BRAC UNIVERSITY

Inspiring Excellence

CSE 422: Artificial Intelligence

Predicting Heart Disease using personal key indicators with the help of Machine Learning

H.M. Sarwer Alam

20301224

hm.sarwer.alam@g.bracu.ac.bd

Table of Contents

Introduction

This project aims to develop a machine-learning model to predict heart disease using personal key indicators. We collect data on personal indicators associated with heart disease, preprocess and engineer the features, and train and evaluate machine learning models. The best model is deployed in a user-friendly interface to provide personalized heart disease assessments and recommendations, enabling individuals to make informed decisions about their health and lifestyle.

Motivation

Heart failure is a significant global health issue, affecting millions of people worldwide and placing a significant burden on healthcare systems. Early detection and intervention are critical to improving patient outcomes and reducing healthcare costs associated with heart failure management. Machine learning (ML) has the potential to revolutionize heart failure detection by leveraging large datasets and advanced algorithms to identify patterns and predict heart failure risk.

The motivation for developing a heart failure detection ML project stems from several key factors. First, heart failure is often asymptomatic or presents with nonspecific symptoms, making it challenging to diagnose in its early stages. ML algorithms can analyze a wide range of clinical data, such as electronic health records, medical imaging, and patient demographics, to identify subtle patterns and indicators of heart failure risk that may not be readily apparent to clinicians.

Second, timely detection of heart failure can enable early intervention and prevent disease progression. ML models can be trained on large datasets to predict heart failure risk with high accuracy, allowing healthcare providers to identify at-risk patients and implement targeted interventions, such as lifestyle modifications, medication adjustments, or referral to specialists, to prevent or mitigate heart failure development.

Third, heart failure imposes a substantial economic burden on healthcare systems, including hospitalization costs, medications, and long-term care. Early detection of heart failure can lead to cost savings by reducing hospitalizations and emergency room visits, optimizing treatment plans, and improving patient outcomes.

In summary, developing a heart failure detection ML project has the potential to revolutionize healthcare by enabling early detection and intervention, improving patient outcomes, and reducing healthcare costs associated with heart failure management. By leveraging the power of ML, we can enhance the accuracy and efficiency of heart failure detection, ultimately leading to better patient care and outcomes.

Data Description

Link:https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease?resource=download

