Chronic Disease Prediction

CIS 5930: Data Mining Project

Ajay Aryan(aa23bf), Ved Bhatt(vb23d), Sanskar Chouhan(sc23bq)

1. Abstract

The main objective of this project is to predict different types of chronic disease using more than 1 machine learning model and analysing their datasets for chronic diseases. Our list of chronic disease includes - diabetes, kidney disease, and heart disease. The results show how timely interventions and individualized preventive efforts can lead to better health outcomes. The model's importance in proactive healthcare techniques and its wider impact on public health are highlighted, along with ethical considerations and privacy safeguards.

2. Introduction

The worldwide increasing frequency of chronic illnesses, such as diabetes, renal disease, and heart disease, calls for innovative approaches to early detection and prevention of these disease. To meet this need, our research uses cutting-edge machine learning algorithms, namely K-Nearest Neighbors (KNN) and Naive Bayes classifiers, to build a predictive model customized for these common chronic illnesses. Using a variety of datasets that include medical and demographic data allows for a thorough understanding of each person's unique health profile.

Selected for their different approaches, the non-parametric instance-based KNN algorithm and the probabilistic classifier Naive Bayes are contrasted to find out how well they predict the onset of particular chronic diseases. This analysis highlights the interpretability and transparency made possible by feature selection techniques in addition to emphasizing predictive accuracy. Outside of the academic setting, our research could be useful in developing proactive and more individualized approaches to managing chronic diseases, as well as informing healthcare strategies and enabling timely interventions. This innovation contributes to better patient outcomes and lessens the overall burden on healthcare systems, which has significant implications for healthcare professionals.

3. Literature Survey

I. This prediction review by Divya Jain, Vijendra Singh examines feature selection and classification techniques for chronic disease prediction. It delves into filter, wrapper, and hybrid methods, emphasizing their impact on model accuracy. The review underscores the evolving landscape of healthcare informatics and highlights potential avenues for future research in chronic disease prediction.[1]

II. The research presents a CNN-MDRP algorithm for healthcare big data, achieving 94.8% prediction accuracy for chronic disease outbreaks. It addresses incomplete data challenges, emphasizing the integration of structured and unstructured data which was conducted by Min Chen, Yixue Hao, Kai Hwang, Lu Wang, Lin Wang.[2]

III. The research investigates Bayesian theory, implementing Naive Bayes and K-Nearest Neighbour (KNN) classifiers for a "credit card approval" dataset. The study explores classification performance, demonstrating the application of Bayesian principles in decision-making under uncertainty. The research done by Mohammed J. Islam, Q. M. Jonathan Wu, Majid Ahmadi, Maher A. Sid-Ahmed.[3]

4. Methodology

- KNN (K-Nearest Neighbour) - K-Nearest Neighbours (KNN) utilizes the innate patterns in a dataset to provide a reliable approach for chronic disease prediction. First, a wide range of datasets containing relevant personal data is gathered, such as demographics and medical history. Careful preprocessing methods are then used to deal with outliers and missing values, guaranteeing the consistency of the dataset. Next, in order to minimize dimensionality and maximize the effectiveness of the model, features are selected with an emphasis on the most significant variables.

The dataset is split into training and testing sets after it is ready. After learning the training set, the KNN algorithm uses the average value of the K nearest neighbors or the majority class to determine the labels to be assigned during predictions. The selection of 'K' and the distance metric have a substantial impact on the accuracy of the model, thus requiring careful consideration. The model's performance on the testing set is measured by evaluation metrics like accuracy and precision, which result in iterative changes to the hyperparameters for the best predictive results. KNN's methodology, which essentially takes a methodical approach to training, evaluating, and preparing data, is what makes it so successful at predicting chronic illnesses.

- Naive Bayes - A unique approach is used to predict chronic diseases using the probabilistic machine learning algorithm Naive Bayes. The first step in the procedure is to put together a large dataset that includes characteristics such as patient demographics, lifestyle factors, and medical history. The Naive Bayes model is trained using this dataset as its basis. Because each feature is assumed to contribute to the outcome independently, the algorithm relies on the assumption of feature independence, which simplifies the probability calculations.

