deep Learning Neural Network Model. *International Journal of Innovative Technology and Exploring Engineering*, *9*(3), 2244–2248. https://doi.org/10.35940/ijitee.c9009.019320

Siddharth. (2021, July 10). *Heart Disease Prediction using KNN -The K-Nearest Neighbours Algorithm*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/07/heart-disease-prediction-using-knn-the-k-nearest-neighbours-algorithm/

Talukdar, J., & Singh, T. P. (2023). Early prediction of cardiovascular disease using artificial neural network. *Paladyn*, *14*(1). https://doi.org/10.1515/pjbr-2022-0107

Thirugnanam, M. (2013). A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL). *International Journal of Computer Applications in Technology*, *68*.

Tick, V. K., Meeng, N. Y., Mohammad, N. F., Harun, N. H., Alquran, H., & Mohsin, M. F. M. (n.d.). Classification of Heart Disease using Artificial Neural Network. *Journal of Physics*, *1997*(1), 012022. https://doi.org/10.1088/1742-6596/1997/1/012022

Uddin, S., Khan, A., Hossain, E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-1004-8

Vardhan, G. H., Reddy, N. S. S., & Umamaheswari, K. M. (2022). Heart disease prediction using machine learning. *International Journal of Health Sciences*, 6(S2), 7804–7813. https://doi.org/10.53730/ijhs.v6nS2.6955

Vora, U. (2022, January 1). Heart disease prediction using Support Vector Machine (SVM). *Medium*. https://utsavvora.medium.com/heart-disease-prediction-using-support-vector-machine-svm-34d8c01c596

Weng, S., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, *12*(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

World Health Organization: WHO. (2019, June 11). *Cardiovascular diseases*. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

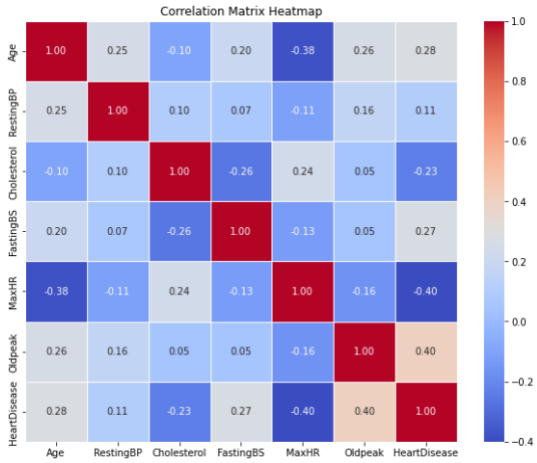
Xing, Y., Wang, J., Zhao, Z., & Gao, A. (2007). Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease. *2007 International Conference on Convergence Information Technology (ICCIT 2007), Gwangju, Korea (South)*, 868–872. https://doi.org/10.1109/ICCIT.2007.204

APPENDIX

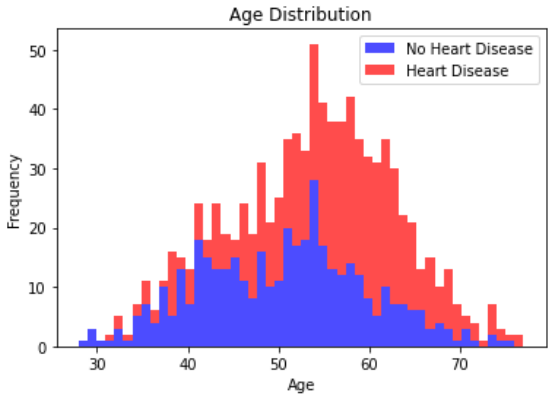
1. DATA EXPLORATION VISUALS

In this section, all of the graphs that were created to explore the data are shown here.