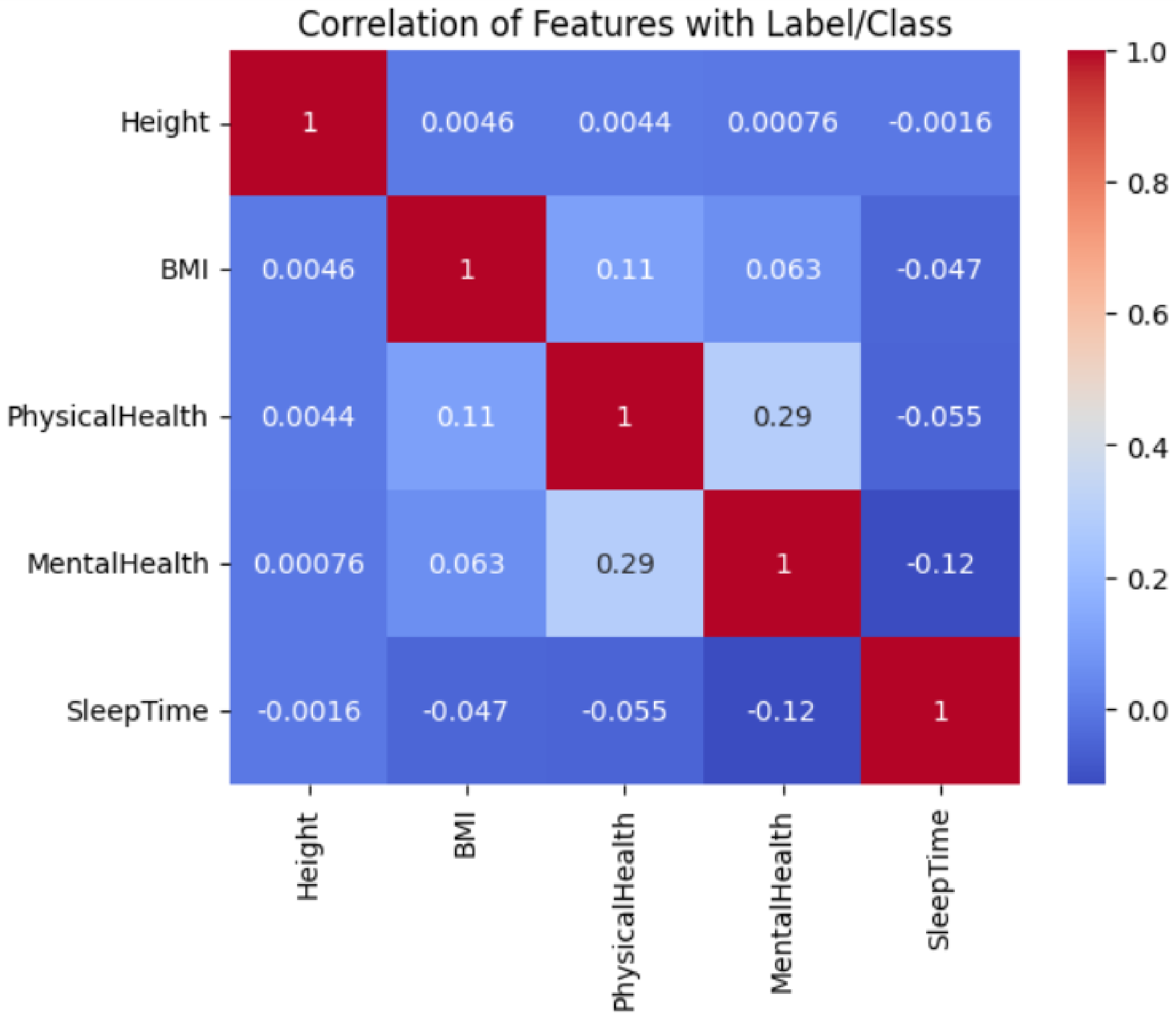
Number of Features: 19

Type of class/label: Categorical and Continuous

Number of data points: 85441

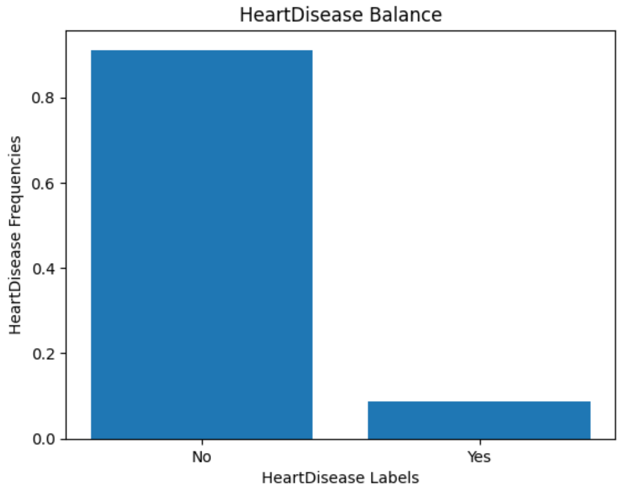
Types of features: 19

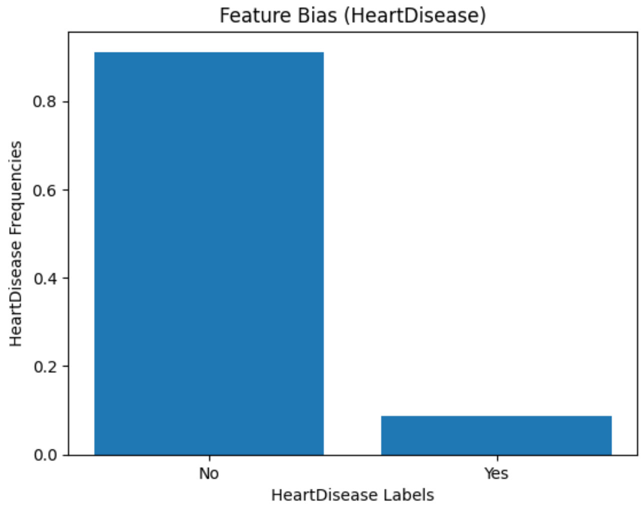
Correlation of the features along with the label/class:



Here, A positive correlation (close to 1) indicates that as the values of a feature increase, the values of the label/class also tend to increase, and vice versa.A negative correlation (close to -1) indicates that as the values of a feature increase, the values of the label/class tend to decrease, and vice versa. A correlation close to 0 indicates little or no linear relationship between the feature and the label/class.

