Accuracy assessment of applied supervised machine learning models on usual data probability distributions

R. Chefira*, S. Rakrak
L2IS: Laboratory of Engineering Informatics and Systems at Faculty of Sciences and Techniques, Cadi Ayyad University, Marrakesh

* chefira.reda@gmail.com ; s.rakrak@uca.ma

Abstract. In this paper, an application analysis of supervised classification techniques on several probability distributions is carried out. Accuracy as well as usual standard metrics have been highlighted to rate the performance of generated learning models. Using data that fit different distributions, we investigated whether the application of a classification method had an optimizing impact on the accurateness of its correlated learning model.

1. Introduction
The Internet of Things has emerged as one of the most exploited areas of research in modern times. This is due to the need for biosensors, capable of collecting data and exchanging them using smart devices, mostly based on cloud solutions. Several areas are inter-related to these smart solutions, such as agriculture automation, security and surveillance, smart homes and cities, machine to machine sensor networks, telemedicine, and healthcare.

In several data-driven fields, optimal decision making is the key issue. Probability theory provides us with the ability to improve decision-making systems by analyzing the in-depth behavior of the data[1]. Systems often used in the medical field tend to optimize the actions performed by practitioners. In doing so, the enormous relevance of machine learning to these industry is second to none[2-3]. However, it is also important to identify some hidden aspects of the data. in the field of telemedicine and healthcare, several applications have been designed to assist the elderly and people suffering from chronic conditions to ensure a real-time medical monitoring of their vital indicators. Early detection and diagnosis of diseases can slow the progression of illness and can lead to a significant reduction in the cost of health services[4]. Hence, Ubiquitous Systems have the capacity to streamline and enhance the access to health services, especially at the end of the ICU or when patients are allocated to home-based care[5].

Wireless body area networks are the most requested architectures in the development of these applications. They assist the medical care team in capturing data using several decision-making indicators closely related to patient’s health conditions. Heart rate, body temperature, glucose, blood pressure, SO2 and so-on are prime examples of widely used life-supporting and life-sustaining measures[6]. These indicators are part of a WBAN-specific typology that periodically exchange the sensed information with a sink. Therefore, medical information can be collected, exchanged, and stored in an optimal architecture for this purpose. Upon classification of these records, and given the sensitivity of this type of data, it is necessary to develop an analytical study of the distributional and probabilistic aspects.
This work therefore required studying three of the most prevalent probability distributions in the IoT environment, namely Normal, Poisson and Uniform distributions. Several studies have highlighted the availability of data fitting those distributions, i.e. traffic management, security, medical information etc. Furthermore, every machine learning technique used for a data classification purpose includes, in an explicit or implicit manner, a few assumed assumptions about the data. The cornerstone of any data analysis or decision making is the application of the appropriate machine learning technique with the aim to obtain optimal results. Data distributions factors found to be influencing decision making have been explored in several studies, closely related to IoT [7][8]. Several studies propose information-sensitive approaches to optimize data aggregation, data collection or the efficiency of their benchmark parameters. Authors in [9] propose an approach to collect optimized information with a predefined sampling probability to ensure performance in terms of data coverage and information amount. A probabilistic-based classification procedure is applied to normal distributions to assess the gravity of the detected medical anomalies in [8]. A multivariate normal distribution was considered to reduce the complexity of performance analysis of WBAN characteristics during walking in [10]. Along these lines, solutions for diabetic patients have been provided, in [11], a KNN-based classification management system proposed, while in [12], authors designed a health care system that facilitates rapid diagnosis.

The overall structure of this paper takes the form of four sections, including this introductory chapter, the second one deals with the experimental design, data distributions, as well as the involved machine learning techniques methodologies. The third chapter highlights the achieved results in terms of accuracy and effectiveness using several standard measures. Finally, a conclusion chapter aims to summarize this study and outlines a broader research perspective.

2. Methodology

IoT solutions generate a very high volume of traffic during the collection and distribution of information. These can be of a periodic or event-driven nature depending on the scope of the system involved [7],[13]. In a context of in-depth data analysis, it is worth having a global view on the data generated by a device belonging to an IOT network. The first step consists of generating sensed data using a connected device and sharing the gathered information through the Internet. Then comes the process of setting up a logical structure to apply different analysis-based approaches. A final step is the exploitation and use of these data in the coming period. This is relevant in providing the data with the ability to create more valuable information for users and researchers.

