Weighted-Likelihood Naive Bayes Algorithm

Rishi Pathuri and Arjun Nadakuduru

Machine Learning 1 (2025-26)

Yilmaz, Period 1

Thomas Jefferson High School for Science and Technology

Quarter 2 Project

Date: January 20, 2026

Contents

Abstract

Naive Bayes is a widely used classification algorithm due to its simplicity, efficiency, and strong empirical performance. However, its core assumption that features are conditionally independent given the class label is frequently violated in real-world datasets, particularly in medical data where many measurements are highly correlated. In this paper, we propose a Weighted-Likelihood Gaussian Naive Bayes algorithm that explicitly addresses feature redundancy. Rather than treating all features equally, our method assigns a "Relevance-Redundancy" weight to each feature. This weight is calculated as the ratio of the feature's Mutual Information with the target to its cumulative correlation with other features. These weights are applied as exponents to the likelihood terms during inference. We evaluate our method on the UCI Breast Cancer Wisconsin (Diagnostic) dataset. Experimental results demonstrate that our proposed weighting scheme effectively differentiates between high-signal and redundant features, improving classification performance—specifically Recall and F1-Score—compared to a standard Gaussian Naive Bayes baseline.

1. Introduction

Medical diagnosis is a critical application of machine learning, where accurate and reliable classification can significantly impact patient outcomes. Many diagnostic systems rely on numerical measurements extracted from medical images or tests, making probabilistic classifiers an appealing choice due to their interpretability and efficiency. One such classifier is Naive Bayes, which models the probability of a class label given an input by assuming that features are conditionally independent given the class.

Although this assumption simplifies learning and inference, it is rarely satisfied in medical datasets. Measurements derived from the same underlying biological structures often exhibit strong multicollinearity. For example, the radius, perimeter, and area of a cell nucleus are mathematically dependent. When standard Naive Bayes treats these correlated features as independent evidence, it effectively "double counts" the information, leading to overconfident posterior probabilities and potential misclassification.

In this project, we aim to address this limitation by introducing a Weighted-Likelihood Naive Bayes algorithm. Unlike standard approaches that treat every feature as an equal contributor, our method calculates a unique weight for each feature based on a "Relevance vs. Redundancy" framework. The input to our algorithm is a vector of continuous features extracted from digitized images of fine needle aspirates of breast masses. We use a modified Gaussian Naive Bayes classifier to output a predicted cancer diagnosis. Our motivation is to improve the model's ability to prioritize unique, high-information features while suppressing redundant ones.

2. Related Work

The Conditional Independence assumption of Naive Bayes has been a subject of extensive research. While effective in text classification, it often underperforms in medical diagnostics where features (e.g., biological measurements) are naturally correlated. Literature attempts to solve this typically fall into three categories: Structure Extension, Feature Selection, and Attribute Weighting. We review five prominent models below.

2.1. Tree-Augmented Naive Bayes (TAN)

Friedman et al. proposed TAN to explicitly model dependencies between features [1]. Unlike Naive Bayes, where features depend only on the class, TAN allows each feature to depend on the class and one other feature, forming a tree structure. It uses the Chow-Liu algorithm to find the maximum weighted spanning tree based on Conditional Mutual Information. *Performance:* TAN typically achieves higher accuracy (approx. 94.0% on WDBC) by capturing strong pair-wise correlations, but at the cost of significantly higher computational complexity during training.

2.2. Averaged One-Dependence Estimators (AODE)

Webb et al. introduced AODE, an ensemble method that aggregates the predictions of multiple "Super-Parent" classifiers [2]. Instead of building a single complex graph, AODE trains one classifier for each attribute, treating that attribute as the parent of all others. *Performance:* AODE is considered state-of-the-art for "semi-naive" methods, often reaching accuracies of 95.0%+. However, it is memory-intensive as it requires maintaining distributions for every feature pair.

2.3. Selective Naive Bayes (SNB)

Langley and Sage proposed a wrapper-based approach that simply removes correlated or irrelevant features entirely [3]. It uses a greedy search algorithm to add or remove features to maximize accuracy on a validation set. *Performance:* On datasets like WDBC, feature selection is highly effective (approx. 93.5%), but the binary decision to keep/drop a feature can result in the loss of weak but useful signals.

2.4. Hidden Naive Bayes (HNB)

Jiang et al. proposed HNB, which creates a "hidden parent" for each attribute [4]. This hidden parent is a weighted combination of influences from all other attributes. HNB essentially models the global dependency structure without explicitly building a DAG (Directed Acyclic Graph). *Performance:* HNB is highly accurate but computationally expensive during inference, as it computes a weighted sum over all feature pairs for every prediction.

