HeartInspect: Heart Disease Prediction of an Individual Using Naïve Bayes Algorithm

Nurbaity Sabri
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Melaka.
nurbaity_sabri@uitm.edu.my

Anis Amilah Shari
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Melaka.
anisamilah@uitm.edu.my

Khyrina Airin Fariza Abu Samah
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Melaka
khyrina783@uitm.edu.my

Mohd Rahmat Mohd Noordin
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Melaka
mrahmat.noordin@uitm.edu.my

Asma Shazwani Shari,
Faculty of Bussiness Management, Universiti Teknologi MARA Cawangan Kelantan.
asma663@uitm.edu.my

Fadhilah Mohd Ishak@Zainudin
Faculty of Bussiness Management, Universiti Teknologi MARA Cawangan Kelantan.
fadhi513@uitm.edu.my

Wan Masnieza Wan Mustapha
Faculty of Bussiness Management, Universiti Teknologi MARA Cawangan Kelantan
masnieza@uitm.edu.my

Muhammad Fudhail Afiq Nor Rozaini Affendi
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Melaka
fudhailafiq11@gmail.com

2023 IEEE 11th Conference on Systems, Process & Control (ICSPC) | 979-8-3503-4086-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSPC59664.2023.10420149

***Abstract*— Heart disease is a serious health issue that contributes significantly to the high death worldwide. Therefore, the creation of a reliable system for heart disease prediction is essential for early intervention and better results. Such technologies can help identify at-risk persons and enable prompt preventive interventions by utilizing cutting-edge algorithms and analysing pertinent data. Nonetheless, predicting heart disease is a difficult endeavour, especially in underdeveloped regions with few diagnostic tools and a shortage of trained medical workers. Moreover, the healthcare sector produces a tremendous quantity of data on cardiac disease, yet these important resources are frequently underutilized when it comes to making well-informed decisions. Additionally, pricey heart diagnostics like electrocardiograms are now out of reach for the typical person due to increased living expenses. This present work suggests a Naive Bayes-based cardiac disease prediction system as a solution to these problems. The system makes use of a dataset that includes information about a person's heart disease status, height, weight, physical health, difficulties walking, age group, physical activity, general health, and sleep duration. For training and testing purposes, the dataset is partitioned 80/20. The dataset is analysed using the Naive Bayes technique to determine the chance of cardiac disease. Despite assuming independence among the characteristics, the system shows promising performance, reaching about 71–73% precision in heart disease prediction. Even though this falls short of higher standards, it is nevertheless a noteworthy accomplishment in light of the difficulties mentioned in the problem statements. In summary, the Naive Bayes algorithm-based heart disease prediction system reported in this work shows promise for predictions that are 71–73% accurate. This method helps address the problems involved with cardiac disease prediction, particularly in environments with limited resources, by making use of the data that is currently available and overcoming resource constraints.**

Keywords*— *Big Data, heart disease, prediction systems, machine learning,data mining, web app, python, naïve bayes.

I. INTRODUCTION

Heart disease is one of the most lethal types of disease that a person can be diagnosed with. It is very dangerous because it refers to a person's heart and the heart is one of the essential organs that makes a person live. People today tend to get caught up in the daily grind of work and other responsibilities while neglecting their health. Subsequently, they are becoming sick more frequently daily because of their hurried lifestyles and disregard for their health. Additionally, most people suffer from cardiac disease. According to data provided by the World Health Organization (WHO), heart-related diseases account for over 31% of all fatalities worldwide [1]. Thus, it is undeniable that the heart disease trend in this current world is a serious matter to be attended to. In relation to this, the prediction and diagnosis of heart disease are exceedingly challenging especially in poor nations due to the uncommon availability of effective diagnostic instruments and a lack of qualified medical personnel [2]. The primary contributing causes of deaths associated with heart disease are inadequate preventative measures and a shortage of qualified or skilled health personnel.Huge amounts of data on heart disease are generated by the healthcare industry, but regrettably these data are not searched for to find insider information for wise decision-making [3].

A step to assist in detecting early prediction trends for heart disease is the solution by developing this present project. The system to be developed is a web-based application for heart disease prediction. The target is to facilitate in early-stage indication of heart disease to make diagnosis easier and more effective. This project is conducted based on data of patients regarding their status of heart disease and health which is acquired from the Kaggle website [4]. The system is developed based on important attributes filtered from the dataset such as the status of heart disease of an individual, their Height, Weight, Physical Health, Difficulty Walking, Age Category, Physical Activity, General Health, and Sleep Time.

