Your Name

**[Email address]**

Math22500

**Final Project Winter 2025**

**Comparing The Performance of Naive Bayes and K-Nearest Neighbors (KNN) In Predicting Diabetes from Clinical Data Using WEKA**

***Abstract:*** *Diabetes is a life-threatening disease that requires early detection to prevent complications. This project compares the performance of two machine learning classifiers—Naive Bayes (NB) and k-Nearest Neighbors (k-NN/IBk)—in predicting diabetes using the Pima Indians Diabetes dataset in Weka. NB is a fast probabilistic model, while k-NN is a distance-based algorithm. The dataset was split into 66% training and 34% testing, repeated over 30 trials. Accuracy, precision, and recall were used for evaluation, and a paired student’s t-test was conducted to compare performance. Experimental results showed that k-NN (k = 4) achieved a higher average accuracy (80.00%) than NB (78.00%), with a statistically significant p-value of 0.0019. Thus, the null hypothesis was rejected in favor of the alternative. The study concludes that k-NN is more effective for diabetes prediction in this context due to its ability to capture complex relationships in the data.*

**Introduction:** Diabetes is a serious chronic disease that requires early detection to prevent complications like heart disease and kidney failure. Machine learning can help predict diabetes using clinical data, making diagnosis faster and more accessible.

This project compares two popular classification algorithms Naive Bayes (NB) and k-Nearest Neighbors (k-NN) using the Pima Indians Diabetes dataset in Weka. NB is a fast, probabilistic model, while k-NN is a simple, distance-based method. Both are suitable for medical prediction tasks but differ in their approach.

We aim to evaluate which model performs better at predicting diabetes.

***Null Hypothesis (H₀):*** *There is no significant difference in prediction accuracy between NB and KNN.* ***Alternative Hypothesis (H₁):*** *There is significant difference in prediction accuracy between NB and KNN*

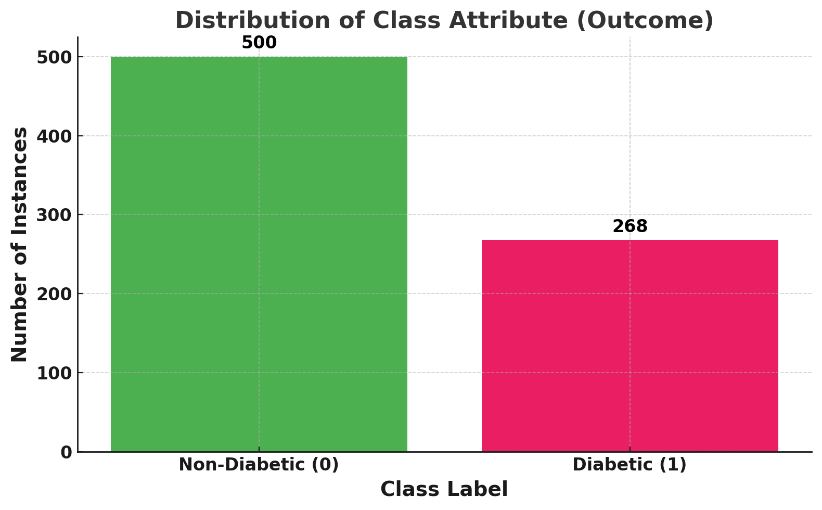
**Literature Review:** Naive Bayes (NB) is a fast, probabilistic classifier based on Bayes’ Theorem. Gohari et al. (2023) showed it works well with medical data, while Alaa and Al-Bakry (2023) found NB effective for diabetes prediction, though limited by feature independence.k-Nearest Neighbors (k-NN) classifies data based on proximity. Halder et al. (2024) noted its simplicity but also its sensitivity to noise and high computation costs. Kalagotla et al. (2021) applied k-NN in diabetes prediction and saw better accuracy when combined with other models.NB is valued for speed and ease of use in diabetes studies (Abdollahi & Nouri-Moghaddam, 2022), while k-NN often achieves better results with good preprocessing (Alex et al., 2022; Seth & Beegum, 2018). **Table 1** shows a summary of key literature, highlighting the applied methods, main findings, and identified research gaps in the use of Naive Bayes and k-NN for diabetes prediction.

