Celal Bayar University Journal of Science

Revealing the Tendency of Analytical Business Enterprises Toward Currency and Prices Via the Naive Bayes Algorithm

Esin Avcı[1*]

[1] Giresun University, Department of Statistics, 28200 Giresun, Türkiye.
* esinavci@hotmail.com
* Orcid No: 0000-0002-9173-0142

Received: January 17, 2025
Accepted: June 6, 2025
DOI: 10.18466/cbayarfbe.1622165

Abstract

Machine learning is a cornerstone of data science, enabling the analysis and prediction of complex data patterns. Among various algorithms, Naive Bayes is a popular probabilistic classifier based on Bayes' theorem, assuming strong independence among features. Its efficiency in handling large datasets, coupled with ease of implementation, makes it a valuable tool in data science workflows. The purpose of this study is to illustrate the research and development trends of Turkish business enterprises based on unit of measure (national currency or US dollars) and price basis measurements by the Naive Bayes algorithm. The result of classifying shows that economic activities show varied preferences for currency and price bases: agriculture, forestry, and fishing, construction, and sewerage, waste management, and remediation activities are split equally between national and US dollar currencies and price types. Manufacturing favors current prices; mining and quarrying prefer national currency and current prices, while services lean towards the US dollar and current prices. Business enterprises with all economic activities prioritize the national currency (67%) and constant prices (67%).

Keywords: Machine learning, Naive Bayes, analytical business enterprise, currency, price.

1. Introduction

Machine learning is a subset of artificial intelligence (AI) focused on developing algorithms that enable computers to learn from data and improve their performance over time without being explicitly programmed [1]. It encompasses a wide range of techniques, including supervised learning, unsupervised learning, and reinforcement learning, to tackle diverse problems such as image recognition, natural language processing, and predictive analytics [2].

One of the foundational algorithms in machine learning is the Naive Bayes classifier, a probabilistic model based on Bayes' theorem. It assumes conditional independence between features, simplifying computations while providing robust performance, especially in text classification and spam filtering tasks [3]. Despite its simplicity, Naive Bayes often achieves competitive accuracy and serves as a baseline for more complex models.

The algorithm's efficiency, interpretability, and versatility make it a cornerstone in machine learning, particularly for applications with high-dimensional data and limited computational resources.

The purpose of this study is to identify the business enterprise R&D (research and development) trend. Therefore, the following provides a quick overview of the Analytical Business Enterprise's R&D (Research and Development).

The Analytical Business Enterprise R&D (Research and Development) sector represents a critical segment of industries focused on innovation, problem-solving, and the advancement of knowledge and technologies. Classified under the International Standard Industrial Classification of All Economic Activities (ISIC) Revision 4, this domain is primarily covered under Division 72 -Scientific Research and Development [4].

This industry involves systematic creative efforts undertaken by specialized organizations to expand scientific understanding and develop new applications, processes, and technologies. Key activities include experimental research, engineering innovation, and applied science development. Enterprises in this sector

play a pivotal role in driving progress across various fields, including biotechnology, information technology, environmental sciences, and industrial engineering [5].

Analytical Business Enterprise R&D is distinguished by its emphasis on data-driven decision-making, quantitative analysis, and strategic experimentation to solve real-world challenges. It serves as a backbone for industries seeking to remain competitive in a rapidly evolving global economy, fostering economic growth and societal advancement [6].

This sector's contributions extend beyond individual organizations, influencing policy, education, and collaboration across industries, positioning R&D as a cornerstone of modern enterprise innovation.

