

## Weight Optimization of Weighted Naive Bayes Classifier for Efficient Classification

Gamzepelin AKSOY<sup>1\*</sup>, Murat KARABATAK<sup>2</sup>

<sup>1</sup> Isparta University of Applied Sciences, Department of Computer Engineering, Isparta/Türkiye.

ORCID: 0000-0002-5328-2983, E-mail: gamzepelinaksoy@isparta.edu.tr

<sup>2</sup> Fırat University, Department of Software Engineering, Elazığ/Türkiye.

ORCID: 0000-0002-6719-7421, E-mail: mkarabatak@firat.edu.tr

(Alınış/Arrival: 02.06.2025, Kabul/Acceptance: 13.06.2025, Yayınlanma/Published: 25.06.2025)

### Abstract

The Weighted Naive Bayes (WNB) classifier is an effective classification method based on the Naive Bayes algorithm; however, determining the appropriate weights within this algorithm remains a significant challenge. The number of optimization based WNB applications in the existing literature is quite limited. Grid Search methods used for weight determination often suffer from high computational costs and an inability to reach the global optimum. Although the Fast Weighted Naive Bayes classifier offers faster solutions, it remains limited in terms of classification performance. Therefore, optimizing weights is of great importance in achieving both computational efficiency and high classification accuracy. In this study, Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) methods were employed to optimize the weights of the WNB algorithm. The proposed GAW-NB and PSOW-NB models were evaluated on five different datasets using 5-fold cross-validation. The experimental results demonstrated that both methods achieved significant improvements in terms of both execution time and classification performance.

**Keywords:** Weighted Naive Bayes, Classification, Genetic Algorithm, Particle Swarm Optimization.

### Ağırlıklı Sade Bayes Sınıflandırıcısında Etkili Sınıflandırma İçin Ağırlıkların Optimizasyonu

#### Özet

Ağırlıklı Naive Bayes (WNB) sınıflandırıcısı, Naive Bayes algoritmasına dayanan etkili bir sınıflandırma yöntemidir ancak bu algorithmada ağırlıkların uygun biçimde belirlenmesi önemli bir problem olarak öne çıkmaktadır. Mevcut literatürde optimizasyon tabanlı WNB uygulamalarının sayısı oldukça sınırlıdır. Izgara arama yöntemi ile yapılan ağırlık belirleme süreçleri, yüksek hesaplama maliyeti ve küresel optimuma ulaşamama gibi dezavantajlara sahiptir. Hızlı Ağırlıklı Naive Bayes sınıflandırıcısı daha kısa sürede çözüm sunsa da sınıflandırma performansı açısından sınırlı kalmaktadır. Bu nedenle, ağırlıkların optimize edilmesi hem zaman verimliliği hem de sınıflandırma doğruluğu açısından büyük önem taşımaktadır. Bu çalışmada, WNB algoritmasındaki ağırlıkların optimizasyonu için Genetik

Algoritma (GA) ve Parçacık Sürü Optimizasyonu (PSO) yöntemleri kullanılmış; geliştirilen GAW-NB ve PSOW-NB modelleri beş farklı veri kümesi üzerinde 5 katlı çapraz doğrulama yöntemiyle test edilmiştir. Deneysel sonuçlar, her iki yöntemin de hem işlem süresi hem de sınıflandırma başarımı açısından anlamlı iyileşmeler sağladığını göstermektedir.

**Anahtar Kelimeler:** Ağırlıklı Sade Bayes, Sınıflandırma, Genetik Algoritma, Parçacık Sürü Optimizasyonu.

## 1. INTRODUCTION

Recent technological advancements have led to a significant increase in data generation across nearly all areas of the digital world. The broader reach of internet access, the integration of mobile devices as an indispensable part of daily life, and innovations such as the Internet of Things (IoT) have caused an exponential growth in data volume. This surge has made the accurate analysis of data and its transformation into meaningful insights inevitable, thereby increasing the importance of fields such as data analytics, Artificial Intelligence (AI), and big data.

Data mining constitutes a foundational methodology employed to extract meaningful patterns and knowledge from large-scale datasets. Core techniques within this domain include association rule mining, clustering, and classification. Among these, association rule mining is particularly instrumental in identifying and elucidating relationships among diverse attributes within databases, thereby revealing the underlying structure of such associations [1].

Clustering aims to generate meaningful data segments by grouping similar records within large-scale databases. While this technique forms groups based on similarities among data points, classification, on the other hand, seeks to assign new data instances to the correct categories using pre-labeled data. Classification is a supervised learning approach that analyzes the relationship between features and class labels in the training data during the model-building phase. Subsequently, this trained model is used to predict the class labels of test data, where only the features are known.