II. APPROACH

Firstly, the title for this project was decided, which is the Heart Disease Prediction of an Individual using Naïve Bayes Algorithm and the project area Big Data Analytic and Data Visualization was determined to suit the stated title. The literature review was done to have a clearer view for the principles of the project and the process of studying as well as looking into previous works like this project, resulting to three chosen articles. Then, the problem statements, objectives and purposes, scope and significances were identified to observe why the system was being made and why it is advantageous to be developed. Figure 1 outlines a streamlined process for health prediction using essential input parameters such as height, weight, physical health, age, and more. Initially, the system calculates the Body Mass Index (BMI) based on the user's height and weight. Subsequently, the workflow connects to a prediction model employing a machine learning algorithm, specifically the Naive Bayes classifier. This model predicts the likelihood of heart disease, leading to a pivotal decision point. If the prediction is negative for heart disease, the process follows one path; alternatively, if positive, it takes a different route. Ultimately, the flowchart encapsulates a clear, systematic approach to health prediction through a machine learning model. The input parameters guide the calculation of BMI and subsequent connection to a prediction model using a machine learning algorithm, the Naive Bayes classifier. This predictive tool efficiently processes the gathered data, providing insights into the probability of heart disease. Based on this prediction, the flowchart branches into distinct outcomes, delineating the end of the predictive process. This succinct flowchart delineates an efficient and structured approach towards health prediction, underlining the potential of machine learning in enhancing healthcare assessments.

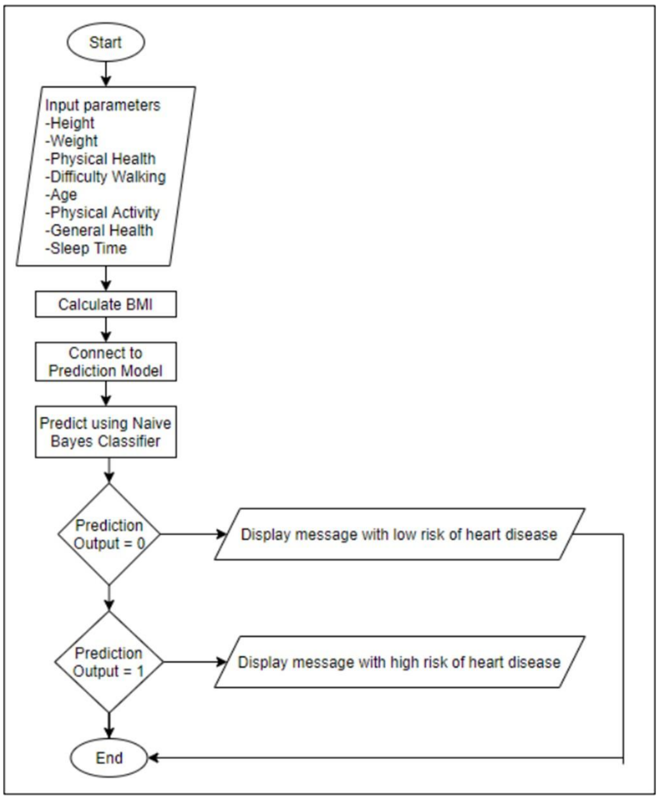
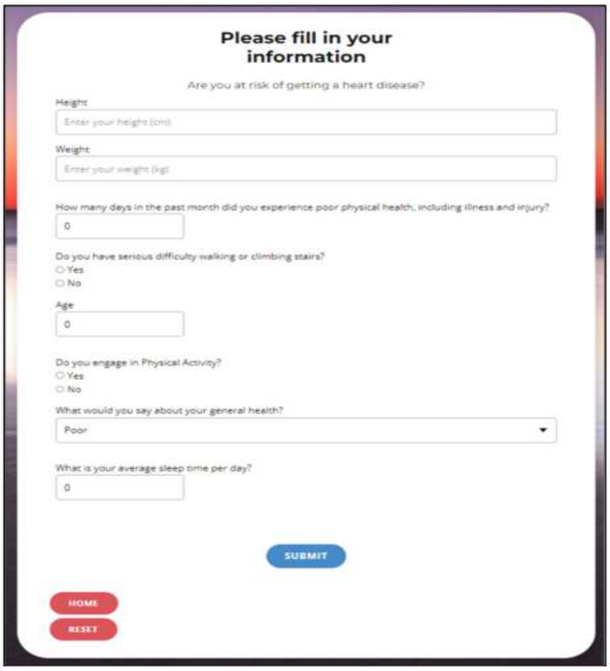