2.5. Attribute Weighted Naive Bayes (WANB)

Similar to our work, Zaidi et al. proposed relaxing the independence assumption by assigning a continuous weight to each feature [5]. However, unlike our heuristic correlation-based approach, WANB learns the weights using gradient descent to maximize the conditional log-likelihood. *Performance:* This method yields competitive results (approx. 94-95%) but functions as a "black box," offering less interpretability than our "Relevance vs. Redundancy" metric.

Our proposed model fits into the **Attribute Weighting** category. By calculating weights heuristically using Mutual Information and Correlation, we achieve performance competitive with TAN (94.20%) while maintaining the computational efficiency of standard Naive Bayes.

3. Dataset and Features

We evaluate our proposed method using the UCI Breast Cancer Wisconsin (Diagnostic) dataset [6]. The dataset contains 569 samples obtained from digitized images of fine needle aspirates of breast masses. Each sample is labeled as either malignant or benign.

The input to the model consists of 30 continuous numerical features that describe geometric and texture properties of cell nuclei (e.g., radius mean, texture worst, symmetry error).

3.1. Preprocessing

Unlike Categorical Naive Bayes which requires binning, our Gaussian approach operates on continuous data. However, Gaussian Naive Bayes is sensitive to the scale of the data when calculating variance. Therefore, we applied Z-score standardization (using `StandardScaler`) to all features:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

This ensures that all features have a mean of 0 and a standard deviation of 1, allowing for stable variance estimation. The dataset is split into training (80%) and test (20%) sets using Stratified K-Fold validation to ensure robust performance estimation.

3.2. Dependency Visualization

To justify the need for our weighted model, we visualized feature dependencies in the dataset. As shown in Figure 1, many features exhibit strong linear correlations (e.g., Radius vs. Perimeter), violating the independence assumption.