Recent empirical studies in economic data classification and business analytics have highlighted both the utility and underexplored potential of the Naive Bayes (NB) classifier. While traditionally considered a baseline model, NB has proven effective in applications involving high-dimensional or categorical data common in domains like credit scoring, customer churn prediction, labor market segmentation, and financial anomaly detection. For instance, Uludağ and Gürsoy (2020) applied NB and KNN algorithms to analyze the financial situations of manufacturing firms using Altman Z-Score parameters. NB achieved success rates between 75% and 86%, demonstrating its utility in financial health assessments [7], Aker and Karavardar (2023) applied NB alongside other machine learning models to predict financial distress in Turkish SMEs. Notably, NB outperformed other models three years prior to financial distress, achieving an accuracy of 92.86%, highlighting its strength in early warning systems [8]. Despite not always achieving the highest accuracy, NB stands out for its speed, interpretability, and robustness in handling imbalanced datasets. However, its use is often limited to benchmark comparisons, with little adaptation to domain-specific needs such as incorporating expert priors or handling ordinal economic variables. Given the growing demand for interpretable and computationally efficient models in real-time decision-making and regulatory contexts, there is a clear research gap and a novel opportunity to revisit Naive Bayes—either through domain-tailored enhancements or as part of hybrid approaches—to support economic intelligence and business analytics more effectively.

This article is organized as follows: Section 2 briefly describes the Machine learning (ML), Naive Bayes algorithm, and types of Naive Bayes Classifiers. Section 3 presents the results of applying the Naive Bayes algorithm. All mentioned analyses were applied in R software by using the "caret," and "klaR" packages. Section 4 presents the conclusion.

2. Materials and Methods

2.1. Machine learning (ML)

Within the field of artificial intelligence (AI), machine learning focuses on creating algorithms that allow computers to learn from data and gradually enhance their performance without explicit programming. It leverages statistical and computational methods to enable algorithms to identify patterns in data and make predictions or decisions based on those patterns. This paradigm shift from rule-based programming to data-driven learning has revolutionized a wide range of fields, including healthcare, finance, transportation, and more [1].

At its core, machine learning involves developing models that generalize well to unseen data. These models are trained on datasets containing input-output pairs, where the learning process involves optimizing a mathematical function to minimize the error between predicted and actual outputs. Based on their learning approach, machine learning techniques can be broadly categorized into three types:

- ✓ *Supervised Learning:* Involves training models on labeled data where both inputs and corresponding outputs are provided. The goal is to map inputs to outputs accurately. Applications include classification (e.g., spam detection) and regression (e.g., predicting housing prices) [9].
- ✓ *Unsupervised Learning:* Focuses on discovering hidden patterns or structures in data without labeled outputs. Techniques like clustering and dimensionality reduction fall under this category, with applications such as customer segmentation and data visualization [10].
- ✓ *Reinforcement Learning:* Involves training an agent to make sequential decisions by interacting with an environment and receiving feedback in the form of rewards or penalties. This approach is widely used in robotics, game playing, and autonomous systems [11].

Machine learning models can also vary based on the algorithms used, ranging from traditional methods like linear regression, decision trees, and support vector machines to advanced techniques like neural networks and ensemble methods. The recent surge in computational power and availability of large datasets has fueled the rise of deep learning, a specialized branch of machine learning that uses neural networks with many layers to solve complex problems such as image recognition, natural language processing, and autonomous driving [12].

However, challenges remain in the field, including issues of data quality, model interpretability, and ethical concerns such as bias and privacy. Addressing these challenges requires careful design, rigorous evaluation, and consideration of societal implications, ensuring that machine learning technologies are both effective and equitable.

Machine learning and the Naive Bayes algorithm are closely linked, as Naive Bayes serves as a foundational technique within the broader domain of machine learning. Specifically, Naive Bayes is a probabilistic classifier that exemplifies supervised learning, a key category of machine learning where models are trained on labeled data to make predictions. It is particularly effective in scenarios involving text classification, spam detection, and sentiment analysis, due to its simplicity and efficiency in handling high-dimensional data [13]. While its assumption of conditional independence among features may not always hold in practice, the algorithm's ease of implementation and scalability make it a valuable tool for learning tasks where quick and interpretable solutions are needed. Furthermore, Naive Bayes illustrates the interplay between statistical principles and machine learning, showcasing how probability theory can inform model design [14].