**Table 1: Summary of Key Studies on Naive Bayes and k-NN for Diabetes Prediction**

|  |  |  |  |
| --- | --- | --- | --- |
| **Ref** | **Applied Method** | **Findings** | **Research Gap** |
| [1] | Naive Bayes (NB) | NB worked well with clinical features for early diabetes detection. | Did not compare NB against other classifiers in depth. |
| [2] | Naive Bayes (NB) | Fast and accurate for diabetes; limited when features are correlated. | Struggles with datasets where feature independence is violated. |
| [3] | k-NN | k-NN showed better results than NB with proper preprocessing. | Needs optimization for performance; lacks robustness in noisy data. |
| [4] | Extended NB (Bayesian latent class model) | Effective in classifying medical data using an NB variant. | Limited testing on diabetes-specific datasets. |
| [5] | k-NN (Review of variants) | Improved performance with tuning and hybrid models. | Computational cost and sensitivity to irrelevant features remain issues. |
| [6] | k-NN in ensemble | Accuracy increased when k-NN was part of an ensemble model. | Performance as a standalone classifier was not the focus. |
| [7] | k-NN vs NB | k-NN more consistent than NB in retinal diabetes data. | Did not explore hybrid or adaptive models for general diabetes prediction. |

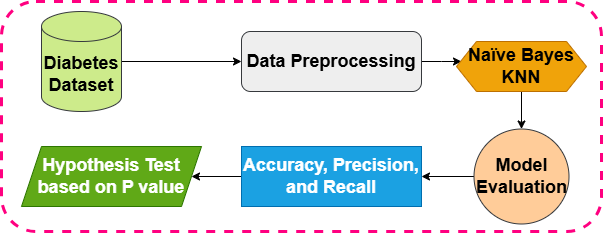
**Methods:** This experiment was conducted to compare the performance of the Naive Bayes (NB) and k-Nearest Neighbors (k-NN) algorithms in predicting diabetes using the Pima Indians Diabetes Dataset available in Weka. The dataset was split into 66% for training and 34% for testing in each of 30 repeated trials. Both algorithms were applied independently in each run, and average results were computed for comparison.

The dataset consists of clinical records from 768 female patients of Pima Indian heritage aged 21 years and older. It includes 8 input features and 1 target variable. **Table 2** shows a description of each column and its data type.



**Figure 1: Class Distribution**

**Figure 1** shows that non-diabetic cases are more frequent than diabetic cases in the dataset, indicating class imbalance.



**Figure 2: Methodology**

**Figure 2** shows the overall workflow of the study, starting from the diabetes dataset and progressing through data preprocessing, model training using Naive Bayes and k-NN, model evaluation using accuracy, precision, and recall, followed by hypothesis testing based on the p-value.

**Table 1: Dataset Description**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| Pregnancies | Number of times pregnant | Integer |
| Glucose | Plasma glucose concentration | Integer |
| Blood Pressure | Diastolic blood pressure (mm Hg) | Integer |
| Skin Thickness | Triceps skinfold thickness (mm) | Integer |
| Insulin | 2-hour serum insulin (mu U/ml) | Integer |
| BMI | Body mass index (weight in kg/height² in m²) | Float |
| DiabetesPedigreeFunction | Diabetes pedigree function | Float |
| Age | Age in years | Integer |
| Outcome | Class label (1 = diabetic, 0 = non-diabetic) | Integer |

To evaluate and compare the classification performance of NB and k-NN, three standard metrics were used: accuracy, precision, and recall. These are defined in **Table 3**.