In the literature, classification problems have been extensively studied, and the algorithms developed for such tasks have found wide applicability across various domains. The Naive Bayes classification algorithm, as one of the statistical classification techniques, is widely favored by researchers due to its computational efficiency and speed. Ravinder et al. emphasized that web data mining plays a crucial role in extracting meaningful information from large-scale textual data, and demonstrated that the Naive Bayes algorithm provides an effective method for text classification and categorization in this context [2]. Jain et al. developed a model for fake news detection by employing natural language processing (NLP) techniques and machine learning methods. In their study, it was observed that the Naive Bayes and Random Forest algorithms demonstrated superior performance compared to other approaches [3]. Arrazyan et al. conducted a study to examine the effectiveness of the Naive Bayes algorithm in predicting diabetes. In their experiments, various training-to-testing data split ratios were assessed, and the highest classification accuracy 88.16% was recorded when 65% of the data was allocated for training and 35% for testing [4].

Research aimed at improving the performance of Bayes classifiers suggests that weighting techniques can offer an effective strategy. In particular, incorporating weighting schemes into

the Naive Bayes classifier can enhance the model's flexibility and lead to improved classification accuracy [5]. However, determining the optimal weight values constitutes a critical optimization problem that directly influences classification accuracy. In the literature, the grid search method is commonly employed for this purpose. Nevertheless, due to its reliance on nested loops, this method incurs high computational costs when applied to large datasets. This challenge underscores the need for more efficient approaches that can reduce processing time while maintaining classification performance. In this context, optimization algorithms offer a promising alternative for the effective determination of weight values [6]. In the study conducted by Lin et al., the Particle Swarm Optimization (PSO) algorithm was employed to optimize the weights in the Weighted Naive Bayes classifier, and its performance was evaluated against the standard Naive Bayes algorithm using multiple benchmark datasets [7].

To date, numerous optimization algorithms have been developed and applied to various types of problems. In particular, metaheuristic optimization techniques have emerged as powerful methods that provide effective solutions to nonlinear problems. Among these techniques, the most well-known and widely used algorithms are the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), both of which have found broad applicability across different domains [8]. Sun et al. proposed the automatic design of CNN architectures for image classification using GA. Their study reported that the GA-based approach outperformed both manual and other automated design methods in terms of classification accuracy, while also requiring lower computational cost [9]. Polap developed an adaptive technique that combines GA with convolutional classifiers for the analysis of microscopy images. The results demonstrated that the GA-based approach achieved 7.5% higher accuracy compared to traditional methods [10].

Particle Swarm Optimization (PSO) has been proposed for clustering problems and tested on two synthetic and five real-world datasets. The findings indicate that the algorithm successfully solves clustering tasks, demonstrating its effectiveness in this domain [11]. Huang developed a novel PSO-based hybrid cluster validity method to address complex clustering and classification problems. In his study, the performance of the proposed method was compared with semi-supervised classification, decision tree, and the Huang index method. The results demonstrated that the proposed approach outperformed existing methods in both clustering and classification tasks [12].

In this study, two methods are presented to improve the speed and efficiency of the classification process: the Genetic Algorithm-based Weighted Naive Bayes (GAW-NB) and the Particle Swarm Optimization-based Weighted Naive Bayes (PSOW-NB). The performance of the developed methods is evaluated by comparing them with the traditional Naive Bayes and Weighted Naive Bayes algorithms. In this context, various datasets from the literature are utilized to analyze accuracy rates and demonstrate the effectiveness of the methods.

The remainder of this paper is organized as follows. In Section 2, the fundamentals of the WNB classifier are introduced. Section 3 describes the optimization process and details the methods used to enhance the performance of the WNB classifier. Section 4 presents the experimental study, including dataset descriptions and implementation details. Section 5 reports and discusses the experimental results obtained through performance evaluations. Finally, Section 6 concludes the paper with a summary of the findings and suggestions for future research.

## **2. WEIGHTED NAIVE BAYES CLASSIFIER**

Bayesian classifiers are among the statistical classification techniques and are preferred by researchers due to their computational efficiency and strong performance [13]. The Weighted Naive Bayes (W-NB) classifier is an enhanced version of the standard Naive Bayes classifier, developed by incorporating weight values into its structure.