Figure 1. System Flowchart

III. Methods and Resources

A. Datasets

CategoricalNB was used because most of the attributes in the dataset consists of categorical varibles. It examines data pre-processing methods, model training procedures, and evaluation measures. To produce precise predictions, the CategoricalNB algorithm, created for categorical data, was used. In order to make predictions, continuous data were converted into categorical variables. Categorical variables were encoded into numerical form using the Sci-kit library's LabelEncoder. With the help of this encoding procedure, categorical data may be represented numerically, making it compatible with the machine learning algorithm that needs numerical inputs. Using the RandomUnderSampler class from the imbalanced-learn package, the approach was used to resolve the class imbalance. Then, by randomly lowering the samples of the majority class, the data were resampled to produce a balanced dataset. The dataset was divided into training and testing sets, with 80% of the data going towards training and 20% going towards testing the prediction model. Then, CategoricalNB function was called to run the dataset and model. Using the RandomUnderSampler class from the imbalanced-learn package, the RandomUnderSampler approach is used to resolve the class imbalance. RandomUnderSampler is a function from the imblearn scikit library. Flask, a popular Python web framework, served as the foundation for developing the heart disease prediction web application. Leveraging its flexibility and ease of use, the developers utilized Flask to create a robust and efficient backend for the web app. This framework allowed for smooth handling of HTTP requests, enabling the application to seamlessly interact with the prediction model. The entire development process took place within Visual Studio Code (VS Code), known for its extensive features and support for Python. VS Code's intuitive interface and debugging capabilities streamline the development, ensuring a well-functioning and user-friendly web application for predicting heart disease risk. Figure 2 displays the intuitive and user-friendly input form of our Heart Disease Prediction Web Application. This interface is meticulously designed to collect crucial health-related data from users, enabling a precise evaluation of their potential risk of heart disease. The form prompts users to input significant parameters such as height, weight, physical health status, physical activity level, age, sleep duration, general health condition, and difficulty in walking, encompassing a comprehensive range of factors essential for an accurate prediction. Each input holds unique importance in assessing the user's heart disease risk, making this interface a vital tool in promoting proactive healthcare management and early risk detection.



Please fill in your information

Are you at risk of getting a heart disease?

Height

Enter your height (cm)

Weight

Enter your weight (kg)

How many days in the past month did you experience poor physical health, including illness and injury?

0

Do you have serious difficulty walking or climbing stairs?

○ Yes
○ No

Age

0

Do you engage in Physical Activity?

○ Yes
○ No

What would you say about your general health?

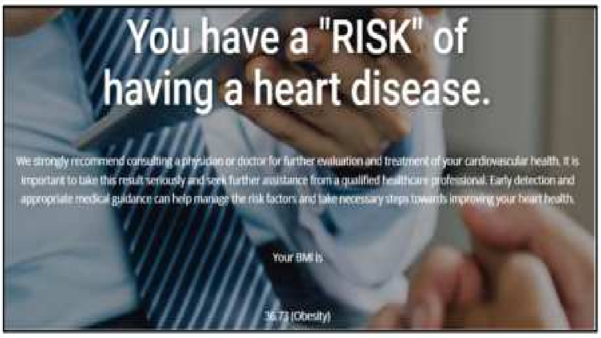
Poor ▾

What is your average sleep time per day?

0

Figure 2. System Main Page

In Figure 3, we present the output page of our Heart Disease Prediction Web Application, which provides users with crucial insights into their heart disease risk assessment. Alongside the user's calculated Body Mass Index (BMI), the page presents a detailed description of the predicted outcome, indicating the likelihood of heart disease based on the provided input parameters. The output furnishes users with actionable information, empowering them to make informed decisions about their cardiovascular health. The inclusion of the BMI value offers an additional reference point, enhancing the comprehensibility and relevance of the predicted outcome. This user-centric design aims to promote health awareness and encourage proactive measures for a healthier lifestyle.



You have a "RISK" of having a heart disease.