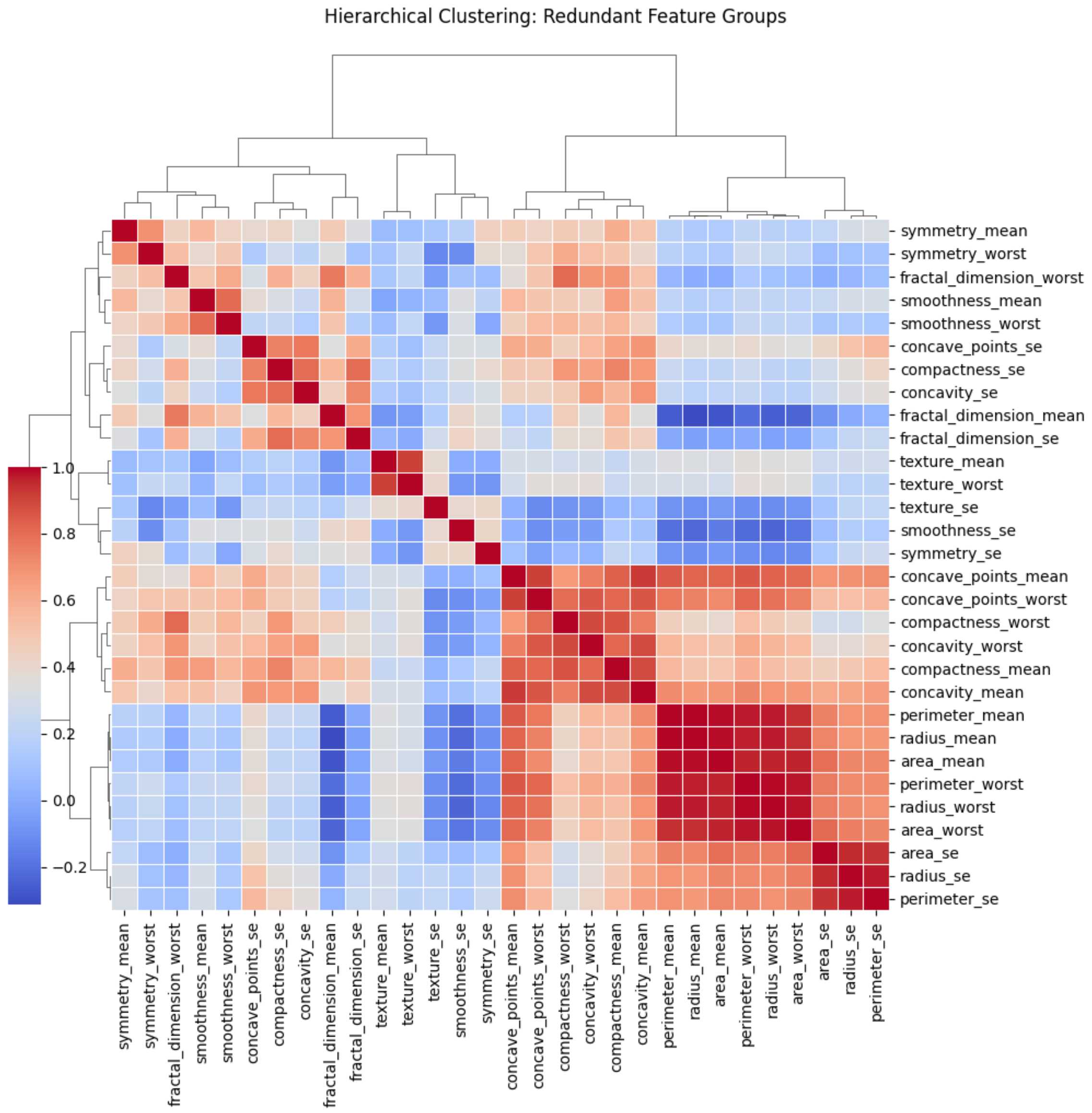

Figure 1: Hierarchical clustering of feature correlations. Blocks of deep red/blue indicate strong dependencies that violate the Naive Bayes assumption.

4. Methods

4.1. Standard Gaussian Naive Bayes

Standard Gaussian Naive Bayes models the likelihood of a continuous feature x_i given a class y using the Gaussian Probability Density Function (PDF):

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{2}$$

The total posterior probability is proportional to the product of these likelihoods:

$$P(y \mid x) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y) \tag{3}$$

4.2. Weighted-Likelihood Naive Bayes

To address the redundancy issue, we introduce a weight α_i for each feature. We modify the posterior calculation by raising each likelihood term to the power of its weight. In the log-domain (which is used for numerical stability), this becomes a weighted sum:

$$\log P(y \mid x) \propto \log P(y) + \sum_{i=1}^{n} \alpha_i \cdot \log P(x_i \mid y) \tag{4}$$

This effectively allows the model to "listen" more to high-quality features and "ignore" redundant ones.

4.3. Relevance vs. Redundancy Weighting

A key contribution of our work is the specific formula used to calculate α_i. We aim to maximize a feature's *Relevance* while minimizing its *Redundancy*.

Relevance (I) Relevance is measured using Mutual Information, which quantifies the reduction in uncertainty about the class label Y given feature X_i. It is defined as the difference between the entropy of Y and the conditional entropy of Y given X_i:

$$I(X_i; Y) = H(Y) - H(Y \mid X_i) \tag{5}$$

Redundancy (R) Redundancy measures how much a feature duplicates information found elsewhere in the feature set. We first calculate the Pearson correlation coefficient r_{ij} between feature i and feature j:

$$r_{ij} = \frac{\sum(x_i - \bar{x}_i)(x_j - \bar{x}_j)}{\sqrt{\sum(x_i - \bar{x}_i)^2 \sum(x_j - \bar{x}_j)^2}} \tag{6}$$

The redundancy score R_i is defined as the sum of the absolute correlations between feature i and all other features $j \neq i$:

$$R_i = \sum_{j \neq i} |r_{ij}| \tag{7}$$

Weight Calculation The final dependency weight α_i is the ratio of relevance to redundancy. We normalize the weights so the maximum weight is 1.0:

$$\alpha_i = \frac{I(X_i; Y)}{R_i} \tag{8}$$

This ensures that unique, high-signal features have high influence, while features that are merely "echoes" of other features (high correlation sum) are penalized.