Biasness/Balanced



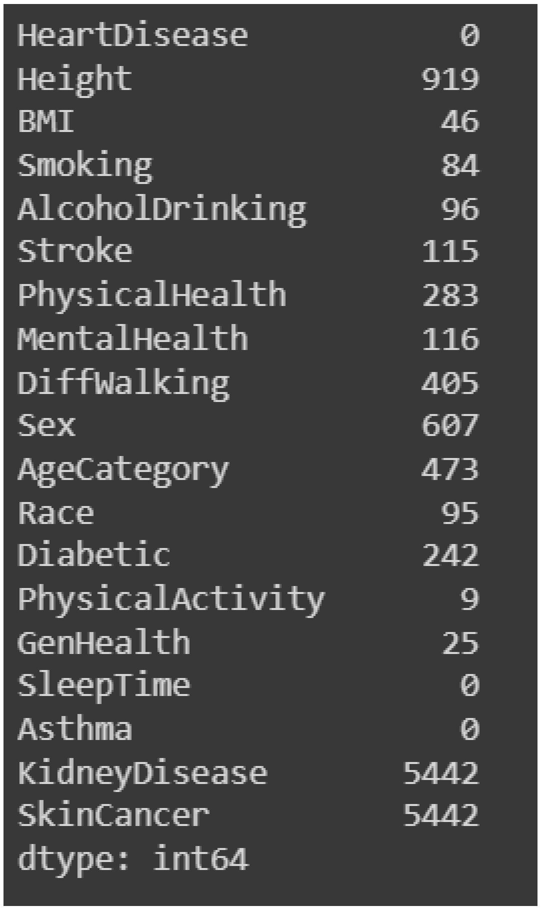


Like this bar chart all the classes' bias/ balance part was checked and the result was the database was biased.

Dataset pre-processing

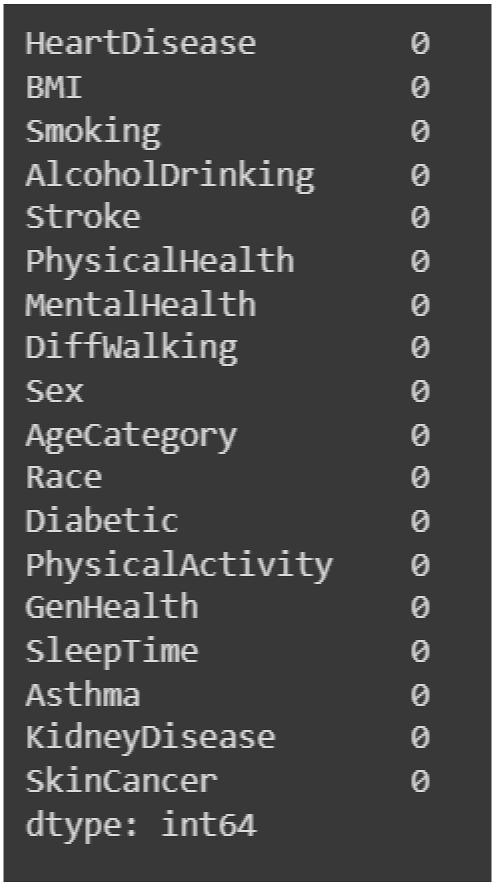
Problem 1 : 16 features had totally 14399 null values.
Solutions: Delete rows

Problem 2: As there is BMI, so height is a extra column which actually needs to calculate BMI and no other uses.
Solution: Delete Column



```
HeartDisease          0
Height              919
BMI                  46
Smoking              84
AlcoholDrinking      96
Stroke              115
PhysicalHealth      283
MentalHealth        116
DiffWalking         405
Sex                 607
AgeCategory         473
Race                 95
Diabetic            242
PhysicalActivity      9
GenHealth            25
SleepTime             0
Asthma                0
KidneyDisease      5442
SkinCancer         5442
dtype: int64
```

Before Pre processing



```
HeartDisease        0
BMI                 0
Smoking             0
AlcoholDrinking     0
Stroke              0
PhysicalHealth      0
MentalHealth        0
DiffWalking         0
Sex                 0
AgeCategory         0
Race                0
Diabetic            0
PhysicalActivity    0
GenHealth           0
SleepTime           0
Asthma              0
KidneyDisease       0
SkinCancer          0
dtype: int64
```

After Pre Processing

Dataset splitting

To train the model data splitting was done as like 80% for training model and 20% for testing.