  The next stage after preparing the dataset is to divide it into training and testing sets. Next, using the former as training data, the Naive Bayes model learns the probabilistic relationships between the features and the target variable, which indicates whether a chronic illness is present or not. Based on observed feature values, the model predicts a patient's probability of having the disease. Using metrics such as accuracy and precision, the model's performance on the testing set is evaluated in the final step. When feature independence assumptions match the features of the data, naive bayes is a useful tool for the prediction of chronic diseases because of its ease of use, effectiveness, and efficiency when working with medical datasets.

- Python Libraries - Python libraries are essential for the prediction of chronic illnesses because they provide a smooth workflow from model evaluation to data manipulation. Data handling is made easier by Pandas' flexible data structures, which enable effective preprocessing of a variety of datasets containing patient data. NumPy, on the other hand, supports numerical operations necessary for jobs like normalization and handling missing values. When combined, they offer a strong basis for data preparation.

  The use of visualization is essential for identifying trends in medical data. Researchers can examine relationships between various health-related features with the help of Matplotlib and Seaborn, which are excellent at producing informative visualizations. Seaborn, which is based on Matplotlib, offers an aesthetically pleasing interface for statistical data visualization. Matplotlib is a comprehensive plotting library.

  A vital tool for developing and accessing predictive models is the Scikit-learn (sklearn) library. It is a preferred tool for researchers forecasting chronic diseases because it incorporates a broad range of machine learning algorithms, such as regression models and classifiers. Additionally, data scientists can evaluate the performance of their predictive models with the help of the model evaluation utilities provided by the library. Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn have all been well integrated to create a potent toolkit for the whole chronic disease prediction process in Python.

5. Implementation

- Load Dataset: Started by using a library such as Pandas to import your dataset. Reading data from a file (such as a CSV file) into a DataFrame for additional analysis is what this step entails.

- Data Investigation: Investigating the dataset to learn about its features and organization. To learn more about the makeup of the dataset, look for missing values, investigate different data types, and compile some basic statistics.

- Data cleaning: Removing null or missing values with appropriate values like mean , median or mode.

- Data conversion: Converting columns to appropriate data type like converting categorical data to numerical if required.

- Information Visualization: Make sense of the data by using visualization tools like Matplotlib and Seaborn to produce plots that highlight distributions, correlations, and patterns. Visualization facilitates comprehension of the connections between various variables.

- Making the Plot Count: To visually represent the distribution of the target variable (chronic disease labels), create a count plot. An overview of the dataset's chronic disease prevalence is given by this plot.

- Method to Obtain Valid Type of Input: Provide a feature that allows users to interactively provide valid categorical input. This function makes sure the values entered match the dataset's categorical features.

- Get Numeric Input Function: Provide a function that allows users to submit numbers. This function ensures accuracy and relevance by validating the input against the dataset's numerical features.

- Function to Forecast Chronic Illness Using User Data: Create a function that uses user input to predict the probability of chronic disease using a pre-trained model. The main predictive element of the system is this function.

- Getting the Data Ready for Prediction: Integrate user input into a model-appropriate format. To do this, the input must be pre-processed to conform to the format used for the first model training.

- Open the dataset and perform preprocessing: Using missing value handling and categorical variable encoding, preprocess the entire dataset. By doing this, you can be sure the dataset is clean and prepared for training.

- Data Scaling and Splitting: Divide the dataset in half, with 80% going toward training and 20% toward testing. To guarantee consistency throughout the model training process, scale numerical features.

- Model Training (Naive Bayes and KNN): Utilizing the training set, teach machine learning models like Naive Bayes and K-Nearest Neighbours (KNN). These models identify links and patterns in the data.

- Assessment of the Model: Utilizing the testing set, assess the trained models. Metrics such as accuracy shed light on how well the models function with unknown data.