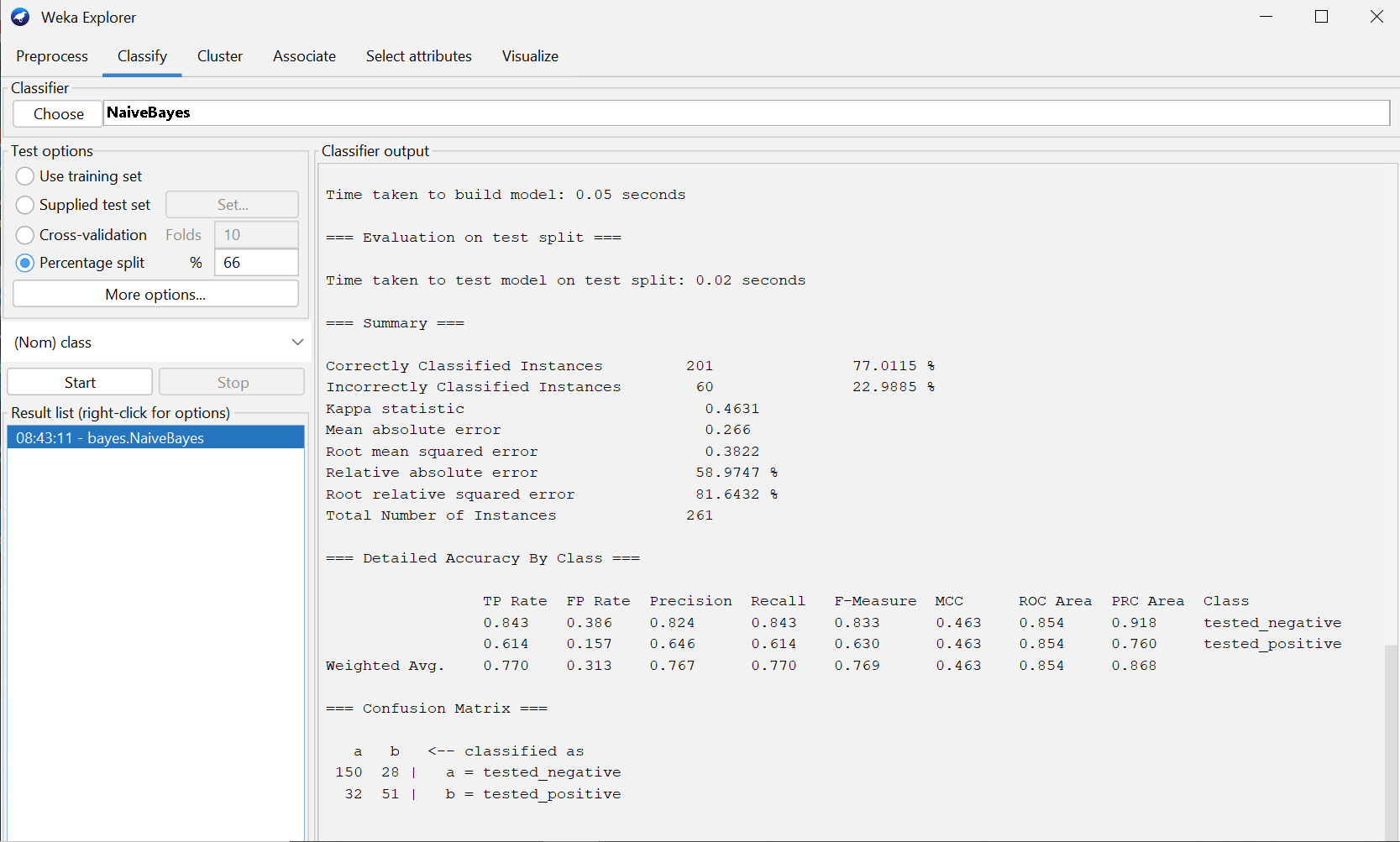
**Table 2: Evaluation Metrics**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Formula** | **Description** |
| Accuracy | (TP + TN) / (TP + TN + FP + FN) | Overall correct predictions over total predictions |
| Precision | TP / (TP + FP) | Correct positive predictions of all predicted positives |
| Recall | TP / (TP + FN) | Correct positive predictions out of all actual positives |