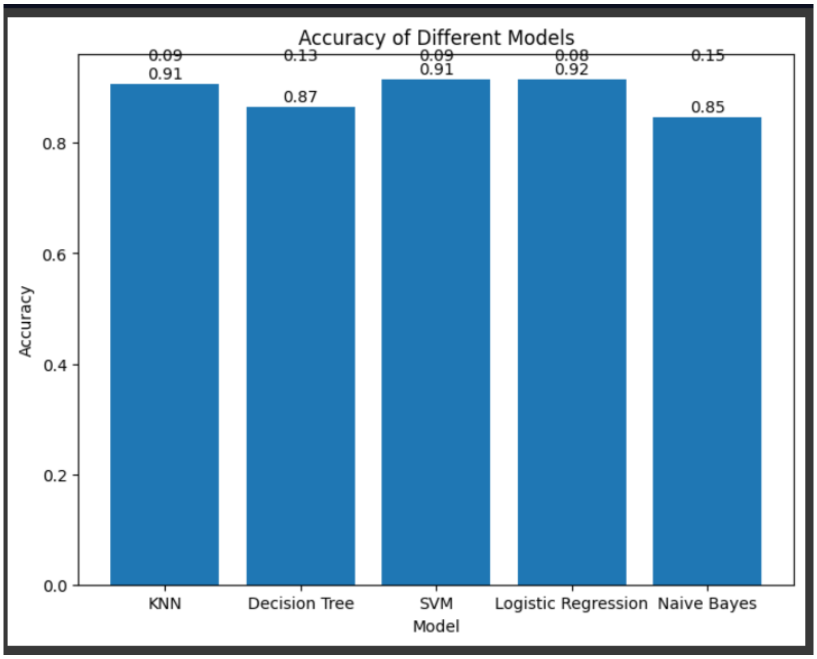
On this basis- Training data set - 61993 and Testing data set - 15499

Model training

Model Name	Accuracy (%)	Error(%)
Logistic Regression	91.57	8.43
Decision tree	86.56	13.44
Kth Nearest Neighbor	90.68	9.32
SVM Model	91.48	8.52
Naive Bayes Classifier	84.57	15.43

From the table, we can see that the Logistic Regression model showed the best performance with 91.57% accuracy and only 8.43% error. On the other hand, the worst performance was given by Naive Bayes Classifier Model which has only 84.57% accuracy and the error was double that of the Logistic Regression model. After that, Kth Nearest Neighbor and SVM Model’s performance was mostly satisfying but the decision tree also had a bad performance with 86.56% accuracy.

Model selection/Comparison analysis



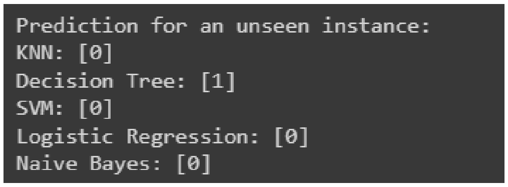
Also, from the Bar comparison we can see that Logistic Regression has the best performance than other models.

Model testing

To taste model, the data was used-

1. Single instance - [35.15, 1, 0, 0, 0, 0, 0, 1, 10, 5, 0, 1, 2, 8, 0, 0, 0]

   Result:

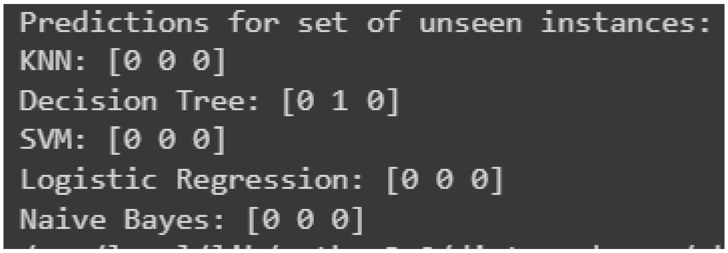
   ```
   Prediction for an unseen instance:
   KNN: [0]
   Decision Tree: [1]
   SVM: [0]
   Logistic Regression: [0]
   Naive Bayes: [0]
   ```

   The result matches with our desired result. All the models gave the right answer except the Decision Tree.

2. Set of Instances- [32.36, 0, 0, 0, 0, 0, 0, 1, 10, 5, 2, 1, 0, 8, 1, 0, 0], [32.61, 1, 0,

0, 0, 0, 0, 0, 9, 5, 0, 1, 1, 5, 1, 0, 1], [26.04, 0, 0, 0, 0, 0, 0, 0, 11, 2, 2, 1, 4, 7, 0, 0, 0]

Result-



```
Predictions for set of unseen instances:
KNN: [0 0 0]
Decision Tree: [0 1 0]
SVM: [0 0 0]
Logistic Regression: [0 0 0]
Naive Bayes: [0 0 0]
```

Here 1st and 3rd was no. So the result is correct but the 2nd's was yes which was made by all models except the decision tree.

