**A picture containing logo

Description automatically generated**

**COMP 8157 Advanced Database Topics**

**Group Project-FINAL REPORT**

**PREDCTIVE ANALYSIS OF HEART DISEASE USING MACHINE LEARNING ALGORITHMS**

**STUDENT NAME:**

Ahmed Abdullah (110068897)

Ahmed Mustafa (110068431)

Alekhya Ryali (110084315)

Karishma Chowdary Cherukuri (110076472)

**GROUP NAME:**

DATA DIVE

**SUBMITTED TO:**

Dr. Shafaq Khan

**1. Abstract:**

Cardiovascular diseases are common these days and they require a precise diagnosis at a precise moment. Every day, the health care sector generates enormous amounts of data about patients and diseases. To effectively extract knowledge from these databases and apply this knowledge for more precise diagnosis and decision-making, machine learning approaches and technologies are available, but it is important to identify the best way to forecast it.

In our research analysis, we have picked a dataset from the 2020 annual CDC survey data that contains 400k adults’ data for predicting factors that lead to heart disease. This project mainly focuses on identifying the variety of variables that are closely related to cardiac problems, analyzing classification ML models with the most recent data, predicting cardiac problems, and finally drawing conclusions based on the accuracy metrics produced by the machine learning techniques.

Classification ML models like Logistic Regression, Naïve Bayes, XGBoost, Decision Tree, and Random Forest were implemented for analyzing the dataset based on the risk factors. After preprocessing and implementation of data on the models, conclusions were drawn based on their performance. The results portray that the highest precision (23%) and recall (78%) scores are achieved with Logistic Regression whereas XGBoost had the highest accuracy (91%).

**2. Introduction and Motivation:**

Prediction utilizing machine learning techniques is the focus of this project. Nowadays, machine learning is widely employed in a variety of business applications, including e-commerce, health care, and many more [1]. The goal of this paper, however, is to predict heart disease using ML techniques.

The most common disease in the world right now is cardiovascular disease. According to estimates around 17.9 million deaths were attributed in 2017 [2]. WHO states that heart disease is the leading cause of death worldwide. Heart disease is one of the common illnesses that many individuals experience in middle or old age, and in many cases, it eventually results in fatal complications [3]. There are a plethora of reasons for the occurrence of Heart attacks and some significant indicators are the presence of diabetes, obesity (high BMI), a lack of physical activity, and excessive alcohol consumption [4]. Hence it is important to predict the occurrence of heart failure to treat the patients in the earlier stages and save human lives.

This report performs the comparison and contrast of different machine learning algorithms on the CDC survey dataset. The final goal of this research is to use the disease-based dataset, apply preprocessing techniques, compare and contrast different machine learning models, and finally analyze the results and the performance of those models using different evaluation metrics.

**3. Literature Review:**

Numerous researchers and authors have investigated, tested, and examined a variety of strategies for predicting heart disease. An evaluation and follow-up system was provided by Otoom et al. [5] for diagnosis and examination of heart disease. Dataset was obtained from Cleveland's UCI repository for the proposed system which contains 303 cases and 76 attributes. After the holdout test, the SVM algorithm yielded the best results.

Various algorithms, including Naive Bayes, Decision Tree, KNN, Logistic Regression, SVM, and ANN, were suggested by Ashok Kumar Dwivedi et al. [6] in his paper. In comparison to other algorithms, Logistic Regression offers greater accuracy.

Using machine learning techniques and a dataset of 500 patients from the Chennai Research Institute, Parthiban et al. [7] used diabetes patients to diagnose cardiovascular disease. Out of which the accuracy of the SVM algorithm was high.

The Naive Bayes algorithm was used by Vembandasamy et al. [8] to diagnose the heart disease prognosis. The dataset, which includes the records of 500 patients, was taken from a facility in Chennai. Naive Bayes provided great accuracy.