We strongly recommend consulting a physician or doctor for further evaluation and treatment of your cardiovascular health. It is important to take this result seriously and seek further assistance from a qualified healthcare professional. Early detection and appropriate medical guidance can help manage the risk factors and take necessary steps towards improving your heart health.

Your BMI is

36.73 (Obesity)

Figure 3. Display Results

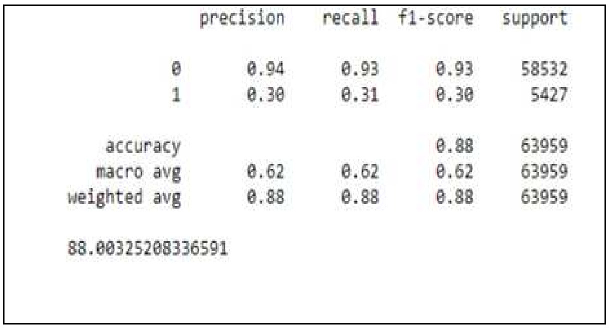
IV. Evaluation

A. Classification Model

From the types of Naive Bayes algorithms, CategoricalNB is used because it is built and suited for categorical variables. The Naïve Bayes algorithm can be helpful in the prediction of heart diseases or any disease since it can manage the missing data greatly and when the independence of attributes is given, it gives this model an edge. It can operate with larger datasets and is quick to train and classify [1].

B. Prediction Model

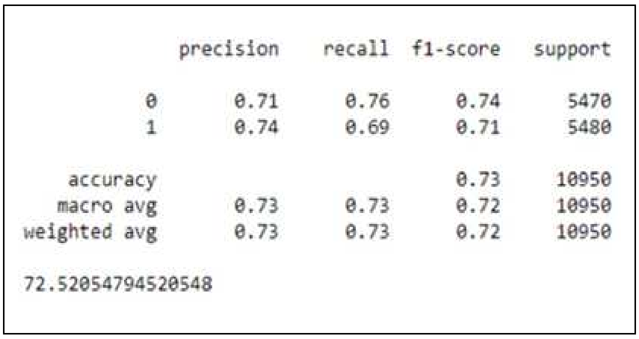
To effectively assess the risk of developing heart disease, it is important to have a comprehensive understanding of the prediction model's accuracy as well as any potential practical implications. Figure 4 shows the results for model prediction model with no process of balancing and sampling the data. Although the accuracy obtained is high, 88%, was not used because the model would have a biased prediction on one instance of the class attribute which is towards the prediction of “No” with the precision of 94% while the “Yes” prediction has a low precision of only 30%.



```
              precision    recall  f1-score   support

           0       0.94      0.93      0.93     58532
           1       0.30      0.31      0.30      5427

    accuracy                           0.88     63959
   macro avg       0.62      0.62      0.62     63959
weighted avg       0.88      0.88      0.88     63959

88.00325208336591
```

Figure 4. Imbalance Model Result

To not let the prediction model be biased towards one prediction class, random sampling method was used to balance both instances of the class attribute. From initially having 292422 instances of individuals not having a heart disease, only 27373 data remain after random sampling is made. This corresponds to the number of individuals with a heart disease from the dataset, which contains 27373 instances. Figure 5 shows the results of model evaluation after the data has been balanced. The accuracy obtained was 72.52%.



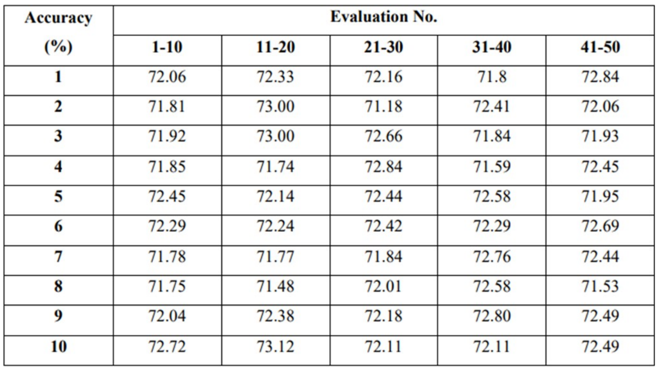
```
              precision    recall  f1-score   support

           0       0.71      0.76      0.74      5470
           1       0.74      0.69      0.71      5480

    accuracy                           0.73     10950
   macro avg       0.73      0.73      0.72     10950
weighted avg       0.73      0.73      0.72     10950

72.52054794520548
```

Figure 5. Balance Model Result

C. Accuracy

Due to the data being randomly sampled for every run and evaluation, therefore a different accuracy was obtained every time. Table 1 shows 50 different model accuracy after 50 evaluations. From the observation, all the accuracies have a small difference range between 71% to 73%. From the accuracy discussion, it is concluded that the accuracy of the completed prediction model is a range from 71% to 73%.

