Comparative Study of Classification Algorithms for Heart Disease Prediction Using Microsoft Azure Machine Learning Studio

EKAAGAR SINGH HARA[1], COL. GURINDER JIT SINGH[2], DR. ARPIT YADAV[3]

[1]*Data Scientist*
[2]*Research Scholar*
[3]*Senior AI Scientist*

***Abstract-* Heart disease continues to be one of the leading causes of death worldwide, making early and accurate prediction crucial for saving lives. This study focuses on comparing different classification algorithms to predict heart disease using Microsoft Azure Machine Learning Studio. We examine popular methods such as Decision Trees, Random Forest, Logistic Regression, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Gradient Boosting to evaluate their performance in predicting heart disease. Using Azure ML Studio's robust tools for data preparation, model training, and evaluation, we assess each model's accuracy, precision, recall, F1-score, and AUC. The goal is to identify which algorithm offers the best balance between predictive power and computational efficiency for heart disease diagnosis. Our findings provide useful insights into the strengths and limitations of these models, helping inform future healthcare applications and decision-making.**

***Index Terms-* classification algorithms, decision trees, gradient boosting, heart disease prediction, healthcare analytics, logistic regression, Microsoft azure machine learning studio, random forest, support vector machines, k-nearest neighbors, model performance evaluation**

I. INTRODUCTION

Heart disease is one of the most pressing global health challenges, responsible for a considerable number of deaths each year. According to the World Health Organization (WHO), cardiovascular diseases (CVDs) are the leading cause of death worldwide, accounting for nearly 18 million lives lost annually. With such an enormous impact on public health, the need for early and accurate diagnosis of heart disease is critical. Machine learning (ML) has become a valuable tool in healthcare for identifying patterns in complex medical data, providing a way to predict and potentially prevent diseases like heart disease. For heart disease prediction, machine learning models can analyze various patient data—ranging from lifestyle factors to clinical measurements like cholesterol levels and blood pressure—to estimate the risk of developing heart disease. These predictive models offer healthcare professionals valuable insights that can support better decision-making and treatment planning. However, not all machine learning algorithms perform equally well in predicting outcomes, and choosing the right model can significantly influence the quality and accuracy of the predictions [1].

This study aims to explore and compare several machine learning algorithms for heart disease prediction using Microsoft Azure Machine Learning Studio. Azure ML Studio is a cloud-based platform that allows researchers to build, deploy, and manage machine learning models efficiently. The algorithms we examine include Decision Trees, Random Forest, Logistic Regression, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Gradient Boosting [2]. Each of these techniques are been widely used in classification tasks, but their performance can vary depending on the data used and the specific problem being addressed. By using Azure ML Studio, we can leverage its powerful data processing, model training, and evaluation tools to conduct a detailed comparison of these algorithms. This platform makes it easier to experiment with different models, allowing us to focus more on understanding their effectiveness rather than dealing with complex technical setups. Our goal is to assess each algorithm's performance using key metrics like accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). These metrics help us understand how well each model identifies heart disease while minimizing false positives and negatives. Additionally, we consider how efficiently each algorithm operates, which is important for real-world applications where time and computational resources are limited. Through this comparative analysis, we aim to identify the most effective machine learning models for heart disease prediction. The insights gained from this research can guide healthcare professionals, data scientists, and institutions in implementing machine learning tools that improve early diagnosis and patient care in the fight against cardiovascular diseases.

II. MICROSOFT AZURE MACHINE LEARNING STUDIO

Microsoft Azure Machine Learning Studio is a cloud-based platform that makes it easy to build, train, and deploy machine learning models without needing to worry about managing infrastructure. It provides a user-friendly interface with drag-and-drop tools, allowing data scientists and developers to experiment with different algorithms and workflows quickly. With its built-in features for data preprocessing, model training, and evaluation, Azure ML Studio helps streamline the machine learning process, making it accessible even to those who are not experts in coding. It also supports seamless integration with other Azure services, making it a powerful solution for developing machine learning applications at scale.

