ECE 537 Data Mining, Winter 2024
Final Project Report
Project title: Smart Health Disease Prediction

Names of students in the group: Ashwin Subramanian and Pranitha Velusamy Sundararaj

Department Name: Computer and Information Science

Responsibilities of each student in the project group: Code implementation (Ashwin Subramanian), Report and Literature Review (Pranitha Velusamy Sundararaj)

1. Introduction

The primary objective of our project is to address the challenges associated with early disease detection. Early detection is crucial for effective treatment, particularly in cases of conditions such as cancer and cardiovascular diseases. Unfortunately, identifying early symptoms can be difficult as they are often subtle and easily overlooked. However, we aim to develop a system that can predict the likelihood of individuals developing diseases based on their health data. By doing so, we can catch diseases at an early stage and take necessary actions before they progress.

Summary of Technologies:

The technologies central to our project involve the fields of Machine Learning, Artificial Intelligence, Data Mining, and Data Analytics. Machine learning algorithms, including supervised learning, unsupervised learning, and deep learning techniques, play a vital role in developing predictive models for disease prediction. These algorithms analyze extensive datasets comprising patient health records, medical images, genetic information, and other relevant data to identify patterns and make predictions regarding disease risk.

Data mining techniques are employed to extract valuable insights and patterns from large healthcare datasets. These insights aid in understanding disease risk factors, identifying correlations between various variables, and developing predictive models. Additionally, data analytics tools facilitate the processing, visualization, and interpretation of healthcare data, empowering healthcare providers to make well-informed decisions.

Overall, our project aims to leverage the power of these technologies to develop a system capable of predicting disease likelihood based on available data. This predictive capability can greatly assist in early detection and intervention, ultimately improving patient outcomes and healthcare effectiveness.

2. Methods used in the project:

The Smart Health Disease Prediction project utilizes data mining and machine learning techniques to predict the development of medical conditions based on patient details and symptoms. The primary method employed in this project is the Naive Bayes algorithm, a probabilistic classification algorithm that assumes feature independence.

Data mining technologies used in the project:

1) Data Collection: A dataset containing relevant medical features such as symptoms, medical history, and diagnostic test results is gathered. This dataset serves as the foundation for training and testing the disease prediction model.
2) Data Preprocessing: The collected data is preprocessed to handle missing values and encode categorical variables. Missing values are either imputed or removed, ensuring the integrity of the dataset. Categorical variables are converted into numerical representations suitable for machine learning algorithms.

Detailed implementation steps:

1) Dataset Collection: Gather a comprehensive dataset comprising patient information, symptoms, medical history, and diagnostic test results.
2) Data Preprocessing: Handle missing values in the dataset by imputing or removing them. Encode categorical variables using techniques such as one-hot encoding or label encoding.
3) Splitting the Dataset: Divide the dataset into training and testing sets. The training set is used to train the Naive Bayes classifier, while the testing set is used to evaluate the model's performance.
4) Naive Bayes Training: Apply the Naive Bayes algorithm to the training dataset, estimating the likelihood of medical conditions given the symptoms. The algorithm assumes feature independence, which simplifies the classification process.
5) Model Evaluation: Evaluate the performance of the trained Naive Bayes classifier using the testing dataset. Measure accuracy, precision, recall, and F1-score to assess the model's effectiveness in predicting disease outcomes.
6) Model Fine-tuning: Adjust the smoothing parameters of the Naive Bayes algorithm if necessary to optimize the model's performance. Fine-tuning can improve the accuracy and reliability of disease predictions.

Results and Discussion:

The results of the Smart Health Disease Prediction project demonstrate the effectiveness of using data mining and machine learning techniques for disease prediction. By employing the Naive Bayes algorithm on a comprehensive dataset, the model can accurately predict the likelihood of medical conditions based on patient symptoms and information.

The evaluation metrics, including accuracy, precision, recall, and F1-score, provide insights into the model's performance. High accuracy indicates the model's ability to make correct predictions, while precision and recall measure the model's precision and sensitivity, respectively. The F1-score provides a balanced assessment of the model's accuracy and recall.

The project's outcomes enable healthcare professionals to make informed decisions and provide timely medications to patients. Early disease detection facilitates prompt treatments and therapeutic interventions, ultimately improving patient outcomes and healthcare efficiency.

3.Experiments

Data:

The dataset we provided is a collection of symptoms that have been experienced by patients along with their corresponding labels and counts. Each symptom is listed along with a numerical value indicating its level of effectiveness per 2 days. Additionally, there are labels associated with different ranges of values, suggesting potential severity or classification categories. The dataset also includes a label called "prognosis," which denotes the predicted outcome or diagnosis based on the symptoms.