Table 1: Accuracy Result



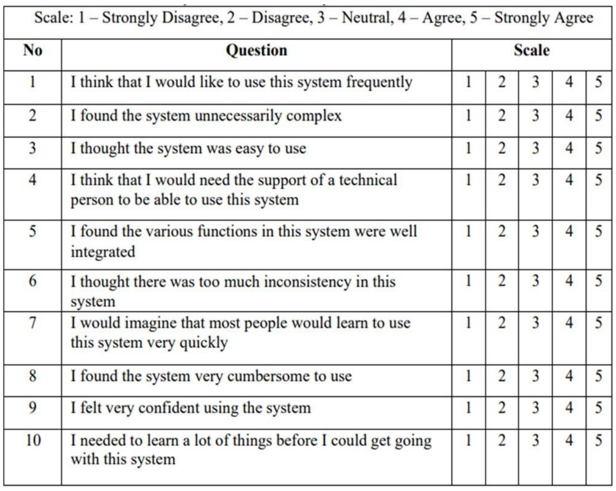
Accuracy (%)	Evaluation No.				
	1-10	11-20	21-30	31-40	41-50
1	72.06	72.33	72.16	71.8	72.84
2	71.81	73.00	71.18	72.41	72.06
3	71.92	73.00	72.66	71.84	71.93
4	71.85	71.74	72.84	71.59	72.45
5	72.45	72.14	72.44	72.58	71.95
6	72.29	72.24	72.42	72.29	72.69
7	71.78	71.77	71.84	72.76	72.44
8	71.75	71.48	72.01	72.58	71.53
9	72.04	72.38	72.18	72.80	72.49
10	72.72	73.12	72.11	72.11	72.49

This system evaluates the usability of the heart disease prediction system using the System Usability Scale (SUS). SUS is a well- known questionnaire-based evaluation method that offers insightful information about the efficacy, efficiency, and user-friendliness of the system.

The goal is to administer SUS to a wide range of users to gain feedback on their opinions and experiences, enabling a thorough analysis of the system's usability. SUS comprises of standardized statements that are evaluated on a 5-point Likert scale and provides analytically useful quantitative data. The evaluation includes users. The deployment of SUS makes it possible to evaluate the heart disease prediction system's strengths and limitations, collect user feedback, and offer insightful suggestions for improvement.

The evaluation's findings are the system's user experience for the heart disease prediction system. Table 2 shows all the questions of the system usability scale questionnaire and table 3 is the results for the usability testing. According to [5], the average score on the SUS is 68%, meaning that if the score is at or above 68%, the system has good system usability. As a result, the system's great score of more than 68%, or 85%, shows that it has good usability. Most of the testers were positive of the system.

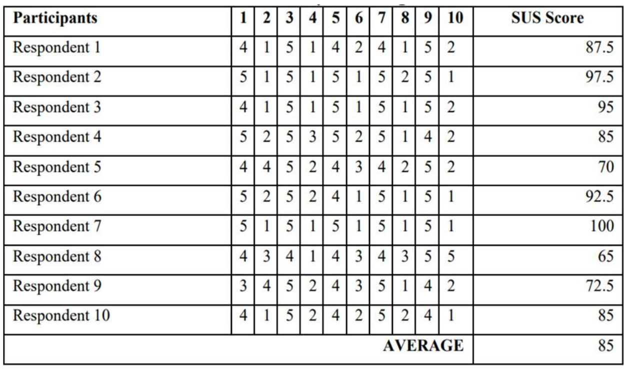
Table 2: System Usability Scale Questionnaire



Scale: 1 – Strongly Disagree, 2 – Disagree, 3 – Neutral, 4 – Agree, 5 – Strongly Agree