III. MACHINE LEARNING ALGORITHMS

- Decision Trees: Decision Trees are a non-parametric supervised learning method used for classification. The algorithm splits the dataset into subsets based on the most significant differentiators, forming a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a class label. While they are easy to interpret and visualize, decision trees can suffer from overfitting, particularly with noisy data.
- Random Forest: Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the class that is the mode of the individual trees' predictions. By averaging multiple decision trees, Random Forest reduces the risk of overfitting compared to a single decision tree and improves generalization. It is highly effective but can be computationally intensive for large datasets.
- Logistic Regression: Logistic Regression is a parametric algorithm used for binary classification. It models the probability that a given input belongs to a particular class by applying the logistic function to a linear combination of the input features. Logistic Regression assumes a linear relationship between the independent variables and the log odds of the outcome, making it suitable for linearly separable data. It is interpretable and efficient but may struggle with non-linear relationships.
- k-Nearest Neighbors (k-NN): k-NN is a non-parametric, instance-based learning algorithm. It classifies a data point based on the majority class among its *k* closest neighbors in the feature space. The algorithm is simple and intuitive but can be computationally expensive, especially as the size of the dataset increases, and its performance is sensitive to the choice of *k* and the distance metric used.
- Support Vector Machines (SVM): SVM is a powerful and versatile classification algorithm that aims to find the optimal hyperplane that maximally separates different classes in the feature space. SVM can handle both linear and non-linear classification by using kernel functions to transform the data into higher dimensions where a linear separation is possible. SVM is effective in high-dimensional spaces, but its performance depends on the choice of the kernel and regularization parameters.
- Gradient Boosting: Gradient Boosting is an ensemble technique that builds models sequentially, where each new model attempts to correct the errors made by the previous models. It combines weak learners, typically decision trees, in a stage-wise fashion to create a strong predictive model. Gradient Boosting algorithms, such as XGBoost and LightGBM, are known for their high accuracy but require careful tuning of hyperparameters and can be slower to train due to their iterative nature.

Each of these algorithms provides unique strengths and weaknesses depending on the nature of the dataset, making it essential to evaluate them based on specific criteria such as accuracy, computational efficiency, and interpretability in different predictive modeling contexts.

IV. METHODOLOGY

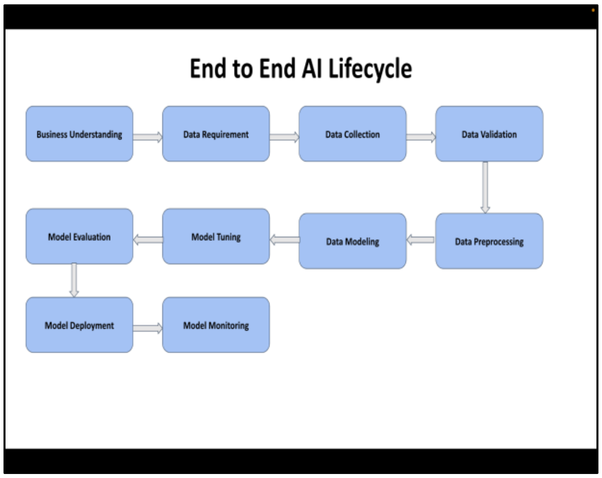

Fig 1: End To End AI Lifecycle

1. Business Understanding: This is the first and most crucial step. For heart disease prediction, the goal is to understand the business or healthcare context—why predicting heart disease is important. In this case, it is to provide doctors and healthcare providers with tools to make early and accurate diagnoses, reducing the risks associated with delayed treatment.
2. Data Requirement: In this phase, we define what kind of data is needed for the project. For heart disease prediction, patient data such as age, gender, cholesterol levels, blood pressure, and lifestyle factors like smoking and exercise habits would be required.
3. Data Collection: Here, the actual data is collected from various sources. In the heart disease prediction case, this could be from hospital records, clinical trials, or public health datasets [3]. Microsoft Azure Machine Learning Studio allows seamless integration with multiple data sources, including on-premises databases, cloud storage, and APIs.
4. Data Validation: Once data is collected, it needs to be validated to ensure it is clean, reliable, and accurate. Any missing or inconsistent values in the patient records need to be addressed. Data validation helps ensure that the subsequent steps in the machine learning workflow produce meaningful results.
5. Data Preprocessing: This step is crucial in preparing the data for modeling. In the case of heart disease prediction, preprocessing might involve normalizing numerical values like cholesterol or blood pressure, converting categorical variables like gender into numeric values, and handling missing or outlier data. Microsoft Azure Machine Learning Studio provides tools to handle these tasks efficiently.
6. Data Modeling: After the data has been prepared, it is time to select and apply machine learning models. For heart disease prediction, algorithms such as Decision Trees, Logistic Regression, Random Forest, and Gradient Boosting can be applied [4]. In Azure ML Studio, you can easily train and test various models to see which performs best in predicting heart disease outcomes.
7. Model Tuning: Once a model has been selected, it may require fine-tuning to improve performance. This step involves adjusting hyperparameters, such as the depth of the decision trees or the number of neighbors in k-NN, to optimize the model's accuracy in predicting heart disease [5].
8. Model Evaluation: In this phase, the model's performance is evaluated using metrics like accuracy, precision, recall, F1-score, and AUC (Area Under the Curve). For heart disease prediction, we evaluate how well the model predicts positive cases of heart disease and its ability to avoid false positives.