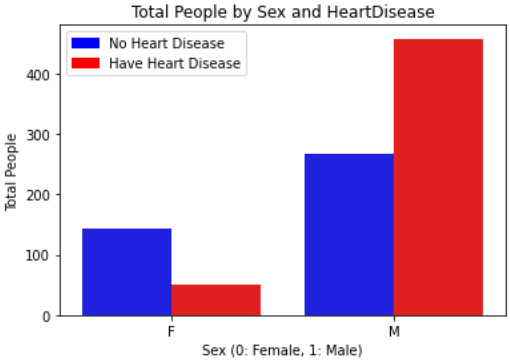
1.1 Correlation Heatmap



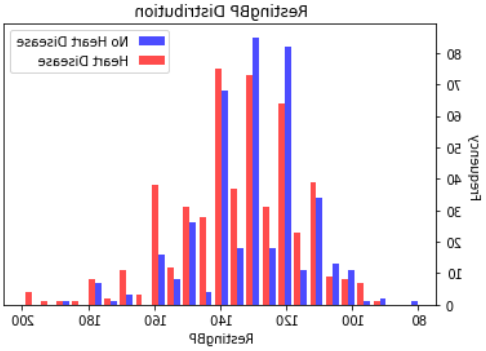
1.2 Age Distribution



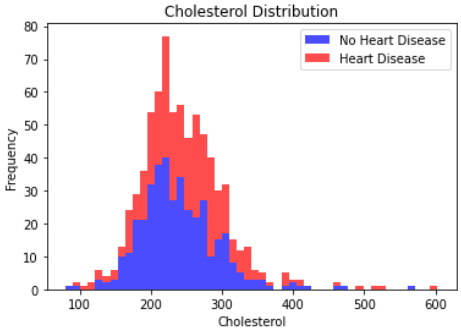
1.3 Total people by sex and Heart Disease



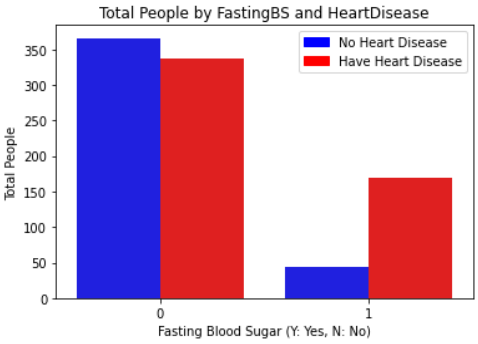
1.4 RestingBP Distribution



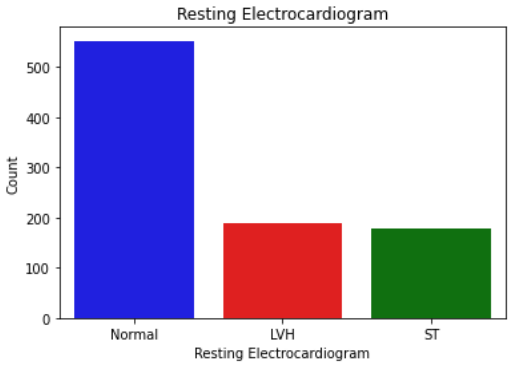
1.5 Cholesterol Distribution



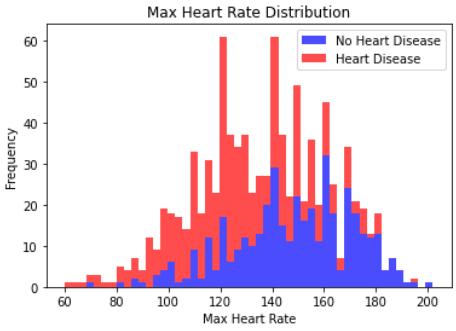
1.6 Total people by fastingBS and Heart Disease