2.1. Experimental design

In this subsection, an end-to-end information management architecture is proposed to describe the process involved using a smart device. Figure 1 illustrates the experimental design. It should be noted that this paper deals solely with the Data aspect of the architecture, addressing both data distribution and data classification steps.
The proposed architecture shown in Figure 1 is depicted as follows:

- **Roles**
  - **Actors**
    - Patient (actor agent)
    - Doctor (actor agent)
  - **Systems**
    - Data collection (System agent)
    - Analysis (System agent)
    - Check Distribution (System agent)
    - Data Classification (System agent)

- **Data collection**
  Using a smart device, a patient can sense its Realtime vital information. Nowadays, many implanted sensors are used to gather biological data such as: heart rate, blood pressure, temperature, gesture and so on. Gathered data are transmitted through wireless technologies (Wi-fi and Bluetooth), and stored in medical-use adapted databases for a given process or task.

- **Analysis**
  While testing hypotheses and assumptions, data analysis is the process of retrieving primary data and turning it into useful information for a decision-making end. To do so, a feature selection is applied. The core principle when using a feature selection technique is to retain the subset of relevant dataset features. Then comes the contribution of a practitioner, whose role is to ensure the inclusion of management rules related to his or her field of expertise. In this case, the doctor can support the architecture by addressing the patient's auscultatory tests, as well as the normal ranges of vital signs.

- **Check distribution**
  A statistical test is carried out to describe whether experiential entries fit a usual data distribution. Goodness of fit was assessed by calculating different statistical measures, namely p-value, s-value and Anderson-Darling, Kolmogorov-Smirnov tests[8]. Data used in this study fit several usual probability distributions, namely:
  - Normal distribution
  - Poisson distribution
  - Uniform distribution

2.2. **Data pre-processing**
The purpose of this sub-section is to explain the data format. The database is a collection of sensed information that is organized so that it can be easily accessed as a decisional-based data source. Datasets
consist of at least 500 decision-making readings, where a single tuple row (entry) \( \{ R \} \) groups three items into a single compound value. The first two numeric variables are the sensed readings \( x \) and \( y \), fitting the same probability distribution. The third one is as a ruleset-based binary decision \( C \). A data row is thus defined by \( R = \{ x, y, C \} \). For every dataset, each of the \( x \) and \( y \) variables is represented by a sequence of one-dimensional vector, stacked together to make a two-dimensional vector \( \{ x, y \} \).

After the data generating step, comes the implementation sequence of several supervised machine learning techniques which typically involves predicting a known outcome or goal. For all datasets featured in this paper, we set the decision rules as follows:

- if \((90 < x < 120) \text{ OR } (60 < y < 80)\) then \( C = 0 \); 0 states for Normal measure.
- else \( C = 1 \); 1 states for Abnormal measure or Anomaly.

This will result of a tuple: \{x-value, y-value, 0\} for normal readings or \{x-value, y-value, 1\} for abnormal reading. The classification process leads us to a process with binary entries. Table 1 summarizes the number of normal and abnormal measures of the generated datasets according to each data distribution approach.

### Table 1. Generated numbers of Normal and Abnormal measures

| Classes         | Normal distribution | Uniform distribution | Poisson distribution |
|-----------------|---------------------|----------------------|----------------------|
| Normal (0)      | 216                 | 192                  | 81                   |
| Abnormal (1)    | 284                 | 308                  | 419                  |

#### 2.3. Assessment of data classification

To evaluate the optimal model among the candidates, it is bound to estimate its performance by means of several measures. True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN) measures variables are used to rate the performance of correct and incorrect predictions. Accuracy, Precision, Recall, F1-score are the criteria commonly used to spotlight the overall effectiveness of any learning models. These are explained as follows:

- The overall model performance:
  - Accuracy = \( \frac{TP + TN}{TP + FP + FN + TN} \)
- The accuracy of the positive predictions:
  - Precision = \( \frac{TP}{TP + FP} \)
- The coverage of the positive sample:
  - Recall = \( \frac{TP}{TP + FN} \)
- The harmonic recall-precision average:
  - F1-score = \( \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \) \( \frac{TP + TN}{TP + FP + FN + TN} \)

Confusion matrix, also known as error matrix, is a layout table that is also used to describe the performance of the generated models. Rows represent the instances in the predicted binary class while each column illustrates the instances in the actual binary class. Table 2 depicts the used standard metrics.