Different ML Architectures using Microsoft Learning Studio

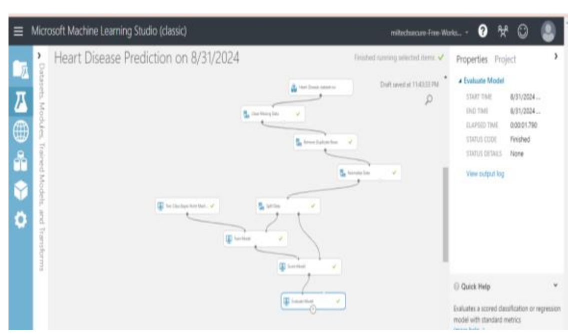

Fig 2: Two Class Bayes Point Layout

Fig 3: Two Class Boosted Decision Tree Layout

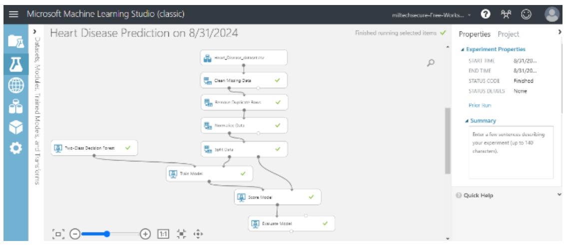

Fig 4: Two Class Decision Forest Layout

Fig 5: Two Class Decision Jungle Layout

V.

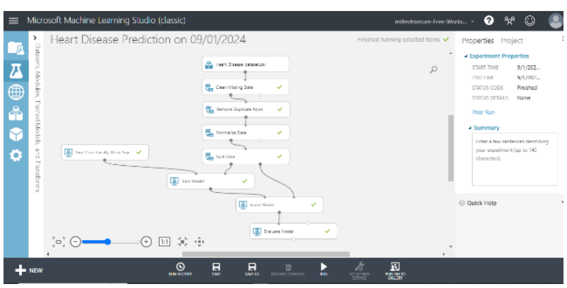

Fig 6: Two Class Locally Deep Support Vector Machine Layout

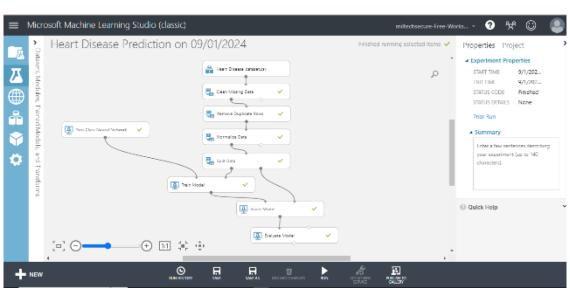

Fig 7: Two Class Neural Network Layout

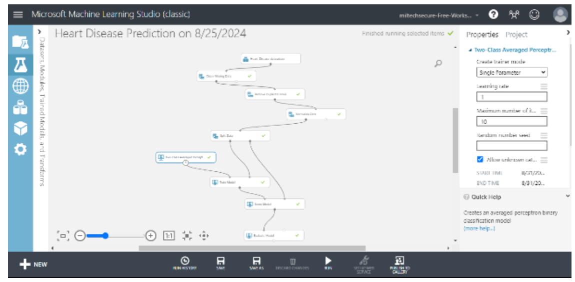

Fig 8: Two Class Average Perceptron Layout

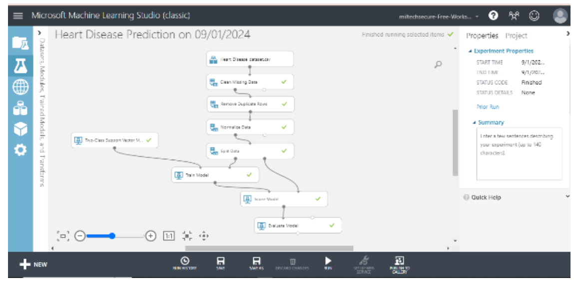

Fig 9: Two Class Support Vector Machine

IV. RESULTS

The results of the heart disease prediction models can be interpreted based on each model's performance metrics. Below are the results of models that will be evaluated and explained:

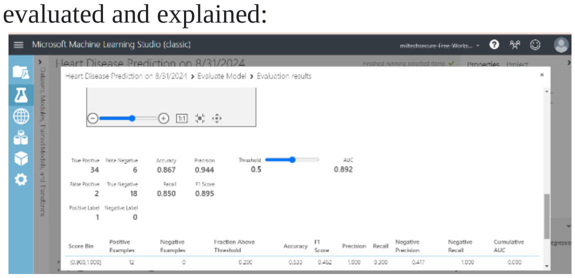

Fig 10: Two Class Bayes Point Results

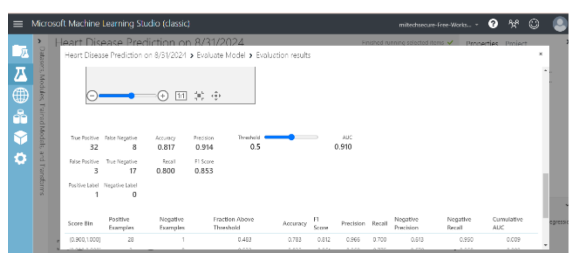

Fig 11: Two Class Boosted Decision Results

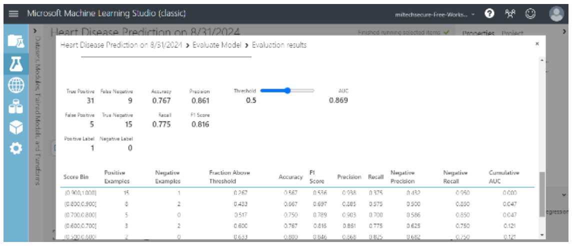

Fig 12: Two Class Decision Forest Results

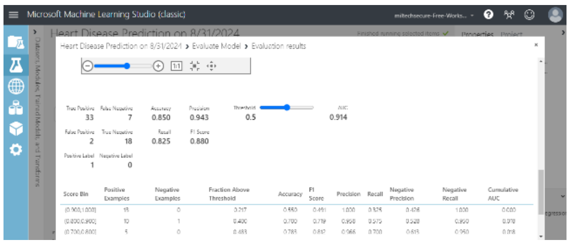

Fig 13: Two Class Decision Jungle Result

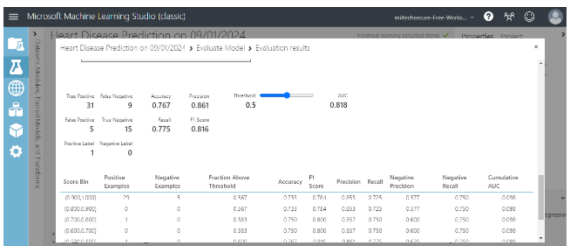

Fig 14: Two Class Locally Deep Support Vector Machine Result

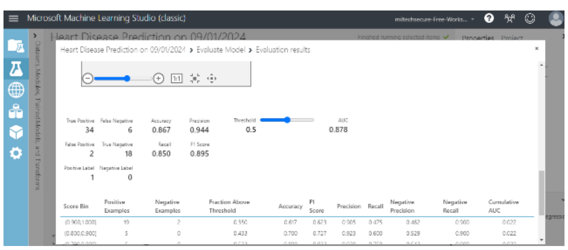

Fig15: Two Class Neural Network Result

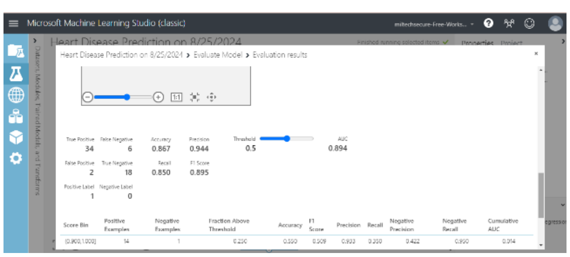

Fig 16: Two Class Perceptron Result

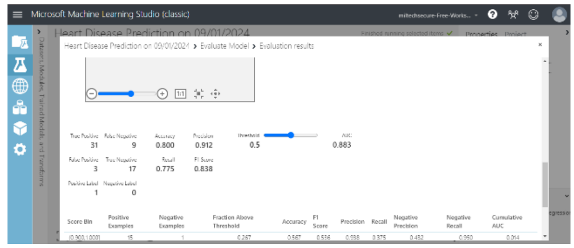

Fig 17: Two Class Support Vector Machine Result

Table 1: Performance of Models

Model Layout Name	Accuracy	Precision	Recall	F1 Score
Two Class Bayes Point	0.867	0.944	0.850	0.895
Two Class Boosted Decision	0.817	0.914	0.800	0.853
Two Class Decision Forest	0.767	0.861	0.775	0.816
Two Class Decision Jungle	0.850	0.943	0.825	0.880
Two Class Locally Deep Support Vector Machine	0.767	0.861	0.775	0.816
Two Class Neural Network	0.867	0.944	0.850	0.895
Two Class Perceptron	0.867	0.944	0.850	0.895
Two Class Support Vector Machine	0.800	0.912	0.775	0.838