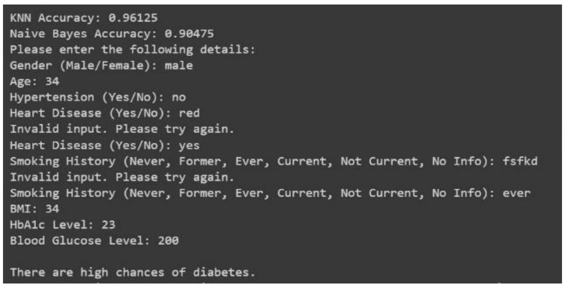
- Making a final projection: Based on user input, use the trained models to generate final predictions. This entails getting predictions from the KNN and Naive Bayes models as well as preprocessing and scaling user input as needed.

6. Result

Diabetes Prediction - The diabetes prediction models produced encouraging results. The K-Nearest Neighbours (KNN) algorithm outperformed the Naive Bayes model, which showed a respectable accuracy of 90.475%, with an impressive accuracy of 96.125%. The models' strong performance in identifying the presence or absence of diabetes in the assessed dataset is demonstrated by the mean accuracy, which was computed at 93.3%.

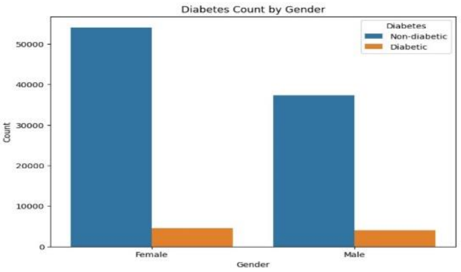
Notably, the likelihood of diabetes was considered high when both models produced accurate predictions, indicating a consensus. On the other hand, inconsistencies between the models indicated a decreased risk of diabetes. This cooperative method takes advantage of both algorithms' strengths to improve prediction reliability.

These findings highlight the potential that machine learning models—KNN and Naive Bayes in particular—have to offer diabetes prediction efforts. A layer of confidence is added to the assessments by emphasizing consensus predictions, which helps healthcare providers make well-informed decisions about patient risk and care strategies. The models' capacity for generalization is bolstered by their high mean accuracy, which gives rise to confidence regarding their usefulness in diabetes-related scenarios for early detection and intervention. It is critical to take these results into account in light of the particular dataset and healthcare setting that the models were created for.

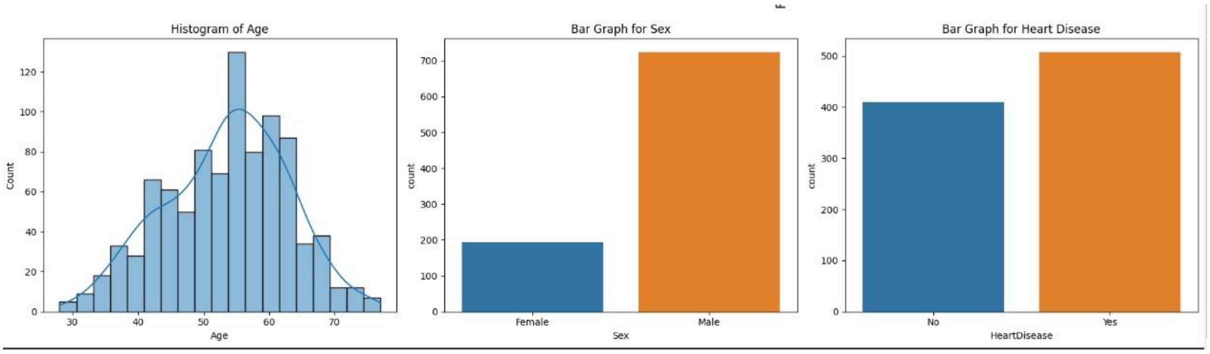


```
KNN Accuracy: 0.96125
Naive Bayes Accuracy: 0.90475
Please enter the following details:
Gender (Male/Female): male
Age: 34
Hypertension (Yes/No): no
Heart Disease (Yes/No): red
Invalid input. Please try again.
Heart Disease (Yes/No): yes
Smoking History (Never, Former, Ever, Current, Not Current, No Info): fsfkd
Invalid input. Please try again.
Smoking History (Never, Former, Ever, Current, Not Current, No Info): ever
BMI: 34
HbA1c Level: 23
Blood Glucose Level: 200

There are high chances of diabetes.
```



