# Accuracy comparison of logistic regression, random forest, and neural networks applied to real MaaS data

Tetiana Reznychenko, Evženie Uglickich, Ivan Nagy

***Abstract*—The paper deals with a comparative analysis of three widely used data analysis methods: logistic regression, random forest, and neural networks. These methods have been evaluated in terms of accuracy, and computational efficiency and applied to different types of data sets, including both simulated and real MaaS data. The study aims to compare the efficiency of each method in classification tasks. The study leads to specific recommendations on which method to use under various circumstances, contributing to the decision-making process in data analysis projects. We have shown that random forests generally provide better accuracy and are resistant to over-training. Neural networks can achieve comparable performance under certain conditions, although at a high computational cost. Logistic regression shows limitations in dealing with complex data structures.**

***Index Terms*—classification algorithms, data analysis, machine learning, Mobility as a Service, random forests**

## I. Introduction

In this era of digitization, an increasing number of life's facets are becoming intertwined with digital advancements. Notably, smart cities, smart mobility (Mobility as a service) [1], the Internet of Things [2] are broadening our horizons, transforming everyday life and introducing innovative methods for managing urban infrastructure and optimizing resources. Furthermore, the collection and analysis of data are central to these improvements, providing invaluable insights in fields such as transportation, healthcare, etc.

Data is an important part of today's world. The volume of data is growing exponentially according to Statista (www.statista.com) and Network World (www.networkworld.com). Data provides valuable information for various fields of activity, as transportation networks, smart cities, medicine, business, etc. Through data analytics, data can be transformed into knowledge, and on their basis to make decisions and understand consumers and their processes. Data analysis is used to identify patterns and anomalies in data sets and to make predictions. In the context of smart cities, for example, data analysis is used to optimize transportation networks, and efficient use of urban resources, which increases the satisfaction of residents, as well as increases the turnover of business and government. Data analysis allows for making rational decisions, which leads to strategy formation, minimizing risks, and maximizing efficiency.

In the field of data science, numerous algorithms are developed to explore data.The algorithms for clustering and classification of discrete data use data mining techniques including methods such as decision trees [3], Bayesian networks [4], neural networks [5], k-nearest neighbors [6], fuzzy rules [7], naive Bayes classifiers [8], [10], genetic algorithms, support vector machines [8], gradient boosting [11], and model-based methods.

The model-based methods include the use of discrete mixture models such as latent class and Rasch mixture models [12], Poisson and negative binomial mixtures [13], mixtures of Poisson regressions [12], [13], mixtures of logistic regressions for binary data [13], Poisson-gamma and beta-binomial models [12], [13] as well as Dirichlet mixtures [14], [15]. The estimation of the mentioned mixtures is solved primarily using the iterative expectation-maximization (EM) algorithm [16].

Each of the algorithms has its strengths and produces its result because of different data analysis strategies. A specific algorithm is better to address certain problems.

The layout of the paper is organized as follows. Section II-A formulates the problem to be solved. Section II-B introduces the overview of the algorithms. The results of experiments with simulated and real data are provided in Section III. Conclusions can be found in Section IV.

## II. Theoretical Background

### A. Problem Formulation

Various data analysis algorithms produce different results with the same data set. This can potentially be misleading for researchers and analysts, as well as making it difficult to decide on the most relevant algorithm for a particular task.

T. Reznychenko is with the Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25, 11000 Prague, Czech Republic (e-mail: tetiana.reznychenko@cvut.cz).

E. Uglickich is with the Department of Signal Processing, Institute of Information Theory and Automation of the CAS, Pod vodárenskou věží 4, 18208 Prague, Czech Republic (e-mail: uglickich@utia.cas.cz).

I. Nagy is with both the Faculty of Transportation Sciences, Czech Technical University, 11000 Prague, Czech Republic, and the Department of Signal Processing, Institute of Information Theory and Automation of the CAS, 18208 Prague, Czech Republic (e-mail: nagy@utia.cas.cz).

The objectives of this research are to evaluate performance and identify the strengths and weaknesses of each method. The main criterion for determining the effectiveness of an algorithm will be accuracy.

### *B. Overview of the algorithms*

**Random Forest** [8] is a method which solves classification, regression, and other tasks. It integrates multiple decision trees. Each tree in the model is treated as a randomly selected subset of features from the entire set. Individual forest trees can identify more important attributes, which increases the overall accuracy of the results and improves the performance of the model. Algorithm description:

*Step 1:* For each tree, a training data set is defined. This is done using a bootstrapping method which randomly selects predictors from the data set.