2.1.1. Naive Bayes Algorithm

The Naive Bayes algorithm is a family of simple yet effective probabilistic classifiers based on Bayes' Theorem. It operates under the assumption of conditional independence, meaning it assumes that the features used to predict the target variable are independent of each other, given the class label. Despite the unrealistic nature of this assumption in many real-world scenarios, the algorithm performs remarkably well for various classification tasks, including spam detection, sentiment analysis, and document categorization.
Bayes' Theorem forms the foundation of the Naive Bayes algorithm, expressed mathematically as:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad \textbf{(2.1)}$$

Here:

- ✓ $P(C|X)$ is the posterior probability of class C given the input features X.
- ✓ $P(X|C)$ is the likelihood of the features given the class.
- ✓ $P(C)$ is the prior probability of the class.
- ✓ $P(X)$ is the evidence or marginal probability of the features.

The algorithm is "naive" because it simplifies the computation of $P(X|C)$by assuming that the features are conditionally independent, allowing $P(X|C)$ to be expressed as the product of individual feature probabilities. For example:

$$P(X|C) = \prod_{i=1}^{n} P(x_i|C) \quad \textbf{(2.2)}$$

where x_i represents the i-th feature in X. This assumption significantly reduces the computational complexity and makes the algorithm scalable even for high-dimensional datasets.

Naive Bayes classifiers are particularly advantageous when working with large datasets or when quick training is required. Despite its simplicity, the algorithm has demonstrated competitive performance, especially in text classification and natural language processing tasks. However, it can struggle in scenarios where the independence assumption is strongly violated or when the dataset contains highly correlated features [13].

Variants of the Naive Bayes algorithm, such as Gaussian, Multinomial, and Bernoulli Naive Bayes, are tailored for specific types of data. For example, Gaussian Naive Bayes is commonly used for continuous data, while Multinomial and Bernoulli Naive Bayes are more suited for discrete or binary data, such as text frequency counts or Boolean feature representations [15].

By combining computational efficiency with effective performance on a wide range of tasks, Naive Bayes remains a cornerstone algorithm in the field of machine learning. Its simplicity also makes it an excellent choice for introducing probabilistic models to newcomers in data science.

2.1.2. Types of Naive Bayes Classifiers

Naive Bayes classifiers come in several variations, each tailored to specific types of data and problem domains. These variants share the common foundation of Bayes' Theorem and the assumption of conditional independence among features but differ in how they model the likelihood $P(X|C)$of the data. The most commonly used types are:

a. Gaussian Naive Bayes:

This variant assumes that the features follow a Gaussian (normal) distribution. It is primarily used for continuous data, where the likelihood of each feature is modeled using the Gaussian probability density function:

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \quad \textbf{(2.3)}$$

Here, μ and σ^2 are the mean and variance of the feature values for each class. Gaussian Naive Bayes is widely applied in scenarios such as medical diagnosis and anomaly detection [14].

b. Multinomial Naive Bayes:

Designed for discrete data, Multinomial Naive Bayes is commonly used for text classification tasks where features represent frequencies or counts, such as word occurrences in a document. The likelihood is calculated based on the frequency distribution of features in each class. This approach has proven effective for problems like spam filtering and document categorization [15].

c. Bernoulli Naive Bayes:

Similar to Multinomial Naive Bayes, this variant is suited for binary or Boolean data, where features take on only two possible values (e.g., presence or absence of a word in a text). Bernoulli Naive Bayes is particularly useful in applications like sentiment analysis and binary text classification tasks. The key distinction between this and the Multinomial model is how zero counts are treated; Bernoulli explicitly accounts for the absence of features [13].

d. Complement Naive Bayes:

This modification of the Multinomial Naive Bayes algorithm is designed to address imbalances in the class distributions, especially in text classification. Complement Naive Bayes calculates the likelihood based on the complement of each class, thereby reducing bias toward the majority class. It is particularly beneficial in scenarios with highly imbalanced datasets [16].

e. Categorical Naive Bayes:

This version is suitable for categorical data, where features represent discrete categories rather than numerical values. It calculates probabilities based on the relative frequency of categories within each class, making it ideal for tasks involving structured categorical datasets [14].

Each type of Naive Bayes classifier is tailored to different data distributions and feature representations, making the algorithm versatile and applicable across a wide range of domains. Despite its simplicity, the method's effectiveness, particularly in text classification and high-dimensional spaces, continues to make it a cornerstone in the machine learning toolkit.

3. Results and Discussion

This study discusses the classifying unit of measure (national currency or US dollars, PPP converted) and price base (constant prices or current prices) of various economic activity groups (agriculture, forestry, and fishing (AFF); construction; sewerage, waste management, and remediation activities (EGSA); manufacturing; mining and quarrying; services; and total-all activities) in Turkey in 2022. 208 business enterprise R&D data by industry is considered.

The dataset used in this study is obtained from ISIC Rev.4 industry-based Analytical Business Enterprise R&D (ANBERD database) https://data-explorer.oecd.org . The OECD's Analytical Business Enterprise Research and Development (ANBERD) database was created to give analysts thorough information on business R&D spending. It displays annual data on industry R&D expenditures. Numerous estimates that expand and supplement national submissions of business enterprise R&D data by industry (primary activity/industry orientation) are included in the ANBERD database.

Table 1 summarizes the economic activities, unit of measure, and price base variables considered.

Table 1. Descriptive statistics of variables

Variables	Category	N (%)
Economic Activity	Agriculture, Forestry, and Fishing (AFF)	4 (2)
	Construction	4 (2)
	Sewerage, Waste Management, And Remediation Activities (EGSA)	4 (2)
	Manufacturing	104 (50)
	Mining and Quarrying	4 (2)
	Services	84 (40)
	Total-all activities	4 (2)
Unit of measure	National currency	104 (50)
	US dollars, PPP converted	104 (50)
Price base	Constant prices	104 (50)
	Current prices	104 (50)

Business enterprises whose R&D data was reported in 2022 in Turkey suggested that most of the economic activities were in manufacturing (50%), as a measure of units these business enterprises use half by half the national currency and US dollars and at the same rate constant and current prices (Table 1).

Other machine learning algorithms like K-Nearest Neighbors (KNN) are not ideal for purely categorical data, because KNN relies on distance metrics (like Euclidean or Manhattan) that assume numeric inputs. These metrics don't make sense for nominal categories without artificial transformations. On the other hand, decision trees and random forests are generally ideal for purely categorical data because they naturally process categorical features by splitting them according to feature values, making them quite suitable without the need for extensive preprocessing. Decision trees and random forests were applied to the data, but their results are not shown because their accuracies did not exceed 50%. Hence, only the Naive Bayes algorithm was used to analyze the data.

The Naive Bayes algorithm was used to classify the economic activities considered on the basis of unit and price. The analyses were carried out with the "caret," and "klaR" packages in the R program.

To use the Naive Bayes algorithm, the data was randomly divided into 75% training and 25% test data. The prior and conditional probabilities obtained from the Naive Bayes classification algorithm are summarized in Tables 2 and 3.

Table 2. Prior probabilities

Variables	Category	Probability
Economic Activity	Agriculture, Forestry, and Fishing (AFF)	0.014
	Construction	0.020
	Sewerage, Waste Management, And Remediation Activities (EGSA)	0.020
	Manufacturing	0.453
	Mining and Quarrying	0.020
	Services	0.446
	Total-all activities	0.027

Based on the prior probabilities, the probability of manufacturing economic activity was 45.3%, while the probability of services was 44.6%.