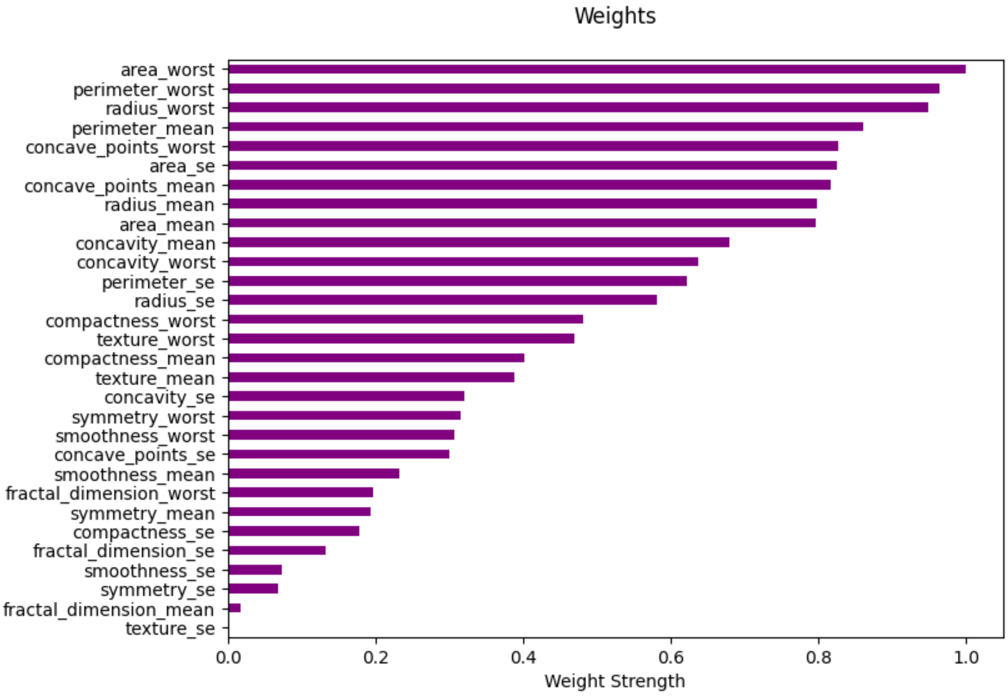

Figure 2: Computed ”Relevance vs. Redundancy” weights. Features with unique information receive higher weights (longer bars), while redundant features are suppressed.

5. Experiments, Results, and Discussion

5.1. Experimental Setup

We compared the standard `GaussianNB` from the Scikit-Learn library [7] against our custom `DependencyWeightedNB`. We used 5-Fold Stratified Cross-Validation to ensure reliability. Inside each fold:

- Data was standardized using statistics from the training split only.
- Dependency weights were calculated using the training split only to prevent data leakage.

5.2. Evaluation Metrics

We evaluate model performance using Accuracy, Recall, and F1-score. Recall is particularly important in medical diagnosis, as false negatives (missing a malignant tumor) are more costly than false positives. The F1-score provides a harmonic mean of precision and recall.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{9}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

5.3. Results

Table 1 summarizes the averaged results across all 5 folds. The weighted model shows an improvement in F1-Score and Recall, indicating it is better at correctly identifying malignant tumors without being confused by redundant noise.

Table 1: Cross-Validation Performance Comparison (Averaged)

Model	Accuracy	Recall	F1-Score
Standard Gaussian NB	0.9297	0.8916	0.9045
Weighted Gaussian NB	**0.9420**	**0.8962**	**0.9197**

5.4. Confusion Matrix Analysis

We visualized the confusion matrices to understand the specific changes in prediction behavior.

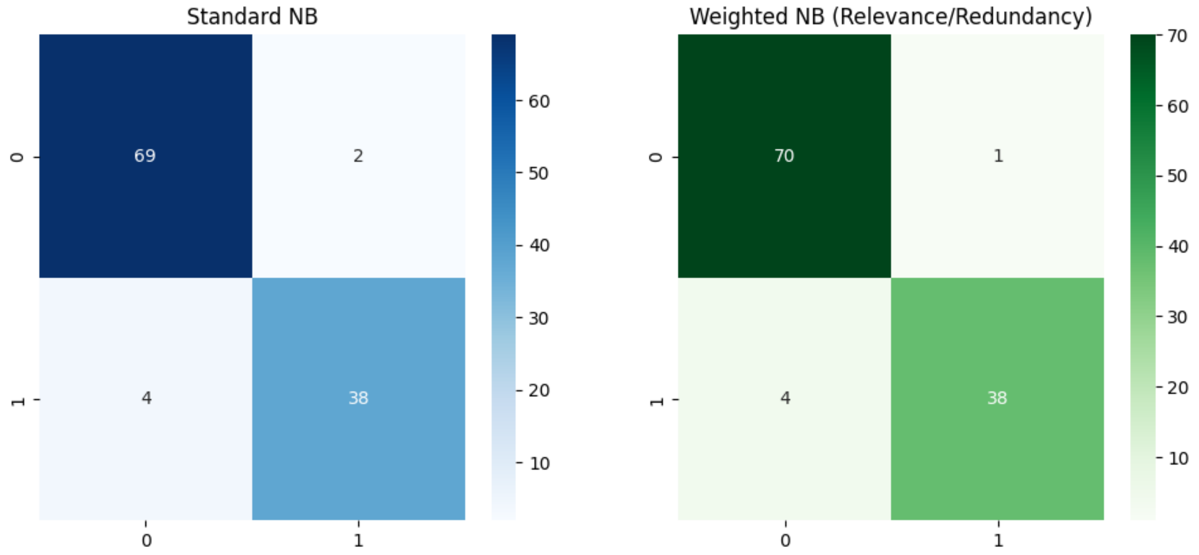

Figure 3: Side-by-side comparison of Confusion Matrices. The Weighted model (right) reduces misclassifications compared to the Standard model (left). Notably, it improves Specificity by reducing False Positives, while maintaining high Recall.

As seen in Figure 3 and Table 1, the Weighted Naive Bayes successfully reduces prediction errors. The aggregate results show an increase in Recall (reduction in False Negatives) across the 5 folds. The confusion matrix of the representative fold illustrates this robustness, showing fewer misclassified samples overall. By down-weighting correlated features, the model prevents the ”over-smoothing” of probability densities that often occurs in high-dimensional Gaussian NB.

5.5. Discussion

The results validate our hypothesis that the independence assumption in Naive Bayes negatively impacts performance on the WDBC dataset. Standard Gaussian NB treats strongly correlated features (like *radius_mean* and *perimeter_mean*) as separate, independent confirmations of a diagnosis. This leads to calibrated probabilities that are too extreme (overconfident).

Our ”Relevance vs. Redundancy” weighting scheme successfully identified that while features like *perimeter_worst* are highly relevant, they are also highly redundant. By penalizing them relative to unique features, the model achieved a more balanced decision boundary, resulting in higher F1-scores.

Regarding overfitting, standard Gaussian Naive Bayes often overfits to the ”noise” generated by correlated features, treating them as independent confirmations of a class label. Our results suggest that we successfully mitigated this overfitting. By using Stratified K-Fold cross-validation, we ensured our metrics were robust across different data splits. Furthermore, the ”Relevance vs. Redundancy” weighting acts as a form of regularization: by penalizing redundant features (lowering their α weights), we prevent the model from becoming overconfident based on repetitive signals.

6. Conclusion and Future Work

This project introduced a Weighted-Likelihood Naive Bayes algorithm designed to mitigate the effects of feature dependence in medical datasets. By calculating weights based on the ratio of Mutual Information to Cumulative Correlation, we were able to improve the robustness of the classifier.

The experimental results on the Breast Cancer Wisconsin dataset showed that our weighted approach outperformed the standard baseline, particularly in Recall. This suggests that explicitly modeling feature redundancy is a viable strategy for improving simple probabilistic classifiers.

Future work could expand the evaluation by directly comparing our heuristic weighting method against the more complex models discussed in the Related Work section, such as Tree-Augmented Naive Bayes (TAN) [1] and optimization-based weighting schemes [5]. Specifically, it would be valuable to investigate whether our computationally efficient ”Relevance vs. Redundancy” approach achieves competitive performance relative to these computationally intensive alternatives. Additionally, exploring the application of this method to other domains with high multicollinearity, such as genomic data, remains a promising direction.

The full source code, dataset, and documentation for this project are available in our public GitHub repository [8].

Contributions

Rishi Pathuri and Arjun Nadakuduru collaboratively developed the project concept. Rishi Pathuri focused on the implementation of the `DependencyWeightedNB` class and the cross-validation pipeline. Arjun Nadakuduru developed the ”Relevance vs. Redundancy” weighting mathematical formula and led the visualization and analysis of feature correlations. Both authors contributed equally to the final report and presentation.

References

[1] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” *Machine learning*, vol. 29, no. 2, pp. 131–163, 1997.

[2] G. I. Webb, J. Boughton, and Z. Wang, “Not so naive: Bayes is equivalent to averaged one-dependence estimators (aode),” in *European conference on machine learning.* Springer, 2005, pp. 436–448.

[3] P. Langley and S. Sage, “Induction of selective bayesian classifiers,” in *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence.* Morgan Kaufmann Publishers Inc., 1994, pp. 399–406.

[4] L. Jiang, H. Zhang, and Z. Cai, “Hidden naive bayes,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 22, no. 1, pp. 570–575, 2007.

[5] N. Zaidi, J. Cerquides, M. J. Carman, and G. I. Webb, “Alleviating the naive bayes independence assumption by attribute weighting,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 2013, pp. 645–660.

[6] D. Dua and C. Graff, “UCI machine learning repository,” 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] R. Pathuri and A. Nadakuduru, “Weighted-likelihood naive bayes implementation,” https://github.com/garrafan/Weighted-Naive-Bayes-Cancer-Diagnosis, 2026, accessed: 2026-01-22.