Conclusion

In conclusion, the results obtained from applying various machine learning algorithms such as Logistic Regression, Decision Tree, Kth Nearest Neighbor, SVM Model, and Naive Bayes Classifier to predict heart disease using personal key indicators are promising. With accuracy ranging from 84.57% to 91.57%, these models demonstrate good performance in predicting heart disease.

Among the models tested, Logistic Regression and SVM Model performed the best with accuracy rates of 91.57% and 91.48% respectively, indicating their potential as effective tools for heart disease prediction. Kth Nearest Neighbor also showed promising results with an accuracy of 90.68%. Although Decision Tree achieved an accuracy of 86.56%, it may require further refinement to improve its predictive performance. Naive Bayes Classifier, with an accuracy of 84.57%, showed the lowest accuracy among the models tested, suggesting that it may not be the most suitable choice for heart disease prediction based on personal key indicators.

These findings highlight the potential of machine learning techniques in predicting heart disease using personal key indicators. However, further research and validation using larger datasets and additional evaluation metrics are necessary to ensure the reliability and generalizability of these models in real-world clinical settings. Nevertheless, the results obtained from this study provide a valuable foundation for future research and practical applications of machine learning in predicting heart disease, which has the potential to significantly contribute to early detection and prevention of this critical health condition.

Future work/Extension

The prediction of heart disease using personal key indicators with the help of machine learning has shown promising results with high accuracy rates obtained from various algorithms. However, there are several potential future works and extensions that can be explored to further improve the accuracy and applicability of these models.

Feature Selection and Feature Engineering: In the current study, a specific set of personal key indicators were used for prediction. Further research can focus on identifying and selecting the most relevant features or personal key indicators that have the highest predictive power for heart disease. This can be achieved through advanced feature selection techniques such as Recursive Feature Elimination, Principal Component Analysis, or domain-specific knowledge-based feature engineering, which can help to improve the accuracy of the models.

Ensemble Methods: Ensemble methods such as Random Forest, Gradient Boosting, or Stacking can be employed to combine the predictions of multiple base models, such as Logistic Regression, Decision Tree, Kth Nearest Neighbor, SVM Model, and Naive Bayes Classifier, to create a more robust and accurate prediction model. Ensemble methods have the potential to leverage the strengths of different models and reduce the weaknesses, leading to improved performance.

Hyperparameter Tuning: Hyperparameters are crucial parameters of machine learning algorithms that control their behavior. Optimizing hyperparameters can significantly impact the performance of the models. Future research can focus on fine-tuning the hyperparameters of the models used in this study to achieve better accuracy. Techniques such as Grid Search, Random Search, or Bayesian Optimization can be employed for hyperparameter tuning.

Validation on Larger Datasets: The accuracy of machine learning models is often affected by the size and quality of the dataset used for training and validation. Future work can involve validating the performance of the models on larger and more diverse datasets to assess their robustness and generalizability in real-world clinical settings. This can help to further validate the findings and ensure that the models can be effectively applied to different populations.

Real-world Implementation and Clinical Validation: Once the models are optimized and validated on larger datasets, real-world implementation and clinical validation can be pursued to assess the practical applicability of these models. This can involve collaborating with healthcare institutions, collecting data from real patients, and conducting prospective studies to evaluate the performance of the models in real clinical scenarios.

Interpretability and Explainability: Another important aspect of machine learning models is their interpretability and explainability. Future work can focus on developing models that are not only accurate but also interpretable and explainable, allowing healthcare professionals to understand the underlying mechanisms and reasoning behind the predictions. This can enhance the trust and acceptance of these models in clinical practice.

In conclusion, while the current study has shown promising results in predicting heart disease using personal key indicators with the help of machine learning, there are several future works and extensions that can be explored to further enhance the accuracy, robustness, and applicability of these models. Continued research and validation in this area have the potential to significantly contribute to early detection and prevention of heart disease, leading to improved patient outcomes and better healthcare decision-making.

References

(n.d.). scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation. Retrieved April 18, 2023, from https://scikit-learn.org/stable/index.html

HPlot a DataFrame using Pandas – Data to Fish. (n.d.). Data to Fish. Retrieved April 18, 2023, from https://datatofish.com/plot-dataframe-pandas/