Table 3. Conditional probabilities

Variables	Category	Unit of measure		Price base	
		National currency	US dollars, PPP converted	Constant prices	Current prices
Economic Activity	**Agriculture, Forestry, and Fishing (AFF)**	0.50	0.50	0.50	0.50
	Construction	0.50	0.50	0.50	0.50
	Sewerage, Waste Management, And Remediation Activities (EGSA)	0.50	0.50	0.50	0.50
	Manufacturing	0.50	0.50	0.46	**0.54**
	Mining and Quarrying	**0.67**	0.33	0.33	**0.67**
	Services	0.46	**0.54**	0.44	**0.56**
	Total-all activities	**0.67**	0.33	**0.67**	0.33

AFF, construction, and EGSA economic activities prefer both national and US dollar currencies equally in terms of unit of measure. A similar situation applies to the price base. While manufacturing business enterprises prefer the current price (54%) to constant prices, they use national and US dollars at the same rate as a unit of measure. Mining and quarrying activity prefers national currency over US dollars, and it also prefers current price over the constant price (67%). Services activity prefers US dollars (54%) over national currency, and it also prefers current price (56%) over the constant price. Business enterprises that have total-all activity prefer national currency to the US dollar and constant price to the current price at the same rate (67%).

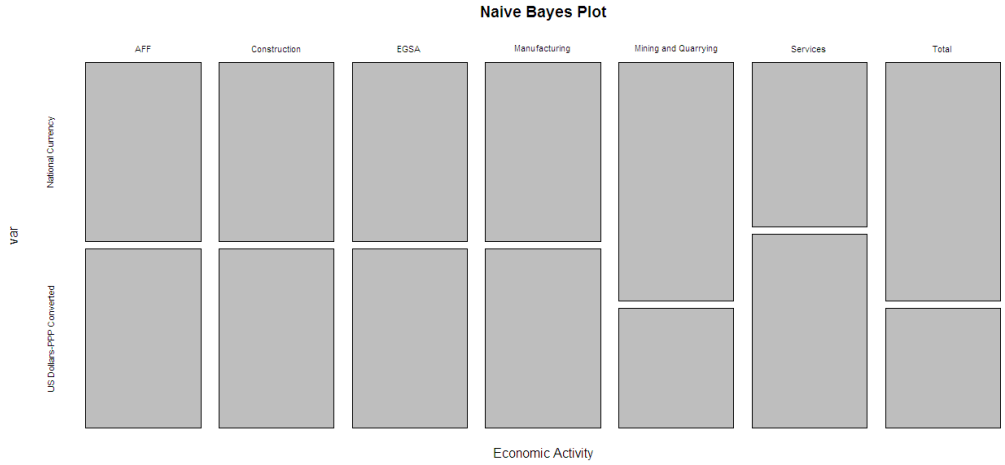
Based on the data in Table 3, the services sector shows a higher preference for reporting in U.S. dollars (54%) and current prices (56%), in contrast to sectors like agriculture or construction, which exhibit an even split across units and price bases. This pattern in services may reflect several economic and policy-related factors.

First, services, especially internationally traded ones like finance, tourism, or IT, are more exposed to global markets, where transactions are often denominated in USD. Using USD in statistical reporting may therefore provide a more stable and globally comparable measure, especially in countries with volatile exchange rates or high inflation. The preference for current prices (as opposed to constant prices) further suggests sensitivity to recent price movements, making these figures more reflective of nominal market values than inflation-adjusted real growth.