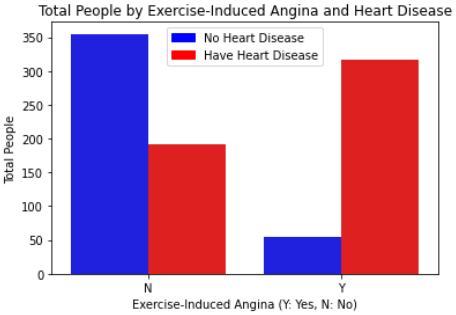
1.7 Resting Electrocardiogram



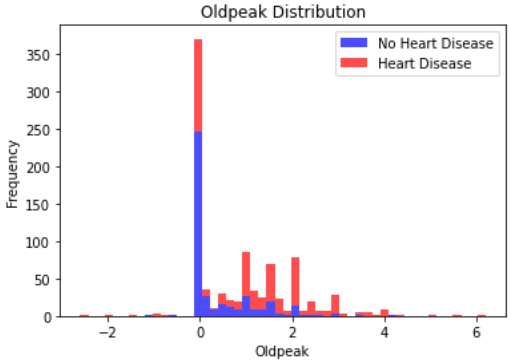
1.8 Max Heart Rate Distribution



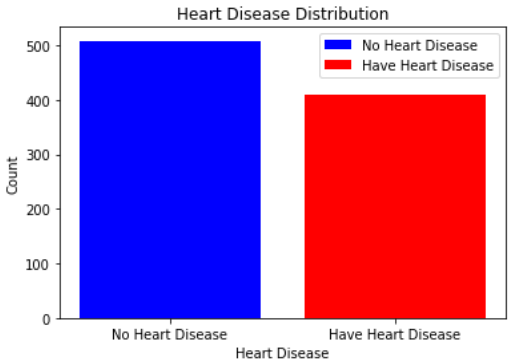
1.9 Total People by Exercise-Induced Angina and Heart Disease



1.10 Oldpeak Distribution

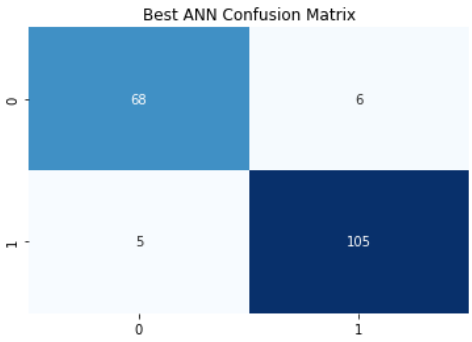


1.11 Heart Disease Distribution

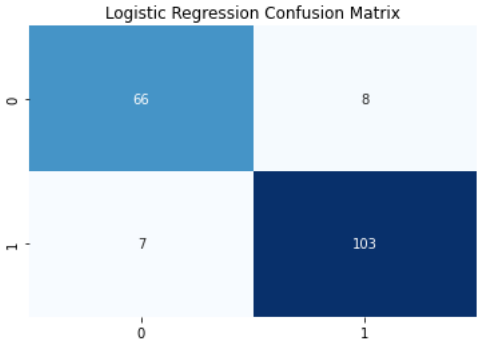


In this section, all of the graphs that have been created to evaluate the models' performances are shown here.

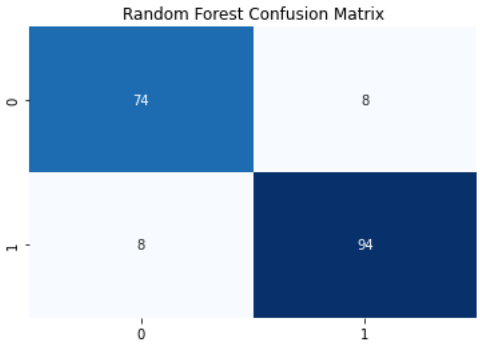
2.1 ANN Confusion Matrix



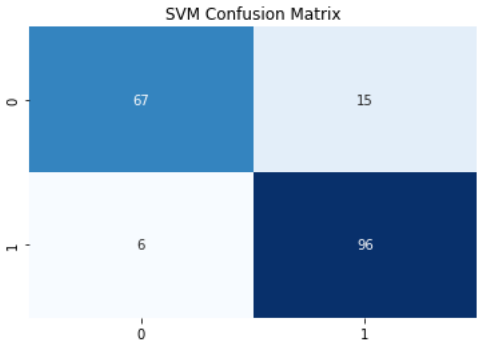
2.2 Logistic Regression Confusion Matrix