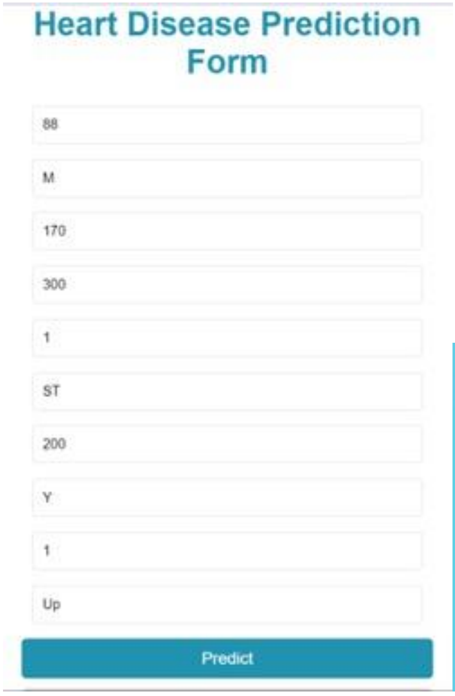
Heart Disease Prediction - The performance of the heart disease prediction models varies; the Naive Bayes model outperforms the K-Nearest Neighbours (KNN) algorithm, which achieves an accuracy of 84.78%, while the KNN algorithm achieves an accuracy of 69.02%. It was decided to rely on the predictions made by the Naive Bayes model for more reliable and accurate results because of the significantly lower accuracy of KNN.



Compared to KNN, the Naive Bayes model's accuracy of 84.78% indicates a more dependable ability to determine whether heart disease is present in the examined dataset. The fact that Naive Bayes predictions are preferred indicates that they perform better in this specific situation.

A poor selection of hyperparameters or difficulties in identifying the underlying patterns in the dataset could be the cause of KNN's reduced accuracy. Thus, selecting the more precise Naive Bayes forecasts strengthens the heart disease prediction model's legitimacy and highlights its potential as an important clinical decision-making tool.

Practically speaking, the Naive Bayes model's higher accuracy emphasizes how good it is at predicting heart disease, underscoring the significance of choosing models that are suited to the particulars of the dataset and the complexities of the prediction task. In healthcare settings where accuracy is crucial, this method guarantees more accurate and dependable results.



Heart Disease Prediction Form

88

M

170

300

1

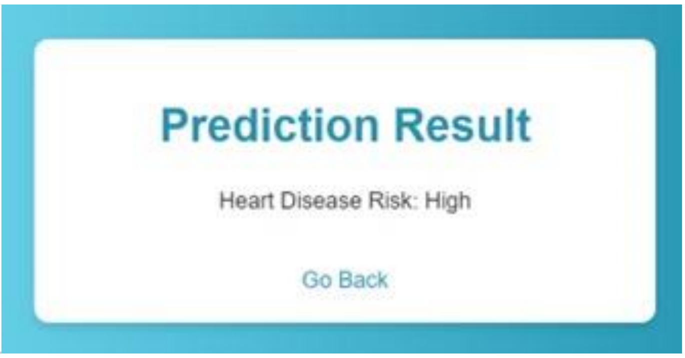
ST

200

Y

1

Up



Prediction Result

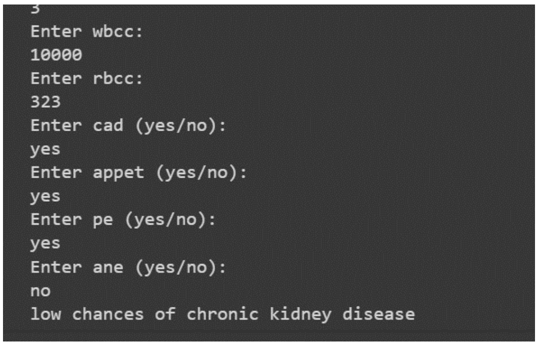
Heart Disease Risk: High

Kidney Disease Prediction - The kidney disease prediction models have shown good performance. The K-Nearest Neighbours (KNN) algorithm outperformed the Naive Bayes model, which showed a respectable accuracy of 92.5%, with an astounding accuracy of 97.5%. The reliability of both models in correctly predicting the presence or absence of kidney disease in the examined dataset is reflected in the mean accuracy, which was computed at a 95% confidence level.