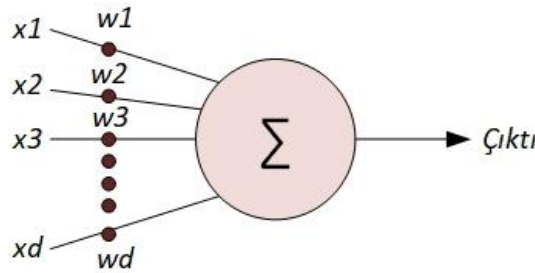
The primary purpose of the weighting function is to modify or enhance the influence of each individual feature on the outcomes. A positive weight function defined on a set A, denoted as  $\omega:A \rightarrow R^+$ , is expressed as a positive function over A. If the weight function is defined as  $\omega(a) = 1$  for all  $a \in A$ , then all elements are assigned equal weight and the system can be regarded as unweighted.

$$\sum_{a \in A} f(a) \tag{1}$$

However, given a weight function  $\omega:A \rightarrow R^+$  the weighted sum is expressed as follows:

$$\sum_{a \in A} f(a)\omega(a) \tag{2}$$

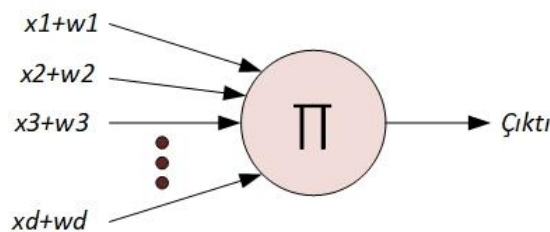
The weighting process within the structure of an artificial neural network is illustrated in Figure 1.



**Figure 2. 1** The weighting process for the summation function

This weighting approach is also applicable to the Naive Bayes classifier. Nevertheless, since the primary operation in the Bayes classifier is multiplication, weights are integrated by adding them to the input values rather than multiplying. The principle behind incorporating these weights into the input values is detailed in Equation 3 and illustrated in Figure 2 [6].

$$\prod_{a \in A} (f(a) + \omega(a)) \tag{3}$$



**Figure 2. 2** The weighting process for the multiplication function

### **3. OPTIMIZATION PROCESS**

Optimization is the process of determining the values that decision variables should take in order to maximize the value of an objective function defined under certain constraints [14]. In the literature, various optimization algorithms have been applied to different types of optimization problems. Optimization techniques are generally categorized into three main groups: analytical methods, heuristic methods, and metaheuristic methods.

Genetic Algorithm and Particle Swarm Optimization are among the most well-known and widely used algorithms within the family of metaheuristic methods. In this study, Genetic Algorithm and Particle Swarm Optimization were employed to optimize the weights of the Weighted Naive Bayes (W-NB) classifier.

#### **3.1. Genetic Algorithm**

Genetic Algorithm (GA) is one of the earliest and most widely used optimization algorithms in the field of evolutionary computation. Inspired by natural evolutionary processes, this algorithm was developed to solve search and optimization problems. Introduced by John Henry Holland in 1975, GA is based on the principle of natural selection. In this algorithm, solutions are obtained through a randomized search within a defined solution space.

Genetic Algorithms simulate natural selection and biological reproduction processes to enable the evolution of the fittest individual over successive generations. Operating without reliance on strict mathematical formulations, this method derives optimal solutions from the most successful individuals of previous generations, and is considered a nonlinear and stochastic process [15]. The ability of genetic algorithms to generate effective solutions in a relatively short amount of time is considered one of their major advantages.

The fundamental steps of the Genetic Algorithm can be summarized as follows [16]:

1. All possible solutions are encoded as binary strings.
2. The initial population is generated from randomly selected sets of candidate solutions.
3. The fitness value of each individual is calculated using a predefined fitness function.
4. Reproduction is performed based on the fitness values of the individuals.
5. The fitness values of the newly generated individuals are evaluated; crossover and mutation operations are applied to produce a new population.
6. All processes are repeated until a predefined number of generations is reached or a stopping criterion is satisfied.
7. Once the number of iterations reaches the stopping criterion, the best individuals are selected according to the fitness function, and the process is terminated.

#### **3.2. Particle Swarm Optimization Algorithm**

Particle Swarm Optimization (PSO) is a widely recognized optimization technique categorized under swarm intelligence algorithms, which draws inspiration from the collective behavior observed in natural phenomena, such as bird flocking or fish schooling. Originally proposed by Kennedy and Eberhart in 1995 [17], PSO operates similarly to Genetic Algorithms by starting with a randomly initialized set of candidate solutions. Unlike Genetic Algorithms, however, PSO does not use crossover or mutation mechanisms. Instead, it guides the search for optimal

solutions by continuously updating particle positions based on each particle’s historical best (personal best) and the overall best position (global best) within the population.