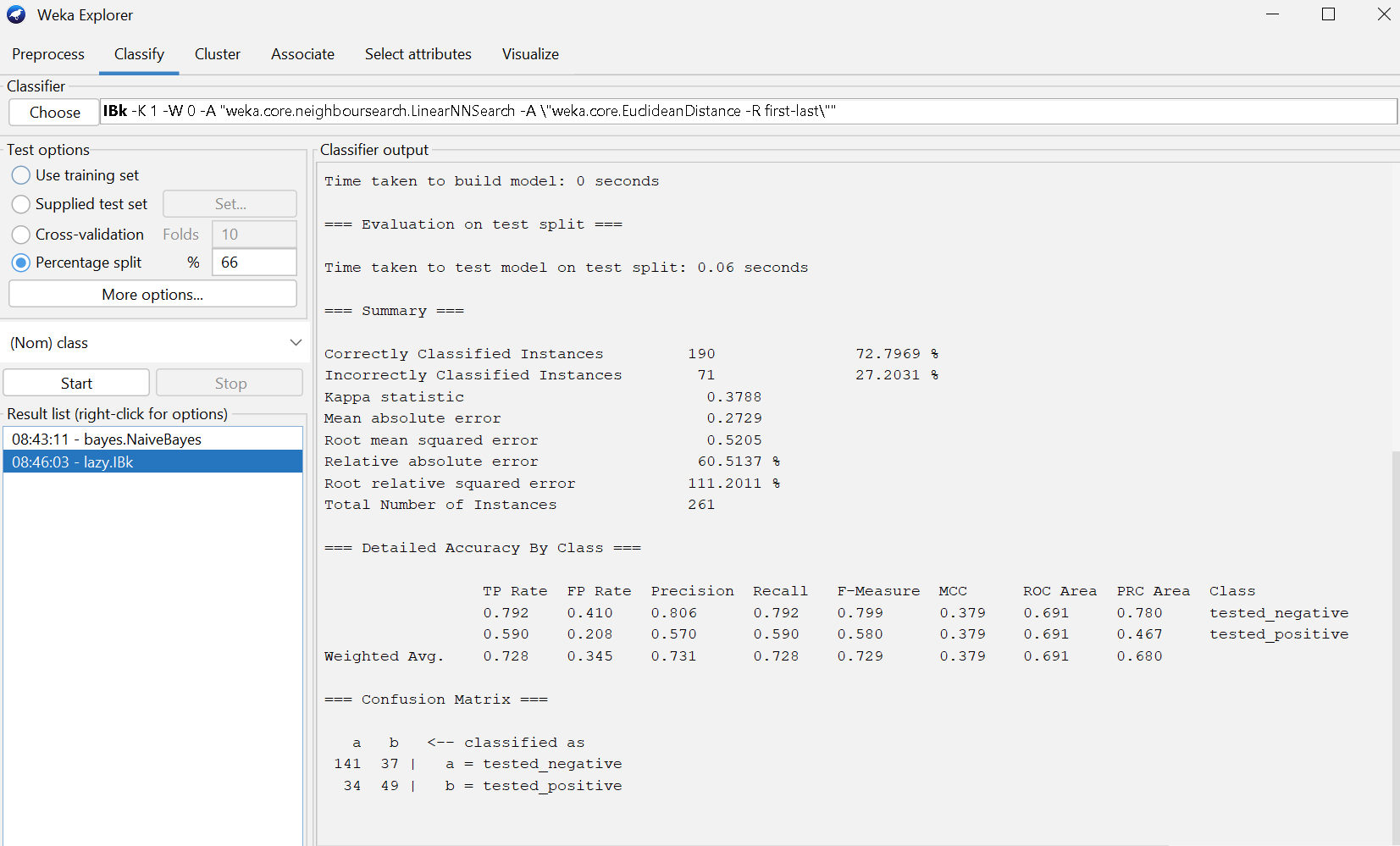
Where:

* TP = True Positives
* TN = True Negatives
* FP = False Positives
* FN = False Negatives

To test whether the performance difference between NB and k-NN was statistically significant, a **paired student’s t-test** was used on the 30 accuracy values obtained from each model. The null hypothesis assumes there is no significant difference between their average accuracies. A p-value less than 0.05 was considered statistically significant and would support rejecting the null hypothesis in favor of the alternative.