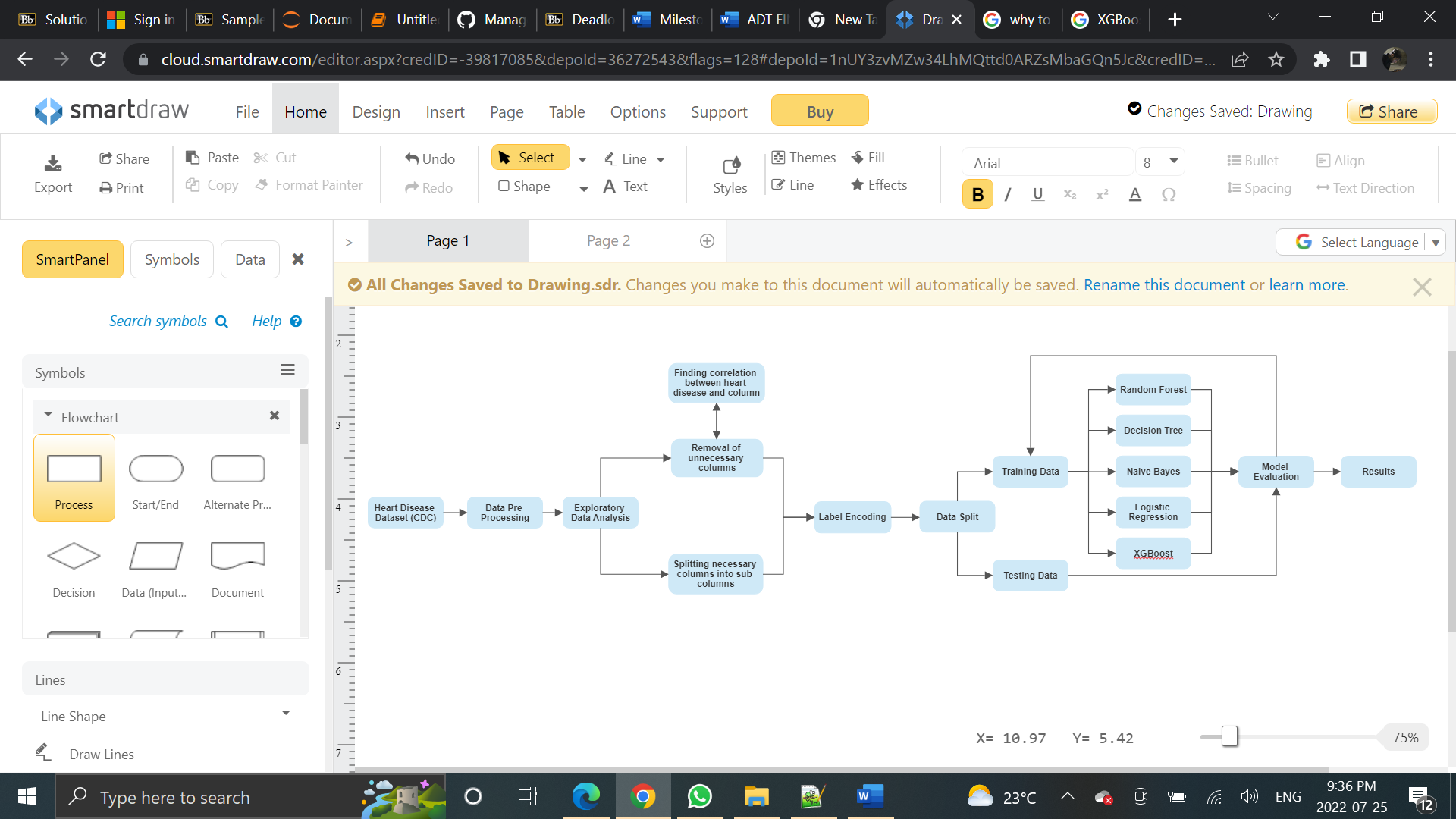
A. Malav et al. [9] provide a highly efficient hybrid algorithmic strategy for the prediction of cardiovascular illness to find and determine unidentified knowledge about heart disease by utilizing a hybrid approach in which both artificial network and Naive Bayes operate together.

The biggest disadvantage of the studies stated above is that they rely on an outdated dataset from the UCI Machine Learning Repository, which was contributed in 1988. Another drawback, in our opinion, is that previous research only used 303 rows of that dataset, which was used to train machine learning models. Our research focuses on creating and evaluating machine learning models utilizing the CDC's most recent data. The dataset is current and is large enough (319,795 rows) for our models to attain a better learning rate.

