DOI: 10.15849/IJASCA.250730.07

Int. J. Advance Soft Compu. Appl, Vol. 17, No. 2, July 2025
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Identifying Optimal Data Distributions for Enhanced Data Modeling in Machine Learning

Yousef Jaradat[1], Mohammad Masoud[1], Ahmad Manasrah[2], Mohammad Alia[3], Khaled M. Suwais[4], and Sally Almanasra[4]

[1] Electrical Engineering, Al-Zaytoonah University of Jordan, Amman, Jordan
{y.jaradat, m.zakaria}@zuj.edu.jo
[2] Mechanical Engineering, Al-Zaytoonah University of Jordan, Amman, Jordan
ahmad.mansrah@ zuj.edu.jo
[3] Cybersecurity Department, Al-Zaytoonah University of Jordan, Amman, Jordan
dr.m.alia@zuj.edu.jo
[4] Faculty of Computer Studies, Arab Open University, Saudi Arabia
{khaled.suwais, s.almanasra }@arabou.edu.sa

Abstract

Understanding how data is distributed is crucial for building accurate models in machine learning and data science projects. In this paper, we explore practical methods to help identify the best-fitting distribution for real-world datasets. We cover visual techniques like histograms and Q-Q plots, as well as statistical tests such as Kolmogorov-Smirnov (KS) and Anderson-Darling (AD). We also look at model evaluation using criteria like Akaike information criterion (AIC) and Bayesian information criterion (BIC) to ensure a good fit. To illustrate these methods, we use the California Housing dataset, showing how wrong assumptions about data distribution can lead to poor model performance. By following the guidelines provided in this paper, data scientists can choose the right distribution, leading to more accurate models, better anomaly detection, and smarter decision-making across different fields.

Keywords: *Density Distribution, Data Modeling, Visual inspection, Statistical tests.*

1 Introduction

In the continuously evolving field of data science, the importance of data cannot be emphasized enough; it is the vital force that drives the development of meaningful discoveries and revolutionary innovations. However, in its original state, data can be a complex and challenging entity, filled with intricate details. This is where the practice of data modeling becomes crucial, allowing us to organize and structure this unrefined digital landscape. Data modeling entails creating a visual representation of an entire information system or its components to demonstrate the connections between data points and structures. This representation may include diagrams, symbols, and written elements to

Received 13 April 2025; Accepted 16 June 2025

illustrate the data and its relationships. In the context of machine learning and data science, data modeling involves developing a model that captures the inherent patterns and connections within a dataset, [1,2]. However, Data modeling is essential for several reasons[3, 4]:

- Enhanced comprehension of data: Data models offer a precise and succinct means of comprehending intricate data configurations. Through visual depiction of entities, attributes, and relationships, data models simplify the comprehension of data significance and context.
- Efficient Communication: Data models help to enable communication among various parties involved, including business analysts, data engineers, and software developers. Establishing a common comprehension of the data model helps to ensure alignment in efforts and minimize misinterpretations.
- Improved Data Accuracy: Data models play a crucial role in identifying and resolving data discrepancies, duplications, and mistakes. Through the establishment of data types, limitations, and validation regulations, data models uphold data reliability and enhance data excellence.
- An effective structure for databases: Data models play a crucial role in the design of efficient databases. Through meticulous planning of data structure and relationships, data models contribute to enhancing database performance and reducing storage needs.
- Efficient Data Incorporation: Data models facilitate the integration of data from various sources by establishing a standardized data structure and mapping data elements across disparate systems. This streamlines the overall data integration process.
- Making well-informed decisions: Thoughtfully constructed data models offer valuable perspectives into business operations and facilitate evidence-based decision-making. Through assessing the connections among data components, data models can uncover concealed patterns and trends that can guide strategic decisions.

Data distribution refers to the dispersion or arrangement of values within a dataset. This analogy likens it to a topography, with peaks and troughs indicating the frequency of specific values. In more formal language, data distribution is commonly depicted using a probability density function (PDF) to illustrate the probability of encountering various values [5,6]. Understanding the distribution of data is the pivotal point on which the balance of success and failure hinges. It is the key that opens the door to meaningful insights, enabling dynamic and agile decision-making. Equipped with this knowledge, we move forward with unwavering confidence, wielding the ability to mold and reshape our data into powerful models that unravel the most complex mysteries, [7]. In fact, model and data density have an impact on the result of the modeling process within a data science project as a proxy data discovery process. Depending on the nature and objective of your project, you can sometimes be compelled to manipulate your data to obtain a different underlying distribution or even generate artificial data to reflect the model you want to use. By selecting a different model to utilize, you will also alter the objective and method for demonstrating, verifying, and explaining your modeling strategy and your data findings, [8].

In this paper, we aim to answer a key question: How can we reliably identify the best data distribution for real-world datasets to improve the accuracy of our models and decision-making? The goal is to provide practical, easy-to-follow steps for choosing the right distribution, using methods like visual inspection, statistical tests, and model evaluation. One of the big challenges in data science is that real-world data doesn't always follow a typical distribution, which makes it tricky to apply certain models effectively. If the distribution is misunderstood, it can lead to poor predictions, ineffective feature engineering, and issues with identifying anomalies. This work is important because it offers a clear approach to tackling these challenges, helping data scientists make better decisions when modeling data, leading to more accurate and reliable outcomes.

The rest of the paper is organized as follows: Section 2 overviews the role of data density distribution in data modeling; Section 3 provides the necessary steps for selecting the best density distribution of a dataset; Dealing with Non-Standard Distributions is examined in section 4; Section 5 provides a demonstration of finding density distribution with the California Housing Dataset; Conclusion is provided in section 6.

2 The Role Of Data Distribution In Data Modeling

In the field of data science, data modeling is essential for developing predictive models, creating insights, and making sound decisions. Understanding the data's underlying distribution is a critical component of data modeling. This distribution, commonly known as a probability density function (PDF), describes how frequently certain values appear in the dataset. In this paper, we'll look at why knowing data distribution is important and how it affects different parts of data modeling in data science projects, [9].

- ***Choosing the Right Model***: Selecting an appropriate statistical model is highly contingent on the data's distribution. Various distributions display unique attributes like mean, variance, and asymmetry. Specifically, in cases where the data conforms to a normal distribution, models like linear regression and Gaussian Naive Bayes are suiTABLE. Conversely, for data with heavy tails, robust models such as quantile regression or t-distributions may be more suitable. Figure 1 shows histogram plots for normal and exponential distributions, illustrating their distinct shapes.

