# AVALIANDO O IMPACTO DA SELEÇÃO DE RECURSOS BASEADA NA CORRELAÇÃO PONTO-BISSERIAL EM CLASSIFICADORES DE APRENDIZADO DE MÁQUINA: UM ESTUDO DE CASO DE DETECÇÃO DE FRAUDE EM CARTÃO DE CRÉDITO

**Cited as:**
A.H. Alkurdi, A., R. Asaad, R., M Almufti, S., & S. Ahmed, N. (2024). Avaliando o impacto da seleção de recursos baseada na correlação ponto-bisserial em classificadores de aprendizado de máquina: um estudo de caso de detecção de fraude em cartão de crédito. ***Revista Gestão & Tecnologia***, *24*, 166–196. https://doi.org/10.20397/2177-6652/2024.v24.2843

**Ahmed A.H. Alkurdi***
Department of Information Technology, Duhok Technical College, Duhok Polytechnic University, Duhok, KRG-Iraq
Department of Computer Science, College of Science, Nawroz University, Duhok, KRG-Iraq
Email: Ahmed.alaa@dpu.edu.krd
ORCID: https://orcid.org/0000-0002-8060-4087
*Corresponding Author: Ahmed A.H. Alkurdi.

**Renas R. Asaad**
Department of Computer Science, College of Science, Nawroz University, Duhok, KRG-Iraq
Department of Technical Informatics, Technical College of Informatics, Akre University for Applied Science, Duhok, KRG-Iraq
Email: Renas.beda89@gmail.com
ORCID: http://orcid.org/0000-0002-1762-662X

**Saman M Almufti**
Department of Computer Science, College of Science, Nawroz University, Duhok, KRG-Iraq
Department of Technical Informatics, Technical College of Informatics, Akre University for Applied Science, Duhok, KRG-Iraq
Email: Saman.almofty@gmail.com
ORCID: https://orcid.org/0000-0002-1843-745X

**Nawzat S. Ahmed**
Department of Information Technology, Duhok Technical College, Duhok Polytechnic University, Duhok, KRG-Iraq
Email: nawzat.ahmed@dpu.edu.krd
ORCID: https://orcid.org/0000-0003-1028-0491

## RESUMO

**Objetivo:** Este artigo examina os fatores que influenciam a conscientização e a adoção das Normas Internacionais de Contabilidade do Setor Público (IPSAS) nas unidades públicas do Vietnã. O objetivo é identificar os principais desafios e impulsionadores que afetam a compreensão e a implementação dessas normas.

**Métodos:** O estudo utiliza uma metodologia de pesquisa, coletando respostas de uma amostra de unidades de serviço público no Vietnã. O questionário foi elaborado para avaliar o nível de conscientização e prontidão dessas unidades para adotar as IPSAS, considerando variáveis como apoio gerencial, treinamento e infraestrutura técnica. Foi realizada uma análise estatística para determinar os fatores mais influentes.

**Resultados:** Os resultados destacam que o apoio gerencial, o treinamento adequado e o acesso à infraestrutura técnica apropriada são cruciais para a implementação bem-sucedida das IPSAS. A falta de conscientização, treinamento insuficiente e limitações de recursos são as principais barreiras

à adoção dessas normas. Unidades públicas com maiores níveis de conscientização e melhor acesso a recursos são mais propensas a implementar as IPSAS com sucesso.

**Contribuição:** O estudo oferece insights valiosos sobre o processo de adoção das IPSAS no setor público do Vietnã. Ele oferece recomendações para melhorar os programas de treinamento, aumentar o apoio gerencial e fortalecer a capacidade técnica das unidades públicas para garantir uma implementação mais suave das normas.

**Conclusão:** A implementação das IPSAS no setor público do Vietnã é influenciada por vários fatores-chave, como conscientização, treinamento e infraestrutura. O fortalecimento dessas áreas pode melhorar significativamente o processo de adoção e aumentar a transparência e a responsabilidade na gestão financeira pública.

**Palavras-chave:** Cartão de Crédito. Fraude. Aprendizado de Máquina. Desempenho Preditivo. Seleção de Recursos Baseada em PBC.

## EVALUATING THE IMPACT OF POINT-BISERIAL CORRELATION-BASED FEATURE SELECTION ON MACHINE LEARNING CLASSIFIERS: A CREDIT CARD FRAUD DETECTION CASE STUDY

## ABSTRACT

**Objective:** This article examines the factors influencing the awareness and adoption of International Public Sector Accounting Standards (IPSAS) in public units in Vietnam. It seeks to identify key challenges and drivers that affect the understanding and implementation of these standards.

**Methods:** The study uses a survey methodology, gathering responses from a sample of public service units in Vietnam. The survey is designed to assess the level of awareness and readiness of these units to adopt IPSAS, considering variables such as management support, training, and technical infrastructure. Statistical analysis was performed to determine the most influential factors.

**Results:** The findings highlight that managerial support, adequate training, and access to proper technical infrastructure are crucial for successful IPSAS implementation. Lack of awareness, insufficient training, and resource limitations are the primary barriers to the adoption of these standards. Public units that have higher levels of awareness and better access to resources are more likely to successfully implement IPSAS.

**Contribution:** The study provides valuable insights into the process of adopting IPSAS in Vietnam's public sector. It offers recommendations for improving training programs, enhancing managerial support, and strengthening the technical capacity of public units to ensure smoother implementation of the standards.

**Conclusion:** The implementation of IPSAS in Vietnam's public sector is affected by several key factors, including awareness, training, and infrastructure. Strengthening these areas can significantly improve the adoption process and enhance transparency and accountability in public financial management.

**Keywords**: Credit Card. Fraud. Machine learning. Predictive performance. PBC-based feature selection.

# EVALUACIÓN DEL IMPACTO DE LA SELECCIÓN DE FUNCIONES BASADA EN CORRELACIÓN BISERIAL PUNTUAL EN CLASIFICADORES DE APRENDIZAJE AUTOMÁTICO: UN ESTUDIO DE CASO DE DETECCIÓN DE FRAUDE CON TARJETAS DE CRÉDITO


## RESUMEN

**Objetivo:** Este artículo examina los factores que influyen en la conciencia y adopción de las Normas Internacionales de Contabilidad del Sector Público (IPSAS) en las unidades públicas de Vietnam. Su objetivo es identificar los principales desafíos y motores que afectan la comprensión e implementación de estas normas.

**Métodos:** El estudio utiliza una metodología de encuesta, recopilando respuestas de una muestra de unidades de servicio público en Vietnam. La encuesta está diseñada para evaluar el nivel de conciencia y preparación de estas unidades para adoptar las IPSAS, considerando variables como el apoyo gerencial, la capacitación y la infraestructura técnica. Se realizó un análisis estadístico para determinar los factores más influyentes.

**Resultados:** Los hallazgos destacan que el apoyo gerencial, la capacitación adecuada y el acceso a una infraestructura técnica adecuada son cruciales para una implementación exitosa de las IPSAS. La falta de conciencia, la capacitación insuficiente y las limitaciones de recursos son las principales barreras para la adopción de estas normas. Las unidades públicas con mayores niveles de conciencia y mejor acceso a recursos tienen más probabilidades de implementar con éxito las IPSAS.

**Contribución:** El estudio proporciona información valiosa sobre el proceso de adopción de las IPSAS en el sector público de Vietnam. Ofrece recomendaciones para mejorar los programas de capacitación, aumentar el apoyo gerencial y fortalecer la capacidad técnica de las unidades públicas para garantizar una implementación más fluida de las normas.

**Conclusión:** La implementación de las IPSAS en el sector público de Vietnam está influenciada por varios factores clave, como la conciencia, la capacitación y la infraestructura. Fortalecer estas áreas puede mejorar significativamente el proceso de adopción y aumentar la transparencia y la responsabilidad en la gestión financiera pública.


**Palabras clave:** IPSAS. Sector público. Vietnam. Normas contables. Implementación, Conciencia.

Editor Científico: José Edson Lara
Organização Comitê Científico
Double Blind Review pelo SEER/OJS
Recebido em 21.09.2022
Aprovado em 15.03.2024

## 1. INTRODUCTION

In the dynamic and evolving landscape of financial transactions, the ubiquitous usage of credit cards has emerged as a double-edged sword. It offers unparalleled convenience for consumers in terms of purchasing goods and services on credit and cash advances. However, this convenience is marred by the escalating challenge of credit card fraud, which has become a significant concern for financial institutions and consumers alike. The multifaceted nature of

credit card fraud, encompassing theft of physical cards, unauthorized use of card details, and fraudulent transactions, has made it a formidable issue to address[1].

A significant hurdle in credit card fraud detection arises from the inherent characteristics of transaction data, which is typically highly imbalanced. The vast majority of transactions are legitimate, overshadowing the relatively minuscule proportion of fraudulent activities. This imbalance presents a substantial challenge in accurately detecting fraud using machine learning techniques. Furthermore, the dynamic nature of transaction behaviors, where fraudsters incessantly modify their methods to mimic legitimate patterns, exacerbates the complexity of accurate fraud detection[1], [2].

Responding to these challenges, financial institutions have increasingly adopted advanced computational methodologies, particularly machine learning algorithms, to bolster their fraud detection capabilities. Techniques such as Logistic Regression, Naive Bayes, Random Forest, and Multilayer Perceptron have been deployed with varying degrees of success. Integral to the efficacy of these techniques is the process of feature selection, which involves identifying and selecting the most relevant variables from the dataset. Effective feature selection is crucial for improving the accuracy and efficiency of machine learning models, reducing overfitting, and minimizing the training time. It also plays a significant role in mitigating the issues posed by skewed data, thereby facilitating more efficient model training[1], [3].

This paper aims to scrutinize the impact of Point-Biserial Correlation-based feature selection on the performance of various machine learning classifiers, particularly in the context of credit card fraud detection. The research intends to assess key performance metrics, including accuracy, precision, and speedup ratio, both pre- and post-application of feature selection. Through this analysis, the paper aspires to contribute substantively to the ongoing efforts in augmenting fraud detection mechanisms, thereby reinforcing the security framework of credit card transactions and safeguarding both financial institutions and their clientele from the adverse effects of fraudulent activities.

Delving deeper into the aspect of feature selection in machine learning, it is imperative to acknowledge its significance, especially in high-dimensional domains such as credit card fraud detection. The process of feature selection plays a pivotal role in managing the complexity of large datasets by discerning between relevant and irrelevant features. The relevance of a feature is determined by its substantial contribution to the predictive accuracy of a model, while irrelevant features add complexity without enhancing performance. The diversity of feature selection methodologies, including filter, wrapper, embedded, and hybrid methods, necessitates a judicious choice of technique tailored to the specific demands of the task at hand[4], [5].

In the specific arena of credit card fraud detection, the selection of an effective feature selection method is of paramount importance due to the intricate and high-dimensional nature of transaction data. The challenge is to identify a method that can effectively discern the nuanced patterns of fraudulent activities amidst a sea of legitimate transactions. Forward feature selection methods, operating within a predefined feature space to identify features that augment classification performance, are particularly pertinent in this context. Recent advancements in feature selection, such as those employing Distance Correlation (DisCo), have shown promise in selecting a limited number of highly effective features from a large pool, thereby enhancing both the efficiency and interpretability of machine learning models in fraud detection scenarios[6], [7].

Correlation analysis, especially Point-Biserial Correlation, is another cornerstone in the realm of machine learning applied to fraud detection. This form of analysis is instrumental in deciphering the relationships between variables in a dataset. Understanding these relationships is crucial for the development of accurate and efficient models. Point-Biserial Correlation is specifically applicable when examining relationships between a binary variable (e.g., fraud/no fraud) and a continuous variable (e.g., transaction amount). It is invaluable in feature selection and optimization of machine learning models in fraud detection contexts, enabling the development of predictive models that are both robust and efficient[8], [9].

Correlation coefficients, central to correlational research, serve as indicators of the strength and direction of relationships between variables. It is essential to recognize, however, that correlation does not equate to causation. The presence of a correlation signifies an association but does not necessarily imply that changes in one variable are the cause of changes in another. In the field of credit card fraud detection, leveraging correlation analysis, and understanding these relationships can markedly enhance the predictive accuracy of machine learning models[9], [10].

The integration of feature selection and correlation analysis forms a fundamental framework for the application of machine learning in credit card fraud detection. The effectiveness of the feature selection techniques directly influences the model's capability to identify fraudulent transactions accurately. Simultaneously, correlation analysis provides deeper insights into the relationships between transactional variables, which is critical for developing sophisticated and reliable fraud detection systems. this paper endeavors to explore these aspects in depth, aiming to contribute significantly to the field of credit card fraud detection[11], [12].

This study aims to highlight the effect of feature selection based on the results of point biserial correlation analysis in the optimization of several machine learning models. The credit card fraud dataset is utilized as a case study to demonstrate the influence of PBC Based feature selection due to the large number of instances and features. The study intends to draw a comparison and underline the key differences of model performance for the dataset pre-feature selection and post-feature selection. The models' performance is evaluated using several evaluation metrics (accuracy, precision, recall, etc.) and required training time. This approach provides a balanced view of the tradeoff predicted between the evaluation metrics and the training speed, which is assumed to be an inverse relationship. i.e. the study anticipates by reducing the volume of data, the model accuracy is slightly lessened while the training time is significantly enhanced.

The intricate patterns and relationships unveiled by this analysis provide insights into the role of PBC-based feature selection in optimizing machine learning models for credit card fraud detection. By presenting an empirical evaluation of this method, the study contributes to the ongoing discourse on the strategic application of feature selection in high-dimensional and imbalanced datasets, aiming to bolster the security mechanisms against credit card fraud in an increasingly digitalized economy.

This paper is composed of a preface and six substantive sections that cohesively present the research conducted. Section 1: Introduction sets forth the paper's premise, outlining the pertinence of credit card fraud detection and the complexity introduced by imbalanced transaction data, advocating for advanced machine learning solutions. Section 2: Literature Review systematically appraises relevant scholarly work, spotlighting the role of Point-Biserial Correlation (PBC) in enhancing machine learning classifiers across diverse fields. Section 3: Model explains the adoption of PBC as a feature selection tool for a gamut of machine learning classifiers and hypothesizes its potential benefits in fraud detection efficiency and effectiveness. Section 4: Dataset describes the comprehensive Credit Card Fraud Detection Dataset 2023, focusing on its key attributes crucial for the research. Section 5: Evaluation Metrics details the metrics employed to assess the classifiers' performance, providing a multi-dimensional view of the impact of PBC-based feature selection. Section 6: Results contrasts the performance of classifiers before and after PBC-based feature selection, showcasing the nuanced impact on efficiency and predictive accuracy. The paper concludes with Section 7: Conclusion, summarizing the findings and underscoring the significant influence of PBC-based feature selection on improving machine learning classifiers for credit card fraud detection, while highlighting the broader implications and contributions of the research.

## 2. LITERATURE REVIEW

The literature review encompasses a range of studies that delineate the application of machine learning across various domains. This review particularly focuses on the role of feature selection, accentuated by Point-Biserial Correlation analysis, in augmenting the efficacy of machine learning classifiers.

In [13] the significance of feature selection is highlighted within a critical healthcare context. The study employs Point-Biserial Correlation analysis to enhance the predictive accuracy of machine learning models, specifically targeting early relapse in multiple myeloma patients. This underscores the essential role of precise feature selection in medical prognostic models, illustrating how it can directly impact patient outcomes.

Expanding the applicability of these methods to environmental science, [14] explores the utilization of Point-Biserial Correlation. This research establishes the method's effectiveness in identifying associations between dichotomous and continuous variables, particularly in discerning the toxicity mechanisms of chemical additives in microplastics. The study provides a detailed and sophisticated viewpoint on the ways in which machine learning and feature selection techniques might aid in comprehending and reducing environmental health hazards.

The importance of feature selection in improving the performance of machine learning classifiers is further emphasized in [15] by Khalid Iqbal and Muhammad Shehrayar Khan. This study explores the field of cyber security, highlighting the significance of Point-Biserial Correlation in enhancing email categorization systems, specifically for the purpose of identifying spam. The study's results demonstrate the wider significance of feature selection in different machine learning applications, emphasizing its essentiality in attaining greater accuracy and efficiency.

Paper [16] examines the primary obstacles and potential drawbacks associated with employing performance indicators and surrogate markers in multi-class machine learning classification within a healthcare context. The paper provides insightful recommendations for more accurate and reliable methodologies in radiation oncology research, reflecting on the broader implications of feature selection and model validation in clinical settings.

Further, the [17] offers a systematic approach to transforming continuous predictors in logistic regression models. The incorporation of the Point-Biserial Correlation coefficient in this process marks a novel approach within the field, aiming to refine model performance. This research not only contributes to logistic regression analysis but also emphasizes the importance of methodical feature transformation in predictive modeling.

The innovative approach for COVID-19 prediction presented in [18] combines feature selection and machine learning techniques to significantly enhance prediction accuracy. This study exemplifies the potential of machine learning in urgent healthcare scenarios, demonstrating how targeted feature selection can drastically improve diagnostic processes.

The use of machine learning classifiers in [19] indicates the potential of these methods in anesthesia drug detection. This study exemplifies how selecting significant PPG waveform features can lead to reliable, non-invasive detection methods, further broadening the scope of machine learning applications in healthcare.

Moreover, [20] showcases the application of machine learning classifiers in psychiatric diagnosis. This study illuminates the crucial role of cortical thickness and subcortical volume as significant features, underlining the importance of feature selection in enhancing the accuracy of psychiatric disorder classifications.

In a similar vein, [21] demonstrates the effectiveness of artificial intelligence and machine learning techniques in hematology. This study underscores the need for precise feature selection in developing efficient screening methods for hematologic malignancies, showcasing the practical implications of these technologies in clinical diagnostics.

[22] presents an innovative approach to identifying risk factors for morbid glomerular hypertrophy using machine learning. The effectiveness of Symbolic Regression via Genetic Programming (SR via GP) in feature selection highlights the method's potential in medical research, particularly in nephrology.

The study [23] uncovers significant disparities in machine learning model performance for diagnosing asymptomatic bacterial vaginosis across different ethnic groups. This research brings to light the importance of considering ethnic-specific features and the necessity for unbiased AI tools in healthcare diagnostics.

Lastly, [24] and "Biserial Miyaguchi–Preneel Blockchain-Based Ruzicka-Indexed Deep Perceptive Learning for Malware Detection in IoMT" highlight the application of machine learning and feature selection in cybersecurity. These studies emphasize the efficacy of machine learning combined with correlation algorithms for feature selection in detecting phishing SMS and malware in IoMT, illustrating the versatility and importance of these methodologies in protecting digital infrastructure.

In summation, the literature review encapsulates the diverse applications of machine learning and the pivotal role of feature selection, particularly through Point-Biserial Correlation analysis, across various domains. This body of work collectively underlines the transformative

impact of these techniques in enhancing the accuracy, efficiency, and reliability of machine learning models, thereby contributing significantly to their respective fields.

## 3. MODEL

this research delves into the effects of employing Point-Biserial Correlation (PBC) as a feature selection tool on several prevalent machine learning classifiers in the realm of credit card fraud detection as detailed in figure 1. The study is pivotal in its endeavor to quantify the extent to which PBC-based feature selection can optimize classifier performance. By focusing on a suite of diverse classifiers, the research provides comprehensive insights into the efficacy of PBC-based feature selection across various algorithmic structures.

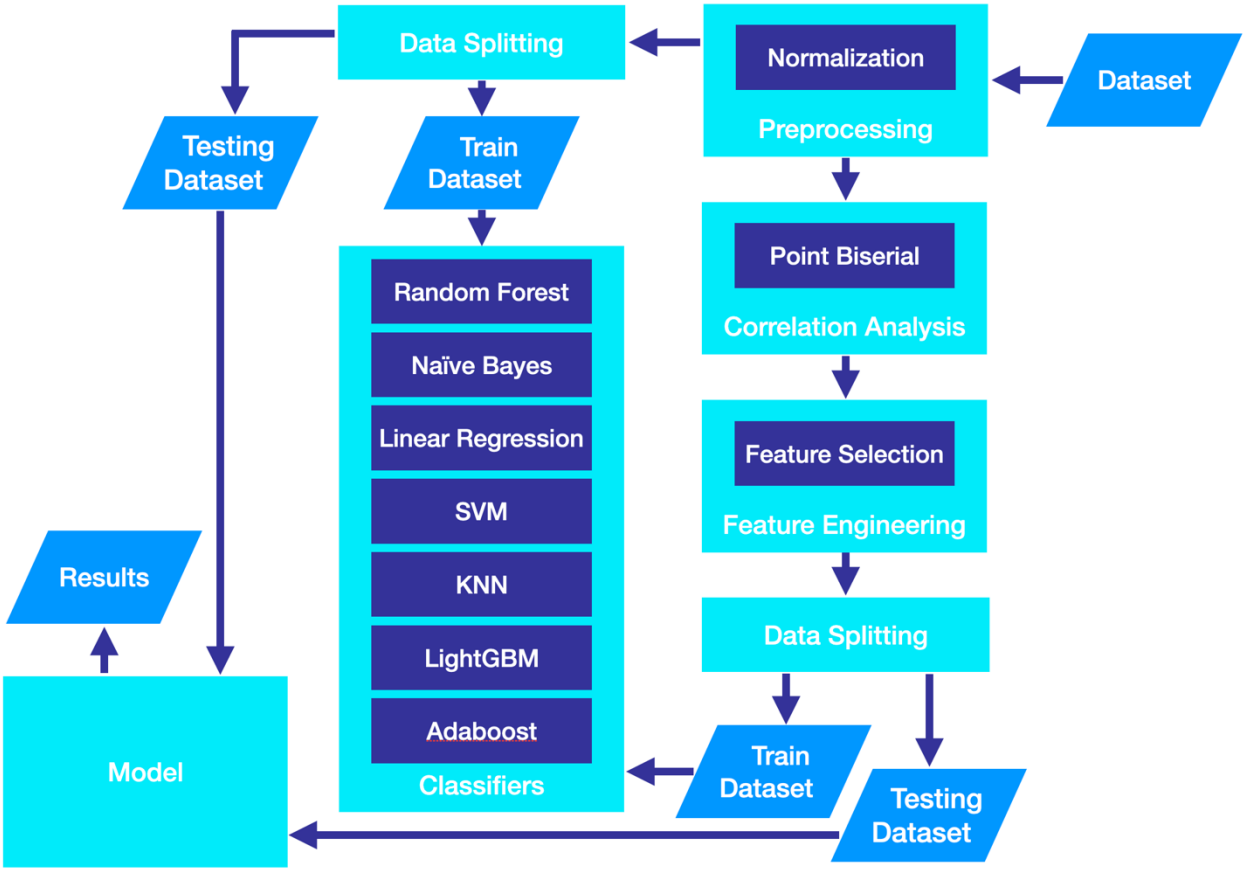


Figure 1: Proposed Model

This paper investigates various machine learning classifiers for fraud detection, such as Random Forest, Naive Bayes, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), LightGBM, and AdaBoost. The selection of these classifiers is determined by their extensive application in fraud detection scenarios, as well as their diverse degrees of intricacy and operating methodologies. The study utilizes a meticulous technique to evaluate the performance metrics of each classifier, including accuracy, precision, and speedup ratio.

This evaluation is conducted in two distinct scenarios: one using the original dataset and the other employing feature selection based on Point-Biserial Correlation (PBC). This comprehensive comparative analysis aims to elucidate the impact of feature selection on the efficiency and effectiveness of each classifier.

The utilization of PBC (Principal Component Analysis) in feature selection has great importance in this study. The Point Biserial Correlation (PBC), a very effective method for assessing the association between a binary variable and a continuous variable, is utilized to methodically identify key factors that have a major impact on the classification of credit card transactions as either fraudulent or valid. The hypothesis posits that employing PBC-based feature selection will yield a more precise selection of features, hence enhancing the classifiers' capacity to accurately and efficiently identify fraudulent behaviors. This theory is predicated on the premise that eliminating superfluous or less significant attributes may diminish interference and computational intricacy, hence potentially enhancing classifier performance metrics.

### 3.1 Point Biserial Correlation

The Point Biserial Correlation (PBC) statistic is employed to analyze the association between a binary variable that has two categories and a metric variable that has a continuous scale. PBC, a modified version of Pearson's product moment correlation, is specifically designed to handle situations where one of the variables involved is dichotomous, often represented by the values 0 and 1. The coefficient r_pb in PBC measures both the strength and direction of the association between these distinct types of variables. This metric is valuable for elucidating the nature of the relationship between binary and continuous variables, unveiling their correlation in various analytical scenarios[25].

Correlational research examines the relationship between different qualities or variables by studying how changes in one are associated with changes in others. Such study is vital in disciplines like nursing and health research, as it is crucial to comprehend the interconnections between factors. It has the capability to forecast events using existing data and knowledge, as well as to analyze the frequency and connections between variables. Correlational research, including PBC, is indispensable in various domains due to its crucial role in analysis and interpretation[26], [27].

An important aspect to remember is that correlation does not imply causation. It merely indicates that two variables are associated, without providing information on the nature of their relationship. This distinction is crucial in interpreting correlation coefficients correctly[28].

In practical applications, the correlation coefficient is significant in linear regression models. A higher correlation coefficient value indicates better prediction accuracy of the dependent variable with the least errors. The Coefficient of Determination ($R^2$) derived from the correlation coefficient provides a measure of how much variation in the dependent variable is explained by the independent variable[10].

Figure 2: Heatmap for Point Biserial Correlation Analysis Results presents a visual representation of the point biserial correlation coefficients calculated between the binary fraud classification variable and continuous/ordinal predictor variables within the credit card fraud dataset. The heatmap elucidates the linear relationship between the class label and individual features, with the color intensity and the accompanying numeric values signifying the strength and direction of these relationships.

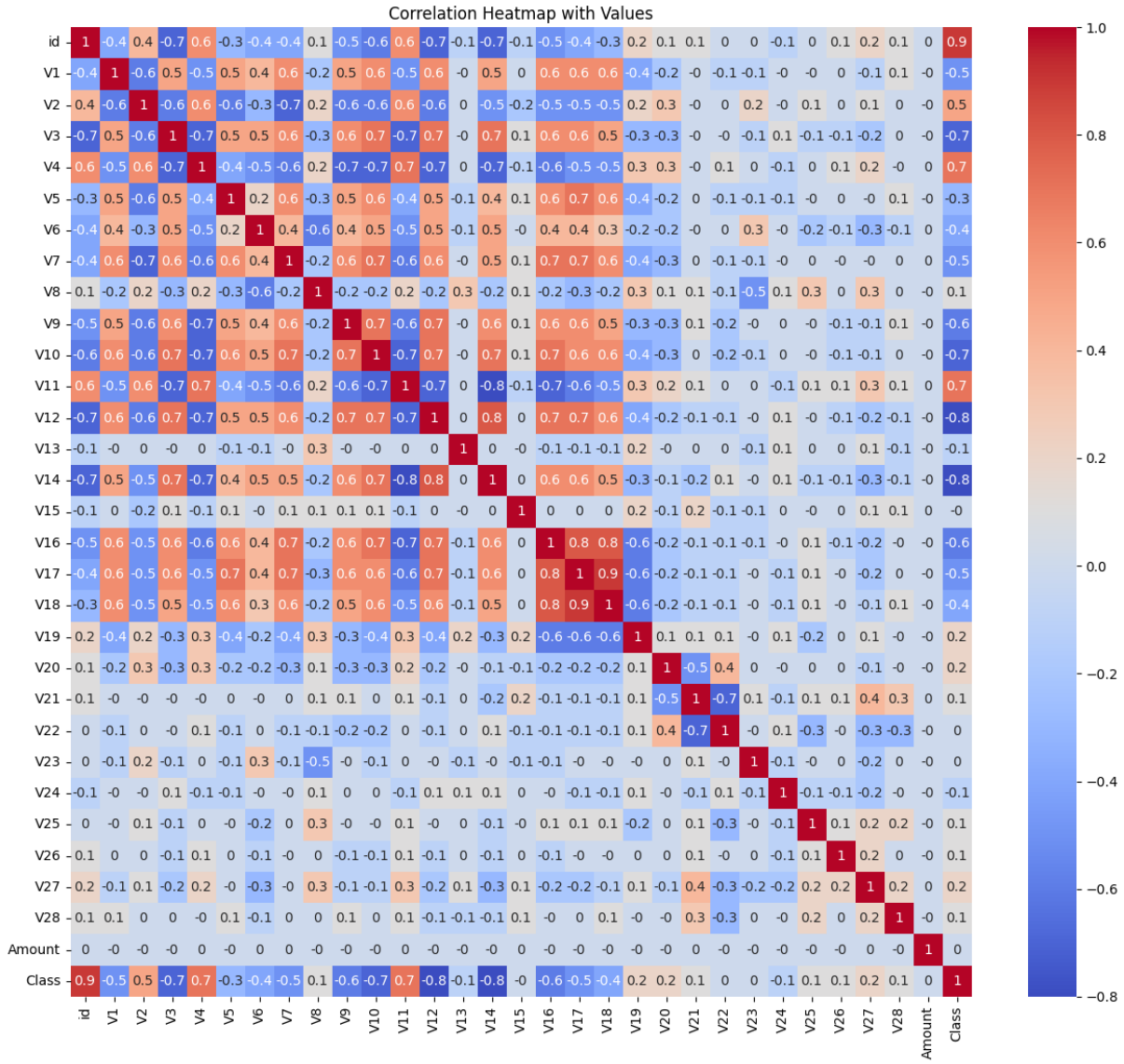


Figure 2: Heatmap for Point Biserial Correlation Analysis Results

The correlations range from -1 to 1, with a value of 1 indicating a perfect positive linear relationship, -1 a perfect negative linear relationship, and 0 no linear relationship. In the context of fraud detection, features with higher absolute values of correlation are indicative of a stronger linear association with the occurrence of fraud.

A cursory analysis of the heatmap reveals a varied landscape of correlation strengths. Certain features exhibit notably high positive correlations with the class variable, as indicated by the dark red squares (e.g., V11, V4, V2), suggesting that these features increase as the likelihood of fraud increases. Conversely, some features display strong negative correlations (e.g., V17, V14, V12), as denoted by the dark blue squares, implying that higher values of these features may be associated with legitimate transactions.

It is noteworthy that some features demonstrate minimal to no correlation with the class variable, as evidenced by the lighter shades, indicating that these variables might have negligible predictive power in differentiating between fraudulent and legitimate transactions. This highlights the potential for dimensionality reduction through feature selection, which could streamline the machine learning models by focusing on the most informative attributes.

The heatmap also informs the preprocessing steps, as the observed correlations can guide the feature engineering process, potentially leading to more robust and interpretable models. For instance, features with high correlations could be prioritized, while those with low correlations might be candidates for exclusion or for more complex transformations to uncover non-linear relationships.

In summary, Figure 2's heatmap provides a critical foundation for feature selection and model optimization in the study's pursuit to enhance credit card fraud detection methodologies. The point biserial correlation analysis stands as a testament to the nuanced and multifaceted nature of the dataset, emphasizing the importance of a strategic approach in handling high-dimensional data for predictive modeling.

### 3.2 Feature Selection

Feature selection, a critical process in data mining and machine learning, involves selecting a subset of relevant features from a dataset for model construction. This process is particularly important in handling high-dimensional data, where it addresses the 'curse of dimensionality' and helps in building simpler, more comprehensible models, enhancing data mining performance, and preparing clean, understandable data. Feature selection can be broadly classified into supervised, unsupervised, and semi-supervised methods based on the availability

of supervision, such as class labels in classification problems. Each category approaches feature relevance and selection criteria differently, based on the nature of the available data and the specific goals of the analysis[7].

The strategies for feature selection can be categorized into wrapper, filter, and embedded methods. Wrapper methods evaluate the quality of selected features based on the predictive performance of a predefined learning algorithm. Filter methods, on the other hand, assess feature importance independently of learning algorithms, relying on characteristics of the data such as feature discriminative ability, correlation, and mutual information. Embedded methods are a tradeoff between filter and wrapper methods, embedding the feature selection process into model learning. These methods aim to maximize relevance and minimize redundancy among features. The relevance of a feature is determined based on its contribution to the target prediction, while redundancy is assessed in terms of replaceability by other features[6], [29].

In practice, given the computational infeasibility of evaluating all possible feature subsets, especially for large datasets, heuristic methods are often employed to find a sufficiently good subset of features. The process typically involves four steps: subset generation, subset evaluation, a stopping criterion, and validation of the results. This approach ensures that a practically efficient subset is chosen without necessarily achieving theoretical optimality. This balance between computational feasibility and effectiveness in feature selection is crucial in practical data mining and machine learning applications[5], [30].

### 3.3 PBC based feature selection

This research introduces a distinctive approach to feature selection. PBC, a variant of Pearson's correlation, is adept at evaluating associations between a binary outcome (e.g., fraud/no fraud) and continuous variables (e.g., transaction amount). This method is crucial for determining the relevance of features, which is a key step in optimizing machine learning models for fraud detection. PBC's effectiveness lies in its ability to discern influential features from high-dimensional data, thus enhancing model accuracy and computational efficiency. By methodically identifying significant attributes, PBC-based feature selection reduces model complexity and interference, potentially improving classifier performance.

In the model, a filtration feature selection approach is performed based on the results of PBC. Features that are highly correlative are merged into a single feature using the mean values of the feature instances. As a results, the dataset is reduced by approximately 38% in volume, which is a considerable amount in term of the fraud dataset.

### 3.4 Classifiers

The suggested model proposed in this study relies on a diverse range of classifiers. The machine learning methods examined in this paper encompass Random Forest, Naive Bayes, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), LightGBM, and AdaBoost.

*Random forest*

The Random Forest classifier is a significant advancement in the field of machine learning, credited to L. Breiman in 2001. It stands out for its versatility and capability in both classification and regression tasks, particularly in scenarios with a high number of variables compared to observations. The core principle of Random Forest involves the aggregation of multiple randomized decision trees, each contributing to the final prediction outcome. This methodology not only exhibits remarkable performance but also offers adaptability for a wide range of large-scale problems and learning tasks, including the generation of variable importance measures[31], [32].

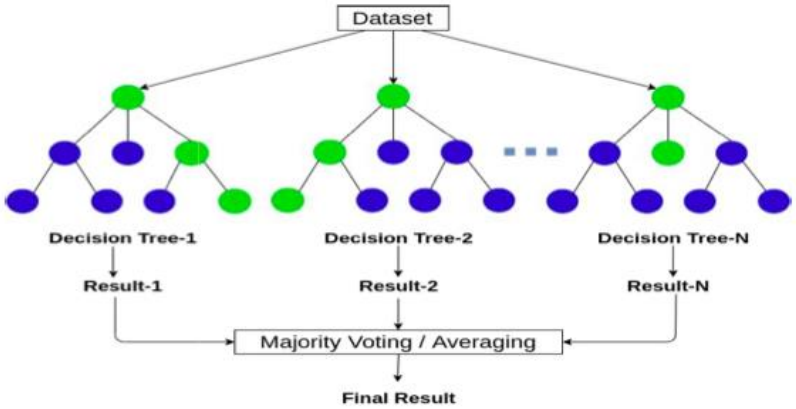


**Figure 3**: Random Forest Classifier [33].

The construction of a Random Forest classifier involves the growth of multiple, distinct decision trees. This process starts by drawing a subset of observations, either with or without replacement, from the original dataset. These selected observations are then exclusively used for building each tree. A key aspect of the algorithm is the split performed at each node of the tree, guided by the CART-criterion and considering a randomly chosen subset of the available directions, termed 'mtry'. The construction of individual trees ceases when each cell contains fewer than a specified 'nodesize' number of points. It's noteworthy that these parameters — the number of sampled data points per tree ('an'), 'mtry', and 'nodesize' — play a crucial role in the algorithm's performance and are adjustable depending on the specific application[34].

A remarkable attribute of the Random Forest algorithm is its resistance to overfitting, even as the number of trees in the forest ('M') increases. This characteristic enables the attainment of more accurate predictions without the risk of overfitting, a common challenge in many machine learning models. However, it's important to balance the computational cost, which increases linearly with the number of trees, against the desired accuracy. Therefore, the choice of 'M' involves a trade-off between computational complexity and the stability of predictions[35].

*Naïve Bayes*

The Naive Bayes Classifier (NBC) is a fundamental probabilistic classification algorithm in machine learning. This classifier is based on applying Bayes' theorem with the naive assumption of conditional independence between every pair of variables. In NBC, for a set of independent variables $X = \{x_1, x_2, \dots, x_n\}$, the posterior probability is constructed for each possible class $C = \{c_1, c_2, \dots, c_n\}$The classification score $P(C|X)$ is then calculated, which is proportional to the product of the prior probability of each class $P(C)$ and the likelihood $P(X|C)$. The final classification is determined by the argument that maximizes this classification score[36], [37].

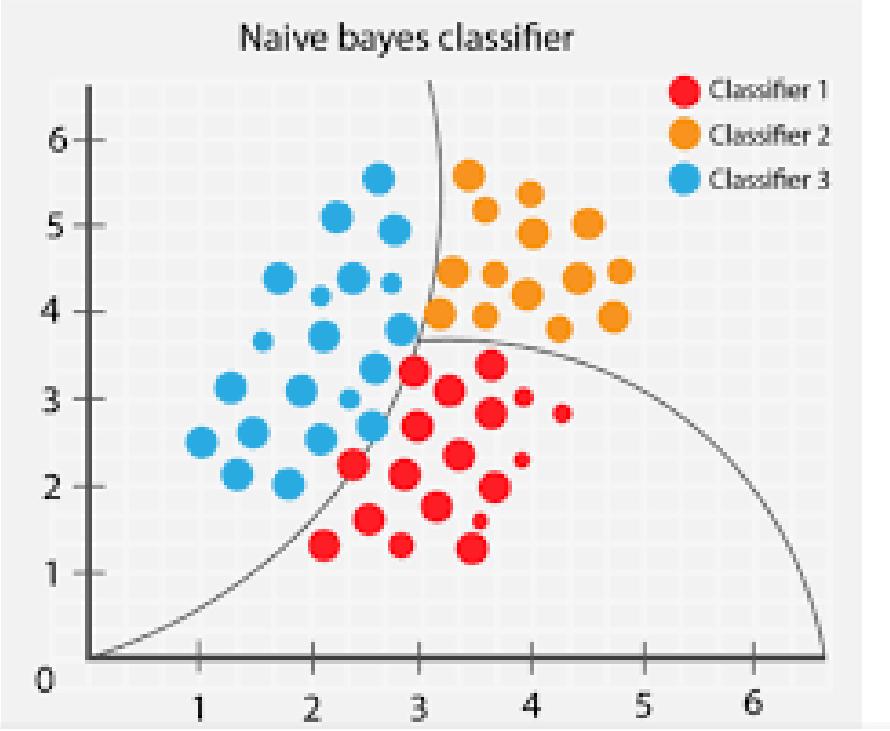


**Figure 4**: Naïve Bayes Classifier [38].

The implementations of the Naive Bayes Classifier vary mostly based on the assumptions they make about the likelihood distribution. In collaborative filtering, the data can be presumed to follow a multinomial distribution, where each user rating is assigned to a specified set of rating values. Nevertheless, in practical scenarios, the calculation of $P((A1 = v1 \cap A2 = v2 \cap \ldots \cap An = vn)|C_i$is proven to be very intricate and results in a substantial increase in computing burden. In order to simplify this situation, it is common to make the simplistic assumption that all attributes are independent of each other. This allows for a more direct calculation by considering simply the multiplication of individual probabilities [39].

The Naive Bayes Classifier is highly effective because to its efficiency and simplicity, particularly when handling discrete and continuous data across several domains. Although the Naive Bayes approach assumes independence, which may not always be valid in real-life situations, it has been effectively utilized in several domains, showcasing its practical utility and resilience [39]. The Naive Bayes Classifier is widely used in machine learning and data mining applications due to its capacity to handle many data sources and problem domains, as well as its computational efficiency.

*Linear Regression*

Logistic Regression is a prevalent statistical modeling technique that finds use in diverse disciplines, such as machine learning. This paradigm is highly efficient for binary classification tasks, where the conclusion is either one of two distinct possibilities. Logistic Regression stands out due to its distinctive approach, as it does not necessitate a linear association between the dependent and independent variables. Instead, it utilizes a combination of continuous and discrete predictors to estimate the likelihood of an outcome, enabling successful modeling of intricate interactions between variables[40].

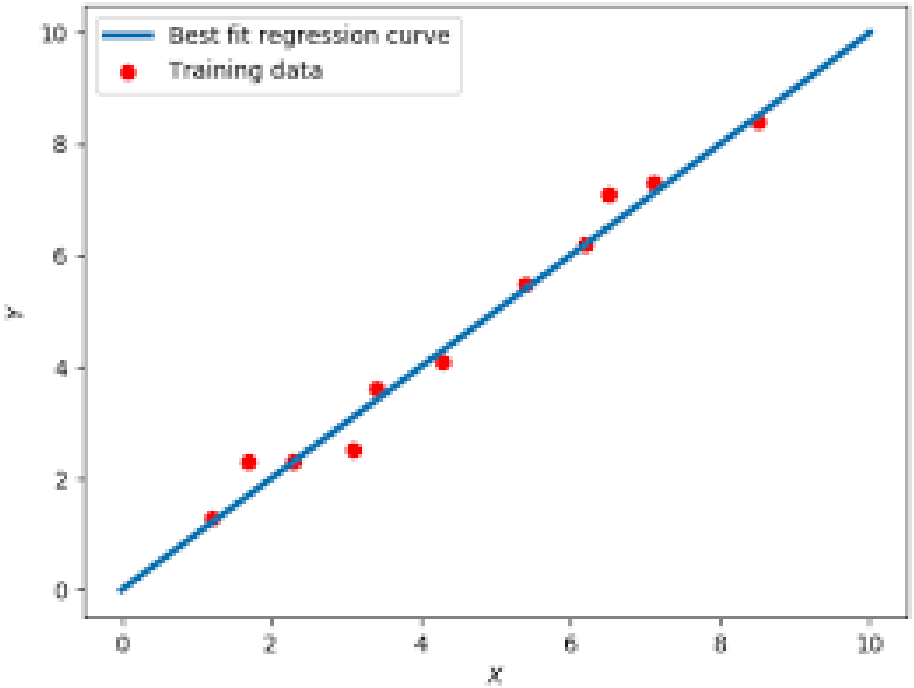


**Figure 5**: Linear Regression Classifier [41].

The logistic regression function calculates the probability of the presence of a specific feature of interest, denoted as S, using a multiple linear function. This function is defined as $Logit(S) = \beta_0 + \beta_1 M_1 + \beta_2 M_2 + \cdots + +\beta_k M_k$, where $M_1, M_2, \ldots, M_k$ represent the predictor values, and $\beta_0, \beta_k, \ldots \beta_k$ are the coefficients or weights assigned to each input variable. The weights are commonly established by a technique called maximum likelihood, which computes the likelihood of witnessing the data based on the model's parameters [42].

In practical applications, logistic regression is applied in a binary classification setting to model the posterior probability $P(y_i|x_i)$ as $\frac{1}{1+e^{(-y_i \omega^T x_i)}}$, where $x_i$ is a training feature vector labeled with $y_i$ and $\omega$ is the parameter vector determined at training time. The effectiveness of logistic regression in such applications is also dependent on its ability to handle binary classification outcomes like true positives, true negatives, false positives, and false negatives, which are crucial for evaluating the performance of the classifier through metrics like specificity or true negative rate (TNR) and sensitivity or true positive rate (TPR)[43].

*SVM*

Support Vector Machines (SVMs are a class of sophisticated machine learning models designed for classification and regression tasks. SVM is fundamentally a supervised learning model that builds on the principle of dividing data points into distinct classes by creating a hyperplane or a set of hyperplanes in a high or infinite-dimensional space. This approach is

instrumental in scenarios where the goal is to categorize data points distinctly, ensuring that different classes are separated by a clear gap that is as wide as possible[44], [45].

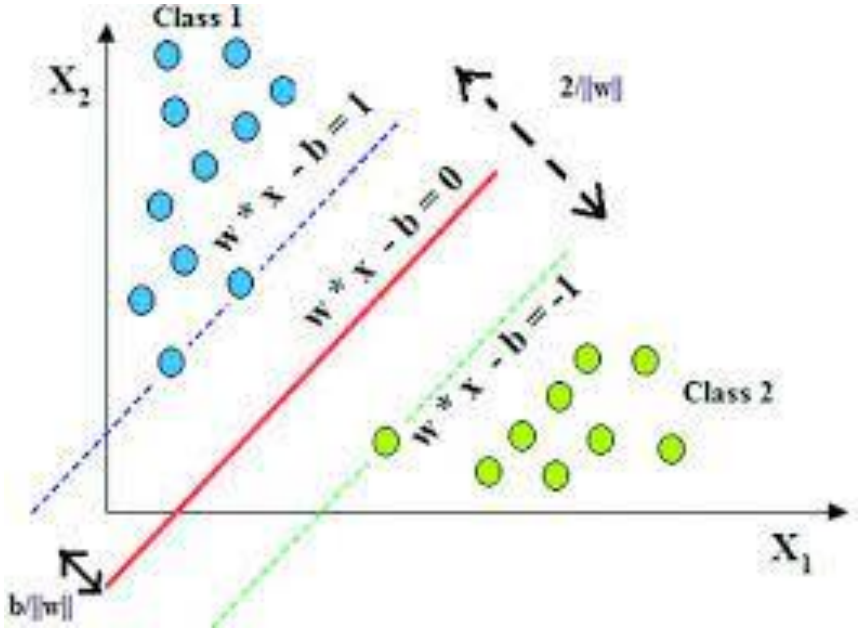


**Figure 6**: SVM Classifier [46].

One of the critical aspects of SVM is its capability to handle both linear and non-linear classification problems. In linear classification, SVM looks for the optimal hyperplane that separates the data points of one class from another with the maximum margin. For non-linear classification challenges, SVM utilizes the kernel trick, effectively transforming the input space into a higher-dimensional space where a linear separator is sufficient. This technique enables SVMs to model complex relationships between the data points and their labels. The power of SVMs lies in their ability to find the best possible boundary or decision surface that can distinguish between different classes of data points[47], [48].

SVM classifiers employ varying weighting values on the cost function to optimize classification outcomes. The adaptability and robustness of SVM are seen in its application across numerous disciplines, such as medicine and agriculture. For instance, when categorizing frequency domain features derived from heart rate variability data, SVM has shown to achieve high classification accuracies, outperforming other machine learning techniques. The adaptability of SVM in diverse applications, coupled with its high accuracy rates, underscores its significance as a powerful tool in the realm of machine learning and pattern recognition[47].

*KNN*

Because of its simplicity and high classification accuracy, the k-Nearest Neighbors (kNN) algorithm is well-known in the disciplines of data mining and statistics. This algorithm works on the principle of proximity, categorizing samples by comparing them to their nearest counterparts in the training set. The kNN classifier improves on this strategy by taking the k closest instances into account and making a classification decision based on the majority rule. The choice of the 'k' value is critical in kNN since it has a significant impact on the classifier's performance. Increasing the value of 'k' can help to reduce the influence of noise in the dataset. However, in order to attain optimal performance, cross-validation processes are frequently used to determine 'k'[49].

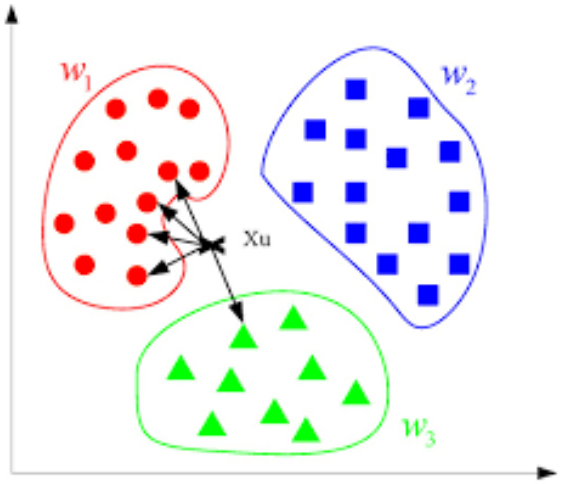


**Figure 7**: KNN Classifier [50].

There is no distinct training step in the kNN method, as there is in model-based approaches, where the model is learned using training data and then used to predict test samples. To determine the nearest neighbors, the classification procedure begins by computing the distance between the test sample and each of the training samples. The majority consensus among these nearest neighbors then determines the classification decision. The lack of a model in this approach allows for a more quick and straightforward categorization procedure, making kNN a viable alternative for a wide range of applications[51], [52].

Furthermore, the kNN algorithm can also be utilized effectively within ensemble methods. Ensemble methods, combining multiple classifiers, can significantly enhance the prediction performance, especially in the presence of non-informative features in the data sets. The ensemble approach using kNN classifiers involves selecting classifiers based on their individual performance and combining them for collective performance on a validation data set. This method leverages the simplicity of kNN while enhancing its robustness and accuracy

through the ensemble framework. The integration of kNN in ensemble methods underlines its adaptability and effectiveness in various complex classification tasks[53].

*LightGBM*

LightGBM is a sophisticated version of Gradient Boosting Decision Tree (GBDT) algorithms. LightGBM, created by Microsoft, is notable for its exceptional efficiency and efficacy in managing extensive data sets and features with high dimensions. The algorithm utilizes a histogram-based approach to expedite the training process, minimize memory usage, and enhance parallel learning, setting it apart from conventional GBDT methods[54], [55].

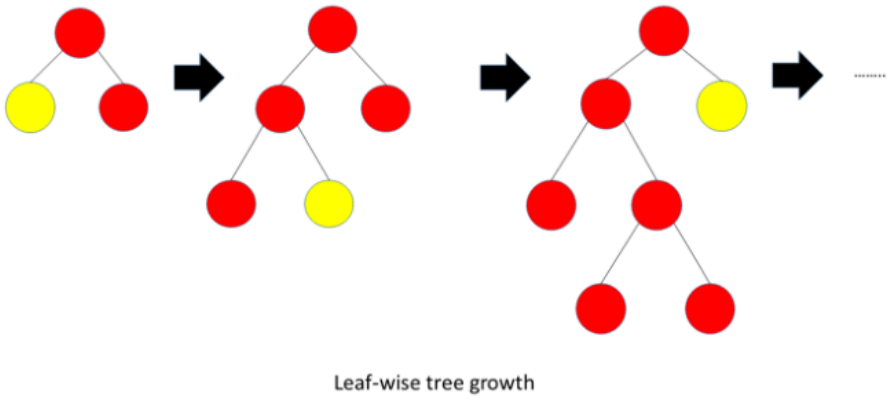


**Figure 8**: Light GBM [56].

LightGBM's basic operation comprises sharing training data over multiple computers and using a local vote mechanism to decide the most significant attributes. This is followed by a global voting process to choose the ultimate selection. The parallel voting decision tree algorithm allows for the efficient processing of large datasets. To improve computation performance while retaining accuracy, LightGBM employs gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) approaches. GOSS improves LightGBM's capacity to identify the most favorable point of division by analyzing variance gain, whereas EFB speeds up training by combining multiple unique features into a smaller number of dense features[55].

Experimental results show that LightGBM consistently outperforms other machine learning models, such as Random Forest, XGBoost, Support Vector Machine, and k-nearest neighbors, in domains such as protein-protein interaction prediction and miRNA classification in breast cancer patients. Because of its great prediction accuracy and efficient handling of

complex datasets, LightGBM is a significant tool in machine learning. It is especially beneficial for large-scale and high-dimensional data analysis[57].

*AdaBoost*

The AdaBoost classifier marks a significant milestone in ensemble learning, known formally as Adaptive Boosting. Esteemed for its straightforward yet effective approach, AdaBoost is adept at constructing a robust classifier through the amalgamation of multiple weak classifiers. This ensemble technique is recognized for its efficiency across various classification tasks[58].

In its operational framework, AdaBoost systematically trains a sequence of weak classifiers. Each classifier in this sequence is designed to focus on samples that were misclassified by its predecessors. By assigning increased weights to these challenging cases, AdaBoost ensures that succeeding classifiers give more attention to these difficult instances. This iterative approach enables the ensemble to adaptively refine its focus on the most challenging elements of the training data. The contribution of each classifier to the final model is weighted according to its accuracy, with greater emphasis placed on the more accurate classifiers during the decision-making process[59].

The notable advantage of AdaBoost is in its capacity to be applied to classification scenarios where the performance of individual classifiers is just slightly superior to random chance. The collective potency of these uncomplicated classifiers generates a composite model with noteworthy robustness. The algorithm's efficiency is attributed to its capacity to exploit the advantages of individual weak classifiers while compensating for their limitations[59].

The algorithm's versatility is further highlighted by its ability to be applied to a wide range of classification procedures, making it suitable for both binary and multi-class classification problems. The emphasis of AdaBoost on effectively addressing challenging cases during training significantly contributes to its exceptional accuracy in various real-world scenarios[60].

## 4. DATASET

The Credit Card Fraud Detection Dataset 2023 [61] significantly contributes to improving the identification of credit card fraud, namely in the creation of specialized algorithms. With a dataset of more than 550,000 transactions made by European cardholders in 2023, this dataset provides a strong foundation for analyzing transactional patterns and detecting instances of fraud.

The dataset is split to training and testing data by a ratio of 0.7. 70% of data volume is employed for model training, while 30% is utilized to test the model's performance. this is insignificant in terms of the main objective of this research. That being said, using the lowest agreed upon ratio to split the data increases the challenge of training models accurately, which is more difficult in terms of the reduced dataset.

This dataset is notable for its incorporation of 28 anonymous traits, which are identified as V1 through V28. These qualities comprise several transactional aspects, which may include the timing, geographical location, and other pertinent details. The precise characteristics of each attribute are not revealed, guaranteeing the confidentiality and anonymity of the data. The utilization of this anonymization method is crucial, as it safeguards the confidentiality and integrity of cardholder data while simultaneously offering a comprehensive dataset for thorough analysis.

Another notable characteristic is the 'Transaction Amount', which is denoted as 'Amount' in the dataset. This function quantifies the financial worth of each transaction and is essential for discerning spending patterns and trends. Comprehending these patterns is crucial for identifying abnormalities that may suggest fraudulent transactions, as spending behaviors frequently exhibit substantial variations in instances of fraud.

Additionally, the dataset contains a 'Fraud Indicator' called 'Class' for every transaction. The binary label categorizes each transaction as either fraudulent (1) or genuine (0). This categorization is very crucial for machine learning applications, particularly within the realm of supervised learning. Machine learning models are trained to distinguish between genuine and fraudulent transactions by employing these labels. This property is crucial in developing algorithms that can effectively and efficiently identify instances of fraud.

The Credit Card Fraud Detection Dataset 2023 is a valuable resource for individuals involved in research, data science, and financial security. The extensive and well-organized data, namely the anonymized characteristics, transaction amounts, and fraud indicators, serve as a solid basis for extracting data-driven insights and utilizing advanced analytical techniques to enhance the security and dependability of credit card transactions.

## 5. EVALUATION METRICS

This study investigates the effectiveness of feature selection through the utilization of Point-Biserial Correlation (PBC) in enhancing the performance of machine learning classifiers. The study incorporates many measures to evaluate the performance of the model both before and after the use of PBC-based feature selection. The collection includes many parameters such

as accuracy, precision, recall, F1-score, receiver operating characteristic area under curve (ROC AUC), specificity, and training time. Each of these metrics offers a fresh viewpoint on the model's performance, enabling a thorough evaluation of the consequences of PBC feature selection.

$$\text{Accuracy} = \frac{(True\ Positives\ +\ True\ Negatives)}{Total\ Samples} \quad (1)$$

$$\text{Precision} = \frac{True\ Positives}{True\ Positives\ +\ False\ Positives} \quad (2)$$

$$\text{Recall} = \frac{True\ Positives}{True\ Positives\ +\ False\ Negatives} \quad (3)$$

$$\text{F1 Score} = \frac{2\ *\ (Precision\ *\ Recall)}{(Precision\ +\ Recall)} \quad (4)$$

$$\text{Specificity} = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad (5)$$

The ROC AUC, or Receiver Operating Characteristic Area Under Curve, quantifies the model's capacity to differentiate between different classes at different thresholds. A greater ROC AUC value signifies superior model performance, indicating its efficacy in differentiating between fraudulent and non-fraudulent transactions[62].

Training Time (s): Training time is an operational metric that indicates the model's efficiency. In practical applications, a shorter training time is often advantageous, particularly when models need to be retrained frequently with new data.

## 6. RESULTS

In this research, the results of various classifiers are analyzed both with and without Point-Biserial Correlation (PBC) based feature selection. The examination of these findings offers valuable perspectives on the efficacy of PBC feature selection in augmenting the performance of machine learning models employed in credit card fraud detection.

**Table1**: Results without Feature Selection PBC based

| *Classifier* | *Accuracy* | *Precision* | *Recall* | *F1-Score* | *Roc Auc* | *Specificity* | *Training Time (s)* |
|---|---|---|---|---|---|---|---|
| *Random Forest* | 0.99 | 0.99 | 1 | 0.99 | 0.99 | 0.99 | 304.34 |
| *Naïve Bayes* | 0.91 | 0.97 | 0.85 | 0.91 | 0.97 | 0.97 | 0.25 |
| *Logistic Regression* | 0.96 | 0.97 | 0.95 | 0.96 | 0.99 | 0.97 | 1.82 |
| *SVM* | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 4772.98 |
| *KNN* | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.07 |
| *LightGBM* | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 1.79 |
| *Adaboost* | 0.97 | 0.97 | 0.96 | 0.97 | 0.99 | 0.98 | 142.32 |

***Table2**: Results after Feature Selection PBC Based*

| *Classifier* | *Accuracy* | *Precision* | *Recall* | *F1-Score* | *Roc Auc* | *Specificity* | *Training Time (s)* |
|---|---|---|---|---|---|---|---|
| *Random Forest* | 0.99 | 0.99 | 1 | 0.99 | 0.99 | 0.99 | 232.93 |
| *Naïve Bayes* | 0.92 | 0.98 | 0.86 | 0.92 | 0.98 | 0.98 | 0.17 |
| *Logistic Regression* | 0.96 | 0.97 | 0.94 | 0.96 | 0.99 | 0.98 | 1.07 |
| *SVM* | 0.99 | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 | 5885.70 |
| *KNN* | 0.99 | 0.99 | 1 | 0.99 | 0.99 | 0.99 | 0.06 |
| *LightGBM* | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.95 |
| *Adaboost* | 0.96 | 0.97 | 0.95 | 0.96 | 0.99 | 0.97 | 82.44 |

The results presented in the study demonstrate a notable impact of Point-Biserial Correlation (PBC) based feature selection on various machine learning classifiers in the context of credit card fraud detection. The comparison of classifier performance metrics before and after feature selection reveals insightful trends and variances.

Since PBC based feature selection is used to reduce the dataset volume by merging highly correlated features. The computational cost of training models is generally reduced. Applying classification algorithms to a smaller subset of data should decrease the training time accordingly. However, model performance accuracy wise may also be slightly affected. This research aims to enhance the computational cost of classifier training through the application of PBC based feature selection, while objectively evaluating the model's accuracy. It is evident in the results that the proposed model, for all but one classifier, achieved significant enhancement to classifier training while preserving the predictive accuracy. A detailed description of the outcomes is discussed hereafter.

Initially, without feature selection, classifiers such as Random Forest, SVM, KNN, and LightGBM exhibited high accuracy, precision, recall, F1-Score, ROC AUC, and specificity, all hovering around 0.99. This indicates a strong ability to differentiate between fraudulent and non-fraudulent transactions. The Naïve Bayes and Adaboost classifiers, while still performing commendably, showed slightly lower metrics in comparison. Particularly noteworthy was the SVM classifier's prolonged training time of 4772.98 seconds, suggesting a computational intensity in processing the full feature set.

Upon implementing PBC-based feature selection, several classifiers experienced a marginal improvement in certain metrics. For instance, Naïve Bayes showed an increase in accuracy, precision, recall, and ROC AUC, alongside a reduction in training time, indicating enhanced efficiency and effectiveness in fraud detection with a reduced feature set. Similarly, Logistic Regression and Adaboost classifiers demonstrated a slight uptick in specificity and a

decrease in training time, implying a more efficient processing capability without significant compromise on performance.

However, some classifiers like SVM saw an increase in training time to 5885.70 seconds, despite maintaining high performance across most metrics. This suggests that while feature selection can streamline the model, the computational complexity for certain algorithms might still be substantial. Conversely, classifiers like Random Forest, KNN, and LightGBM maintained their high performance across all metrics while benefiting from reduced training times, highlighting the effectiveness of PBC-based feature selection in optimizing these models for both accuracy and efficiency.

In analyzing the efficiency gains from Point-Biserial Correlation (PBC) based feature selection across various classifiers, speedup ratios were calculated. These ratios, representing the training time ratio before and after feature selection, serve as a metric for efficiency improvement. For instance, the Random Forest classifier exhibited a 31% improvement in training efficiency, evidenced by a speedup ratio of approximately 1.31. Naïve Bayes showed a remarkable 47% increase in training speed. Logistic Regression and LightGBM demonstrated substantial efficiency gains, with improvements of 70% and 88%, respectively. However, an anomaly was observed in the SVM classifier, where efficiency decreased, as indicated by a speedup ratio of about 0.81. Additionally, the KNN and Adaboost classifiers showed improvements of 17% and 73%, respectively, highlighting the diverse impacts of PBC-based feature selection on different machine learning models.

These speedup ratios underscore the impact of PBC-based feature selection on the computational efficiency of the classifiers, with most classifiers demonstrating significant improvements in training speed, except for the SVM classifier, which showed a decrease in efficiency.

The study reveals that PBC-based feature selection generally enhances the efficiency of machine learning classifiers in credit card fraud detection, as evidenced by reduced training times, without significantly compromising their predictive accuracy and other performance metrics. This indicates that PBC-based feature selection is a valuable tool in optimizing fraud detection systems, particularly in scenarios where computational resources and time are critical factors.

## 7. CONCLUSION

The examination of Point-Biserial Correlation-based feature selection has uncovered its substantial impact on the computational efficiency and performance metrics of different

machine learning classifiers in identifying credit card fraud. The investigation indicates that classifiers such as Random Forest, SVM, KNN, and LightGBM, which initially demonstrated strong performance across all measures, can improve their efficiency by reducing training times through feature selection. Naïve Bayes and Logistic Regression demonstrated slight enhancements in certain metrics and decreased training times, suggesting improved skills in detecting fraud with a simplified set of features. Nevertheless, the SVM classifier demonstrated a rise in the duration of training, indicating that feature selection may not consistently decrease the computational complexity across various algorithms. The observed speedup ratios following feature selection provide evidence of enhanced efficiency in the majority of classifiers. This study validates that utilizing PBC-based feature selection is a valuable advantage in enhancing fraud detection systems, especially in situations when computational resources and time are crucial.

The paper's contribution is manifold. It provides an empirical assessment of the impact of PBC-based feature selection on classifier performance, thus offering insights into the potential of this method to enhance model accuracy and training efficiency. It contributes to the methodological discourse by comparing the efficiency gains across various classifiers, thus informing the selection of appropriate models in practice. Furthermore, the paper adds to the literature on machine learning in fraud detection by demonstrating the applicability of PBC in high-dimensional, imbalanced datasets. Finally, it serves as a bridge between theoretical underpinnings and practical implementations, offering guidance to practitioners in the financial industry on optimizing fraud detection systems.

## REFERENCE

D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, “Credit Card Fraud Detection - Machine Learning methods,” in *2019 18th International Symposium INFOTEH-JAHORINA, INFOTEH 2019 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., May 2019. doi: 10.1109/INFOTEH.2019.8717766.

S. Misra, V. O. Matthews, A. Adewumi, O. S. Covenant University (Ota, IEEE Nigeria Section, and Institute of Electrical and Electronics Engineers, *Proceedings of the IEEE International Conference on Computing, Networking and Informatics (ICCNI 2017) : 29-31 October, 2017, Covenant University, Canaanland, Ota, Ogun State, Nigeria*.

V. N. Dornadula and S. Geetha, “Credit Card Fraud Detection using Machine Learning Algorithms,” in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 631–641. doi: 10.1016/j.procs.2020.01.057.

R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of Biomedical Informatics*, vol. 85. Academic Press Inc., pp. 189–203, Sep. 01, 2018. doi: 10.1016/j.jbi.2018.07.014.

A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Jul. 2015, pp. 1200–1205. doi: 10.1109/MIPRO.2015.7160458.

R. Das, G. Kasieczka, and D. Shih, "Feature Selection with Distance Correlation," Nov. 2022, [Online]. Available: http://arxiv.org/abs/2212.00046

J. Li *et al.*, "Feature selection: A data perspective," *ACM Computing Surveys*, vol. 50, no. 6. Association for Computing Machinery, Dec. 01, 2017. doi: 10.1145/3136625.

D. Kornbrot, "Point Biserial Correlation," in *Wiley StatsRef: Statistics Reference Online*, Wiley, 2014. doi: 10.1002/9781118445112.stat06227.

E. Curtis, C. Comiskey, and O. Dempsey, "Importance and use of correlational research," *Nurse Res*, vol. 23, no. 6, pp. 20–25, Jul. 2016, doi: 10.7748/nr.2016.e1382.

N. J. Gogtay and U. M. Thatte, "Principles of Correlation Analysis," 2017.

M. Tanner *et al.*, "Introduction to Multivariate Analysis Analysis of Failure and Survival Data The Analysis and Interpretation of Multivariate Data for Social Scientists The Analysis of Time Series-An Introduction, Sixth Edition Bayes and Empirical Bayes Methods for Data Analysis, Second Edition Bayesian Data Analysis, Second Edition."

[B. Verhulst and M. C. Neale, "Best Practices for Binary and Ordinal Data Analyses," *Behav Genet*, vol. 51, no. 3, pp. 204–214, May 2021, doi: 10.1007/s10519-020-10031-x.

A. S. Kubasch *et al.*, "Predicting Early Relapse for Patients with Multiple Myeloma through Machine Learning," *Blood*, vol. 138, no. Supplement 1, pp. 2953–2953, Nov. 2021, doi: 10.1182/blood-2021-151195.

J. Jeong and J. Choi, "Development of AOP relevant to microplastics based on toxicity mechanisms of chemical additives using ToxCast™ and deep learning models combined approach," *Environ Int*, vol. 137, Apr. 2020, doi: 10.1016/j.envint.2020.105557.

K. Iqbal and M. S. Khan, "Email classification analysis using machine learning techniques," *Applied Computing and Informatics*, 2022, doi: 10.1108/ACI-01-2022-0012.

A. Chatterjee, M. Vallières, and J. Seuntjens, "Overlooked pitfalls in multi-class machine learning classification in radiation oncology and how to avoid them," *Physica Medica*, vol. 70, pp. 96–100, Feb. 2020, doi: 10.1016/j.ejmp.2020.01.009.

M. Chang, R. J. Dalpatadu, and A. K. Singh, "Selection of Transformations of Continuous Predictors in Logistic Regression," in *Advances in Intelligent Systems and*

*Computing*, Springer Verlag, 2018, pp. 443–447. doi: 10.1007/978-3-319-77028-4_58.

S. Subash Chandra Bose, A. Vinoth Kumar, A. Premkumar, M. Deepika, and M. Gokilavani, "Biserial targeted feature projection based radial kernel regressive deep belief neural learning for covid-19 prediction," *Soft comput*, vol. 27, no. 3, pp. 1651–1662, Feb. 2023, doi: 10.1007/s00500-022-06943-x.

S. G. Khalid, S. M. Ali, H. Liu, A. G. Qurashi, and U. Ali, "Photoplethysmography temporal marker-based machine learning classifier for anesthesia drug detection," *Med Biol Eng Comput*, vol. 60, no. 11, pp. 3057–3068, Nov. 2022, doi: 10.1007/s11517-022-02658-1.

W. Yassin *et al.*, "Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis," *Transl Psychiatry*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41398-020-00965-5.

S. Syed-Abdul *et al.*, "Artificial Intelligence based Models for Screening of Hematologic Malignancies using Cell Population Data," *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-61247-0.

Y. Ushio *et al.*, "Machine learning for morbid glomerular hypertrophy," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-23882-7.

C. Celeste *et al.*, "Ethnic disparity in diagnosing asymptomatic bacterial vaginosis using machine learning," *NPJ Digit Med*, vol. 6, no. 1, Nov. 2023, doi: 10.1038/s41746-023-00953-1.

G. Sonowal, "Detecting Phishing SMS Based on Multiple Correlation Algorithms," *SN Comput Sci*, vol. 1, no. 6, Nov. 2020, doi: 10.1007/s42979-020-00377-8.

Y. Cheng and H. Liu, "A short note on the maximal point-biserial correlation under non-normality," *Br J Math Stat Psychol*, vol. 69, no. 3, pp. 344–351, Nov. 2016, doi: 10.1111/bmsp.12075.

D. G. Bonett, "Point-biserial correlation: Interval estimation, hypothesis testing, meta-analysis, and sample size determination," *British Journal of Mathematical and Statistical Psychology*, vol. 73, no. S1, pp. 113–144, Nov. 2020, doi: 10.1111/bmsp.12189.

"USEFULNESS OF CORRELATION ANALYSIS Samithambe Senthilnathan." [Online]. Available: https://ssrn.com/abstract=3416918https://ssrn.com/abstract=34169182https://ssrn.com/abstract=3416918

D. Mustafa Abdullah, A. Mohsin Abdulazeez, and A. Bibo Sallow, "Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 141–149, May 2021, doi: 10.48161/qaj.v1n2a58.

B. Taha Chicho, A. Mohsin Abdulazeez, D. Qader Zeebaree, and D. Assad Zebari, "Machine Learning Classifiers Based Classification For IRIS Recognition," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 106–118, May 2021, doi: 10.48161/qaj.v1n2a48.

R. Rajab Asaad, "Review on Deep Learning and Neural Network Implementation for Emotions Recognition," *Qubahan Academic Journal*, vol. 1, no. 1, pp. 1–4, Feb. 2021, doi: 10.48161/qaj.v1n1a25.

A. Parmar, R. Katariya, and V. Patel, "A Review on Random Forest: An Ensemble Classifier," in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 26, Springer Science and Business Media Deutschland GmbH, 2019, pp. 758–763. doi: 10.1007/978-3-030-03146-6_86.

A. Chaudhary, S. Kolhe, and R. Kamal, "An improved random forest classifier for multi-class classification," *Information Processing in Agriculture*, vol. 3, no. 4, pp. 215–222, Dec. 2016, doi: 10.1016/j.inpa.2016.08.002.

K. I. Taher, A. M. Abdulazeez, and D. A. Zebari, "Data Mining Classification Algorithms for Analyzing Soil Data," *Asian Journal of Research in Computer Science*, pp. 17–28, May 2021, doi: 10.9734/ajrcos/2021/v8i230196.

A. A. H. Alkurdi, "Enhancing Heart Disease Diagnosis Using Machine Learning Classifiers," *Fusion: Practice and Applications*, vol. 13, no. 1, pp. 08–18, 2023, doi: 10.54216/FPA.130101.

G. Biau and E. Scornet, "A Random Forest Guided Tour," Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.05741

P. Valdiviezo-Diaz, F. Ortega, E. Cobos, and R. Lara-Cabrera, "A Collaborative Filtering Approach Based on Naïve Bayes Classifier," *IEEE Access*, vol. 7, pp. 108581–108592, 2019, doi: 10.1109/ACCESS.2019.2933048.

J. Karandikar, T. McLeay, S. Turner, and T. Schmitz, "Tool wear monitoring using naïve Bayes classifiers," *International Journal of Advanced Manufacturing Technology*, vol. 77, no. 9–12, pp. 1613–1626, Apr. 2015, doi: 10.1007/s00170-014-6560-6.

K. Chaudhuri, "Building Naive Bayes Classifier from Scratch to Perform Sentiment Analyses." 2023. Accessed: Dec. 01, 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2022/03/building-naive-bayes-classifier-from-scratch-to-perform-sentiment-analysis/

F. J. Yang, "An implementation of naive bayes classifier," in *Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018*, Institute of Electrical and Electronics Engineers Inc., Dec. 2018, pp. 301–306. doi: 10.1109/CSCI46756.2018.00065.

A. Prabhat and V. Khullar, "Sentiment classification on big data using Naïve bayes and logistic regression," in *2017 International Conference on Computer Communication and Informatics, ICCCI 2017*, Institute of Electrical and Electronics Engineers Inc., Nov. 2017. doi: 10.1109/ICCCI.2017.8117734.

P. Date and T. Potok, "Adiabatic quantum linear regression," *Sci Rep*, vol. 11, no. 1, p. 21905, Nov. 2021, doi: 10.1038/s41598-021-01445-6.

Y. Yang and M. Loog, "A Benchmark and Comparison of Active Learning for Logistic Regression," Nov. 2016, doi: 10.1016/j.patcog.2018.06.004.

L. Dong, J. Wesseloo, Y. Potvin, and X. Li, "Discrimination of Mine Seismic Events and Blasts Using the Fisher Classifier, Naive Bayesian Classifier and Logistic

Regression," *Rock Mech Rock Eng*, vol. 49, no. 1, pp. 183–211, Jan. 2016, doi: 10.1007/s00603-015-0733-y.

S. Kumar, S. Mishra, P. Khanna, and Pragya, "Precision Sugarcane Monitoring Using SVM Classifier," in *Procedia Computer Science*, Elsevier B.V., 2017, pp. 881–887. doi: 10.1016/j.procs.2017.11.450.

C. Venkatesan, P. Karthigaikumar, A. Paul, S. Satheeskumaran, and R. Kumar, "ECG Signal Preprocessing and SVM Classifier-Based Abnormality Detection in Remote Healthcare Applications," *IEEE Access*, vol. 6, pp. 9767–9773, Jan. 2018, doi: 10.1109/ACCESS.2018.2794346.

A. Vinayagam *et al.*, "A random subspace ensemble classification model for discrimination of power quality events in solar PV microgrid power network," *PLoS One*, vol. 17, no. 1, p. e0262570, Jan. 2022, doi: 10.1371/journal.pone.0262570.

A. S. Manek, P. D. Shenoy, M. C. Mohan, and K. R. Venugopal, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier," *World Wide Web*, vol. 20, no. 2, pp. 135–154, Mar. 2017, doi: 10.1007/s11280-015-0381-x.

D. Mustafa Abdullah and A. Mohsin Abdulazeez, "Machine Learning Applications based on SVM Classification A Review," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 81–90, Apr. 2021, doi: 10.48161/qaj.v1n2a50.

A. Murugan, S. A. H. Nair, and K. P. S. Kumar, "Detection of Skin Cancer Using SVM, Random Forest and kNN Classifiers," *J Med Syst*, vol. 43, no. 8, Aug. 2019, doi: 10.1007/s10916-019-1400-8.

G.-F. Fan, Y.-H. Guo, J.-M. Zheng, and W.-C. Hong, "Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting," *Energies (Basel)*, vol. 12, no. 5, p. 916, Mar. 2019, doi: 10.3390/en12050916.

S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans Neural Netw Learn Syst*, vol. 29, no. 5, pp. 1774–1785, May 2018, doi: 10.1109/TNNLS.2017.2673241.

H. Rashid Abdulqadir, A. Mohsin Abdulazeez, and D. Assad Zebari, "Data Mining Classification Techniques for Diabetes Prediction," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 125–133, May 2021, doi: 10.48161/qaj.v1n2a55.

A. Gul *et al.*, "Ensemble of a subset of kNN classifiers," *Adv Data Anal Classif*, vol. 12, no. 4, pp. 827–840, Jan. 2018, doi: 10.1007/s11634-015-0227-5.

C. Chen, Q. Zhang, Q. Ma, and B. Yu, "LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion," 2019.

D. Ge, J. Gu, S. Chang, and J. H. Cai, "Credit card fraud detection using lightgbm model," in *Proceedings - 2020 International Conference on E-Commerce and Internet Technology, ECIT 2020*, Institute of Electrical and Electronics Engineers Inc., Apr. 2020, pp. 232–236. doi: 10.1109/ECIT50008.2020.00060.

A. bin Asad, R. Mansur, S. Zawad, N. Evan, and M. I. Hossain, "Analysis of Malware Prediction Based on Infection Rate Using Machine Learning Techniques," in *2020*

*IEEE Region 10 Symposium (TENSYMP)*, IEEE, 2020, pp. 706–709. doi: 10.1109/TENSYMP50017.2020.9230624.

D. Wang, Y. Zhang, and Y. Zhao, "LightGBM: An effective miRNA classification method in breast cancer patients," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Oct. 2017, pp. 7–11. doi: 10.1145/3155077.3155079.

A. Subasi *et al.*, "Sensor based human activity recognition using adaboost ensemble classifier," in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 104–111. doi: 10.1016/j.procs.2018.10.298.

Y. Zhao, L. Gong, B. Zhou, Y. Huang, and C. Liu, "Detecting tomatoes in greenhouse scenes by combining AdaBoost classifier and colour analysis," *Biosyst Eng*, vol. 148, pp. 127–137, Aug. 2016, doi: 10.1016/j.biosystemseng.2016.05.001.

Shram Sadhana Bombay Trust College of Engineering and Technology, IEEE Computer Society, Institute of Electrical and Electronics Engineers. Bombay Section, and Institute of Electrical and Electronics Engineers., *ICGTSPICC 2016 : International Conference on Global Trends in Signal Processing, Information Computing and Communication : proceedings : 22-24 December 2016, Jalgaon, Maharashtra, India*.

N. ELGIRIYEWITHANA, "Credit Card Fraud Detection Dataset 2023." Sep. 2023. Accessed: Oct. 01, 2023. [Online]. Available: https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023

N. Asaad Zebari, A. A. H. Alkurdi, R. B. Marqas, and M. Shamal Salih, "Enhancing Brain Tumor Classification with Data Augmentation and DenseNet121," *Academic Journal of Nawroz University*, vol. 12, no. 4, pp. 323–334, Oct. 2023, doi: 10.25007/ajnu.v12n4a1985.