**4. Proposed Model/ Implementation Details:**

As mentioned above, this paper serves the purpose of demonstrating a relative study of different ML algorithms on a big set of data. We have chosen a subset of disease diagnosis as our study of interest; predicting heart disease to be more precise.

**4.1 Workflow of Our Model:**



**Figure.1.** Workflow

**4.1.1 Dataset:**

We have used a public dataset [10] from the 2020 annual CDC survey data of 400k adults related to their health status. CDC [11] is the US’ leading science-based, data-driven, service organization that protects the public’s health. According to CDC, heart disease is one of the leading causes of death among different races in America and hence it becomes essential to harness the power of ML and find useful and life-saving trends in the data.

The dataset is from 2020, dated as of the most recent by CDC, and consists of 401,958 rows and 279 columns. Most columns include questions asked to patients like "Do you have serious difficulty walking or climbing stairs?" or "Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]" [10]. For the purpose of this research, we are using a subset of the mentioned dataset that has already been compressed and pre-cleaned. Hence, this new subset consists of 319,796 rows with 18 important factors/columns.

**4.1.2 Data Pre-Processing:**

In data pre-processing, we have done exploratory data analysis using **Pandas data frame**. Pandas is a Python library used for preprocessing of data. We have analyzed the dataset using the **matplotlib library** of Python to plot some charts to have a visual insight into our columns. We have dropped some of the columns that were not affecting the result of heart disease whereas we have split some columns into sub-columns to normalize the columns. After cleaning the dataset, we have then encoded the columns using one hot encoding which converts the categorical value into numeric value so that it is ready for machine learning algorithms.

**4.1.3 Splitting of Data:**

The data is then split into training and testing data respectively with the ratio of 80:20 percentage using the sklearn library of Python. Training data is then sent to all machine learning algorithms one by one, and their classification report was generated using sklearn.metrics library of Python.

**4.1.4 Model Evaluation:**

The models are then evaluated on test data using a classification report and confusion matrix and were then sent back to the train data phase if the results are not satisfactory.

**5. Results:**

**5.1 Feature Engineering:**

**BMI:**

We divided BMI data into 5 categories according to the health industry specification USA. These categories with their conditions are as follows:

|  |  |  |
| --- | --- | --- |
| **Category** | **Condition** | **Count** |
| Underweight | 0 – 18.5 | 5114 |
| Normal | 18.5 – 24.9 | 95134 |
| Overweight | 24.9 – 29.9 | 114752 |
| Obese | 29.9 – 34.9 | 62172 |
| Extremely Obese | 34.9 – 100 | 42623 |

**Table.1.** BMI Data

Chart, bar chart

Description automatically generated **Figure.2.** BMI Categories

From the above table, we can see that most of the people in our dataset are overweight and that BMI has a linear relationship with heart disease. Hence, we will break down this column into 5 columns to get better results from our model.

**Sleep Time:**

Like BMI, Sleep Time was also divided into 3 categories: Low (0 – 6) , Normal (6 – 9) and High (9 – 24). It was found that people in the high category had the most chances of getting a heart disease.

Chart, bar chart

Description automatically generated

**Figure.3.** Sleep Time

**Age Category:**

We can observe that as the age increases, chances of getting a heart disease also increases. Furthermore, since this column has too many values, we have replaced them using the mean of the value range. For instance, 18-24 becomes 21 and so on.

Chart, bar chart

Description automatically generated

**Figure.4.** Age Category

**5.2 Dropping Columns:**

**Mental Health:**

Mental Health was dropped from the analysis since it had no effect on the predicting class. Furthermore, it had 205,401 values as 0 and 17,373 as 30 and was very imbalanced.

**Alcohol Drinking:**

Only 4 percent of people were found to have heart disease who were also heavy drinkers hence we dropped this column.