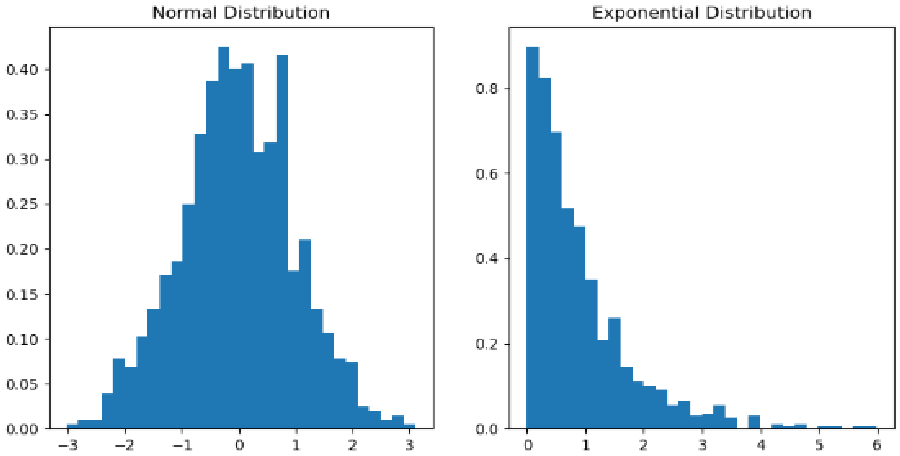

Figure 1: Histogram plots for normal and exponential distributions

- ***Feature Engineering and Transformation***: Comprehending the distribution of data plays a crucial role in informing the process of feature engineering and transformation [10]. In cases of skewed data, techniques such as taking the logarithmic function or applying the Box-Cox transformation can be utilized to normalize the distribution, thereby bolstering the performance of models that adhere to the assumption of normality. Likewise, when dealing with multimodal data, discerning the subpopulations within can result in more effective feature representation. Figure 2 shows how log transformation can normalize a skewed distribution.

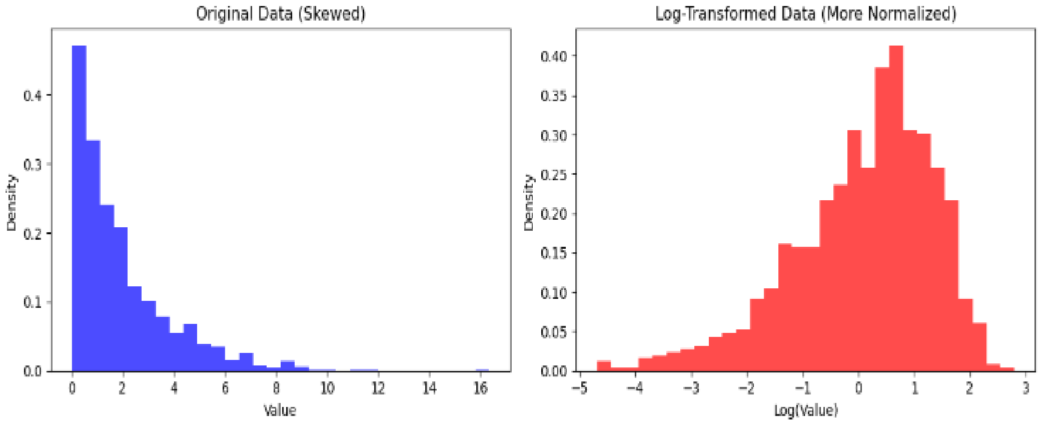

Figure 2: Normalization skewed distribution using log transformation

- ***Anomaly Detection***: Detection of anomalies depends greatly on understanding the usual patterns within the data. By creating a model of the data's distribution, we are able to set a standard for anticipated values. Any deviations from this standard can be identified as possible anomalies, which are frequently suggestive of mistakes, deception, or infrequent occurrences, [10,11]. Figure 3 shows an outlier (anomaly) of normal data.

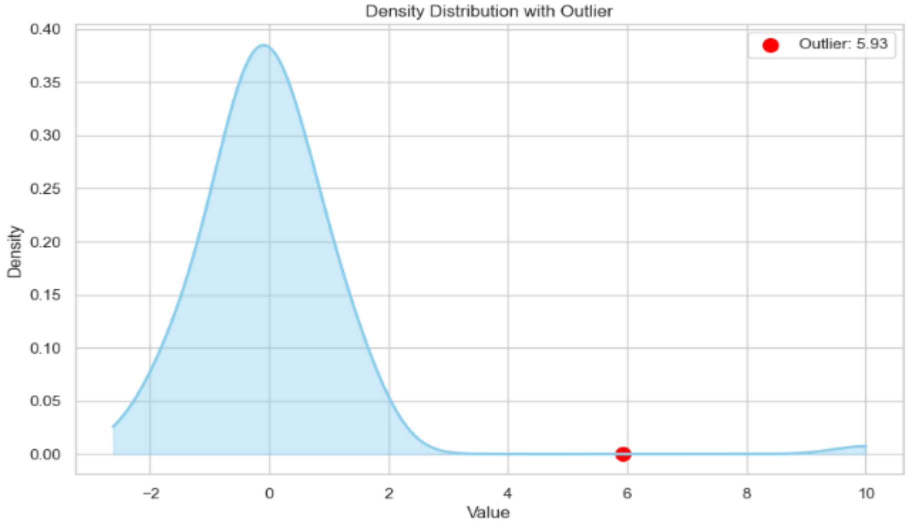

Figure 3: Normal data distribution with an outlier

- ***Data Generation and Simulation***: In situations where there is a lack of data, creating artificial data that replicates the original distribution is highly beneficial. This allows for the training and testing of models in a variety of circumstances, ultimately improving their resilience. Figure 4 depicts the generation of data from a known distribution.

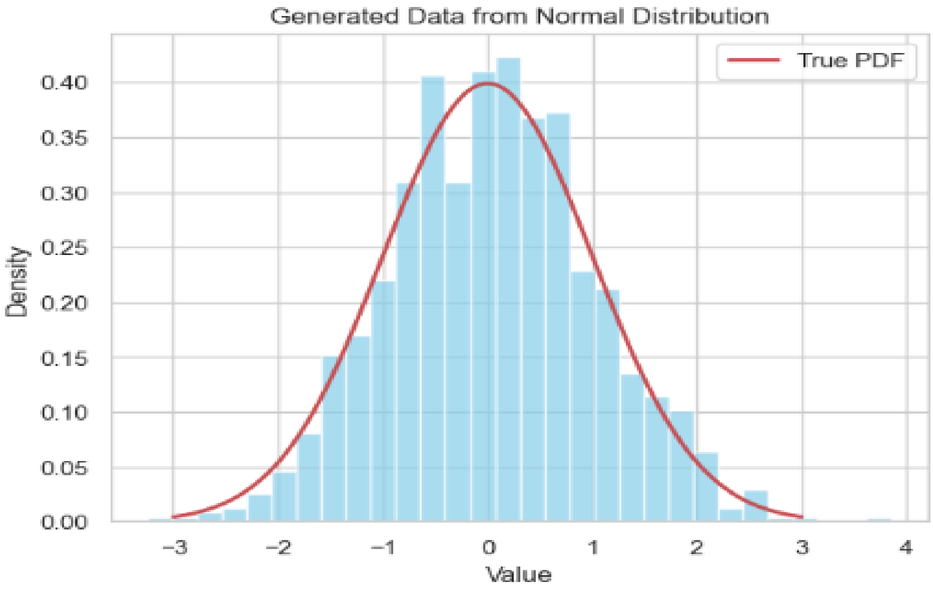

Figure 4: Generation of normally distributed data

- ***Probability Estimation and Decision Making:*** Understanding the dispersion of data enables us to estimate the likelihood of particular occurrences. In risk assessment, for example, comprehending the spread of potential losses can guide risk management tactics.