The steps of the Particle Swarm Optimization (PSO) algorithm are outlined as follows [18]:

1. The initial positions and velocities of the particles are randomly generated in a d-dimensional search space.
2. The optimization fitness function, defined over d variables, is evaluated for each particle.
3. The fitness value of each particle is compared with its personal best value. If the current fitness is better than the personal best, the personal best is updated and the current position is recorded as the new personal best.
4. The fitness value is also compared with the global best value in the population. If the current fitness is better than the global best, the global best is updated accordingly, along with the index and value of the corresponding particle.
5. The velocities and positions of particles are updated according to the standard PSO update equations. These updates allow the particles to move through the search space toward better solutions. Velocity and position updates are performed using Equation (4) and Equation (5), respectively.

$$V_{id} = V_{id} + c_1 * rand() * (P_{id} - X_{id}) + c_2 * rand() * (G_{id} - X_{id}) \quad (4)$$

$$X_{id} = X_{id} + V_{id} \quad (5)$$

6. Iterations continue from Step 2 until convergence criteria are met or the maximum number of generations is completed.

## 4. EXPERIMENTAL STUDY

In this study, the GA and PSO methods were applied to optimize the weights of the Weighted Naive Bayes (W-NB) classifier. Two variants, namely the Genetic Algorithm-based Weighted Naive Bayes Classifier (GAW-NB) and the Particle Swarm Optimization-based Weighted Naive Bayes Classifier (PSOW-NB), were tested on five distinct datasets. The performance of the proposed methods was evaluated by comparing them with each other. All datasets were obtained from the UCI Machine Learning and Intelligent Systems Repository [19]. The characteristics of the datasets used in the experimental study are presented below.

### 4.1. Tic-Tac-Toe Dataset

Tic-tac-toe is one of the oldest and most popular logic-based games. The game is played on a 3x3 board using the symbols “x” and “o”. The Tic-tac-toe dataset encodes all possible board configurations under the assumption that the "x" player always makes the first move. The dataset contains 9 features and a total of 960 instances. The attributes of the Tic-tac-toe dataset are presented in Table 1. In the dataset, “x” represents the x-player, “o” represents the o-player, and “b” denotes a blank space [20].

**Table 1.** Attribute information of the Tic-Tac-Toe dataset

Attributes	Values
------------	--------

Top-left	o, b, x
Top-middle	o, b, x
Top-right	o, b, x
Middle-left	o, b, x
Middle-middle	o, b, x
Middle-right	o, b, x
Bottom-left	o, b, x
Bottom-middle	o, b, x
Bottom-right	o, b, x
Class	Positive, Negative

## 4.2. Post-Operative Patient Dataset

This dataset is designed to classify patients' postoperative conditions and assist in determining the appropriate hospital unit for referral following surgery. The postoperative period, which begins after a major surgical intervention, is often marked by complications such as hypothermia and is considered a critical phase in patient recovery. The dataset comprises 8 features and includes 90 instances. The detailed attributes of the dataset are provided in Table 2 [21].

**Table 2.** Feature details of the dataset for postoperative patient classification

Attributes	Values
Internal Temperature (°C)	High (>37), Medium (36–37), Low (<36)
Surface Temperature (°C)	High (>36.5), Medium (35–36.5), Low (<35)
Oxygen Saturation (%)	Excellent (≥98), Good (90–98), Medium (80–90), Poor (<80)
Last Blood Pressure Measurement	High (>130/90), Medium (90/70–130/90), Low (<90/70)
Surface Temperature Condition	Stable, Moderately Stable, Unstable
Internal Temperature Condition	Stable, Moderately Stable, Unstable
Blood Pressure Stability	Stable, Moderately Stable, Unstable
Perceived Comfort Level	0–20 (numerical value)
Class	I (Patient referred to Intensive Care Unit), S (Patient ready for discharge), A (Patient referred to general hospital floor)

## 4.3. The Qualitative Bankruptcy Dataset

The Qualitative Bankruptcy dataset is developed to estimate the likelihood of corporate bankruptcy based on a range of qualitative risk factors. The dataset includes 7 features and 250 instances. Detailed attribute information is provided in Table 3. The features are categorized into the following risk dimensions [22]:

- **Industrial Risk:** Evaluates factors such as government regulations, international treaties, market rivalry intensity, supply chain stability, macroeconomic fluctuations, competitiveness in domestic and global markets, and product lifecycle stages.
- **Management Risk:** Assesses managerial competency and stability, the relationship between management and ownership, effectiveness of human resource strategies, business growth strategies, operational outcomes, feasibility of strategic plans, and likelihood of success.
- **Financial Flexibility:** Measures the firm's capacity to access and utilize direct, indirect, and alternative financial resources.
- **Credibility:** Reflects the firm's credit history, reliability of financial disclosures, and strength of its relationships with financial institutions.
- **Competitiveness:** Examines market positioning, depth of core competencies, and uniqueness of business strategies adopted.