List of columns before and after feature engineering:

A screenshot of a computer

Description automatically generated with low confidence

**Figure.5.** Columns before feature extraction

Table

Description automatically generated

**Figure.6.** Columns after feature extraction

**5.3 Model Comparison:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Random Forest** | **Decision Tree** | **Naïve Bayes** | **Logistic Regression** | **XGBoost** |
| **Accuracy** | 74.5% | 80% | 82.4% | 75.1% | 91% |
| **Recall** | 13% | 37% | 52% | 78% | 9% |
| **Precision** | 36% | 19% | 26% | 23% | 55% |

**Figure.7.** Model comparison

**6. Limitations and Challenges:**

* One of the limitations is regarding our dataset that the predicting class is imbalanced, hence classic machine learning modeling approach did not give out the best results and we had to work with weights to achieve good results.
* Furthermore, due to the imbalanced predicting class, we couldn’t evaluate our model based on just accuracy. Hence, precision (calculating true positives) was used to evaluate our models.
* We were able to achieve the highest recall value of 78% and the highest precision value of 55%. While this is sufficient, it can further be improved using techniques that aren’t mentioned in this paper and hence we consider this as a limitation as well.
* The dataset at hand was huge and due to time constraints, we were only able to work with a few classification techniques. Hence, if we had more time, techniques such as clustering could have been used to make clusters of attributes that contributed significantly to predicting the probability of having a heart disease.

**7. Conclusion and Future Work**:

While it’s important to assess and build models based on the medical data of patients, it’s also important to take the lifestyles of these people/patients into consideration. In this paper, five ML classification algorithms, namely, Random Forest, Logistic Regression, Naïve Bayes, Decision Trees, and XGBoost were applied to the CDC Survey dataset and prediction models have been developed to predict heart disease. The CDC dataset has questions related to the lifestyle of people as well their medical conditions. Logistic Regression performed the best in our case since we are using precision and recall evaluating our models’ performances. It was able to achieve a staggering recall of 78% and a precision of 23% while its accuracy was also at 75.1%.

In the future, various heart disease datasets from health data repository can be combined, and the best performing classification model using contemporary machine learning models can be outlined. Furthermore, other machine learning techniques such as clustering and deep learning (DL) can be applied to this dataset to understand which variable has more effect on the dependent variable by making clusters of data whereas DL can help us in improving our prediction by using optimized weights for our imbalanced predicting class.

# References

|  |  |
| --- | --- |
| [1] | R. Nichenametla, "https://www.researchgate.net/publication/326733163\_Prediction\_of\_Heart\_Disease\_Using\_Machine\_Learning\_Algorithms". |
| [2] | S. K. A. R. S. R. K. Y. T. a. A. .. R. T. K. Kartikh, "Implementation of a Heart Disease Risk Prediction Model Using Machine Learning," *Computational and Mathematical Methods in Medicine,* vol. 2022, pp. 1-14, 2022. |
| [3] | C. B. C. a. S. CarolinJeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked,* vol. 16, no. 100203, pp. 1-9, 2019. |
| [4] | Z. S. Md Manjurul Ahsan, "Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review," 13 December 2021. [Online]. Available: https://arxiv.org/abs/2112.06459. |
| [5] | E. J. T. Q. Y. a. K. C. G. K. C. Tan, "A hybrid evolutionary algorithm for attribute selection in data mining," *Expert Systems with Applications,* no. 36, pp. 8616-8630, 2008. |
| [6] | A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications,* vol. 29, no. 10, pp. 685-693, May 2018. |
| [7] | G. Parthiban and S. K. Srivatsa, "Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients," *International Journal of Applied Information Systems,* vol. 3, no. 7, pp. 25-30, August 2012. |
| [8] | E. D. a. R. S. K. Vembandasamy, "Heart Diseases Detection Using Naive Bayes Algorithm," *International Journal of Innovative Science, Engineering & Technology,* vol. 2, no. 9, pp. 441-444, September 2015. |
| [9] | K. K. a. P. K. A. Malav, "PREDICTION OF HEART DISEASE USING K-MEANS and ARTIFICIAL NEURAL NETWORK as HYBRID APPROACH to IMPROVE ACCURACY," *International Journal of Engineering and Technology,* vol. 9, no. 4, pp. 3081-3085, August 2017. |
| [10] | "Personal Key Indicators of Heart Disease," 2022. [Online]. Available: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease. |
| [11] | "Centres for Disease Control and Prevention," [Online]. Available: https://www.cdc.gov/. |