3 Steps For Selecting The Best Density Distribution Of A Dataset

Selecting the appropriate distribution is an important step in data modeling. The choice has a considerable impact on the accuracy of your model, the validity of your projections, and the efficacy of statistical testing, [13,14]. The steps below show how to choose the best density distribution of a dataset.

3.1 Visual Inspection

a. ***Histograms***: Histograms offer a visual depiction of the distribution of data. By creating a histogram, we can obtain an initial understanding of the shape of the data's distribution (e.g., bell-shaped for normal distribution, right-skewed for lognormal distribution). Figure 5 shows the power of visual depiction (histogram) of data to determine the underlining density distribution of a dataset.

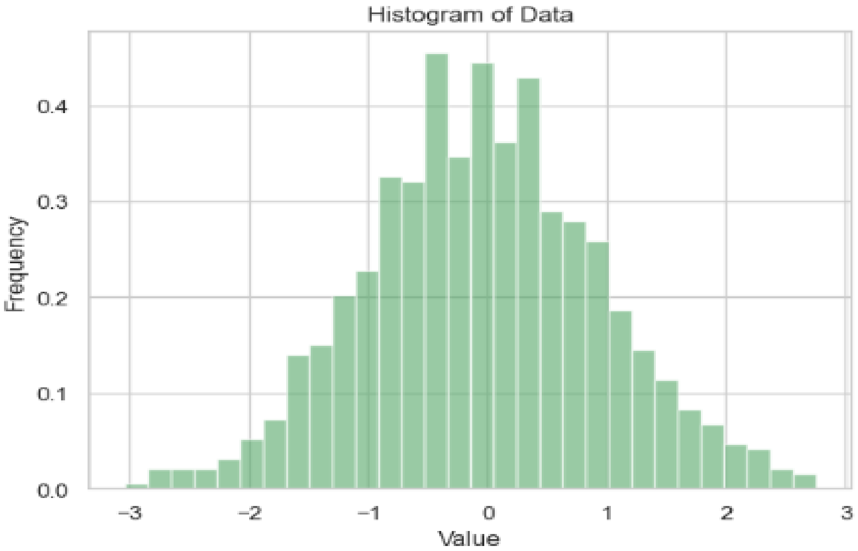

Figure 5: Visual inspection of data using histogram

b. ***Q-Q Plots (Quantile-Quantile Plots)***: Q-Q plots entail comparing the quantiles of your data with those of a theoretical distribution. When the points fall near a straight line, it indicates a favorable fit of the distribution. Figure 6 shows the power of visual depiction (Q-Q plot) of data to determine the underlining density distribution of a dataset.

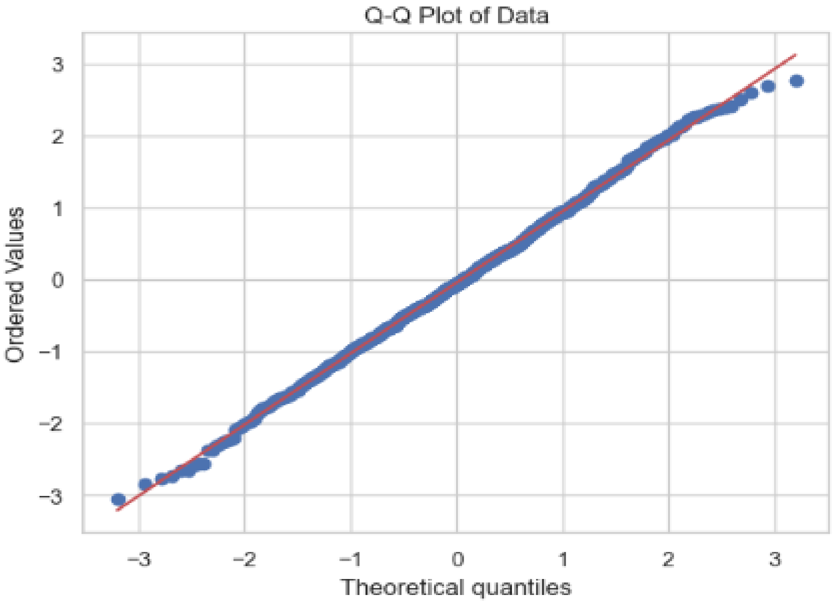

Figure 6: Visual inspection of data using Q-Q plot

3.2 Statistical Tests

a. ***Goodness-of-Fit Tests***: Goodness-of-fit tests are statistical methods used to evaluate the degree of compatibility between a given data set and a particular probability distribution. These tests play a crucial role in determining how well a specific distribution model fits the observed data. Commonly employed goodness-of-fit tests include: the renowned *Kolmogorov-Smirnov (KS) test*, which examines the largest difference between the empirical distribution function and the theoretical distribution function; *the Anderson-Darling (AD) test*, which focuses on the weighted squared difference between the observed and expected cumulative distribution functions; and *the chi-squared test*, which evaluates the deviation between the observed and expected frequencies. By subjecting your data to these tests, you can gain insights into the degree of agreement or discrepancy between the empirical data and the theoretical distribution, enabling you to make informed decisions in various fields such as finance, biology, and engineering. Figure 7 shows that the minimum (KS) statistic determines the distribution with the best fit. The distribution with the lowest KS statistic is deemed the best fit since it shows the smallest divergence between the actual and theoretical distributions. Since the data in Figure 7 was generated from a normal distribution, we expect the normal distribution to be the best fit. However, the results will confirm this by showing a lower KS statistic and a higher p-value for the normal distribution compared to the others.

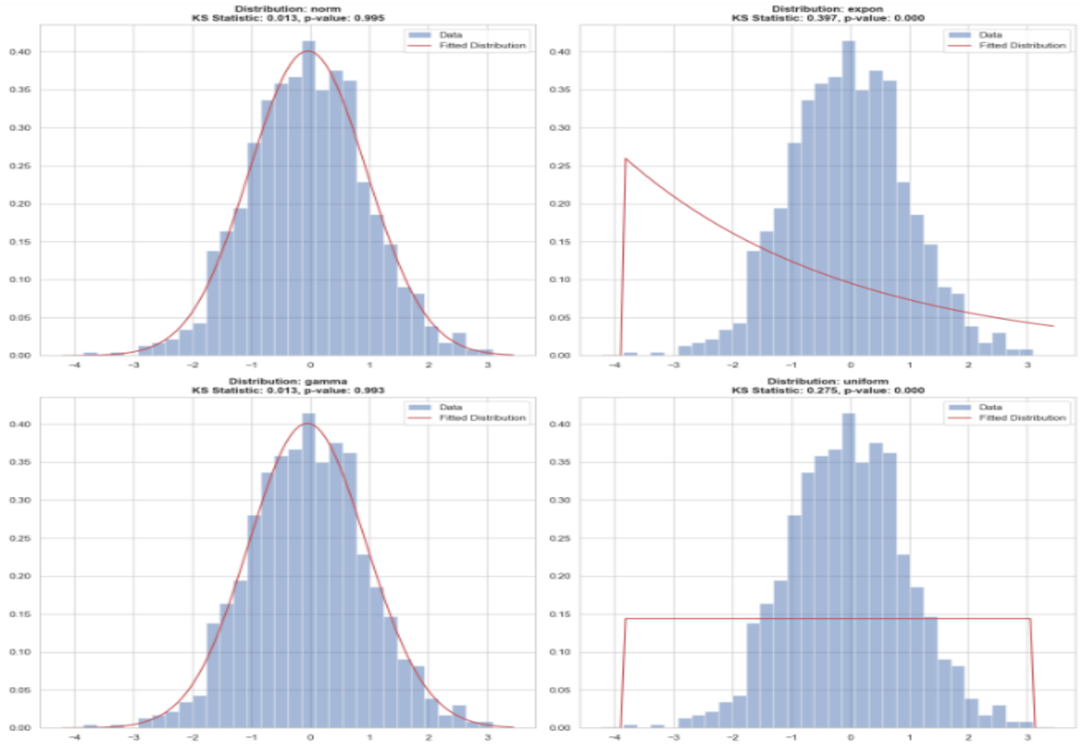

Figure 7: Kolmogorov-Smirnov test statistic and p-value

b. ***Information Criteria***: *Akaike Information Criterion (AIC)* and *Bayesian Information Criterion (BIC)* are widely used measures to compare the fit of various statistical models and distributions. These criteria aim to assess the goodness of fit by considering both the model complexity and the quality of the fit. AIC and BIC values are computed based on the likelihood function and penalize complex models, encouraging the selection of simpler models. Lower AIC and BIC values indicate better fits and suggest that the selected model provides a more accurate representation of the underlying data. These criteria play a significant role in model selection and can guide researchers in choosing the most appropriate model for their analysis. Figure 8 shows the AIC and BIC information criteria for fitting normally generated data. The lowest values of AIC and BIC indicate the underlying density function of the data. Since the data in Figure 8 was generated from a normal distribution, we expect the normal distribution to be the best fit. However, the results will confirm this by showing a lower AIC and BIC for the normal distribution compared to the others.

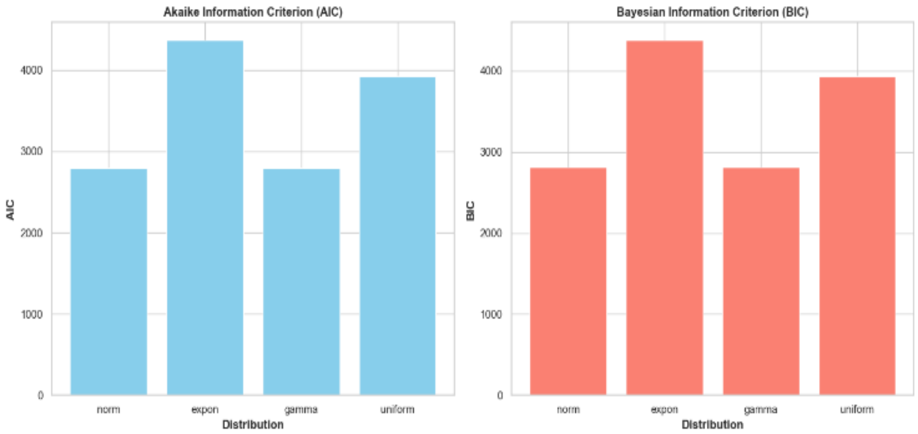

Figure 8: AIC and BIC information criteria

3.3 Domain Knowledge

The nature of the data itself can suggest a likely distribution. For instance, count data might follow a Poisson distribution, where the average value of events occurring in a fixed interval of time or space is known, while waiting times might follow an exponential distribution, which characterizes the time between events in a Poisson process. Utilizing domain knowledge can frequently provide valuable initial guidance on which distributions to test and explore further in order to gain deeper insights and understanding. It allows us to narrow down our choices and focus our analysis on the most appropriate distribution models, ensuring more accurate and meaningful interpretations of the data. By leveraging our understanding of the context and variables at play, we can make informed decisions and tailor our statistical analysis to effectively capture the underlying patterns and characteristics of the data. This approach enhances our ability to draw meaningful conclusions and make insightful predictions, contributing to the advancement of knowledge in various fields and applications.

The choice of methods for identifying the most appropriate data distribution is based on both the nature of the data and practical considerations. We often start with visual tools like histograms and Q-Q plots because they give a simple, clear snapshot of the data’s shape. These visuals help us quickly spot things like skewness, multiple peaks, or patterns that deviate from the norm—common traits in real-world datasets. When we need a more detailed assessment, we turn to statistical tests like the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests. These tests provide a formal way to measure how well the data matches a theoretical distribution. The KS test is great for larger datasets and continuous data, while the AD test is more sensitive to extreme values, making it useful when the data has outliers or unusual patterns. Finally, we use tools like AIC and BIC to make sure we’re not overcomplicating things. These criteria help balance the model’s complexity with how well it fits the data, especially when we’re working with datasets that have lots of variation or intricate patterns. By focusing on these specific traits—like skewness, outliers, or the need to compare multiple distributions—we make sure we’re picking the right method for the job.

In short: we don’t just pick methods randomly; we choose based on what the data tells us. The right method leads to the right model, and the right model leads to better results.

4 Dealing With Non-Standard Distributions

When data deviates from a standard distribution, it can create difficulties for statistical analysis and modeling. In situations where data does not adhere to a standard distribution, comprehending the underlying structure is essential for efficient modeling and analysis. These situations may encompass, [15,16]:

- **Multi-Modal Distributions**: Data that contains multiple peaks may not adhere to a standard single-peak distribution model, such as the normal or exponential distributions.
- **Heavy-Tailed Distributions**: Data that exhibits heavy-tailed behavior is more likely to generate extreme values when compared to standard distributions such as the normal distribution.
- **Skewed Distributions**: Data that displays substantial skewness is not suiTABLE for symmetric distributions such as the normal distribution.
- **Truncated Distributions**: Data that is truncated at specific points may not conform to any well-known distribution.
- **Mixture of Distributions**: data that is produced from a combination of various distributions.
- **Periodic Distributions**: data that exhibits recurring patterns, like time series data that displays seasonal fluctuations.
- **Outlier-Rich Distributions**: Data that contains a significant number of outliers.

Figure 9 shows some examples of data distributions that deviate from known standard distributions

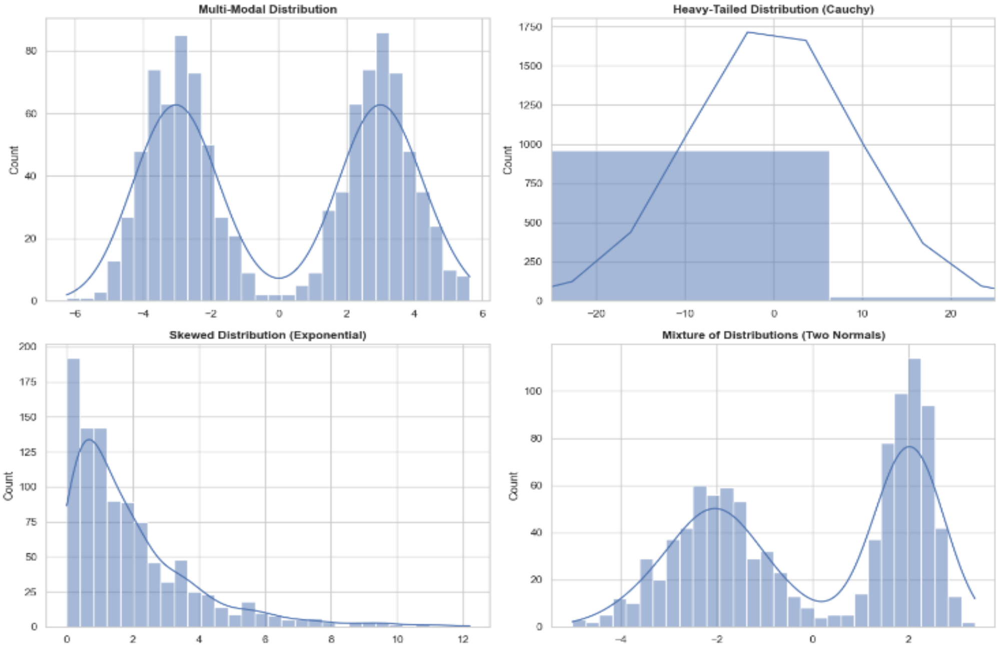

Figure 9: Non-standard distributions examples

To effectively handle data with non-standard distributions, a wide range of techniques come into play. These approaches are specifically designed to address the challenges posed by datasets that deviate from the norm. Here are some techniques to deal with such data, [17,18]:

4.1 Transformations

Data alternation or transformation can improve its conformity to a standard distribution. Examples of such alterations include taking the Log Transformation, Square Root Transformation. Box-Cox Transformation, Yeo-Johnson Transformation, Reciprocal Transformation, Rank Transformation, Quantile Transformation, Z-Score Normalization, Min-Max Scaling and Robust Scaling transformation. Figure 10 depicts a number of transformation techniques applied to some data drawn from exponential distribution [19].

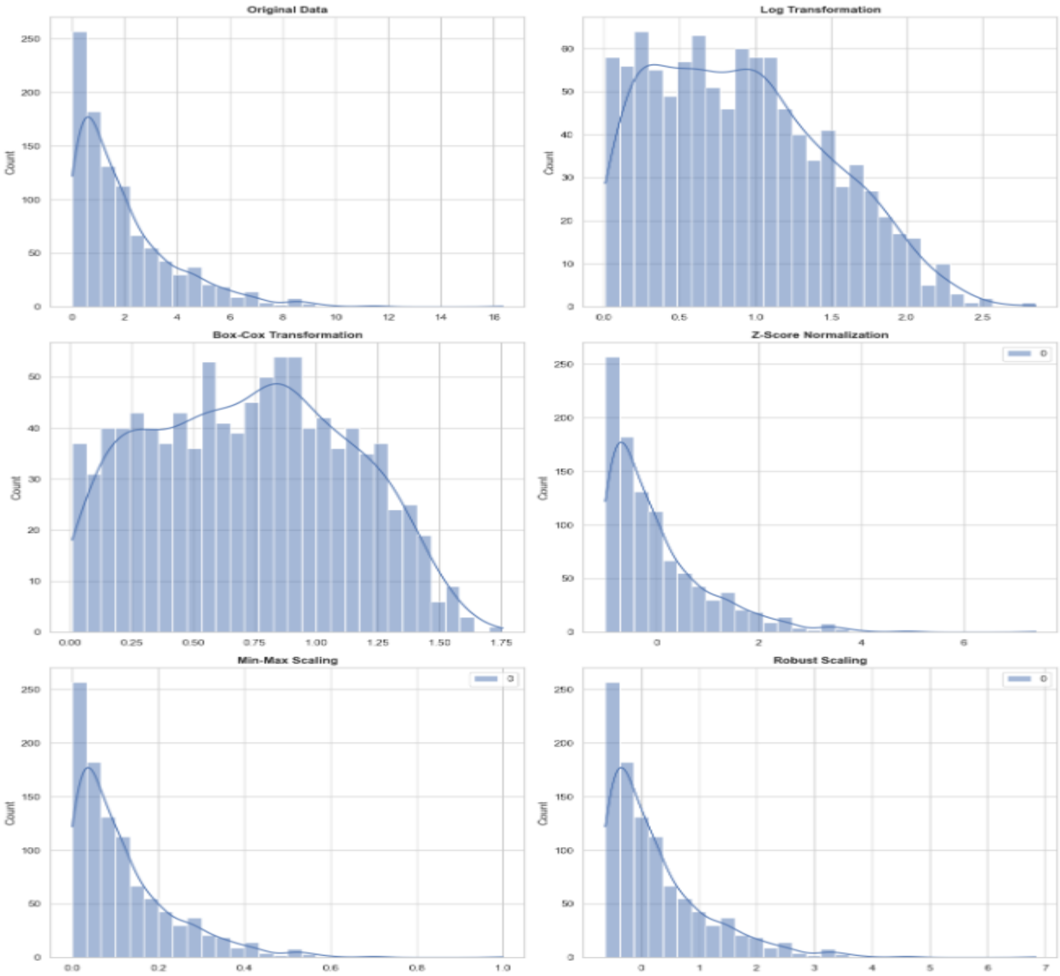

Figure 10: Transformation techniques for non-standard distribution data

4.2 Mixture Models

These utilize a combination of various standard distributions to represent intricate data patterns. Mixture models are effective tools for representing complex data patterns. Some of the common mixture models are: Gaussian Mixture Model (GMM), Dirichlet Process Mixture Model (DPMM), Hidden Markov Models (HMMs), Mixture of Experts (MoE), and Latent Dirichlet Allocation (LDA). Figure 11 shows some of examples of mixture models applied to data with non-standard distributions.

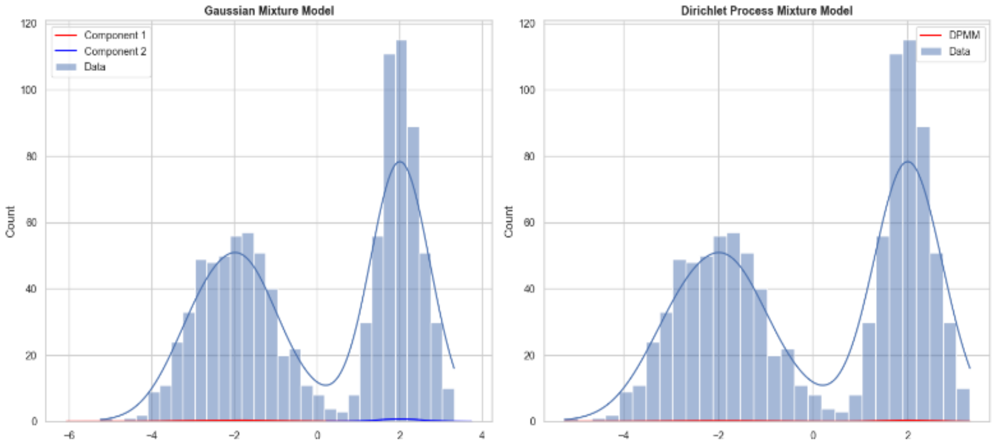

Figure 11: Mixture models examples

4.3 Non-Parametric Methods

Versatile methods for calculating the probability density function of data are available without the need to make assumptions about a particular distribution. These methods include: Kernel Density Estimation (KDE), Histogram, Nearest Neighbor Density Estimation, Rank-Based Methods, Bootstrap Method, Parzen Windows, and Spline Density Estimation. These non-parametric methods offer versatile and strong tools for estimating the probability density function of data without requiring a predefined parametric form. Each method has advantages and is appropriate for a variety of data and analysis situations. These strategies allow data scientists to obtain deeper insights into the underlying distribution of their data, resulting in more accurate and robust models and analyses [20]. Figure 12 shows examples of using non-parametric methods for density estimation.

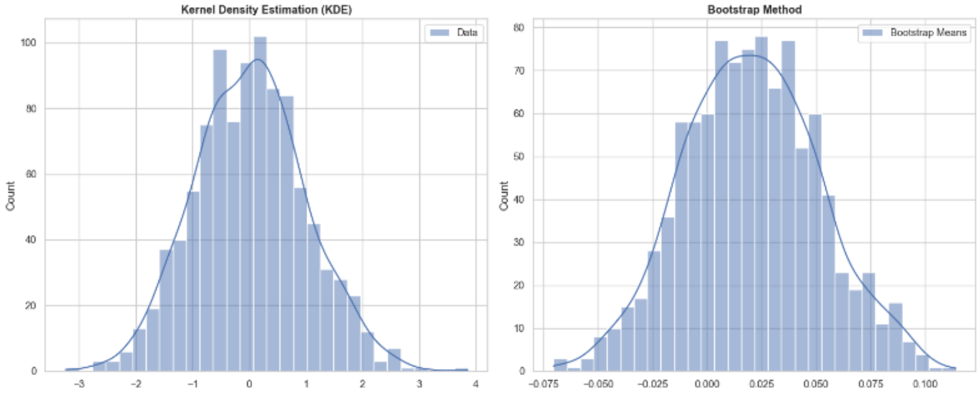

Figure 12: Non-Parametric models examples

4.4 Custom Distributions

If one possesses extensive expertise in a particular field, it is possible to devise a tailored distribution that precisely captures the characteristics of the dataset. Let's assume we have domain knowledge that suggests our data is best modeled by a piecewise function, as shown in Figure 13.

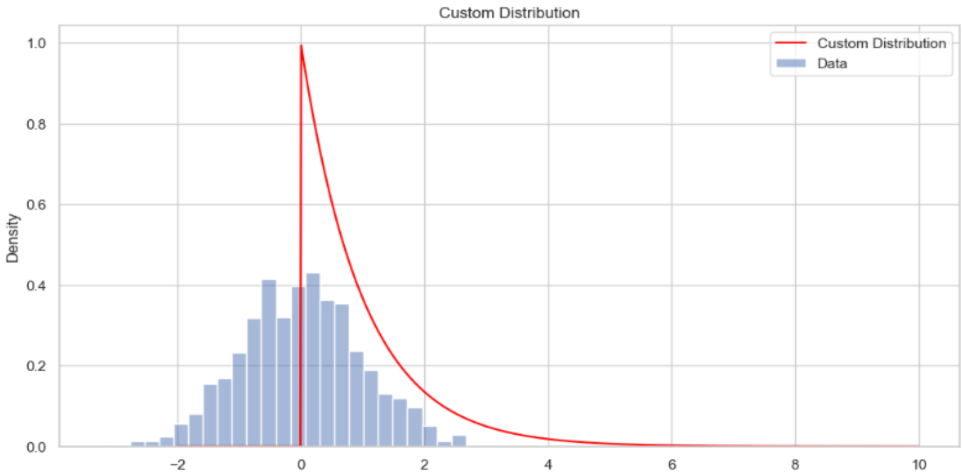

Figure 13: Custom distribution based on expertise of the field

5 Demonstration With The California Housing Dataset

In this section, a demonstration of the importance of finding the underlying density distribution is applied to a data science project on the well-known dataset of California housing dataset. The California housing dataset covers data on many aspects of houses in several California districts, such as median house value, median income, housing median age, total rooms, total bedrooms, population, households, latitude, longitude. This dataset is frequently used in regression tasks, with the purpose of predicting the median house value based on other variables.

Figure 14 demonstrates that the Median House Value ("MedHouseVal") in the California Housing dataset is not normally distributed. This is evident in the corresponding Q-Q plot, which significantly deviates from the diagonal line. However, after applying a log-normal transformation, Figure 15 shows an improved histogram that more closely resembles a normal distribution. The improved Q-Q plot, with reduced deviation from the diagonal line, further supports this.

To assess the impact of this transformation, a linear regression model was applied to both the original and transformed "MedHouseVal" data. The mean squared error (MSE) was then calculated for each. Figure 16 clearly illustrates a dramatic decrease in MSE for the transformed variable, from 0.56 to 0.05. This highlights the importance of aligning data with the assumptions of machine learning and data science models to improve their performance [21].

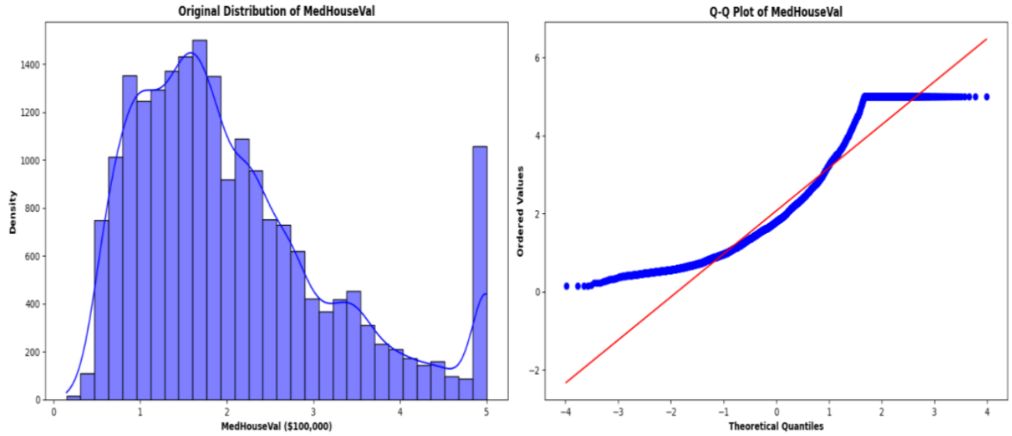

Figure 14: PDF and Q-Q plot of the original variable

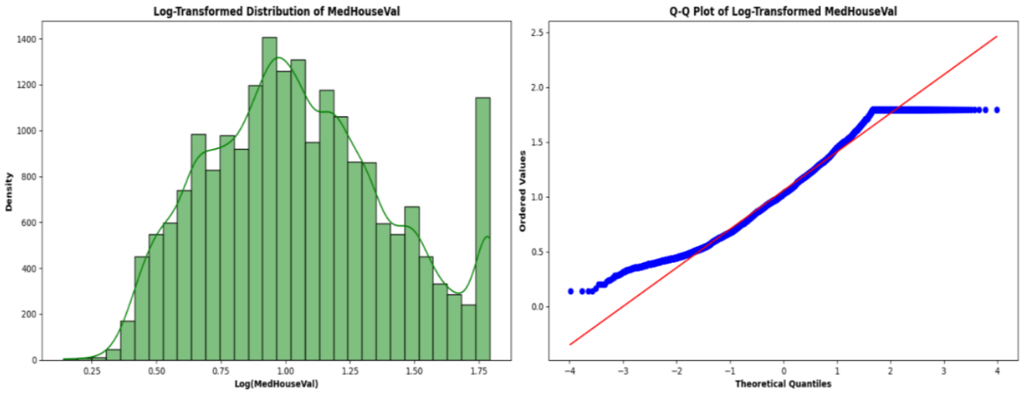

Figure 15: PDF and Q-Q plot of the transformed variable

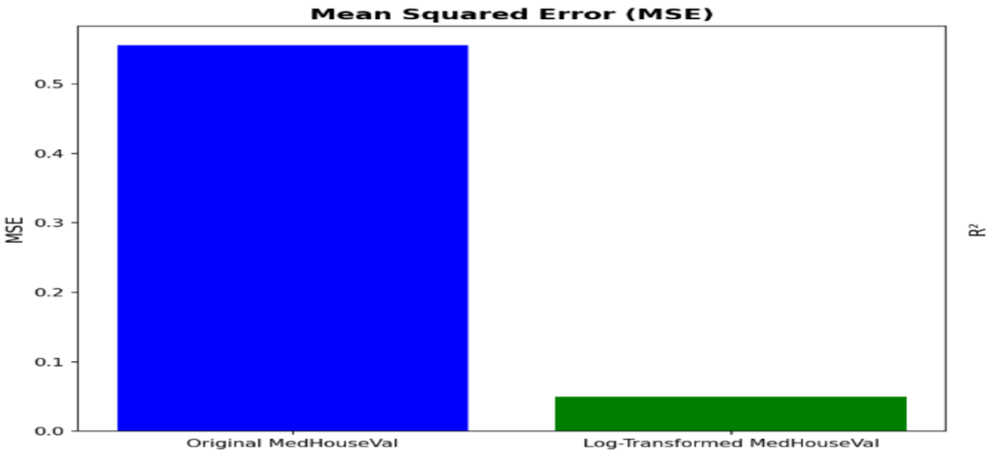

Figure 16: Regression MSE of the original and transformed variable

6 Conclusion

Understanding how data is distributed plays a key role in building accurate models in data science. In this paper, we explored practical methods like histograms, Q-Q plots, and statistical tests such as Kolmogorov-Smirnov and Anderson-Darling to help identify the best-fitting distributions. We also looked at how tools like AIC and BIC help balance model complexity with fit. By applying these techniques to the California Housing dataset, we showed how choosing the right distribution can improve model accuracy, feature engineering, and anomaly detection.

The importance of these findings goes far beyond the specific dataset we used. Knowing the right distribution helps data scientists avoid mistakes in model selection, leading to more reliable results in various fields. This paper emphasizes the need to consider both the characteristics of the data and the context in which it's being used.

Looking ahead, there's exciting potential for future research. It would be interesting to explore how machine learning could be used to automate distribution selection. There's also room to test these methods on larger, more complex datasets, or those with unusual distributions and noise. Another exciting area is integrating these techniques into real-time data systems, which could open new doors for dynamic modeling and faster decision-making. By refining these approaches, we can continue improving the accuracy and impact of data-driven insights across different industries.

ACKNOWLEDGMENT

The authors would like to thank the Arab Open University and Al-Zaytoonah University of Jordan for providing the necessary scientific research supplies to implement this work.

FUNDING INFORMATION

The authors extend their appreciation to the Arab Open University for funding this work through AOU research fund no. (AOUKSA-524008).

References

[1] Wickham, H., Çetinkaya-Rundel, M., and Grolemund, G., *R for data science*, 2023.

[2] Matheus, R., Janssen, M., and Maheshwari, D., Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities, *Government Information Quarterly*, 2020.

[3] Whang, S., Roh, Y., Song, H., and Lee, J., Data collection and quality challenges in deep learning: A data-centric ai perspective, *The VLDB Journal*, 2023.

[4] Maharana, K., Mondal, S., and Nemade, B., A review: Data pre-processing and data augmentation techniques, *Global Transitions Proceedings*, 2022.

[5] Jaradat, Y., Masoud, M., Jannoud, I., & Zaidan, D., The impact of nodes distribution on energy consumption in wsn. *In 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, (pp. 590-595), IEEE, 2019.

[6] Jaradat, Y., Masoud, M., & Al-Jazzar, S., A comparative study of the effect of node distributions on 2D and 3D heterogeneous WSN. *International Journal of Sensor Networks*, 33(4), 202-210, 2020.

[7] Talská, R., Menafoglio, A., Hron, K., Egozcue, J., et al., Weighting the domain of probability densities in functional data analysis, *Wiley Online Library*, Stat, vol. 2020..

[8] Van de Schoot, R., Depaoli, S., King, R., Kramer, B., et al., Bayesian statistics and modelling, Nature Reviews, 2021.

[9] Nandi, G. and Sharma, R., Data Science fundamentals and practical approaches: understand why data science is the next, 2020.

[10] Mehra, P., and Ahuja, M. S., Machine Learning based Anomaly Detection for High

Dimensional Time-Series data in Linear Neural Network. *International Journal of Advances in Soft Computing and its Applications*, vol 15, Issue 1, pp. 14 – 29, 2023.

[11] Khairunissa, J., Wahjuni, S., Soesanto, I., and et. al., Multi-Object Tracking Algorithm for Poultry Behavior Anomaly Detection. *International Journal of Advances in Soft Computing and its Applications*, vol 15, Issue 1, pp. 159 – 176, 2023.

[12] Hawashin, B., Althunibat, A., Kanan, T., AlZu'bi, S., & Sharrab, Y., Improving arabic fake news detection using optimized feature selection. *In 2023 international conference on information technology (ICIT)*, (pp. 690-694). IEEE, 2023.

[13] Law, A. M., How to build valid and credible simulation models, *2022 Winter Simulation Conference (WSC)*, 2022.

[14] Ahsan, M., Mahmud, M., Saha, P., Gupta, K., Effect of data scaling methods on machine learning algorithms and model performance, *Technologies*, 2021.

[15] Sun, J. and Xia, Y., Pretreating and normalizing metabolomics data for statistical analysis, Genes *& Diseases*, 2024

[16] Nolan, J.P., Univariate stable distributions, Springer *Series in Operations Research and Financial Engineering*, Springer, 2020.

[17] Tredennick, A. T., Hooker, G., Ellner, S. P., and Adler, P. B., A practical guide to selecting models for exploration, *inference, and prediction in ecology*, Ecology, 2021.

[18] Haslbeck, J., Ryan, O., Robinaugh, D.J., et al., Modeling psychopathology: From data models to formal theories, *psychological methods*, 2022.

[19] Samarah, T., Hindieh, A., Daoud, M., and Almaini, M., Intelligent Voice Search Strategies for Digital Marketing Transformation, *International Journal of Advances in Soft Computing and its Applications*, Volume 17, Issue 1, Pages 338 – 354, 2025.

[20] Rifada, M., Suliyanto, N., Tjahjono, E., and Kesumawati, A., The logistic regression analysis with nonparametric approach based on Local scoring algorithm (case study: Diabetes Mellitus Type II cases in Surabaya of Indonesia), *International Journal of Advances in Soft Computing and its Applications*, Volume 10, Issue 3, Pages 168 – 178, 2018.

[21] Masoud, M., Jaradat, Y., Rababa, I., and Manasrah, A., Turnover Prediction using Machine Learning: Empirical Study, *International Journal of Advances in Soft Computing and its Applications*, Volume 13, Issue 1, Pages 193 – 207, 2021.

BIOGRAPHIES OF AUTHORS

Yousef Jaradat is an IEEE senior member and a professor of electrical and computer engineering at Al-Zaytoonah University of Jordan. He received his PhD from New Mexico State University, New Mexico, USA, in 2012. His research interests include wireless networks, network modeling and simulation, A.I and machine learning, computer security and quantum computing.

Mohammad Masoud is a professor of electrical Engineering at Al-Zaytoonah University of Jordan. He received his PhD in Communication Engineering and Information Systems from Huazhong University of Science and Technology (HUST), Wuhan, China in 2012. He is a reviewer in many computer and communication journals. His research interests include computer network measurements, network security, machine learning, Software Defined Networking (SDN), embedded systems, control theory and Cyber Physical Systems (CPS).

Ahmad Manasrah is currently an associate professor of Mechanical Engineering at Al- Zaytoonah University of Jordan. He received his PhD degree from The University of South Florida. He was a research assistant and a member of Rehabilitation Engineering and Electromechanical Design Lab at the USF. He is also a member of ASHRAE, Jordan. His interests include Renewable Energy, Smart Energy Technology Mechanical Control, and Education.

Prof. Dr. Mohammad A. Alia, his Ph.D. from Universiti Sains Malaysia (USM), Penang, in 2008. His research interests include public key cryptosystems, fractals, image processing and steganography, wireless networks, and machine learning.

From 2009 to 2019, he held several administrative positions at the Faculty of Science and Information Technology. He later served as the Dean of Scientific Research and Innovation at Al-Zaytoonah University of Jordan for four years. Throughout his career, he has chaired numerous academic and institutional committees both within the university and at the national level. He has also supervised several postgraduate theses in computer science and related fields.

Khaled M. Suwais is a university professor of Information Security and Cryptography at the Arab Open University, Saudi Arabia. He received his Ph.D. in Cryptography and Information Security from Universiti Sains Malaysia in 2009. His research focuses on cryptographic algorithms, game theory, artificial intelligence, and parallel computing, with notable contributions to stream ciphers, static taint analysis, and evolutionary models for cooperative behavior. He has published extensively in peer-reviewed journals. He has led funded projects on cybersecurity, steganography, and machine learning applications. Prof. Suwais has received multiple awards. He serves as a reviewer for several high-impact journals.

Sally Almanasra is an Associate Professor specializing in Software Engineering and Artificial Intelligence. She holds a Ph.D. from Universiti Sains Malaysia (2014) and an M.Sc. from Al-Balqa Applied University (2007), both with honors. Her research focuses on AI, game theory, parallel computing, and information security, with applications in optimization, and machine learning. Dr. Almanasra has published extensively in peer-reviewed journals and conferences, contributing to advancements in adaptive algorithms, intelligent systems, and computational methods. She has led several funded research projects exploring AI applications in big data, mobile technologies, and evolutionary models. Her collaborative work spans interdisciplinary areas including IoT, computer vision, and Arabic language processing.