- **Operational Risk:** Covers aspects such as procurement stability and diversity, production efficiency, operational continuity, market demand, product and sales diversification, pricing strategies, and distribution network efficiency.

**Table 3.** Attribute information of the Qualitative Bankruptcy dataset

Attribute	Values
Industrial Risk	Positive, Average, Negative
Management Risk	Positive, Average, Negative
Financial Flexibility	Positive, Average, Negative
Credibility	Positive, Average, Negative
Competitiveness	Positive, Average, Negative
Operational Risk	Positive, Average, Negative
Decision	Bankrupt, Non-bankrupt

#### 4.4. The Mammographic Mass Dataset

The Mammographic Mass dataset is designed to differentiate between benign and malignant breast lesions. It may serve as a valuable tool for classifying the malignancy of mammographic masses using BI-RADS attributes and patient age information. After removing instances with missing values, the dataset comprises a total of 825 samples, including 400 benign and 425 malignant cases. Detailed information regarding the six attributes included in the dataset is provided in Table 4 [23].

**Table 4.** Attribute information of the Mammographic Mass dataset

Attribute	Values
BI-RADS Assessment	Ordinal values from 1 to 5
Age	Patient's age (integer)
Shape	Mass shape: 1 = Round, 2 = Oval, 3 = Lobular, 4 = Irregular (nominal)
Margin	Mass margin: 1 = Circumscribed, 2 = Microlobulated, 3 = Obscured, 4 = Ill-defined, 5 = Spiculated (nominal)
Density	Mass density: 1 = High, 2 = Iso-dense, 3 = Low, 4 = Fat-containing
Class	Malignant = 1, Benign = 0

#### 4.5. The Breast Cancer Dataset

This dataset was developed to support decision-making regarding the administration of radiation therapy in breast cancer patients, based on specific clinical and diagnostic features. It comprises 680 instances and includes 9 attributes. Detailed information about the features of the breast cancer dataset is provided in Table 5.

**Table 5.** Attribute information of the Breast Cancer dataset

Attribute	Values
Age	10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99
Menopause	Below 40, 40 and above
Tumor Size	0–4, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59
Involved Nodes	0–2, 3–5, 6–8, 9–11, 12–14, 15–17, 18–20, 21–23, 24–26, 27–29, 30–32, 33–35, 36–39
Node-Caps	No, Yes

Degree of Malignancy	1,2,3
Breast	Left, Right
Breast Quadrant	Left-upper, Left-lower, Right-upper, Right-lower, Central
Radiation	No, Yes
Class	no-recurrence-events, recurrence-events

#### 4.6. Weighted Naive Bayes Process

In the Weighted Naïve Bayes (W-NB) classifier, the number of weights corresponds to the number of attributes in the dataset. In the conventional Global W-NB approach, these weights are typically determined using the grid search method. In this study, the grid search technique was utilized to identify the optimal weight values, with all weights constrained within the range  $[0, 1]$ . However, as the number of attributes increases, the computational complexity of W-NB also grows due to the nested loop structure inherent in grid search. To mitigate this computational burden, a step size of 0.2 was adopted for weight intervals. Nevertheless, this method remains computationally expensive, especially with larger datasets. Therefore, the use of optimization algorithms is proposed as an effective alternative to enhance efficiency in determining weight values.

#### 4.7. Genetic Algorithm Process

In this study, each weight in the Weighted Naïve Bayes (W-NB) classifier was defined as a decision variable within the Genetic Algorithm (GA) framework. These decision variables were constrained to lie within the range  $[0, 1]$ . The classification accuracy served as the objective function of the algorithm, with the primary aim being to maximize this accuracy. To initiate the optimization, the initial population was randomly generated. Each individual in the population was encoded using 8-bit binary strings. These binary strings were used to represent real-valued feature weights by scaling their integer equivalents to the  $[0, 1]$  range. This encoding scheme allowed binary representations to be effectively used in a continuous optimization space. The fitness value of each individual was calculated based on its classification accuracy, which reflects the proportion of correctly classified samples over the total number of samples, including both positive and negative predictions.

