International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

An Enhancement of the Gaussian Naive Bayes Algorithm Applied to Air Quality Classification

Merlinda C. Binalla[1], Maisie Allena F. Villanueva[2]

[1,2]Bachelor of Science in Computer Science, Computer Science Department, Pamantasan ng Lungsod ng Maynila

Abstract

The Gaussian Naive Bayes Algorithm is a machine learning technique based upon the Bayes Theorem. It is commonly used for classification tasks to calculate the likelihood of events. This study developed an enhanced GNB algorithm to classify the air quality in Pamantasan ng Lungsod ng Maynila. The enhancement made in this study sought to increase the classification performance of the traditional GNB against zero frequency issues. The zero-frequency problem is an inherent limitation of the conventional GNB due to the algorithm's reliance on multiplying probabilities. It occurs when a feature value is absent from the training data. The Parzen-Rosenblatt Window method was applied to address the issue and increase the algorithm's stability against the problem. OpenWeather-AQI and USA-AQI datasets were used to evaluate the algorithm. The algorithm's accuracy improved from 71.77% to 74.16% (2.39%) in the OpenWeather-AQI dataset. In comparison, the other dataset showed a 5.26% improvement, increasing from 59.33% to 64.59%. These results showcase how the enhanced GNB algorithm outperforms the traditional one. Thus, the Enhanced GNB Algorithm effectively addresses the zero-frequency problem, increasing classification accuracy and demonstrating its potential as a reliable method for assessing air quality.

Keywords: Machine Learning, Gaussian Naïve Bayes, Bayes theorem, Zero-frequency, Air pollution, Air Quality Index, Parzen-Rosenblatt Window method, Philippines

Chapter 1
INTRODUCTION

1.1 Background of the Study

In the context of modernization, environmental and health protection is often a collateral that experts consider as the human civilization pursues to evolve. One of the main issues is the worsening of air quality not only in the Philippines but around the world. An estimated 4.2 million premature deaths, which has linkage to exposure to chromic air pollution in 2019 was reported by the World Health Organization (WHO), 89% of this number is reported to be people living in low to middle income countries (WHO, 2022). Aside from the adverse health effects of exposure to bad air quality, studies have shown a connection between air pollution and cognitive outcomes, particularly a decrease in school performance of those students who were exposed to chronic low-level traffic-related air pollution (Gardin, T.N & Requia, W. J., 2023).

In order to monitor and mitigate the continuous exacerbation of the air quality WHO has released a guideline that determines the threshold for key air pollutants that would cause significant risk to humans. The said guideline aims to present suggestions in terms of monitoring air quality as well as a valid evidence-informed supporting document when curating environment related legislation and policies (WHO, 2021). Accordingly, the United States of Environmental Protection Agency (EPA) have also issued their own air quality guide as well as a system for tracking air quality (EPA, 2024). Therefore, this paper aims to classify the air quality of Pamantasan ng Lungsod ng Maynila (PLM) and its nearby areas using the enhanced Gaussian Naive Bayes algorithm.

The Gaussian Naive Bayes (GNB) algorithm is a machine learning technique based upon the Bayes theorem, which estimates the likelihood of an occurrence based on previous knowledge of conditions that may be relevant to the event. Mainly used for classification tasks, this algorithm classifies continuous data that follows a Gaussian distribution (Kamble & Dale, 2022). By analyzing its attributes or features, GNB predicts the probability that a newly introduced data point is considered part of a specific classification class. GNB functions are predicated assuming that the attributes of a given data point are mutually exclusive and conform to a Gaussian or normal distribution. While this assumption may not always correspond with practical situations, GNB frequently achieves competitive classification accuracy due to its computational efficiency and inherent advantages. This assumption enhances GNB's overall computational efficacy by reducing the complexity of the calculations required to predict class probabilities (Anand et al., 2022).

In 2023, a comparative analysis of different machine learning algorithms was conducted by a researcher named Salim Lahmiri to identify the most optimal model for predictive maintenance and fault classification in electric drive trains. One of the classification models implemented was Gaussian Naive Bayes, and the outcomes demonstrate that the algorithm operates with minimal computational effort. On the other hand, its performance was hindered by dependability issues and faulty assumptions that required improvement. Compared to alternative models, the model achieved a comparatively low accuracy rate of 71.3%. Furthermore, it demonstrated a significant degree of bias and accuracy variance.

Through this investigation, the study's findings will provide valuable insights into the practical application of GNB in assessing air quality conditions of Pamantasan ng Lungsod ng Maynila and nearby areas, thereby contributing to a comprehensive understanding of its utility in safeguarding student well-being and ascertaining whether the air quality fosters an environment conducive to students' academic success.

1.2 Naive Bayes Classifier

The Naive Bayes Classifier is a probabilistic classifier that inherently considers each feature independently (Poola et al., 2023). It uses the Bayes Principle, particularly the Naive Bayes Classification formula, to classify using the following equation.

$$P(C|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \frac{P(C)\prod_{i=1}^{n} \quad P(X_i = x_i|C)}{P(X_1 = x_1,\ X_2 = x_2, \ldots, X_n = x_n)}$$

Where C is each class name, $X_1, X_2, \ldots, X_n$ represents the features of each data point. On the numerator, $P(C)$ acts as the prior probability of class C, $P(X_i = x_i|C)$represents the likelihood of each feature considering the given class C,$\prod_{i=1}^{n} \quad P(X_i = x_i|C)$ gives the product of the likelihoods for each feature and lastly on the denominator $P(X_1 \ = x_1, X_2 = x_2, \ldots, X_n = x_n)$ represents the normalizing constant or

the evidence (Gayathri., B. M., & Sumathhi, C. P., 2016).

1.3 Gaussian Naive Bayes Classifier

The Gaussian Naive Bayes Classifier is a variation of a Naive Bayes Classifier which is used when dealing with continuous data or classifying values that are assumed to be distributed according to the Gaussian distribution (Cinar, 2023). Therefore, when calculating $P(X_i = x_i|C)$ or the likelihood of each feature given the class C, the Gaussian (normal) distribution is used.