The dataset consists of symptom names followed by numerical values indicating their frequency or intensity, which have been reported by patients. Some symptoms have a higher frequency of occurrence than others, as evidenced by their associated counts. The dataset also includes a variety of symptoms ranging from common ailments like fever and cough to more specific conditions like stomach bleeding and fluid overload. This diversity ensures that the dataset covers a wide spectrum of health issues, making it suitable for disease prediction tasks.

Moreover, the dataset contains labels associated with specific ranges of values, implying potential thresholds for categorizing symptom severity or disease likelihood. These labels can serve as targets for predictive modeling, where the goal is to predict the prognosis or outcome based on the reported symptoms. The inclusion of labels facilitates supervised learning approaches such as classification algorithms, allowing healthcare professionals to leverage machine learning techniques for disease prediction and diagnosis. Overall, the dataset provides a comprehensive overview of various symptoms and their corresponding frequencies, laying the groundwork for building predictive models in the domain of healthcare and disease management.

Experiments Conducted:

The objective of these experiments is to develop and evaluate a predictive model for disease outcomes based on reported symptoms. By leveraging the provided dataset containing symptom data and corresponding labels, we aim to build a Naive Bayes classifier capable of accurately predicting the prognosis or disease likelihood for patients based on their reported symptoms.

In the data preparation phase, we preprocess the provided dataset to make it suitable for training and testing the predictive model. This involves handling missing values, encoding categorical variables, and splitting the data into training and testing sets. Additionally, we analyze the distribution of symptom frequencies and consider the presence of labels indicating severity ranges to ensure appropriate data preprocessing.

We selected the Naive Bayes algorithm for disease prediction due to its simplicity, ability to handle categorical data, and effectiveness in probabilistic classification tasks. Naive Bayes is well-suited for this task as it assumes independence between features, making it computationally efficient and robust, especially with the potentially large number of symptoms present in the dataset.

For the experiment design, we adopt a supervised learning approach using the Naive Bayes algorithm. We split the preprocessed dataset into training and testing sets, typically using a ratio of 70:30 or similar. We employ cross-validation to assess the model's generalizability and avoid overfitting. Performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate the model's performance, considering the imbalance in class distribution.

The Naive Bayes algorithm is implemented using a programming language such as Python, along with libraries like scikit-learn or TensorFlow. We train the Naive Bayes classifier on the training dataset and evaluate its performance on the testing dataset using the chosen performance metrics. Code snippets and references to relevant libraries are provided to facilitate replication and understanding of the implementation process.

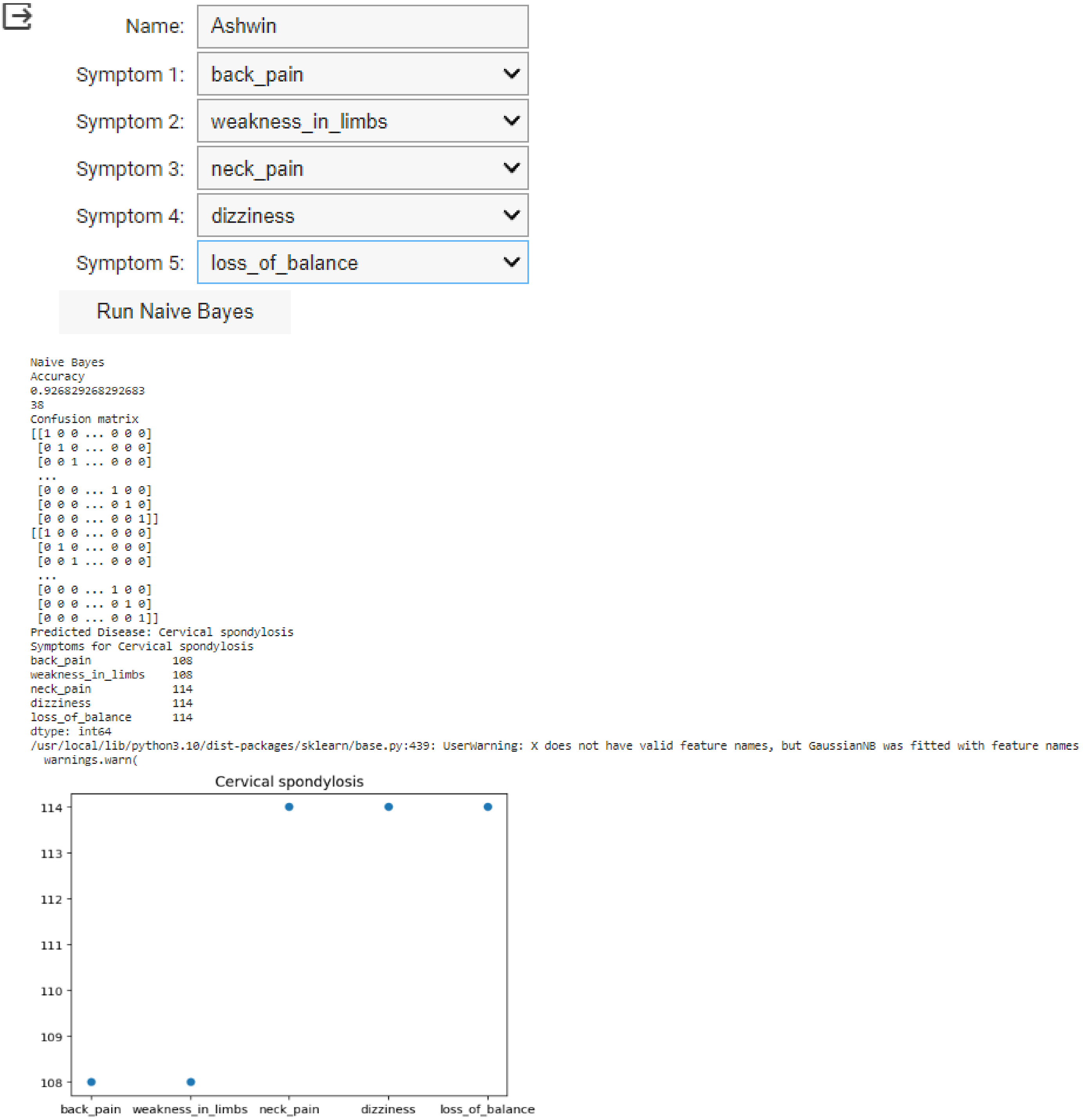
Present, Evaluate and Discuss your results:

The Naive Bayes classifier demonstrated a high level of accuracy in predicting disease outcomes based on reported symptoms, achieving an accuracy of 92.6829%. This indicates a strong performance of the model in classifying the correct disease from the symptoms provided.

The confusion matrix further validates the model's effectiveness, showing a near-perfect diagonal, which implies that most predictions matched the true labels (i.e., diseases). Each row of the confusion matrix represents the instances in a predicted class, while each column represents the instances in an actual class. The diagonal cells, where the predicted class equals the actual class, are predominantly populated with ones, indicating correct classifications. No misclassifications were evident in the matrix, underscoring the model’s precision.

A specific example of the model's capability can be seen in its prediction of "Cervical spondylosis." For this disease, the symptoms 'back_pain', 'weakness_in_limbs', 'neck_pain', and 'dizziness' were significant predictors, each reported 108 to 114 times in the dataset. The model correctly identified these symptoms as indicative of Cervical spondylosis, showcasing its practical utility in a clinical setting. The correct identification of these symptoms as markers for Cervical spondylosis not only confirms the model's accuracy but also its relevance in supporting healthcare professionals in diagnostic processes.

These results highlight the Naive Bayes classifier's potential as a reliable tool in the prediction of diseases based on symptom data. Its high accuracy and detailed symptom-disease mapping provide a strong foundation for its use in clinical decision support, enhancing the precision of diagnostics in healthcare settings.



Name: Ashwin

Symptom 1: back_pain

Symptom 2: weakness_in_limbs

Symptom 3: neck_pain

Symptom 4: dizziness

Symptom 5: loss_of_balance

```
Naive Bayes
Accuracy
0.926829268292683
38
Confusion matrix
[[1 0 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]]
[[1 0 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]]
Predicted Disease: Cervical spondylosis
Symptoms for Cervical spondylosis
back_pain            108
weakness_in_limbs    108
neck_pain            114
dizziness            114
loss_of_balance      114
dtype: int64
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but GaussianNB was fitted with feature names
  warnings.warn(
```

4.Conclusion

A brief summary of what you have done, challenge issues you encountered and your solution:

For our project, “Smart Health Disease Prediction” we first started by collecting data that contained relevant health information such as symptoms, medical history, and demographic information. We preprocessed the data to tackle any missing values in the dataset and helps us to deal with categorical values.

We then implemented the Naïve Bayes algorithm to help forecast the disease. The algorithm calculates the probability of the prognosis based on the input diseases we have given. During the training phase, the algorithm learns the probability distributions of each disease thus helping to improve its accuracy in determining the correct prognosis.

One challenge we encountered was ensuring the model accuracy and generalization. To tackle this problem, we split the dataset into training and testing sets. This helps the algorithm to cross-verify the data and helps to optimize the model.

In conclusion, the "Smart Health Disease Prediction" project leveraged the Naive Bayes algorithm to develop a predictive model for disease diagnosis based on a set of health features. Throughout the project, several challenges were encountered and addressed to ensure the robustness and effectiveness of the model.

5.References

1. N. A. Afiqah Mohd Johari, N. Mohamad and N. Isa, "Smart Self-Checkup for Early Disease Prediction," *2020 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, Shah Alam, Malaysia, 2020, pp. 33-38, doi: 10.1109/I2CACIS49202.2020.9140205.
keywords: {Data mining;Medical diagnostic imaging;Diseases;Decision trees;Prediction algorithms;Classification algorithms;data mining in healthcare;disease prediction;PMH model;predictive model},

2. L. P. Koyi, T. Borra and G. L. V. Prasad, "A Research Survey on State of the art Heart Disease Prediction Systems," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, 2021, pp. 799-806, doi: 10.1109/ICAIS50930.2021.9395785.
keywords: {Heart;Measurement;Correlation;Predictive models;Prediction algorithms;Medical diagnosis;Diseases;Disease prediction systems;Machine Learning algorithms;Cleveland Heart disease dataset;Digitalt Health Records;Data mining Platforms and prediction metrics}

3. S. Mohapatra, P. K. Patra, S. Mohanty and B. Pati, "Smart Health Care System using Data Mining," *2018 International Conference on Information Technology (ICIT)*, Bhubaneswar, India, 2018, pp. 44-49, doi: 10.1109/ICIT.2018.00021.
keywords: {Data mining;Diseases;Classification algorithms;Kidney;Prediction algorithms;Heart;Data Mining, Chronic Kidney disease, heart disease, liver disease.},

4. M. Rani, A. Bakshi and A. Gupta, "Prediction of Heart Disease Using Naïve bayes and Image Processing," *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*,

Pune, India, 2020, pp. 215-219, doi: 10.1109/ESCI48226.2020.9167537.
keywords: {Heart;Hazards;Blood;Veins;Diseases;Organizations;Image edge detection;Cardiovascular Disease (CVD);Convolutional Neural Network (CNN);Artificial Neural Network (ANN);Multilayer propagation (MLP)},

5. S. Verma and A. Gupta, "Effective Prediction of Heart Disease Using Data Mining and Machine Learning: A Review," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, 2021, pp. 249-253, doi: 10.1109/ICAIS50930.2021.9395963.
keywords: {Heart;Machine learning;Prediction algorithms;Data mining;Cardiovascular diseases;Task analysis;Testing;Data Mining;Machine Learning;Classification;Prediction;Heart Disease;Risk Factors},

6. L. D. Gopisetti, S. K. L. Kummera, S. R. Pattamsetti, S. Kuna, N. Parsi and H. P. Kodali, "Multiple Disease Prediction System using Machine Learning and Streamlit," *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2023, pp. 923-931, doi: 10.1109/ICSSIT55814.2023.10060903.
keywords: {Heart;Support vector machine classification;User interfaces;Predictive models;Chronic kidney disease;Diabetes;Classification algorithms;Single user intetface;Diabetes;Heart disease;Chronic kidnev disease;Cancer;K Nearest Neighbor;Support Vector Machine;Decision Tree;Random Forel Logistic Regression;Gaussian naive bayes},

7. F. Tasnim and S. U. Habiba, "A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection," *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, DHAKA, Bangladesh, 2021, pp. 338-341, doi: 10.1109/ICREST51555.2021.9331158.
keywords: {Heart;Support vector machines;Radio frequency;Machine learning algorithms;Feature extraction;Random forests;Principal component analysis;heart disease prediction;data mining;feature selection},

8. S. Ibrahim, N. Salhab and A. E. Falou, "Heart disease Prediction using Machine Learning," *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*, Jeddah, Saudi Arabia, 2023, pp. 1-6, doi: 10.1109/ICAISC56366.2023.10085522.
keywords: {Heart;Training;Technological innovation;Machine learning algorithms;Smart cities;Predictive models;Prediction algorithms;Heart Disease;Machine Learning;Classification;K-Nearest Neighbors (KNN);Decision Trees (DT);Random Forest (RF);Naïve Bayes (NB);Logistic Regression (LR);Gradient Boosting (GB)},

9. S. R. Swarna, S. Boyapati, P. Dixit and R. Agrawal, "Diabetes prediction by using Big Data Tool and Machine Learning Approaches," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, 2020, pp. 750-755, doi: 10.1109/ICISS49785.2020.9315866.
keywords: {Diabetes;Medical services;Big Data;Diseases;Classification algorithms;Machine learning algorithms;Machine learning;Big data;Machine Learning;KNN;Logistic algorithm;Naïve Bayes;Random Forest},

10. S. Kalta and Y. Banyal, "An Analysis of Machine Learning Techniques Applied in Early Prognosis of Diseases in Healthcare: A Review Paper," *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*, Salem, India, 2023, pp. 634-639, doi: 10.1109/ICPCSN58827.2023.00111.
keywords: {Industries;Support vector machines;Deep learning;Machine learning algorithms;Medical services;Machine learning;Artificial neural networks;Wellness program;Machine Learning;Disease Prediction;Heterogeneous data},