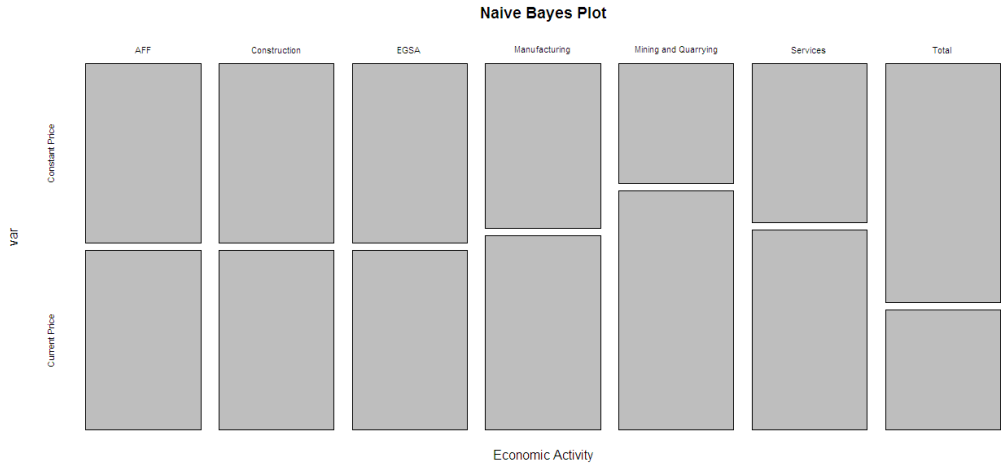
This pattern aligns with how services operate: they tend to be less capital-intensive than goods-producing sectors and more price-sensitive in the short run, so their reported values are more likely to fluctuate with policy shifts (e.g., currency devaluation, tax changes, or deregulation) and inflationary pressures. Hence, the combined use of USD and current prices in services reporting may offer a more practical and timely view of the sector's value and international competitiveness in dynamic economic environments.

Naive Bayes plots that display the marginal probability of predictor variables according to the category are displayed in Figure 1.

Celal Bayar University Journal of Science
Volume 21, Issue 4, 2025, p 62-69
Doi: 10.18466/cbayarfbe.1622165



(a)



(b)

Figure 1. Naive Bayes plots: (a) Unit of measure. (b) Price base

The size of the boxes in each category of economic activity that are considered is proportional to the marginal probability (Figure 1).

The model achieves an accuracy of 65%, meaning it correctly classifies 65 out of every 100 examples overall. While this suggests moderate performance, it's important to look beyond accuracy, especially in cases where class imbalance may be present. The precision of 0.68 indicates that when the model predicts a positive class, it is correct about 68% of the time, suggesting a relatively low false positive rate. The recall of 0.63 shows that the model is able to identify approximately 63% of all actual positive cases, meaning it misses about 37% of them. The F1 score of 0.66, which is the harmonic mean of precision and recall, reflects a balanced trade-off between identifying positive cases accurately and not producing too many false positives. Overall, while the accuracy is somewhat modest, the solid precision and recall suggest the model is reasonably good at identifying positives without being too prone to error (Table 4).

Table 4. Evaluation metrics

Recall	Precision	F1 score	Accuracy
0.63	0.68	0.66	0.65

4. Conclusion

Machine learning has revolutionized data-driven decision-making by enabling systems to learn patterns and make predictions without explicit programming. Among its algorithms, the Naive Bayes classifier stands out for its simplicity, efficiency, and effectiveness, particularly in tasks like text classification and spam filtering. Naive Bayes works very well in real life, even though it is based on the idea that features are independent. It is a reliable and computationally efficient way to deal with large datasets and high-dimensional data. Its balance of simplicity and power makes it a fundamental tool in the machine learning toolkit.

The trend of several economic activities is examined in this study using unit and price-based metrics. Using the Naive Bayes technique, this trend was exposed.

Preferences for currency and price base usage vary significantly across different economic activities. AFF, construction, and EGSA activities exhibit equal preferences for national and US dollar currencies, as well as for current and constant prices. Manufacturing enterprises predominantly favor the current price while maintaining an equal preference for national and US dollar currencies. Mining and quarrying activities demonstrate a clear preference for the national currency and the current price, while services activities lean towards the US dollar and the current price. Conversely, enterprises engaged in total-all activities prioritize the national currency over the US dollar and prefer the constant price over the current price, both at a notable rate of 67%. These patterns highlight the diverse economic preferences shaped by activity specific contexts.

One of the key advantages of this research lies in its methodological simplicity and computational efficiency, making it particularly suitable for categorical datasets common in economic and policy analysis. Additionally, the study provides new insights into activity specific reporting behaviors that may reflect underlying macroeconomic factors such as inflation sensitivity, exchange rate volatility, and global market exposure. However, the model's predictive accuracy of 65% and its reliance on the assumption of feature independence represent notable limitations, especially given the potential interdependencies among economic indicators. Moreover, the relatively small sample size for some activity groups may limit generalizability. Future research could improve classification performance by employing more reported factors. Furthermore, exploring the integration of expert informed priors or domain knowledge into the Naive Bayes framework could enhance its applicability for economic intelligence and policy making.

Author's Contributions

Esin Avcı: Analyzed and interpreted the findings in addition to drafting and writing the manuscript.

Ethical and Informed Consent for Data Used

Ethical approval and informed consent for data use were not required for this research, as the study did not involve human subjects, personal data, or sensitive information. The data utilized were obtained from publicly available and anonymized sources, and all aspects of the research adhered to ethical standards and legal requirements.

References

[1]. Mitchell, TM. Machine Learning. McGraw Hill , Maid enhead, United Kingdom, international student edition. 1997; pp. 1-13.

[2]. Russell, S, Norvig, P. Artificial Intelligence: A Modern Approach (4th ed.). Pearson Education Limited, United Kingdom. 2020; pp. 853-882

[3]. Manning, CD, Raghavan, P, Schütze, H. Introduction to Information Retrieval. Cambridge: Cambridge University Press. 2008; pp. 80-274.

[4]. United Nations. International Standard Industrial Classification of All Economic Activities (ISIC), Rev.4. Retrieved from https://unstats.un.org. (accessed at 25.12.2024).

[5]. OECD. Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development. OECD Publishing. Retrieved from https://www.oecd.org. (accessed at 25.12.2024).

[6]. World Economic Forum. The Future of Jobs Report 2020. Retrieved from https://www.weforum.org. (accessed at 25.12.2024).

[7]. Uludağ, O., Gürsoy, A. 2020. On the financial situation analysis with knn and Naive Bayes classification algorithms. *Journal of the Institute of Science and Technology*; 10(4): 2881-2888. (https://doi.org/10.21597/jist.703004)

[8]. Aker, Y., Karavardar, A. 2023. Using machine learning methods in financial distress prediction: sample of small and medium sized enterprises operating in Turkey. *Ege Academic Review*; 23(2): 145-162. (https://doi.org/10.21121/eab.1027084)

[9]. Bishop, CM. Pattern Recognition and Machine Learning. Springer. Berlin. 2006. pp. 20-30.

[10]. Hastie, T, Tibshirani, R, Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. California. 2009. pp. 485

[11]. Sutton, RS., & Barto, AG. Reinforcement Learning: An Introduction (2nd ed.). MIT Press, Cambridge, Massachusetts. 2018. pp. 1-25.

[12]. LeCun, Y, Bengio, Y, Hinton, G.2015. Deep learning. *Nature*; 521(7553): 436-444.

[13]. Zhang, H. The Optimality of Naive Bayes. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, USA, 2004, pp 312–317.

[14]. Murphy, KP. Machine Learning: A Probabilistic Perspective. MIT Press. Cambridge, Massachusetts. 2012. pp. 1-31.

[15]. McCallum, A, & Nigam, KA. Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on Learning for Text Categorization, Madison, Wisconsin, 1998, pp 41-48.

[16]. Rennie, JDM, Shih, L, Teevan, J, & Karger, DR. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington, DC USA 2003, pp 616–623.