No	Question	Scale				
1	I think that I would like to use this system frequently	1	2	3	4	5
2	I found the system unnecessarily complex	1	2	3	4	5
3	I thought the system was easy to use	1	2	3	4	5
4	I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
5	I found the various functions in this system were well integrated	1	2	3	4	5
6	I thought there was too much inconsistency in this system	1	2	3	4	5
7	I would imagine that most people would learn to use this system very quickly	1	2	3	4	5
8	I found the system very cumbersome to use	1	2	3	4	5
9	I felt very confident using the system	1	2	3	4	5
10	I needed to learn a lot of things before I could get going with this system	1	2	3	4	5

Table 3: Usability Testing Result



Participants	1	2	3	4	5	6	7	8	9	10	SUS Score
Respondent 1	4	1	5	1	4	2	4	1	5	2	87.5
Respondent 2	5	1	5	1	5	1	5	2	5	1	97.5
Respondent 3	4	1	5	1	5	1	5	1	5	2	95
Respondent 4	5	2	5	3	5	2	5	1	4	2	85
Respondent 5	4	4	5	2	4	3	4	2	5	2	70
Respondent 6	5	2	5	2	4	1	5	1	5	1	92.5
Respondent 7	5	1	5	1	5	1	5	1	5	1	100
Respondent 8	4	3	4	1	4	3	4	3	5	5	65
Respondent 9	3	4	5	2	4	3	5	1	4	2	72.5
Respondent 10	4	1	5	2	4	2	5	2	4	1	85
AVERAGE											85

V. Conclusion

The heart disease prediction system project has, in the end, effectively accomplished all its goals. By designing an intuitive and user-friendly interface, the first goal, to develop a web-based application for the prediction system, was completed. The second goal was to create a functional and easily accessible system by concentrating on the web-based application. Finally, the app's usability and functionality underwent a thorough evaluation, which supported its effectiveness in correctly forecasting heart disease. The study was successful in producing a dependable and effective method for predicting heart disease by achieving these goals. The heart disease prediction system addresses several critical problem statements. Firstly, the challenge of accurately predicting and diagnosing heart disease, particularly in resource-limited settings, has been addressed through the development of an accessible web application. Secondly, the abundance of heart disease data generated by the healthcare industry can now be leveraged for wise decision-making, thanks to the system's utilization of this valuable information. Lastly, the system tackles the issue of expensive cardiac tests by providing an affordable and practical alternative for individuals to assess their risk of heart disease. By offering a web application, utilizing data for prediction modelling, and ensuring accessibility and affordability, the heart disease prediction system effectively solves these problems, making it a valuable tool in the healthcare domain. Future research will focus on creating output categories with greater nuance than a simple "yes" or "no" prediction to improve the heart disease prediction system. The inclusion of risk categories like low, medium, high, and very high risk would give consumers more in-depth knowledge about their risk levels for heart disease, empowering them to make wise decisions about their health. To accomplish this, the prediction model can be updated, and the output generation method can be changed to place people in the right risk groups based on their predicted probability or pertinent characteristics. The system would provide thorough insights into a person's level of heart disease risk by including these revised risk categories, allowing for better risk assessment and tailored recommendations.

ACKNOWLEDGMENT

The corresponding author can be contacted through anisamilah@uitm.edu.my.

REFERENCES

[1] Katarya, R., & Meena, S. K. (2021). Machine learning techniques for heart disease prediction: A comparative study and analysis. Health and Technology, 11(1), 87–97. https://doi.org/10.1007/s12553-020-00505-7

[2] Yahaya, L., David Oye, N., & Joshua Garba, E. (2020). A comprehensive review on heart disease prediction using data mining and machine learning techniques. American Journal of Artificial Intelligence, 4(1), 20. https://doi.org/10.11648/j.ajai.20200401.12

[3] Islam, M. T., Rafa, S. R., & Kibria, M. G. (2020, December 19). Early prediction of heart disease using PCA and hybrid genetic algorithm with k-means. ICCIT 83 2020 - 23rd International Conference on Computer and InformationTechnology,Proceedings. https://doi.org/10.1109/ICCIT51783.2020.9392655

[4] Mohaimin, Md.M., (2022). Heart disease EDA + prediction. [Data File]. Retrieved November 3, 2022,from https://www.kaggle.com/code/ mushfirat/heartdisease-eda-prediction/notebook.

[5] Alathas, H. (2018). How to measure product usability with the system usability scale (SUS) score. UX Planet. Retrieved July 18, 2023, from https://uxplanet.org/how-to-measureproduct-usability-with-the-systemusability- scale-sus-score-69f3875b858f