Subsequent generations were formed using the Roulette Wheel selection method, which probabilistically favors individuals with higher fitness values, thereby increasing their likelihood of contributing to the next generation. In the crossover phase, 60% of the population underwent recombination. For this purpose, randomly selected 8-bit individuals were paired, and a single-point crossover was applied by exchanging 4-bit segments between them. This process aimed to produce offspring with improved fitness characteristics inherited from their parents. Following crossover, a mutation operation was performed by flipping a single bit in selected individuals to introduce genetic diversity and prevent premature convergence. The stopping criterion was defined as no improvement in the fitness value over 2000 consecutive iterations. The control parameters used in this GA process were selected based on recommendations by De Jong and are presented in Table 6 [24].

**Table 6.** Genetic Algorithm parameters based on De Jong's recommendations

Control Parameters	Value
Population Size	100
Crossover Rate	0.60

#### 4.9. Particle Swarm Optimization Process

In the PSO-based optimization, the weights of the WNB are treated as particles within the swarm. Each particle represents a complete weight vector, with each dimension corresponding to a feature weight constrained in the range [0, 1]. The objective of employing PSO is to maximize the classification accuracy. Initially, 100 particles were randomly generated for each weight. While the inertia (approach) velocity can be initialized randomly in general, all initial velocities were set to zero in this study. The PSO parameters  $c_1$  and  $c_2$  were both set to 2, whereas the random coefficients  $r_1$  and  $r_2$  were drawn from a uniform distribution in the range [0, 1]. Fitness was measured as the ratio of correctly classified instances to the total number of instances. During the optimization process, particles updated their positions by considering both their personal best solutions and the global best solution found by the swarm. The algorithm was iterated for 1000 generations to determine the optimal weight values. This entire optimization process was repeated 50 times, and the average of the classification accuracies obtained across all runs was computed.

### 5. EXPERIMENTAL RESULTS

In this section, the classification performances of the standard Naive Bayes, Weighted Naive Bayes, and the Weighted Naive Bayes models optimized via Genetic Algorithm (GAW-NB) and Particle Swarm Optimization (PSOW-NB) are compared across five benchmark datasets. All experiments were conducted using MATLAB. The performance metrics were calculated by averaging the classification results obtained from each fold of the 5-fold cross-validation. This validation strategy was employed to enhance the robustness and generalizability of the evaluation. The detailed fold-wise classification results for each dataset are presented in Table 7.

**Table 7.** Comparison Of NB, WNB, GA-WNB and PSO-WNB Results

Database Name	NB	W-NB	GAW-NB	PSOW-NB
<b>Tic-Tac-Toe Dataset</b>				
1. Fold	69.79	77.60	79.16	78.48
2. Fold	75.00	80.20	81.77	81.44
3. Fold	72.39	77.60	77.60	79.11
4. Fold	68.75	75.00	75.52	75.55
5. Fold	67.70	73.95	74.47	75.59
<b>Accuracy</b>	<b>70.72</b>	<b>76.87</b>	<b>77.70</b>	<b>78.03</b>
<b>Post-Operative Patient Dataset</b>				
1. Fold	44.44	83.33	83.33	83.33
2. Fold	55.55	55.55	55.55	67.22
3. Fold	66.66	77.77	72.22	78.70
4. Fold	61.11	73.61	72.22	72.22
5. Fold	77.77	77.77	77.77	82.03
<b>Accuracy</b>	<b>61.11</b>	<b>73.61</b>	<b>72.21</b>	<b>76.70</b>
<b>Qualitative Bankruptcy Dataset</b>				
1. Fold	98.00	100.00	100.00	100.00
2. Fold	100.00	100.00	100.00	100.00
3. Fold	100.00	100.00	100.00	100.00
4. Fold	100.00	100.00	100.00	100.00
5. Fold	100.00	100.00	100.00	100.00
<b>Accuracy</b>	<b>99.60</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>

<b>Mammographic Mass Dataset</b>				
1. Fold	87.34	86.06	86.66	87.27
2. Fold	80.72	86.06	86.06	86.66
3. Fold	84.33	82.42	83.63	83.03
4. Fold	83.13	89.09	89.09	89.69
5. Fold	80.72	88.48	88.48	88.48
<b>Accuracy</b>	<b>83.25</b>	<b>86.42</b>	<b>86.78</b>	<b>86.69</b>
<b>Breast Cancer Dataset</b>				
1. Fold	97.05	99.26	99.26	99.26
2. Fold	93.38	98.52	98.52	98.52
3. Fold	94.11	96.32	97.05	97.79
4. Fold	97.05	98.52	98.52	98.52
5. Fold	97.79	98.52	98.52	98.52
<b>Accuracy</b>	<b>95.88</b>	<b>98.23</b>	<b>98.37</b>	<b>98.52</b>

As presented in Table 7, the GAW-NB and PSOW-NB classifiers achieved higher classification accuracy than the standard NB and W-NB models across all five datasets. On the Tic-Tac-Toe dataset, the PSOW-NB model achieved the highest accuracy at 78.03%, which is 7.31% higher than NB and 1.16% higher than W-NB. In the Post-Operative Patient dataset, PSOW-NB reached 76.70%, outperforming NB by 15.59%. In the Mammographic Mass dataset, the highest accuracy was obtained by the GAW-NB model with 86.78%, which is 3.53% higher than NB and slightly better than PSOW-NB. For the Breast Cancer dataset, PSOW-NB achieved 98.52% accuracy, improving upon NB by 2.64%. In the Qualitative Bankruptcy dataset, all models except NB reached 100% accuracy.

Despite the differences in dataset sizes, structures, and class distributions, both optimization-based approaches consistently produced higher accuracy values across all experiments. These findings demonstrate that integrating metaheuristic algorithms into the W-NB framework not only improves classification accuracy but also enhances the generalizability of the models and their ability to deliver consistent results across different datasets.

## 6. CONCLUSION

The Weighted Naive Bayes (W-NB) classifier is an enhanced version of the standard Naive Bayes algorithm, offering improved classification performance through the incorporation of feature weights. However, the reliance on the grid search method in W-NB leads to significantly increased computational time, particularly when dealing with high-dimensional datasets. In such cases, determining the optimal set of weights may become computationally infeasible.

In this study, two metaheuristic optimization algorithms Genetic Algorithm and Particle Swarm Optimization were employed to optimize the weights of the W-NB classifier. The proposed GAW-NB and PSOW-NB classifiers aimed to achieve higher classification accuracy than the standard Naive Bayes classifier while reducing the computational cost associated with the traditional W-NB approach.

All experiments were carried out on five benchmark datasets using 5-fold cross-validation. The results demonstrated that both GAW-NB and PSOW-NB consistently outperformed the NB and W-NB classifiers in terms of classification accuracy. For example, PSOW-NB achieved the highest overall accuracy of 98.52% on the Breast Cancer dataset, while GAW-NB yielded the best performance of 86.78% on the Mammographic Mass dataset. On the Qualitative Bankruptcy dataset, all models except NB reached 100% accuracy. These findings provide

strong evidence that integrating metaheuristic optimization techniques into the W-NB framework can lead to the development of fast and highly accurate classifiers.

However, some limitations should be acknowledged. The proposed methods were tested on relatively small and clean datasets. The effectiveness and scalability of the approaches for large-scale or noisy datasets remain to be examined. Furthermore, since metaheuristic algorithms rely on stochastic processes, results may vary across runs, and care should be taken to avoid premature convergence or overfitting.

Future studies may explore alternative optimization algorithms or hybrid approaches to further enhance the efficiency, robustness, and scalability of the W-NB classifier, especially in applications involving high-dimensional and complex data environments.

## 7. ACKNOWLEDGEMENTS

This article was written as a part of master's thesis titled "Optimization of the weights of Weighted Naive Bayesian Classifier" at Firat University. Thesis no: 527473.

## REFERENCES

- [1] Aggarwal CC, Yu PS. Data Mining Techniques for Associations, Clustering and Classification. In: Zhong N, Zhou L, editors. *Methodol. Knowl. Discov. Data Min.*, Berlin, Heidelberg: Springer; 1999, p. 13–23. [https://doi.org/10.1007/3-540-48912-6\\_4](https://doi.org/10.1007/3-540-48912-6_4).
- [2] Ravinder B, Seeni SK, Prabhu VS, Asha P, Maniraj SP, Srinivasan C. Web Data Mining with Organized Contents Using Naive Bayes Algorithm. 2024 2nd Int. Conf. Comput. Commun. Control IC4, 2024, p. 1–6. <https://doi.org/10.1109/IC457434.2024.10486403>.
- [3] Jain J, Upadhyay SK, Nayak SK. Analyzing the Effectiveness of Machine Learning Algorithms in detecting Fake News. *Comput. Commun. Intell.*, CRC Press; 2025.
- [4] Arrayyan AZ, Setiawan H, Putra KT. Naive Bayes for Diabetes Prediction: Developing a Classification Model for Risk Identification in Specific Populations. *Semesta Tek* 2024;27:28–36. <https://doi.org/10.18196/st.v27i1.21008>.
- [5] Karabatak M. A new classifier for breast cancer detection based on Naïve Bayesian. *Measurement* 2015;72:32–6. <https://doi.org/10.1016/j.measurement.2015.04.028>.
- [6] Aksoy G, Karabatak M. Performance Comparison of New Fast Weighted Naïve Bayes Classifier with Other Bayes Classifiers. 2019 7th Int. Symp. Digit. Forensics Secur. ISDFS, 2019, p. 1–5. <https://doi.org/10.1109/ISDFS.2019.8757558>.
- [7] Lin J, Yu J. Weighted Naive Bayes classification algorithm based on particle swarm optimization. 2011 IEEE 3rd Int. Conf. Commun. Softw. Netw., 2011, p. 444–7. <https://doi.org/10.1109/ICCSN.2011.6014307>.
- [8] Tian Z, Fong S, Tian Z, Fong S. Survey of Meta-Heuristic Algorithms for Deep Learning Training. *Optim. Algorithms - Methods Appl.*, IntechOpen; 2016. <https://doi.org/10.5772/63785>.

- [9] Sun Y, Xue B, Zhang M, Yen GG, Lv J. Automatically Designing CNN Architectures Using the Genetic Algorithm for Image Classification. *IEEE Trans Cybern* 2020;50:3840–54. <https://doi.org/10.1109/TCYB.2020.2983860>.
- [10] Połap D. An adaptive genetic algorithm as a supporting mechanism for microscopy image analysis in a cascade of convolution neural networks. *Appl Soft Comput* 2020;97:106824. <https://doi.org/10.1016/j.asoc.2020.106824>.
- [11] Cura T. A particle swarm optimization approach to clustering. *Expert Syst Appl* 2012;39:1582–8. <https://doi.org/10.1016/j.eswa.2011.07.123>.
- [12] Huang KY. A hybrid particle swarm optimization approach for clustering and classification of datasets. *Knowl-Based Syst* 2011;24:420–6. <https://doi.org/10.1016/j.knosys.2010.12.003>.
- [13] Kotsiantis SB. *Supervised Machine Learning: A Review of Classification Techniques* 2007.
- [14] Dagdia, Zaineb Chelly, Mirchev, Miroslav. *Optimization Problem - an overview* 2020.
- [15] Alhijawi B, Awajan A. Genetic algorithms: theory, genetic operators, solutions, and applications. *Evol Intell* 2024;17:1245–56. <https://doi.org/10.1007/s12065-023-00822-6>.
- [16] Gen, Mitsuo, Cheng, Runwei. *Foundations of Genetic Algorithms*. *Genet. Algorithms Eng. Optim.*, John Wiley & Sons, Ltd; 1999, p. 1–52. <https://doi.org/10.1002/9780470172261.ch1>.
- [17] Jain M, Saihpal V, Singh N, Singh SB. An Overview of Variants and Advancements of PSO Algorithm. *Appl Sci* 2022;12:8392. <https://doi.org/10.3390/app12178392>.
- [18] Eberhart, Shi Y. Particle swarm optimization: developments, applications and resources. *Proc. 2001 Congr. Evol. Comput.* *IEEE Cat No01TH8546*, vol. 1, 2001, p. 81–6 vol. 1. <https://doi.org/10.1109/CEC.2001.934374>.
- [19] Home - UCI Machine Learning Repository n.d. <https://archive.ics.uci.edu/> (accessed January 30, 2025).
- [20] Aha D. Tic-Tac-Toe Endgame 1991. <https://doi.org/10.24432/C5688J>.
- [21] Sharon Summers LW. Post-Operative Patient 1991. <https://doi.org/10.24432/C5DG6Q>.
- [22] Kim M-J, Han I. The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. *Expert Syst Appl* 2003;25:637–46. [https://doi.org/10.1016/S0957-4174\(03\)00102-7](https://doi.org/10.1016/S0957-4174(03)00102-7).
- [23] Elter M. Mammographic Mass 2007. <https://doi.org/10.24432/C53K6Z>.
- [24] De Jong K. Learning with genetic algorithms: An overview. *Mach Learn* 1988;3:121–38. <https://doi.org/10.1007/BF00113894>.