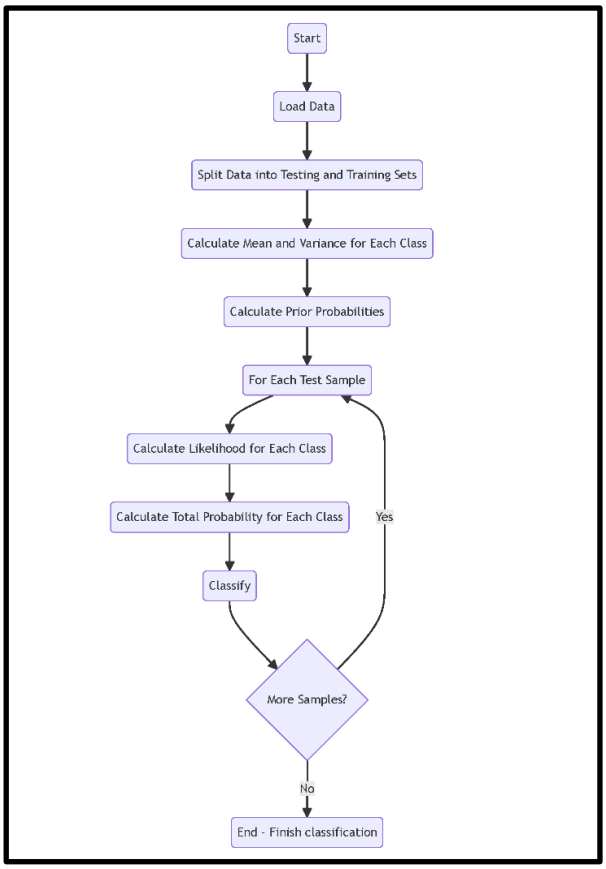
$$P(X_i = x_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} exp(-\frac{(X_i - \mu_C)^2}{2\sigma_C^2})$$

Where X represents the variable, C as the class, μ for the mean and σ as the standard deviation (Kamel, H., Abdulah, D., & Al-Tuwaijari J. M., 2019).

1.3.1 Gaussian Naive Bayes Flowchart

This section of the paper discusses the figure which illustrates the process flow of the traditional Gaussian Naive Bayes Algorithm, in order to have a visual representation and overview of the said algorithm.

Figure 1. Original Gaussian Naive Bayes Algorithm



The process starts by loading the dataset, which is then split into training and testing sets. During the training phase, the mean and variance for each class is computed to be used for calculation of the prior probability of each class, which wraps up the training phase of the model. Afterwards, the testing phase

begins, where for each data point in the testing set, the likelihood of the data considering each class is computed using the Gaussian probability function. The likelihood computed is then used to obtain the total probability of each class and lastly the class which has the highest posterior probability is assigned as the class of the new data point. This process is repeated for each sample in the training set until none are left which completes the algorithm.

1.3.2 Gaussian Naive Bayes Pseudocode

Considering the overview given in the previous section, this part of the study further specifies the process of the algorithm, by providing a detailed outline of its process through a pseudocode in order to convey the fundamental ideas and procedures of the Gaussian Naive Bayes Algorithm.

(1) Data Collection

(2) Data Pre-processing

(a) Incomplete Data Removal

(b) Data Transformation

(c) Data Normalization

(3) Data Split (Training Dataset and Testing Dataset)

(4) Training Phase

(a) For each class C in Y_{train}

Calculate the prior probability $P(C) = \frac{N_C}{N}$

Where N_C is the number of samples in class C and N is the total number of samples in the training set.

For each feature X_i in X_{train}

1) Calculate the mean $\mu_{i,C}$ of feature X_i in class C.

2) Calculate the variance $\sigma_{i,c}^2$ of feature X_i in class C.

(b) End For

(5) Testing Phase

(a) For each sample X_{test} to classify

For each class C

1) Initialize $P(X_{test}|C) = P(C)$

2) For each feature X_i in X_{test}

a) Calculate the Gaussian likelihood $P(X_i = x_i|C) = \frac{1}{\sqrt{2\pi\sigma_{i,C}^2}} exp(-\frac{(X_i-\mu_{i,C})^2}{2\sigma_{i,C}^2})$

b) Multiply $P(X_i = x_i|C)$ and $P(X_{test}|C)$

3) End For

4) Calculate the posterior probability $P(X_{test}|C)$ class C given the test sample X_{test}.$P(X_{test}|C) = P(C) \times \prod_{i=1}^{n} \quad P(X_i = x_i|C)$

End For

Classify X_{test} with the highest posterior probability $P(X_{test}|C)$

(b) End For

(6) Evaluation Phase

1.4 Statement of the Problem

Considering the traditional Gaussian Naive Bayes' wide usage as a classification algorithm, there are still gaps within the algorithms which hinder its performance.

Zero probability issues can occur when encountering feature values absent in training data, affecting model reliability. The zero-probability problem is an inherent issue occurring when estimating the likelihood of rare events. When there is no representation for an instance or priori knowledge there is no evidence to base the probability estimate, leading to the zero-probability issue (Witten & Bell, n.d).

1.5 Objective of the Study

This study aims to develop an enhanced Gaussian Naive Bayes algorithm that enhances its classification accuracy and alleviate the issue of zero probability results when using the algorithm of the dataset by implementing the Parzen-Rosenblatt Window method into the dataset.

1.6 Significance of the Study

This study aims to be beneficial:

To the PLM community. The aim of this study is to deepen students' understanding of the detrimental effects of polluted surroundings on their academic performance. Gaining awareness of these impacts should help in advocating for healthier learning environments. Accordingly, this provides teachers with vital insights for adapting their teaching techniques and curriculum delivery in response to changing environmental conditions. Furthermore, increased awareness of air quality concerns can lead professors to advocate for improving air quality in educational settings, enhancing their students' general well-being and academic success. Lastly, this also benefits the stakeholders of PLM, including the employees that consider PLM as their workplace.

To the Environmental and public health agencies. Government and non-government organizations could use the enhanced algorithm for more reliable air quality monitoring. This would enable them to develop targeted strategies to mitigate pollution sources, improve public health responses, and enact policy changes.

To future researchers. This study establishes the foundation for future research endeavors. Future researchers could expand on this study to explore different approaches, incorporate new features and data sources, and address issues in the field of air quality monitoring as well as contributing valuable insights regarding the use of Gaussian Naive Bayes as a classifier algorithm.

1.7 Scope and Limitations

The study focuses on examining and addressing the current issues with the Gaussian Naive Bayes Classifier, namely the algorithm’s inability to counter zero probability issues leading to an overall deterioration of performance. Considering this, the study seeks to alleviate the said issues by proposing the integration of Parzen-Rosenblatt Window Method to the current algorithm to mitigate and further improve the algorithm’s overall performance as a classifier in terms of four metrics, namely accuracy, precision, recall and f1 score.

In terms of application, the study aims to classify air quality data procured from OpenWeatherMap website specifically using GPS coordinates around the Pamantasan ng Lungsod ng Maynila to be characterized based on Air Quality Index (AQI) to provide an informative document regarding the university’s capacity to provide a suitable learning environment for its students.

The study’s findings may be limited in the context of classifying continuous data, which is the type of data to be gathered considering air quality and the result may also be influenced by extraneous factors including but not limited to availability of quality data. Performance metrics including but not limited to time

complexity and space complexity that are not included in the four previously mentioned metrics would also not be discussed in the result, therefore the proposed enhancement's adverse or beneficial effect would not be included in the discussion. Despite this, the study intends to provide valuable and quality insights that would significantly contribute to further research particularly in the context of mitigating the issues addressed on the Gaussian Naive Bayes classifier algorithm

1.8 Definition of Terms

Enumerated below are the terms and their definitions considering their usage in the study.

Accuracy - refers to the overall correctness of the model.

Air Quality Index (AQI) - is the system used to warn the public when air pollution is dangerous. It is a measurement of air pollutant concentrations and their associated health risks.

Air Quality Monitoring - refers to the collection and measurement of ambient air pollution samples.