Remarkably, a consensus approach was used in which the collective prediction was deemed more reliable if both models produced accurate predictions indicating chronic kidney disease (CKD). When there was a disagreement among the models, the majority prediction was taken as true. This method increases the assessments' level of confidence and improves the overall predictability of the results.

The collaborative decision-making process and the notably high mean accuracy highlight the potential of these machine learning models, especially KNN and Naive Bayes, in aiding kidney disease prediction efforts. The 92.5% accuracy of Naive Bayes and the 97.5% accuracy of KNN demonstrate how well these algorithms can identify patterns in the dataset and generate precise predictions.

In clinical settings, this consensus-driven predictive model, which favour CKD predictions when both models concur, may be useful in helping medical professionals identify and treat patients at an early stage. The benefit of machine learning models for kidney disease prediction is further highlighted by their high accuracy rates and consensus-based decision-making processes.



```
3
Enter wbcc:
10000
Enter rbcc:
323
Enter cad (yes/no):
yes
Enter appet (yes/no):
yes
Enter pe (yes/no):
yes
Enter ane (yes/no):
no
low chances of chronic kidney disease
```

7. Conclusion

In summary, the features of the dataset and the precision of the selected machine learning models play a critical role in the chronic disease prediction project's success. Prediction reliability is primarily determined by the quality and representativeness of the dataset, which includes elements like feature diversity, completeness, and relevance. Whether the models employ Naive Bayes, K-Nearest Neighbours (KNN), or other algorithms, their accuracy directly reflects their capacity to identify patterns in the dataset. It is critical to understand that the accuracy metrics are not only a good indicator of how well the model performs, but also deeply entwined with the details of the training and testing data. The project's success thus depends on the symbiotic relationship between the features of the dataset and the model's ability to generalize and produce precise predictions, highlighting the significance of meticulous data curation and model training for significant outcomes in the prediction of chronic diseases.

8. Future Work

The development of models to increase their robustness and suitability for use in healthcare should be the main goal of future research in chronic disease prediction. To create more complete models, one approach is to integrate multi-modal data, such as lifestyle and genetic data. Deep learning approaches, such as neural networks, have the potential to improve predictive accuracy by identifying complex patterns in large datasets. It is important to investigate longitudinal studies as a means of monitoring changes in health indicators over time and gaining understanding of how diseases progress.

Furthermore, creating models with enhanced interpretability guarantees openness in comprehending the variables impacting forecasts, which promotes confidence among medical professionals. Early warning systems integrated into real-time monitoring systems may help make timely interventions possible. For a more thorough risk assessment, behavioural and socioeconomic factors must be included in predictive models. Models are applicable in a variety of healthcare settings because their generalizability is ensured through validation across diverse populations. Proper integration of predictive models into clinical workflows requires cooperation between data scientists and healthcare providers. Techniques that protect privacy, such as federated learning, can improve data security. Predictive tools could be seamlessly incorporated into standard clinical procedures with the help of user-friendly interfaces designed specifically for healthcare professionals. Future studies in these fields will improve chronic illness prediction models, which will ultimately lead to advancements in personalized medicine and preventive healthcare.

9. References

[1] A. M. Alhassan and W. M. N. Wan Zainon, "Review of Feature Selection, Dimensionality Reduction and Classification for Chronic Disease Diagnosis," in *IEEE Access*, vol. 9, pp. 87310-87317, 2021, doi: 10.1109/ACCESS.2021.3088613.

[2] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

[3] M. J. Islam, Q. M. J. Wu, M. Ahmadi and M. A. Sid-Ahmed, "Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers," 2007 International Conference on Convergence Information Technology (ICCIT 2007), Gwangju, Korea (South), 2007, pp. 1541-1546, doi: 10.1109/ICCIT.2007.148.

[4] https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/

[5] https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[6] https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease