Comparing the Performance of Naive Bayes and k-Nearest Neighbors (k-NN) in Predicting Diabetes from Clinical Data Using Weka

Due date: April 8, 2025

Introduction

Diabetes is a chronic disease that affects millions of people worldwide. Early detection and management are crucial to prevent complications such as heart disease, kidney failure, and vision loss. Machine learning algorithms have shown promise in predicting diabetes based on clinical data.

This individual project aims to compare the performance of **Naive Bayes and k-Nearest Neighbors (k-NN)** machine learning algorithms in predicting diabetes using **Weka**. The null hypothesis to be tested is:

"There will be a significant difference in the accuracy of diabetes prediction between the Naïve Bayes and k-NN machine learning algorithms."

Expected Learning Outcomes

- Gain hands-on experience with **Weka's classification algorithms**.
- Understand the strengths and weaknesses of **NB vs. k-NN.**
- Learn how to **preprocess and analyze medical datasets**.
- Develop skills in **model evaluation and hypothesis testing**.

Materials and Methods

1. **Dataset**:
    - The dataset will be obtained from a publicly available repository, such as the **Pima Indians Diabetes Dataset** available in Weka. This dataset contains clinical data from patients diagnosed with diabetes and healthy individuals.
    - The dataset includes features such as:
        number of pregnancies, plasma glucose concentration, Blood pressure, skin thickness, insulin level, BMI (Body Mass Index), age

2. **Machine Learning Algorithms**:
    - **Naive Bayes**: A probabilistic classifier based on Bayes' theorem.
    - **k-Nearest Neighbors (k-NN)**: A non-parametric method that classifies data points based on the majority class among their k-nearest neighbors (found in IBk category)

3. **Experiment Design**:
    - You will design an experiment in which you repeat the process of building each machine learning model **30 times** based on different random samples of the dataset.
    - For each repeat:
        - **66%** of the data will be used for training.
        - **34%** of the data will be used for testing.
    - The performance of the two algorithms will be evaluated based on the **average accuracy** of the 30 models built.
4. **Performance Metrics**:
    - **Accuracy**: The percentage of correctly classified instances.
    - **Precision**: The ratio of true positives to the total predicted positives.
    - **Recall**: The ratio of true positives to the total actual positives.
5. **Statistical Analysis**:
    - The performance of the two algorithms will be compared using a **student's t-test** to test the hypothesis.
    - The **p-value** will be calculated to determine if there is a significant difference in accuracy between the two algorithms.

Results

The results of the statistical analysis will be presented in a table showing the mean, standard deviation, and p-value for each machine learning algorithm test. The p-value will indicate whether there is a significant accuracy difference between the performance of the two algorithms.

Translating the p-value resulting from the statistical analysis

- If the p-value is less than 0.05, we can reject the null hypothesis and conclude that there is a significant difference in the performance of the two algorithms.
- If the p-value is greater than 0.05, we fail to reject the null hypothesis and conclude that there is no significant difference in performance between the algorithms.

Tools and resources

1. **Weka**: A machine learning platform for data preprocessing, classification, and evaluation.
2. **Pima Indians Diabetes Dataset**: Available in Weka.
3. **Sheridan Library**: To select papers related to Naive Bayes, k-NN, and diabetes prediction using machine learning.
4. **MS Excel**: For conducting statistical analysis.

Deliverables

1. A Formal Report:
    - Introduction:
        - Describe the importance of detecting diabetes.
        - Relate the previous point to diabetes since that is the dataset we are using.
        - Describe briefly why Naive Bayes and k-NN are good choices for this problem.
        - Write the null hypothesis and explain why you expect there will be different performance.
    - Literature Review:
        - What is Naive Bayes and how does it work?
        - What is k-Nearest Neighbors (k-NN) and how does it work?
        - Using Naive Bayes for predicting diabetes.
        - Using k-NN for predicting diabetes.
    - Methods:
        - Describe the experiment design.
        - Describe the data, clarify the data source, and explain the different columns in the dataset and their data types.
        - Describe the metrics: Accuracy, Precision, and Recall.
        - Describe the data analysis (your hypothesis test method) — the student's t-test.
    - Results, Analysis, Discussion, and Conclusion:
        - Insert a table or chart that includes the mean accuracy and p-value for each machine learning algorithm.
        - Describe the results, including the test hypothesis results.
        - Explain why the hypothesis is accepted or rejected.
        - Interpret your results.

2. Weka Output / Excel Files:
    - Include the Weka output files / the Excel file used to conduct your analysis.
      This file should show your work with the comments/explanation of your steps

3. A Three-Minute Video:
    - A three-minute video to present your work, including the purpose, the hypothesis, the method, the results and the conclusion. You can prepare a slide deck to summarize each section of your report and record your screen while presenting it. However, you must also appear in the recorded video.

Grading Criteria for the Project

Category	Weight (%)	Description
Introduction (Including Hypothesis Statement)	10%	Clearly describes the importance of detecting diabetes, explains why NB and k-NN are good choices, and presents a well-defined hypothesis.
Literature Review	15%	Discusses how NB and k-NN work, their applications in diabetes detection, and references relevant research papers.
Methods (Experiment Design & Data Description)	20%	Describes the dataset (source, features, data types), explains data preprocessing steps, and details the experiment design, including how the models are built and evaluated.
Metrics and Data Analysis	15%	Explains the performance evaluation metrics (accuracy, precision, recall) and the statistical hypothesis test (e.g., t-test).
Results & Discussion	20%	Includes a well-structured table or chart comparing the mean accuracy of the models and p-value results. Provides a clear interpretation of the hypothesis test outcome (acceptance or rejection).
Deliverables Quality (Excel/ Weka Files)	10%	The submitted files (Excel) must be well-documented, with comments explaining each step. Weka models and dataset should be included.
Presentation Video (3-Minute Summary)	10%	The video should clearly explain the project purpose, methodology, results, and conclusion, with the presenter appearing in the video.

Total: 100%