F1 Score - refers to the mean of precision and recall.

Gaussian Naive Bayes - is a type of classification algorithm working on continuous normally distributed features that is based on the Naive Bayes algorithm.

Parzen-window Method - also known as Parzen-Rosenblatt window method, is a non-parametric approach to estimate a probability density function p(x) for a specific point p(x) from a sample p(xn) that doesn't require any knowledge or assumption about the underlying distribution.

Precision - proportion of positive predictions that were actually correct.

Recall - refers to the proportion of positives that were correctly identified.

World Health Organization (WHO) - World Health Organization, specialized agency of the United Nations established in 1948 to further international cooperation for improved public health conditions.

Zero Probability - Probability is a number that represents the likelihood of an event to occur. A probability of zero means there will be no occurrences of the event.

Chapter 2

REVIEW OF RELATED LITERATURE

The following chapter discusses and cites related literature that provides informative insights regarding the concepts and issues that are present in the study.

2.1 Related Literature

2.1.1 Health Impacts of Air Pollution

In the contemporary age of urbanization, awareness regarding the exacerbating condition of our environment has broadened, with the majority of the population having the knowledge that our world is exposed to constant pollution. Despite this, efforts to improve and stabilize environmental conditions remain unbalanced compared to the anthropogenic pollutants produced daily that aggravates our environment (Kelly F.J. & Fussell, J.C, 2015).

The adverse effects of pollution, particularly air pollution on a human have long been studied. Researchers have concluded that continuous exposure to bad air quality has a direct negative effect on different human functions. Firstly, its effect on a human's physical health, wherein evidence has found chronic exposure to bad air quality to be linked with cardiovascular and respiratory diseases. Secondly, it also contributes negatively to the human brain and its nervous system which triggers a shift in the physiology of humans with heightened risks of mental health issues such as depression, cognitive impairments, schizophrenia among others. Lastly, it was also found to be a contributor as a cause to human psychological stress and

anxiety wherein as a coping mechanism against a polluted environment, an individual would opt to stay indoors due to a polluted environment specifically when there are visible indications of pollution (Al Ahad, M. C, 2024).

With aggravating outdoor conditions and limited clean and healthy space for human's the idea of learning comes into perspective. In order to provide solutions to current problems, research is used as a basis to determine the outcome and possible effectiveness of recommended solutions. Therefore, a safe space to learn is deemed as necessary to support and provide an optimal environment to nurture learning and development of students particularly the youth. Despite this, researchers have concluded that learning institutions including schools do not offer adequate conditions to cultivate learning with poor classroom air quality heightening the risk and exposure of students to respiratory diseases (Sadrizadeh, S. et al., 2022).

Considering this, efforts to mitigate and monitor air quality levels have been improving throughout the years. The recent pandemic has shown that the lockdown, particularly the great reduction in the use of vehicles and industrial activities have shown a positive impact on the air quality of the atmosphere (Kaloni, D. et al., 2022). The use of machine learning as a method to classify and predict air quality has been rampant due to its efficiency and performance in generating results given an extensive dataset (Imam, M., et al., 2023). Machine learning algorithms such as Support Vector Machines, Decision trees, K-nearest neighbor and Naive Bayes are popular machine learning algorithms used to predict and classify air quality data. One of the studies showed that using naive bayes as a classifier renders an accuracy of 86.663 percent when used as an air quality classifier (Gupta, N. S., et al., 2022).

2.1.2 A Review of the Existing Gaussian Naïve Bayes Algorithm

The Gaussian Naïve Bayes Classifier is a straightforward yet powerful approach that often yields accurate and reliable models, even when working with small datasets. Its strength lies in simplifying complex predictive modeling problems by assuming conditional independence between features. This means that each feature's influence on the outcome is independent of other features. (Rice, D. M., 2014) However, while it excels in many scenarios, it does show limitations in specific applications.

Lahmiri S. (2023) assessed the performance of Gaussian Naive Bayes and various machine learning algorithms in identifying issues with electric motor trains. Despite the model's simplicity, the study revealed that Gaussian Naïve Bayes exhibited the lowest accuracy rate of 71.3% in comparison to other models, including decision trees and support vector machines. Its inability to precisely estimate class probabilities is attributed to the performance drop, which increases bias (error rate) and variance, especially in complex defect diagnosis tasks. Thus, while the method exhibits significant effectiveness in various situations, its limitations are evident in particular industrial applications where superior algorithms may outperform it.

Denis F. et al. (2006) elaborated on the extensive utilization of Naïve Bayes classifiers in machine learning, especially for text classification applications. The researchers emphasized the algorithm's straightforwardness in generating models and its efficacy in complex scenarios (Domingos & Pazzani, 1997). However, the study pointed out a significant limitation: the assumption of attribute independence often doesn't hold in real-world data, where attributes may display interdependencies. This issue is exacerbated by the presence of classification noise—errors or inconsistencies in the data—which can violate the independence assumption and lead to overfitting, reducing the model's ability to generalize.

Therefore, while Naïve Bayes models, including Gaussian Naïve Bayes, can perform well in certain conditions, their limitations—particularly in handling noisy or correlated data—underscore the need for

preprocessing techniques like noise reduction and feature selection. In contexts where noise or attribute interdependency is prevalent, alternative methods may be required to ensure better model performance.

2.1.3 Common Problems of the Gaussian Naïve Bayes Algorithm

2.1.3.1 Zero Probability

Zero probability is the state in which particular feature values noted during prediction are not reflected in the training set. The model assigns a 0 probability when it tries to calculate the probability for these unobserved values, therefore producing the zero overall posterior probability for that class. This renders the forecast incorrect or unreliable, as the algorithm perceives the outcome as impossible. Moreover, the problem may arise when the approach confronts extreme outliers or values that deviate from the established Gaussian distribution. (Di Paola et al., 2018)

In the study by Naiem Sarah et al., (2023) they further noted this zero-frequency problem due to the nature of the Gaussian Naive Bayes algorithm's reliance on multiplying probabilities. When a feature value is absent from the training data, this multiplication yields a zero probability, resulting in inaccurate classification. Their analysis emphasized that the presumption of feature independence, which is seldom valid in actual datasets, exacerbates the issue, rendering the approach less successful in practical applications.

Another study that demonstrates the zero-probability problem inherent to Naive Bayes algorithms, particularly in text classification tasks, was done by He and Ding (2007). The researchers improved the reliability of the Naive Bayes classifier by utilizing smoothing techniques, including Laplace, Linear, and Witten-Bell Smoothing, to resolve the zero-probability issue, a notable limitation of the model. By including even the unseen features during training, such techniques have markedly enhanced the model's stability and accuracy.

Furthermore, a research by Chandra et al. (2007) acknowledges that in the absence of training cases for a certain attribute and class, the conditional probability for that attribute is rendered zero, hence preventing the model from predicting the class of such instances. Instead of depending on conventional techniques like Laplace smoothing, the researchers devised a new way for calculating probabilities to fix the found problem in the algorithm. Their method identifies occurrences in which the likelihood of an attribute is 0 and applies this to a class probability computation.

2.1.4 Enhancements of the Gaussian Naïve Bayes Algorithm

In the study "A Gaussian-Bernoulli Mixed Naive Bayes Approach to Predict Students' Academic Procrastination Tendencies in Online Mathematics Learning," authors Godinez, C.D., and Lomibao, L. improved the conventional Gaussian Naive Bayes model by integrating a Bernoulli component to accommodate the heterogeneous characteristics of the data.

Gaussian Naive Bayes is generally proficient for continuous data; nevertheless, in this case, certain features were discrete, which aligned more suitably with a Bernoulli distribution. The Gaussian-Bernoulli Mixed Naive Bayes model was created by integrating these two methodologies to better efficiently manage both continuous and categorical data. This hybrid model improved the prediction of procrastination tendencies by accurately capturing the subtleties in student behavior patterns.

The improved model attained 85% accuracy and a Kappa score of 82%, suggesting good predictive ability. This model is presented as a useful tool for educators to identify pupils at risk of procrastinating early, allowing interventions to offset the detrimental effects of academic delays.

Another study by Naeim et al. (2023) underlines the general risk that Distributed Denial of Service (DDoS)

attacks in cloud computing pose for data and financial losses suffered by consumers and service providers alike. Several machine learning techniques have been applied, especially meant to increase the efficacy of the Gaussian Naïve Bayes classifier, thereby lowering the risks associated with it.

Although this classifier is well regarded for its cost-effectiveness and speed, it could further enhance its statistical capabilities, particularly its reliance on multiplication. The zero-frequency issue and feature independence assumptions might result in inaccurate classifications. The article presents an innovative paradigm that addresses the above challenges. The framework uses an iterative feature selection technique to find sets of highly independent features. It uses metrics like Pearson Correlation Coefficient, Mutual Information, and Chi-squared. These enhancements increase categorization accuracy and resilience, demonstrating significant progress in protecting cloud infrastructure against DDoS attacks.

2.1.5 Other Applications of Parzen-Rosenblatt Window Method

2.1.5.1 Medical Image Segmentation

A research paper titled "Level Set Approach Based on Parzen Window and Floor of Log for Edge Computing Object Segmentation in Digital Images" (Rebouças et al., 2021) offers a novel approach for image segmentation in medical pictures that combines the Parzen Window method with Floor of Log (FLog) clustering. This approach is utilized in the study to identify areas of interest (ROIs) in datasets on stroke, lung, and skin disorders. This improves the segmentation procedure for edge computing scenarios. The researchers employed the Parzen window approach for non-parametric density estimation, which allowed for precise ROI detection by conforming to local pixel intensity distributions. This method not only improves segmentation accuracy with high sensitivity (98.57%) and accuracy (98.77%), but it also maintains low processing costs and fast convergence.

2.1.5.2 Hyperparameter Optimization

The research study "Light Gradient-Boosting Machine Algorithm with Tree-Structured Parzen Estimator for Breast Cancer Diagnosis" (Omotehinwa et al., 2023) suggests making use of machine learning to enhance breast cancer detection using the Wisconsin Diagnostic Breast Cancer dataset. The study uses the Light Gradient-Boosting Machine method, the Borderline-SMOTE (Synthetic Minority Oversampling Technique) algorithm, and the Tree-Structured Parzen Estimator (TPE) to tune hyperparameters. Using the Parzen Window approach within the Tree-Structured Parzen Estimator for effective hyperparameter tuning improves the model's ability to identify breast cancer. Its function is crucial in improving model performance by systematically exploring and optimizing hyperparameters, allowing for faster convergence and better classification results. This modification improves the model's accuracy to 99.12%, 100% specificity, and 100% precision when compared to the baseline model.

2.1.5.3 Multi-class Classification

In the research paper "Parzen windows for multi-class classification" (Pan et al., 2008), the authors utilize the Parzen Window technique for multi-class classification tasks, calculating probability density functions to partition the input space into several classes. The study introduces a learning method with Parzen windows that aims to minimize misclassification error by assessing the surplus error relative to the optimal Bayes classifier. The Parzen Window method is beneficial because of its non-parametric characteristics, allowing for precise and adaptable categorization of multidimensional situations and many classes without requiring assumptions about the data distribution. Furthermore, the methodology guarantees robust error convergence rates, and the choice of window width is critical for improving classification efficacy.

Chapter 3
METHODOLOGY

This chapter entails the research methodologies particularly the tools and techniques used in the study. This will serve as a guide for understanding the research methods used and evaluating the validity of the study's findings regarding the enhancements made into the algorithm.

3.1 Research Design

3.1.1 Data Acquisition

The first phase in enhancing the Gaussian Naive Bayes to apply it for air quality classification is to acquire the dataset from relevant and credible resources to ensure quality results, as the dataset influences the overall accuracy of the algorithm (Roh et al., 2019). For this study, the dataset is acquired from OpenWeather API. Open weather is a team of data scientists and IT experts based in London, United Kingdom that specializes in deep weather data science (OpenWeather, n.d). The data is retrieved from OpenWeather's Air Pollution API which provides current and historical forecast of air pollution data namely the air quality index (AQI), polluting gasses specifically *Carbon monoxide* (CO), *Nitrogen monoxide* (NO), *Nitrogen dioxide* (NO_2), *Ozone* (O_3), *Sulphur dioxide* (SO_2), and particulate matters ($PM_{2.2}$ and PM_{10}) and History API which provide hourly weather data. The data taken was from November 30, 2023 until November 24, 2024.

3.1.2 Data Pre-Processing and Cleaning

After extraction, the researchers pre-processed the dataset by applying data deletion and data transformation. 12 dates between November 30, 2023, and November 24, 2024, were found with missing data values. These missing data values were handled by deleting their respective rows. The data extracted also includes an hourly monitoring of the air quality, the researchers took the mean values for every 8 hours resulting in a 3 representation for each day, resulting in 1044 data samples. Lastly, the feature Air Quality Index (AQI) is rounded up to produce a discrete index. This finalizes the first dataset which includes discrete AQI that follows Open Weather's Air Pollution Concept. From this dataset, a second dataset is constructed by converting the values to follow the Index level scale in the USA.

3.1.3. Data Normalization

Data normalization is applied through the use of StandardScaler which uses the equation $z = \frac{(x-\mu)}{\sigma}$ where μ represents the mean and σ as the standard deviation of the training samples (SciKitLearn, n.d.). Data normalization aims to standardize the features of the dataset to use a common scale without misinterpreting the data. It avoids problems caused by datasets with features that have wide ranges of values from each other by creating new values that still retains the distribution and ratio of the dataset (Microsoft, 2024).

3.1.4 GNB Classification

After applying SMOTE and ENN to the dataset the enhanced classifier is initiated which uses the Parzen Window Method to calculate the likelihood for each feature of the class instead of the traditional Gaussian Distribution. Parzen Window is a nonparametric density estimation technique that observes the surrounding data points near point x to estimate its density function (Zambom & Dias, 2012). The Parzen-Rosenblatt Window Method Formula is as follows:

$$p_n(x) = \frac{1}{n}\sum_{i-1}^{n} \quad \frac{1}{h^d}\phi\left[\frac{x - x_i}{h_n}\right]$$

Where $p_n(x)$ represents the estimated probability density function considering point x. n as the number of data points. h and d are the bandwidth and dimensionality, respectively. x_i represents the indexed data point, and ϕ is the kernel function, or specifically for this study, the Gaussian Kernel function. The Gaussian Kernel Function is as follows:

$$\phi(\mu) = \frac{1}{\sqrt{2\pi}} exp(-\frac{\mu^2}{2})$$

Where $\phi(\mu)$ is the Gaussian Kernel, $\frac{1}{\sqrt{2\pi}}$ acts as the normalization for the Gaussian function, and the $exp(-\frac{\mu^2}{2})$ represents the weight of each data point.

3.1.5 Evaluation Method

We measured the algorithm's performance using a confusion matrix, which determines the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) data samples that the algorithm classified. These are then used to calculate the four performance metrics to evaluate the algorithm.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Accuracy calculates the number of data samples that were correctly classified out of the total number of samples in the dataset to determine the correctness of the algorithm during its classification.

$$Precision = \frac{TP}{TP + FP}$$

Precision calculates the number of data samples correctly classified as positive out of all the predicted positive data samples to determine the likelihood that the algorithm correctly classifies a positive instance.

$$Recall = \frac{TP}{TP + FN}$$

Recall calculates the number of data samples correctly classified as positive out of all the actual positive data samples to determine the correct number of instances the algorithm would correctly classify a positive instance.

$$F1\ score = \frac{Precision\ \times\ Recall}{Precision\ +\ Recall}$$

F1 score combines precision and recall by providing a standardized basis for comparison, which is helpful for datasets with uneven class distributions.

3.2 Conceptual Framework of the Existing GNB

This section of the paper arranges the key concepts of the Gaussian Naive Bayes into a conceptual framework seen in the figure below which also includes and highlights the phases in the algorithm where the identified problems discussed in this study occurs.

Figure 2. Conceptual Framework of the Existing GNB

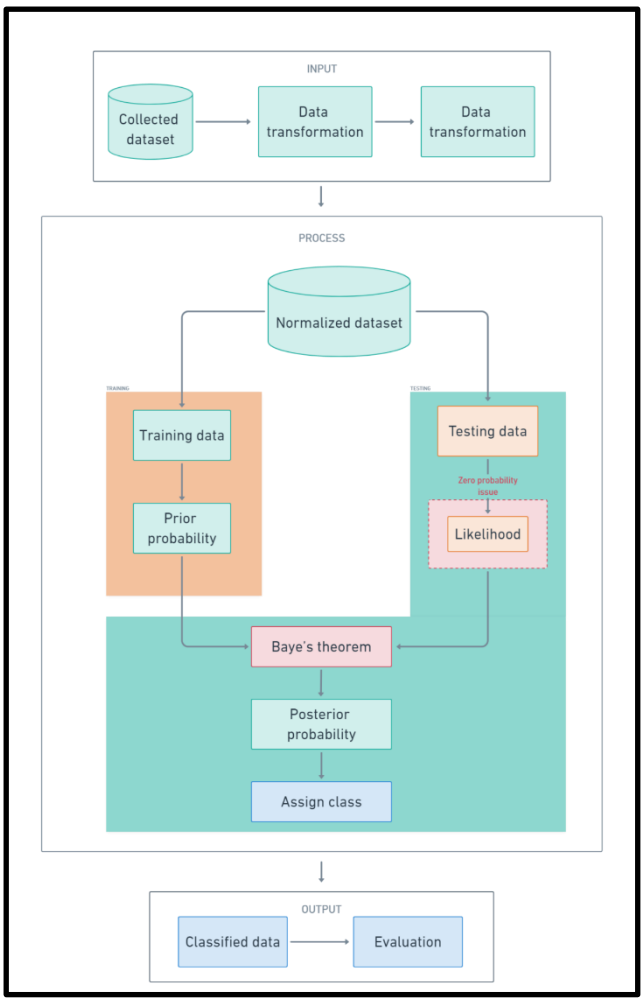


Figure 2 outlines the process of the existing Gaussian Naive Bayes algorithm. The first phase being the Input phase where data is collected and pre-processed, this is also where statement of the problem number 3 persists, pertaining to noisy datasets. The second phase involves the process, which first splits the dataset into training and testing data. The training data is used to calculate the prior probability of each class which is substituted into the Bayes Theorem alongside the calculated likelihood from the testing data which is where statement of the problem number 2 occurs, referring to zero probability instances. Once the likelihood and prior probability is acquired they are substituted into the Bayes Formula to calculate for the posterior probability which inhibits the third statement of the problem, pertaining to the assumption of feature independence by the algorithm. Lastly, this leads to the output phase, where the data is classified to their respective class and finally the algorithm is evaluated.

3.3 Proposed Algorithm

This section discusses the pseudocode and framework of the proposed enhancement to the Gaussian Naive Bayes algorithm. This included the integration of the methods mentioned in the objectives into the GNB algorithm.

3.3.1 Enhanced Gaussian Naive Bayes Pseudocode

(1) Data Collection

(2) Data Pre-Processing
(a) Incomplete Data Removal
(b) Data Transformation
(c) Data Normalization
(3) Data Split (Training Dataset and Testing Dataset)
(4) Training Phase
(a) Model Training

(Objective)

Initialize Parzen Gaussian Naive Bayes with a bandwidth parameter.
Define the fit method for the resampled dataset

1) For each class compute the class prior probability
2) For each feature initialize and fit the Parzen window density estimator from the feature values belonging to the class.
3) Store fitted Parzen window in the class' feature density list

(5) Testing Phase
(a) For each test sample initialize the probabilities of each class
For each class start with the logarithm of the class prior

(Objective)

1) For each feature use the stored Parzen window density estimator to calculate the log-density for the feature.
2) Add the log-density to the class' probability

(b) Classify the test sample to the class with highest posterior probability
(6) Evaluation Phase
(a) Calculate the accuracy
(b) Generate classification report showing the f1 score, precision and recall.

3.3.2 Conceptual Framework of the Enhanced Gaussian Naive Bayes

Figure 3. Conceptual Framework of the Enhanced GNB

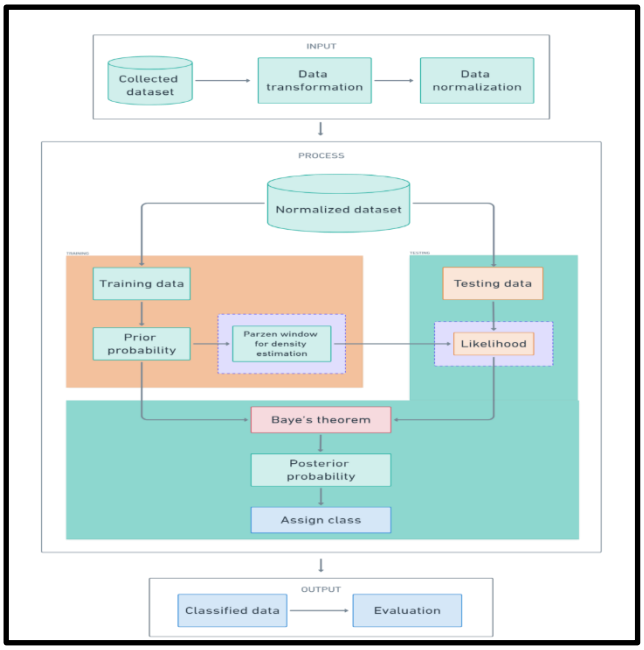


Figure 3 illustrates the conceptual framework of the enhanced Gaussian Naive Bayes, which incorporates the Parzen-Rosenblatt Window method to solve the zero-probability issue inherent within the Gaussian Naive Bayes algorithm. The framework is divided into input, process, and output. The input section involves data collection, transformation, and normalization to prepare the data for processing and classification.

The Process section includes the modifications made to the algorithm. The first step involves splitting the dataset into 20% testing and 70% training set. Afterwards, the prior probability for each class is calculated. Then, the Density Estimation for each feature considering its class is calculated using the Parzen-Rosenblatt Window method, which is fitted to the values of its corresponding feature and stored for later use in calculating likelihood during the testing phase.

The testing phase uses the density estimation computed during the training phase to calculate the likelihood of each data sample considering its class. The computed likelihood and prior probability (from the training phase) are used on Bayes' theorem to calculate the posterior probability of each data sample. The highest posterior probability will determine the class of the data sample. The output phase involves the classified dataset and the evaluation of the algorithm's performance using the four metrics (accuracy, precision, recall, and F1-score).

3.3.3 Parzen-Rosenblatt Window Method for Zero Probability

The Parzen-Rosenblatt Window Method was integrated for calculating the feature density estimation of the dataset, this is used during the calculation of the likelihood of the data for each class. The incorporation of the Parzen-Rosenblatt Window method provides a continuous estimate of the density ensuring that the likelihood would rarely or never equate to zero. The Parzen-Rosenblatt Window used in this study utilized a gaussian kernel which ensures that every instance contributes to the density considering a weight that decreases as the distance increases. Therefore, even when a test point is located in an area not explored during the training phase, the surrounding data influences that calculation therefore preventing zero probabilities. Hence, alleviating the zero probability issue inherent to the traditional GNB.

3.4 System Requirements

3.4.1 Integrated Development Environment

The researchers utilized Google Colaboratory as the testing environment for this study. Google Colab is a free Jupyter notebook that is accessible on the internet, it offers a cloud-based virtual environment that enables the researchers to use the environment's resources and access to its supported libraries.

3.4.2 Libraries

NumPy (numpy) - Used for numerical computations such as calculating distances in the Parzen window density estimation, and performing array operations for data processing.

Pandas (pandas) - Used for handling and manipulating datasets, particularly for splitting features and target variables, iterating over rows for prediction, and ensuring compatibility with custom models.

Matplotlib (matplotlib.pyplot) - Used to plot a heatmap for the confusion matrix.

Seaborn (seaborn) - Used to enhance the visualization of the confusion matrix with a detailed and color-coded heatmap.

Scikit-learn (sklearn):

train_test_split - To split the dataset into training and testing sets for model evaluation.

accuracy_score - To calculate the classification accuracy of the model.

classification_report: To generate a detailed performance report for each class.

confusion_matrix - To create a confusion matrix summarizing the model's performance.
StandardScaler - To normalize features, ensuring they have a mean of 0 and a standard deviation of 1.
Imbalanced-learn (imblearn) - Used for handling imbalanced datasets:
Custom Classes:
ParzenWindow - Implements a Gaussian Parzen window method for density estimation. It is used in the custom Gaussian Naive Bayes model to calculate probabilities for each feature.
ParzenGaussianNB - A custom Gaussian Naive Bayes classifier that uses Parzen windows for density estimation instead of assuming a standard Gaussian distribution for features.

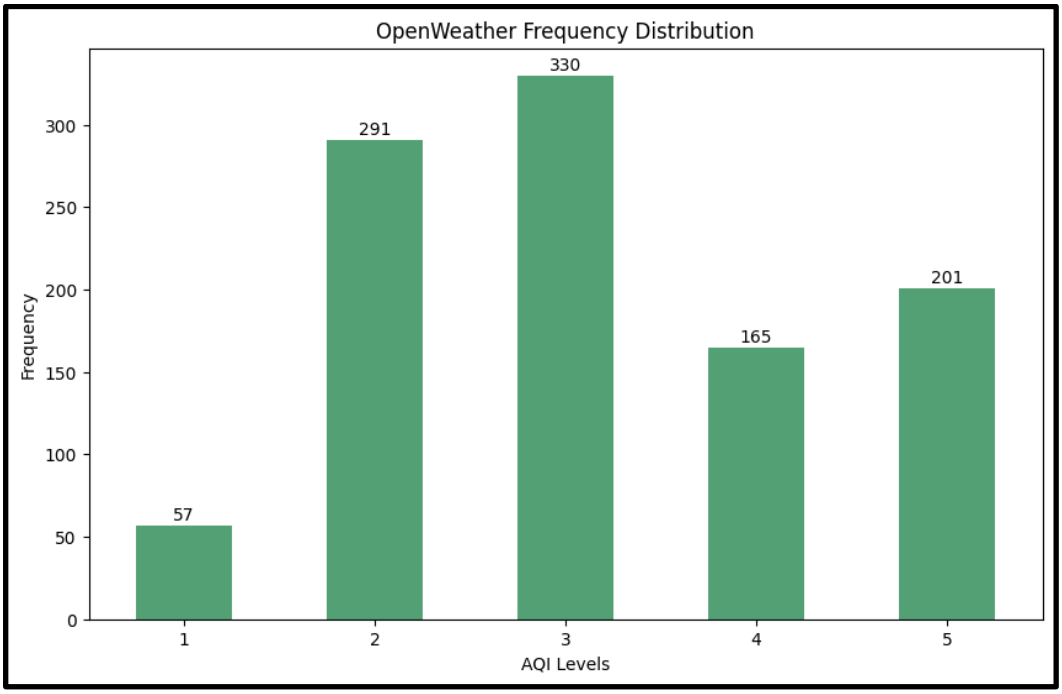
Chapter 4
RESULTS AND DISCUSSION

This chapter summarizes and discusses the results acquired from implementing the proposed methods of this study and aligns them on the research objectives along with contextualizing them within existing literature.

4.1 Results

Figure 4 showcases the frequency distribution of the OpenWeather-AQI dataset entailing the air quality status in Pamantasan ng Lungsod ng Maynila. The criteria used to qualify the quality of air is seen on Appendix A. The figure shows that AQI level 3 (Moderate) contains the most data samples, with 330 data samples belonging to the Moderate air quality group. AQI level 2 (Good) has the second most data samples, with 291. It is followed by AQI level 5 (Very Poor) with 201 data samples, then AQI level 4 (Poor) with 165 samples, and lastly, AQI level 1 (Good) has the lowest number of data samples, with 57 data samples classified.

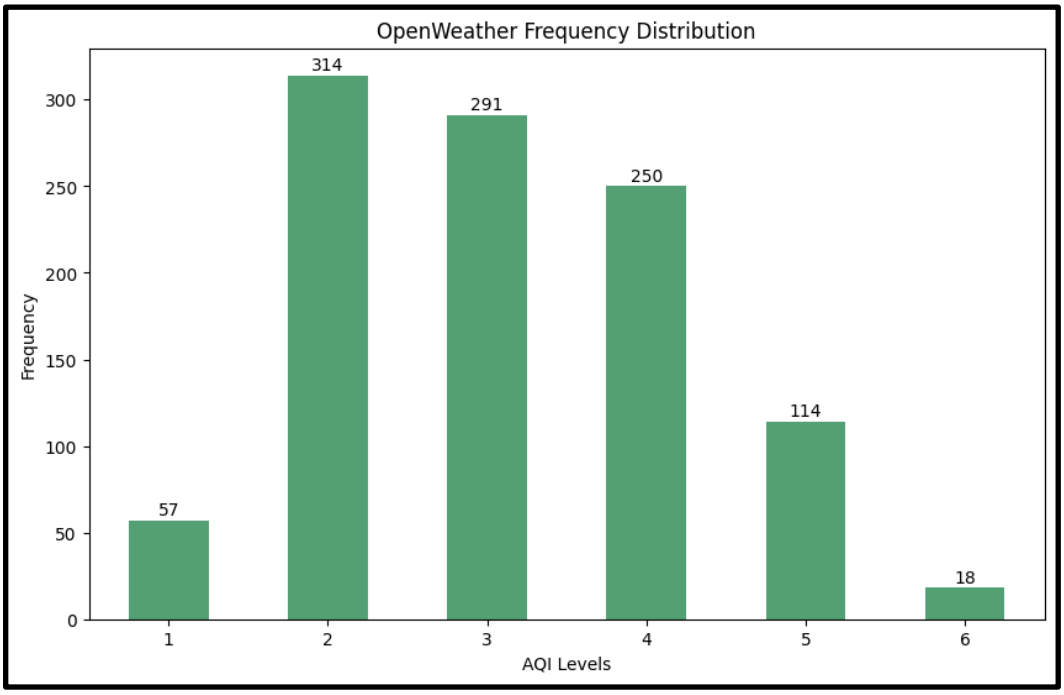
Figure 4. OpenWeather AQI Dataset Frequency Distribution



The frequency distribution above gives an overview of the 2023-2024 air quality data in Pamantasan ng Lungsod ng Maynila. AQI level 3 has the most dates showcasing that the air quality in PLM is moderate for 110 days during the time frame. Consequently, PLM has experienced air quality deemed very poor for the general public, specifically its students and the community that uses the University as its environment for working and studying for 67 days. Furthermore, the figure showcases that the best or fair air quality contained the least data samples interpreting to a worsened air quality experienced by the community over the days.

Figure 5 entails the distribution of dates considering the air quality scale used in the USA, seen in appendix A The figure below includes only 6 of the 7 classes indicated in the USA AQI scale. The dataset contained no samples from the 7th class, which refers to Very Hazardous air quality. Hence, it was removed from the graph. The six classes are ranked based on their index ranging from 0 - 500, and those that would result in an index greater than 500 would be considered Very Hazardous. Class 2 contains the most data samples with 314 data samples or 104 dates classified as Moderate, class 3 (Unhealthy for sensitive groups) followed with 291 data samples or 97 days, followed by class 4 (Unhealthy) which had 250 data samples, then class 5 (Very Unhealthy) contained 114 data samples, class 1 (Good) had 57 data samples. Lastly, class 6 had 18 data samples or 6 dates considered Hazardous.

Figure 5. USA AQI Dataset Frequency Distribution



The USA AQI dataset contains an imbalanced data distribution among the six classes, with 3 classes having 100 or more data samples compared to the other classes. In contrast, class 6 only contained 18 data samples. This indicates that, even after using the USA scale for AQI, the air quality in Pamantasan ng Lungsod ng Maynila was categorized as "Moderate" for most days from November 2023 to November 2024. This is followed the AQI level 3-5, showcasing that the distribution is slightly skewed to the right, which leans towards a worsened air quality experienced by the community over the days

After applying the proposed method to enhance the traditional Gaussian Naive Bayes, the results are recorded in Table 4.1.1. The table shows the performance comparison between the traditional Gaussian Naive Bayes and the enhanced Gaussian Naive Bayes when used on an 80% training set and 20% testing set. The last column of the table shows a check (✓) when the proposed algorithm achieved a better score in terms of the four metrics, a dash (-) if the performance was the same, while it would show a cross (✗) if the traditional algorithm performed better. The accuracy increased by 2.39% when the enhanced GNB was used on the OpenWeather AQI, while an improvement of 5.26% was noted on the USA AQI dataset and consistently outperformed the traditional GNB in terms of the other 3 metrics.

Table 4.1 Performance Comparison of the traditional GNB and GNB with Parzen-Rosenblatt Window Method on a 20/80 Data Split

Dataset Name	Performance Metrics	Traditional GNB	GNB with Parzen-Rosenblatt	Proposed Win?
OpenWeather AQI	Accuracy	71.77%	74.16%	✓
	Precision	73.00%	75.00%	✓
	Recall	72.00%	74.00%	✓
	F1-score	72.00%	74.00%	✓
USA AQI	Accuracy	59.33%	64.59%	✓
	Precision	58.00%	65.00%	✓
	Recall	59.00%	65.00%	✓
	F1-score	58.00%	63.00%	✓

Table 4.2 summarizes the results when the enhanced and traditional algorithm is used on the same datasets, but with a 70% training set and 30% testing set. The results showed that even after decreasing the training set the enhanced algorithm still performed better with a 2.23% gain in terms of accuracy compared to the traditional algorithm.

Table 4.2 Performance Comparison of the traditional GNB and GNB with Parzen-Rosenblatt Window Method on a 30/70 Data Split

Dataset Name	Performance Metrics	Traditional GNB	GNB with Parzen-Rosenblatt	Proposed Win?
OpenWeather	Accuracy	70.70%	72.93%	✓
	Precision	72.00%	74.00%	✓

AQI	Recall	71.00%	73.00%	✓
	F1-score	71.00%	73.00%	✓
USA AQI	Accuracy	58.60%	61.46%	✓
	Precision	59.00%	62.00%	✓
	Recall	59.00%	61.00%	✓
	F1-score	58.00%	60.00%	✓

4.2 Findings

The results in Table 4.1 highlight the effectiveness of using the Parzen-Rosenblatt Window Method to address the "zero-probability" issue inherent in the traditional Gaussian Naive Bayes (GNB) algorithm. This enhancement significantly improves the algorithm's performance across the four metrics for both the OpenWeather AQI and USA AQI datasets.

For the OpenWeather AQI dataset, the enhanced GNB improved classification accuracy by 2.39%, from 71.77% by the traditional GNB to 74.16% by the GNB with Parzen-Rosenblatt. Precision also improved from 73.00% to 75.00%, while recall and F1-score increased from 72.00% to 74.00%, demonstrating an improvement of 2% with the three metrics.

The USA AQI dataset contains a higher class imbalance, which increases the chances of zero-probability issues. Despite this, the enhanced GNB demonstrated significant gains across all metrics. Accuracy increased by 5.26% from 59.33% to 64.59%. Precision had a substantial 7% improvement from 58.00% to 65.00%. Similarly, recall improved from 59.00% to 65.00%. Lastly, the F1 score also increased from 58.00% to 63.00%. The Parzen-Rosenblatt Window Method successfully alleviates the zero-probability issue by providing a smoother probability density estimation, ensuring that no probability would equate to zero. The results highlight the reliability of the enhanced GNB in both datasets, particularly in more complex datasets like the USA AQI dataset, which increases the instances of zero-probability problems where traditional GNB struggles to maintain performance.

The results presented in table 4.2 compares the performance of the traditional Gaussian Naive Bayes (GNB) algorithm with the enhanced GNB incorporating the Parzen-Rosenblatt Window method, applied to the two datasets under a 30/70 data split configuration. For the OpenWeather AQI dataset, the enhanced GNB demonstrated notable improvements across all performance metrics. The accuracy increased by 2.23%, rising from 70.70% to 72.93%, indicating the enhanced model's ability to make more correct predictions. Precision, recall and F1-score also improved by 2%, from 72%, 71% and 71% to 74%, 73% and 73% respectively.

The USA AQI dataset posed greater challenges due to its higher complexity and potential class imbalance. Despite this, the enhanced GNB delivered significant performance gains. Accuracy improved by 2.86%, from 58.60% to 61.46%. Precision increased by 3%, demonstrating the enhanced model's effectiveness in reducing false positives. Recall and F1-score also improved by 2%, indicating better sensitivity and overall classification performance. These results show the enhanced algorithm's capability to manage imbalanced and noisy datasets more effectively than the traditional approach.

4.2 Summary of Results

The enhanced Gaussian Naive Bayes (GNB) algorithm outperformed the traditional GNB, with consistent improvements across all evaluation metrics, including accuracy, precision, recall, and F1-score. For the OpenWeather AQI dataset, the enhanced algorithm increased accuracy by 2.39% and 2.23% for the 20/80 and 30/70 data split, respectively. The enhancements were even more pronounced for the USA AQI dataset, where accuracy improved by 5.26% to 64.59%, and precision, recall, and F1-score saw increases of 3%, 2%, and 2%, respectively. These results highlight the enhanced GNB's effectiveness in handling more challenging datasets with imbalanced and multi-class distributions where zero-probability issues commonly occur.

Chapter 5
CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

The research study entitled **"An Enhancement of the Gaussian Naive Bayes Algorithm Applied to Air Quality Classification"** discovered the following:

1. The integration of the Parzen-Rosenblatt Window method successfully resolved the zero-probability issue inherent in the Gaussian Naive Bayes (GNB) algorithm, resulting in improved classification accuracy and reliability.
2. The enhanced GNB algorithm outperformed the traditional version across all evaluation metrics, with accuracy gains of 2.39% and 5.26% for the OpenWeather AQI and USA AQI datasets, respectively.
3.

5.2 Recommendation

Based on the successful implementation of this research, the following recommendations are proposed to further enhance the algorithm's performance and extend its applicability:

1. **Broaden Dataset Diversity**: Future studies should consider using additional air quality datasets from various regions to validate the enhanced GNB algorithm's performance across diverse environments and conditions.
2. **Incorporate Advanced Preprocessing Techniques**: Explore advanced feature selection and dimensionality reduction methods to optimize dataset quality and computational efficiency further.
3. **Extend Application Scenarios**: Test the algorithm's application in other classification problems, such as medical diagnoses or financial risk assessments, to evaluate its versatility and reliability in different domains.

REFERENCES

1. Anand, M., KiranBala, B., Srividhya, S., Kavitha, C., Younus, M., & Rahman, H. (2022, June 17). *Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer*. Journal of Mobile Information Systems. https://doi.org/10.1155/2022/2436946
2. Cinar, A. (2023, December). *Multi-Class Classification with the Gaussian Naive Bayes Algorithm*. Journal of Data Applications, 2, 1-13. https://doi.org/10.26650/JODA.1389471
3. Di Paola, G., Bertani, A., De Monte, L., & Tuzzolino, F. (2018). A brief introduction to probability. *Journal of Thoracic Disease*, *10*(2), 1129–1132. https://doi.org/10.21037/jtd.2018.01.28
4. EPA. (2024). A*ir Monitoring, Measuring and Emissions Research*. https://www.epa.gov/air-research/air-monitoring-measuring-and-emissions-research

5. Gardin, T. N., & Réquia, W. J. (2023, June 1). *Air quality and individual-level academic performance in Brazil: A nationwide study of more than 15 million students between 2000 and 2020*. Environmental Research. https://doi.org/10.1016/j.envres.2023.115689
6. Gayathri, B. M., & Sumathi, C. P. (2016, August). *An Automated Technique using Gaussian Naïve Bayes Classifier to Classify Breast Cancer*. https://www.ijcaonline.org/archives/volume148/number6/25761-2016911146/
7. Hosein, P., & Baboolal, K. (2022, May 31). *Bayes Classification Using an Approximation to the Joint Probability Distribution of the Attributes.* https://arxiv.org/pdf/2205.14779T
8. Kamble, V., & Dale, M. (2022, January 1). *Machine learning approach for longitudinal face recognition of children*. Elsevier eBooks. https://doi.org/10.1016/b978-0-323-85209-8.00011-0
9. Kamel, H., Abdulah, D. Al-Tuwaijari, J. M. (2019). *Cancer Classification Using Gaussian Naive Bayes Algorithm*. https://ieeexplore.ieee.org/document/8950650
10. Naiem, S., Khedr, A., Idrees, A., Marie, M. (2023). *Enhancing the Efficiency of Gaussian Naïve Bayes Machine Learning Classifier in the Detection of DDOS in Cloud Computing*. IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/document/10302279
11. Poola, R. G., Pl, L. & Sankar, S. (2023, June). *COVID-19 diagnosis: A comprehensive review of pre-trained deep learning models based on feature extraction algorithm.* https://www.sciencedirect.com/science/article/pii/S2590123023001470
12. Roh, Y., Heo, G., & Whang, S. E. (2019, October). *A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective.* https://www.researchgate.net/publication/336352908_A_Survey_on_Data_Collection_for_Machine_Learning_A_Big_Data_-_AI_Integration_Perspective
13. World Health Organization. (2021, September 22). *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. https://www.who.int/publications/i/item/9789240034228
14. World Health Organization. (2022, December 19). *Ambient (outdoor) air pollution.* https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health
15. Zambom, A. Z., & Dias, R. (2012, December). *A Review of Kernel Density Estimation with Applications to Econometrics*. https://arxiv.org/pdf/1212.2812