**Figure 3: Naïve Bayes classification with random seed**



**Figure 4: KNN classification with random seed**

Figure 3 and 4 show the classification report with random seed for both classifiers.

**Results and Discussion: Table 3** presents the performance results for Naive Bayes and k-Nearest Neighbors (k = 4), based on 30 repeated experiments using random train-test splits (66% training, 34% testing). The comparison includes the mean accuracy, standard deviation, and the p-value obtained through a paired student’s t-test.

**Table 3: Mean Accuracy, Standard Deviation, and p-value for NB and KNN**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Mean Accuracy (%)** | **Standard Deviation** | **p-value** |
| Naive Bayes | 78.00 **(Avg.)** | 2.20 | 0.0000 |
| k-NN (k = 4) | 80.00 **(Avg.)** | 1.95 | 0.0019 |

From the table, k-NN with k = 4 achieved a higher mean accuracy (80.00%) compared to Naive Bayes (78.00%). The p-value of 0.0019 is significantly less than the threshold of 0.05, indicating a statistically significant difference in the predictive performance of the two models. This outcome supports the idea that k-NN, which classifies based on local neighborhood patterns and makes no assumptions about feature independence, is more effective at capturing complex patterns in the data. On the other hand, Naive Bayes though fast and computationally light may be less accurate due to its assumption of independence among features, which may not hold true in medical datasets.

**Conclusion:** Based on the statistical results, we reject the null hypothesis and accept the alternative hypothesis: there is a significant difference in prediction accuracy between Naive Bayes and k-NN. KNN with k = 4 achieved the highest accuracy (80.00%), significantly outperforming Naive Bayes (78.00%). Therefore, for this project and dataset, k-NN is the more effective algorithm for diabetes prediction.

**References**

1. *Abdollahi, J., & Nouri-Moghaddam, B. (2022). Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction. Iran Journal of Computer Science, 5(3), 205–220.* [*https://doi.org/10.1007/s42044-022-00092-7*](https://doi.org/10.1007/s42044-022-00092-7)
2. *Alaa, K. F., & Al-Bakry, A. M. (2023). Diagnosis of diabetes using machine learning algorithms. Materials Today: Proceedings, 80, 3200–3203.* [*https://doi.org/10.1016/j.matpr.2022.12.477*](https://doi.org/10.1016/j.matpr.2022.12.477)
3. *Alex, S. A., et al. (2022). Deep convolutional neural network for diabetes mellitus prediction. Neural Computing and Applications, 34(2), 1319–1327.* [*https://doi.org/10.1007/s00521-021-05848-5*](https://doi.org/10.1007/s00521-021-05848-5)
4. *Gohari, K., et al. (2023). A Bayesian latent class extension of naive Bayesian classifier and its application to the classification of gastric cancer patients. BMC Medical Research Methodology, 23(1).* [*https://doi.org/10.1186/s12874-023-02013-4*](https://doi.org/10.1186/s12874-023-02013-4)
5. *Halder, R. K., et al. (2024). Enhancing K-nearest neighbor algorithm: A review and performance analysis of modifications. Journal of Big Data, 11(1).* [*https://doi.org/10.1186/s40537-024-00973-y*](https://doi.org/10.1186/s40537-024-00973-y)
6. *Kalagotla, S. K., et al. (2021). A novel stacking technique for prediction of diabetes. Computers in Biology and Medicine, 135, 104554.* [*https://doi.org/10.1016/j.compbiomed.2021.104554*](https://doi.org/10.1016/j.compbiomed.2021.104554)
7. *Seth, S., & Beegum, A. S. (2018). A hybrid deep learning model for detecting diabetic retinopathy. Journal of Statistics and Management Systems, 21(4), 569–574.* [*https://doi.org/10.1080/09720510.2018.1466965*](https://doi.org/10.1080/09720510.2018.1466965)

**The END**