*Step 2:* Constructing the set of decision trees. Each tree is created based on randomly selected subsets of features.

*Step 3.* Each tree gives a prediction or classification depending on the problem. In classification tasks, the class that is more likely to be predicted by the trees is selected as the final class. In regression tasks, the predictions from all trees are averaged to get the final prediction.

*Step 4:* The resulting prediction is the final answer in the random forest.

**Neural networks** [8]. The mathematical model of a neural network for one layer can be expressed:

$$\alpha = \sigma(Wx + b), \tag{1}$$

where:

- $\alpha-$ activation level;
- $\sigma-$ activation function for the level;
- $W-$ weight matrix for the level;
- $x-$ activation of the previous level;
- $b-$ displacement vector for the level;

For a multilayer network, the process is repeated and the output of one layer becomes the input for the next layer:

$$\alpha^{[n]} = \sigma^{[n]}(W^{[n]}x^{[n-1]} + b^{[n]}), \tag{2}$$

where: $n$ means the number of layers in the network.

The neural network model is trained by minimizing a loss function (e.g., cross-entropy for classification tasks) using optimization techniques such as stochastic gradient descent.

Probabilistic neural networks are a type of neural network based on the method of dynamic decay adjustment [9]. In PNN, each training sample is represented as a point in a multidimensional space, and the task of the network is to estimate the probability density function for each class based on the samples provided. PNN are trained on labeled data using a method called dynamic decay adaptation, which uses constructive training of probabilistic neural networks as the main algorithm. This algorithm generates rules based on numerical data, where each rule is defined as a multidimensional Gaussian function. These functions are adjusted by two thresholds - theta minus and theta plus - to prevent rule conflicts between different classes. Each Gaussian function is characterized by a central vector and a standard deviation that is adjusted during training to include only non-conflicting instances. Selected numeric columns from the input data serve as input for training, while additional columns are used for classification targets. A single column containing class information or multiple numerical columns indicating the degree of class membership from 0 to 1 are selected.

**Logistic regression** [17], as shown in Equation (3), is a linear regression in which we estimate the likelihood of the modeled variable falling in either of 2 response categories. The mathematical model of logistic regression (4) is based on the logistic function:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_n \cdot x_n \tag{3}$$

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}, \tag{4}$$

where:

- $P(y = 1|x)-$ probability that $y$ is equal to 1 for a given vector of input variables;
- $\beta_0, \beta_1, \ldots, \beta_n-$ model coefficients;
- $x_1, \ldots, x_n-$ explanatory variables.

An extended version of logistic regression is multinomial regression in which the target variable can take two or more values.

### *C. Analytical tool*

To conduct experiments, a platform for data analysis as KNIME (www.knime.com) was used. This includes a graphical interface for visually creating analytical flows and integrates machine learning and data analysis techniques. KNIME provides a set of tools for classification and includes various algorithms such as random forest, neural networks, and logistic regression. The platform offers tools for evaluating the performance of classification models, including error matrices, accuracy, completeness, and F-measure.

## III. Experiments

The purpose of this section is to evaluate and compare the effectiveness of the three different methods in the context of specific conditions and tasks. This includes analyzing the accuracy, efficiency, speed, and other relevant characteristics of each method. Experiments are conducted on both simulated and real data. Simulated data is used to create a controlled environment where the performance of each of the three methods can be accurately evaluated without the influence of external variables and noise inherent in real data. This data allows the experiment parameters to be fine-tuned. Real data includes complexities and uncertainties commonly found in the real world, such as noise, missing values, and inaccuracies. This helps to determine the practical value of each method under real-world conditions.

### A. Description of data sets

The experiment part aims to apply classification algorithms to data sets using the KNIME Analytics Platform. Two configurations were used to accomplish the goal.

All datasets were divided into training and testing sets. The training data were used to train the model, enabling it to learn and adapt to the complexities inherent in the data. Specifically, 70% of the data was designated as the training set. The testing data, on the other hand, were used for model validation, ensuring that the model's predictions are reliable and can generalize well to new, unseen data. This set comprised the remaining 30% of the data. To enhance the robustness of our findings, the data for each experiment were shuffled. This shuffling process helps to prevent the model from merely memorizing the order of the data, promoting a more genuine learning and generalization capability.

*1) Simulated data set:* The data set contains $30,000$ observations with 12 explanatory variables, each of which can take the values $1$, $2$, or $3$. The target variable is binary, taking values of $1$ or $2$. The first configuration was randomly generated in Scilab (see www.scilab.org).

*2) Real data set:* The second configuration is a real data set that was collected to investigate the behavior of commuters using a hypothetical Mobility as a service (MaaS) [1]. MaaS is an innovative concepts trying to cope with the current challenges such growing numbers of private vehicles, increasing emission levels in cities, etc. The purpose of the study was to determine the factors contributing to the willingness of residents to use MaaS. This paper [1] provided insights into the characteristics and attitudes of potential MaaS users to understand the willingness to use.

Data were collected with a questionnaire using a 5-point Likert scale (Joshi *et al*., 2015), and takes discrete values. The sample included 6,405 commuters in Germany, the United Kingdom, Poland, and the Czech Republic. The collected survey consisted of 2 parts:

- The first part consisted of socio-economical information, stated preference questions, and travelers' beliefs [1].
- The second part was a choice experiment where each of the participants answered 16 times to the hypothetical situation (in each there were 6 travel modes to choose, with different prices and travel times) where they had to choose the preferred mode.

The second part of the study was selected for analysis, which included a sample of 37,104 responses provided by the participants. There are twenty-four variables which contain information on gender, household size, etc. The main variables used in the analysis include:

- Territory is a variable that represents the size of the agglomeration in which the respondent resides. It is categorized into five levels, denoted by the numbers 3 through 7, where 3 indicates a city with a population ranging from 10,000 to 99,999 people, and 7 denotes a megacity with a population exceeding 3,000,000 individuals. In the original dataset, the territory takes values from 1 to 7. In our experiments, values such as 1 and 2 were omitted.

- Gender indicates the gender of the respondent. It is coded as 1 is male, 2 is female, or 3 is diverse (this category is designed to include individuals who do not exclusively identify as male or female).

- Age represents the age of the respondent, ranging from 18 to 71 years.

- Household size is a variable of the number of people living in the respondent's household. It has five categories, represented by the numbers 1 through 5, likely corresponding to the range from a single-person household to a household of five or more individuals.

- Income describes the income level of the household. It is divided into six categories, coded as 1 through 6. This variable is used to assess the economic status of the household.

- Education categorizes the education level of the respondent. It is indicated by the numbers 1 through 6, each number representing a different level of educational attainment.

- Employer is a variable which is the type of employment. It is categorized into eight levels, denoted by the numbers 1 through 8.

The target variable contains information about price or traveler's choice (more details will be given in Section III. D.)

### B. Setup in KNIME

Our analysis was conducted using the KNIME Analytics Platform (Fig. 1). The following subsections detail the specific configurations applied:

*1) Logistic regression:* one hundred integrations were set to achieve convergence of the model without undue strain on computational resources. Uniform regularization was applied to ensure that all model coefficients are affected equally, which helps to prevent overfitting. Stochastic gradient descent was used, which allows efficient finding of optimal model parameters by randomly selecting samples for each gradient step.

*2) Random forest:* one hundred number of trees were selected. Two unconstrained tree depths were set, as well as a minimum number of samples in leaf one.

*3) Probabilistic neural network:* For handling missing data, "Incorp" was selected, which generates rules with missing values if no replacement value has been found during the learning process. On the "Advanced" panel, two settings were configured as "shrink after commit" (a new rule is reduced in such a way to avoid conflicts with other rules from different classes) and "use the class with max coverage" (which means using only the class with the maximum degree of coverage of the target columns). Theta Minus was 0.2 (this defines the upper boundary of activation for conflicting rules), and theta plus was 0.4, which defined the lower boundary of activation for non-conflicting rules.

Each experiment was conducted 10 times to ensure statistical robustness. Subsequently, we calculated the average values for each method based on these 10 repetitions. The

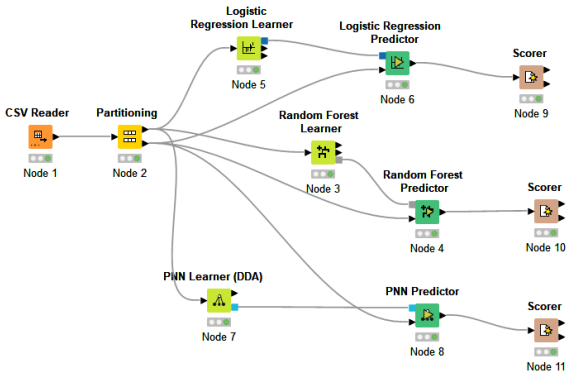


Fig. 1. Simulation of the classification model in KNIME

experiments were carried out on a computer equipped with Windows 11 Pro, and KNIME Analytics Platform 5.1.2.

### C. Experiments with simulated data

**Experiment with the full set of explanatory variables.** The experiment used a simulated data set of 12 explanatory variables.

*Results:*

1) Logistic regression: accuracy reached 0.696.
2) Random forest: accuracy reached 0.934.
3) PNN: the accuracy was 0.697.

PNN had a slightly higher accuracy than logistic regression at 0.697. Random forest showed a high accuracy of 0.934, and was particularly effective for this data set.

*Execution time:*

- Logistic regression: 3.7 seconds.
- Random forest: 3.53 seconds.
- PNN: 4 minutes 28 seconds.

These results indicate a significantly better performance of random forest compared to the other two methods on the full set of explanatory variables.

**Experiment with a simplified set of explanatory variables** that used the same data set but reduced the number of explanatory variables to 6.

*Results:*

1) Logistic regression: the accuracy remained unchanged at 0.696.
2) Random Forest: accuracy decreased to 0.893.
3) PNN: accuracy increased slightly to 0.88.

Reducing the number of explanatory variables resulted in worse results for the random forest, but had minimal impact on the performance of the logistic regression and improved the performance of the PNN.

*Execution time:*

- Logistic regression: 4.6 seconds
- Random forest: 4.4 seconds.
- PNN: 37.0 seconds.

The execution times of the methods show significant differences. Logistic regression and random forest are both very fast, executing in approximately 4.6 and 4.4 seconds respectively. On the other hand, PNN requires substantially more time, at 37.0 seconds. This disparity highlights the need to balance between computational efficiency and model complexity depending on the specific requirements of the task.

### D. Experiments with real data

**Experiment analyzing transportation choices.** The experiment evaluated the effect of 21 explanatory variables on the choice of transportation type among the following options: public transportation, MaaS (Mobility as a Service), and a personal car. The target variable took three values: 1 (choice of public transportation), 2 (choice of MaaS), and 3 (choice of personal car).

*Results:*

1) Logistic regression: 0.81.
2) Random forest: 0.818.
3) PNN: 0.812.

It is interesting to note that in this experiment all three methods showed similar accuracy, which may indicate that the data are homogeneous and have good separability regardless of the chosen classification method. Random Forest slightly outperforms the other models with an accuracy of 0.818, but the differences are not significant.

*Execution time:*

- Logistic regression: 8.12 seconds.
- Random forest: 9.95 seconds.
- PNN: 53.17 seconds.

However, it is worth noting that the execution of the PNN algorithm took more time compared to the other methods, which may be a critical factor when choosing a method to implement in real-world settings. Logistic regression took the least amount of time (8.12 seconds), making it preferable for scenarios that require fast data processing.

**Experiment on comparing transportation service costs.** The second experiment analyzed the influence of 23 explanatory variables on the choice of a more economical transportation option between public transport and MaaS. The target variable ”price” took two values: 1 (public transportation is cheaper than MaaS) and 2 (MaaS is cheaper than public transportation).

*Accuracy results for the different methods:*

1) Logistic regression: 0.813
2) Random forest: 0.81
3) PNN: 0.813

In the experiment, logistic regression and PNN demonstrated identical accuracy. Random forest had slightly lower accuracy, but the difference was minimal.

*Execution time:*

- Logistic regression: 27.31 seconds.
- Random forest: 36.02 seconds.
- PNN: 3 minutes 59 seconds.

The execution time for all three methods increased compared to the first experiment, especially for PNN, which took four times longer.

### E. Discussion

The results obtained from the experiments show that random forest showed higher accuracy than other methods because the method creates a lot of trees and uses bootstrapping, which helps to reduce the risk of overtraining, but its result is affected by the quantity features, unlike logistic regression.

Logistic regression is a powerful generalized linear model, but it may not be able to cope with data in which there are specific relationships between features. This limitation makes it less effective than more sophisticated models. It is to be noted that in the fourth experiment, where the target variable took a binary value, then logistic regression performed better than in the third experiment.

Neural Network performs well on classification tasks when a sufficient amount of training data is available and the data are sufficiently structured. However, it is more computationally intensive than the other methods, due to a more complex model and a larger number of parameters.

## IV. Conclusion

We focused on comparing three methods, and to determine the most effective approach to classification in tasks with two different data sets: simulated and real. The key aspects of our study were to evaluate the efficiency of the approach, and the ability to handle different types of data.

The goals were successfully achieved. We analyzed each of the methods under different conditions of their application and identified the key strengths and weaknesses of each approach. The most important contribution of our study was the confirmation of the high performance of the random forest method in both configurations, making it the preferred choice for handling complex data sets with a large number of explanatory variables.

Nevertheless, our study also identified several open problems and areas for future research. In particular, it is worthwhile to further analyze the importance of the variables, especially to determine which variables were most influential in each experiment. In addition, neural networks have been shown to require significant computational resources, especially when dealing with large amounts of data, and this is also of interest to further explore their potential.

In conclusion, the results of our study emphasize the importance of a comprehensive approach to method selection, taking into account the specifics of the problem, the amount and types of data, and the computational capabilities. We are confident that our findings will help data analytics in their aim to select optimal algorithmic solutions.

## V. Acknowledgements

This work was supported by the StorAIge project and the corresponding Czech institutional support project No. 8A21009. This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 101007321. The JU receives support from the European Union's Horizon 2020 research and innovation program and France, Belgium, Czech Republic, Germany, Italy, Sweden, Switzerland, Turkey.

The project was partially supported by the project TAČR FW06010535.

## References

[1] M. Matowicki, M. Amorim, M. Kern, P. Pecherkova, N. Motzer, O. Pribyl Understanding the potential of MaaS – An European survey on attitudes. Travel Behaviour and Society, 2022, 27: 204-215.

[2] Kolhe, R. V., William, P., Yawalkar, P. M., Paithankar, D. N., Pabale, A. R. (2023). Smart city implementation based on Internet of Things integrated with optimization technology. Measurement: Sensors, 27, 100789.

[3] S.L. Salzberg. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1994. Mach Learn 16, 235-240.

[4] R. E. Neapolitan. Learning Bayesian Networks. Pearson, 2019.

[5] C. C. Aggarwal. Neural Networks and Deep Learning: A Textbook. Springer, 2018.

[6] T. M. Cover and P. E. Hart, Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967, 13(1): 21-27.

[7] M. R. Berthold, B. Wiswedel, T. R. Gabriel. Fuzzy logic in KNIME – modules for approximate reasoning. International Journal of Computational Intelligence Systems, 2013, 6(1): 34-45.

[8] D. Forsyth. Applied Machine Learning. Springer, 2019.

[9] M. R. Berthold, J. Diamond. Constructive training of probabilistic neural networks. Neurocomputing. 1998, 19(1-3), 167-183.

[10] S. M., Basha, D. S. Rajput, R. K. Poluru, S. B. Bhushan, S. A. K. Basha. Evaluating the performance of supervised classification models: decision tree and naïve Bayes using KNIME. International Journal of Engineering & Technology, 2018, 7(4.5): 248-253.

[11] T. Hastie , R. Tibshirani, J. Friedman. "The Elements of Statistical Learning". Springer, 2009.

[12] A. Agresti. Categorical Data Analysis. 3rd Ed. John Wiley & Sons, 2012.

[13] P. Congdon. Bayesian Models for Categorical Data. John Wiley & Sons, 2005.

[14] N. Bouguila, W. Elguebaly. Discrete data clustering using finite mixture models. Pattern Recognition, 2009, 42(1): 33-42.

[15] Y. Li, E. Schofield, M. Gönen, A tutorial on Dirichlet process mixture modeling. Journal of Mathematical Psychology, 2019, 91: 128-144.

[16] M. R. Gupta, Y. Chen. Theory and Use of the EM Method. (Foundations and Trends(r) in Signal Processing). Now Publishers Inc., 2011.

[17] D. W. Hosmer, S. Lemeshow. "Applied Logistic Regression". 2nd Ed. Wiley-Interscience, 2000.

[18] A. Agresti. "An Introduction to Categorical Data Analysis". 3rd Ed. Wiley, 2018.