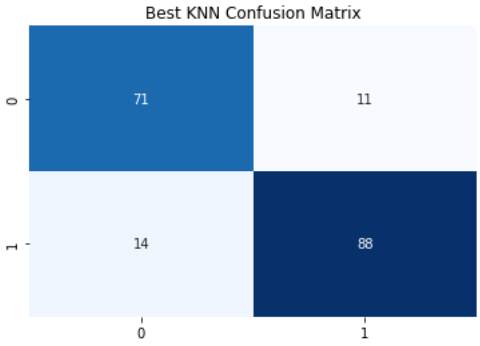
2.3 Random Forest Confusion Matrix



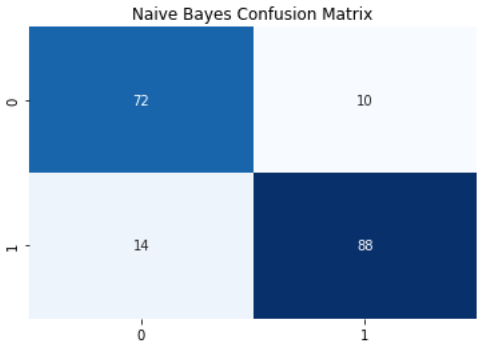
2.4 SVM Confusion Matrix



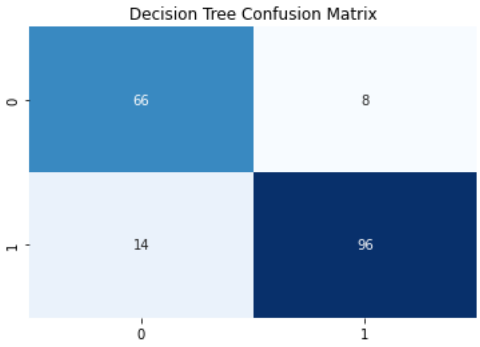
2.5 KNN Confusion Matrix



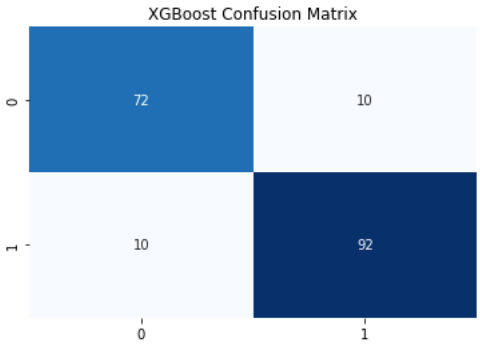
2.6 Naïve Bayes Confusion Matrix



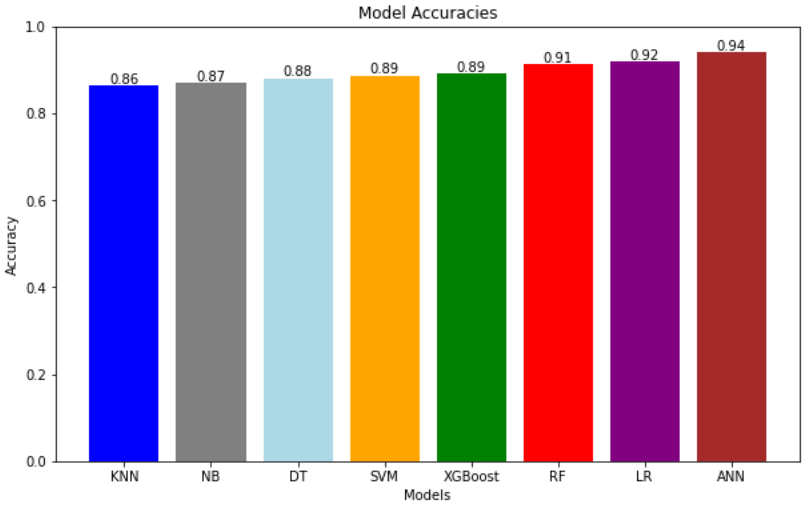
2.7 Decision Tree Confusion Matrix



2.8 XGBoost Confusion Matrix



2.9 Model Accuracies Bar Chart



2.10 Model Accuracies Table

Model Accuracy Score %					
	Model	This study	(Parmar, 2020)	(Emil, 2023)	(Bashir et al., 2019)
1	ANN	94.02	-	-	-
0	Logistic Regression	91.85	85.25	92.39	82.56
2	Random Forest	91.30	85.15	89.13	84.17
7	XGBoost	89.13	-	-	-
3	SVM	88.59	81.97	90.76	84.85
6	Decision Tree	88.04	-	74.46	82.22
5	Naïve Bayes	86.96	85.25	90.76	84.24
4	KNN	86.41	90.16	90.22	-

3. HEART DISEASE PREDICTION PYTHON CODE

For the python code please refer to GitHub under the repository “Heart-Disease-Prediction-Dec-2023”. Or use the following link: https://github.com/PjotrOtten/Heart-Disease-Prediction-Dec-2023/tree/main