In this study, we compared several machine learning algorithms to predict heart disease using Microsoft Azure Machine Learning Studio. The models we tested include Two Class Bayes Point, Boosted Decision Tree, Decision Forest, Decision Jungle, Locally Deep Support Vector Machine (SVM), Neural Network, Perceptron, and Support Vector Machine (SVM). To evaluate how well each model performed, we looked at four key metrics: accuracy, precision, recall, and the F1 score, as shown in Table 1.

Three models—Two Class Bayes Point, Neural Network, and Perceptron—stood out, each achieving the highest accuracy of 86.7%, with a precision of 94.4%, recall of 85.0%, and an F1 score of 89.5%. These results indicate that these models strike a great balance between identifying true positive cases (precision) and covering most of the actual cases (recall), making them reliable for predicting heart disease.

The Decision Jungle model also performed well, with an accuracy of 85.0%, a precision of 94.3%, a recall of 82.5%, and an F1 score of 88.0%. While it slightly trailed the top three models, it still showed strong predictive capabilities.

Boosted Decision Tree and Support Vector Machine (SVM) models had moderate performance, with accuracies of 81.7% and 80.0%, respectively. Both had high precision (91.4% and 91.2%), but their recall rates were lower, resulting in F1 scores of 85.3% and 83.8%.

On the lower end were the Decision Forest and Locally Deep SVM models, both with an accuracy of 76.7%, precision of 86.1%, recall of 77.5%, and F1 scores of 81.6%. These models were less effective at making accurate predictions and balancing precision and recall compared to the others.

CONCLUSION

In conclusion, this study provides valuable insights into the effectiveness of different machine learning models for predicting heart disease using Microsoft Azure Machine Learning Studio. The Two Class Bayes Point, Neural Network, and Perceptron models stood out as the top performers, with each achieving an accuracy of 86.7%, along with high precision (94.4%) and a well-balanced recall (85.0%). These models proved to be the most reliable options for accurately identifying heart disease, making them strong candidates for use in healthcare settings where both precision and recall are critical.

The Decision Jungle model also showed promising results, with an accuracy of 85.0%, making it a viable alternative to the top three models. However, models like Boosted Decision Tree and Support Vector Machine (SVM), while maintaining high precision, had lower recall rates, which could limit their ability to catch all cases of heart disease. Lastly, Decision Forest and Locally Deep SVM had lower overall performance, making them less suitable for this specific application.

Overall, this study suggests that the Bayes Point, Neural Network, and Perceptron models offer the most reliable framework for predicting heart disease. These models could be useful in real-world clinical applications, helping healthcare providers make more accurate diagnoses. Future research could build on these findings by testing these models with larger datasets or integrating them with other diagnostic tools to further improve accuracy and usability in healthcare environments.

REFERENCES

[1] Almulihi, A. H., Alharthi, A. S., Almulihi, A. H., & Almulihi, H. H. (2023). Artificial Intelligence in Healthcare: A Comprehensive Review. Science, 2(5), 1-13. https://doi.org/10.24017/science.2023.2.5

[2] Hasan, R., & Faridi, M. A. (2021). Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction. ITM Web of Conferences, 40, 03007. https://doi.org/10.1051/itmconf/20214003007

[3] Jiang, Y., Jiang, H., Ding, Y., Wang, Y., Bai, C., & Xue, Y. (2023). Applications of artificial intelligence in intelligent biomedicine: A comprehensive review. Intelligent Biomedicine, 3, 100100. https://doi.org/10.1016/j.ibmed.2023.100100

[4] Alharthi, A. S., Almulihi, A. H., Almulihi, A. H., & Almulihi, H. H. (2022). A Comprehensive Review of Artificial Intelligence in Healthcare. Healthcare, 10(6), 1137. https://doi.org/10.3390/healthcare10061137

[5] Amiri, M., Ghiyasvandian, S., Manoochehri, H., & Ebadi, A. (2023). Concept Analysis of Nursing Care Behaviors: A Hybrid Model. Frontiers in Health Informatics, 12, 402. http://dx.doi.org/10.30699/fhi.v12i0.402