### Table 2. Confusion matrix

| Predicted (0) | Predicted (1) |
|--------------|--------------|
| Actual (0)   | TP           | FP           |
| Actual (1)   | FN           | TN           |

#### 2.4. Applied supervised machine learning techniques

The supervised learning techniques employed in this investigation are listed as below:
The chosen techniques relied on three fundamental points:

- The ability to use either classification solvers or regression solvers, or both.
- Scale-sensitive and non-scale-sensitive aspects of these algorithms
- Model bias and variance analysis based on single and ensemble classifiers

Table 3 summarizes the selection criteria for the techniques involved in this study:

| Technique       | Solver type                     | Scaling-sensitive | Class limits |
|-----------------|---------------------------------|-------------------|--------------|
| SV: (SVM+SVR)   | Classification + Regression     | Non-sensitive     | No           |
| LDA             | Classification + Regression     | Non-sensitive     | 2            |
| NB              | Classification                   | Non-sensitive     | No           |
| KNN             | Classification + Regression     | Sensitive         | No           |
| LR              | Regression                      | Non-sensitive     | No           |
| XGBoost         | Classification + Regression     | Non-sensitive     | No           |

The respect of such criteria will undoubtedly make it possible to diversify the options for learning techniques while remaining unbiased as to whether one algorithm has an advantage over another.

2.5. Cross validation
Cross validation is a practical tool for estimating the robustness of machine learning models. It is frequently employed to benchmark and choose a model for a predictive modelling assignment. This approach includes a k parameter that refers to the number of sample groups in which the generated datasets are to be partitioned. In this work, the value of K has been set to five. The value of K is chosen to ensure that each train/test group of data samples is large enough to be significantly representative of the data set. Given that the datasets used in this study are composed of 500 measurements, the value of k=5 will allow us to have 5 data samples of 100 measurements each, which seems satisfactory to us. The aim is to ensure a less biased and less optimistic valuation of the model efficiency, especially during the training phase, which is very critical for the final learning outcomes. It consists first of randomly shuffling the datasets, then split all the observations into k groups of equal size, testing, training, and evaluating the generated groups individually, and finally, summarizing the models performances using the standard evaluation scores.

3. Results
In this section, the results obtained were grouped under the form of different plots, namely the scatter plot of the dataset, the scatter plot of normal and abnormal measures, the data distribution representation of each variable, the training and cross validation scores plots, and finally a time simulation complexity table designed to evaluate the time-duration of each dataset throughout the classification process.

3.1. Normal Distribution
Figure 2 and figure 3 show the scatterplots of the first dataset generated based on the Normal distribution and the dataset variables, binary classified as normal and abnormal, respectively.
Figure 2. Scatterplot of the dataset readings based on the Normal distribution

Figure 3. Scatterplot of the normal and abnormal readings

Figures 4 and 5 illustrate the data distribution as well as histogram representations of the dataset according to the normal distribution of sensed reading values of variables x and y respectively.

Figure 4. Normal distribution of the x readings variable

Figure 5. Normal distribution of the y readings variable

3.2. Poisson Distribution

Figure 6 and figure 7 show the scatterplots of the second dataset generated based on the Poisson distribution and the dataset variables, binary classified as normal and abnormal, respectively.
Figure 6. Scatterplot of the dataset readings based on the Poisson distribution

Figure 7. Scatterplot of the normal and abnormal readings

Figure 8 and figure 9 illustrate the data distribution as well as histogram representations of the dataset according to the Poisson distribution of sensed reading values of variables x and y respectively.

3.3. Uniform Distribution

Figures 10 and 11 show the scatterplots of the second dataset generated based on the Uniform distribution and the dataset variables, binary classified as normal and abnormal, respectively.
Figures 12 and 13 illustrate the data distribution as well as histogram representations of the dataset according to the Uniform distribution of vectors x and y respectively.

3.4. Supervised learning results
Figures 14, 15 and 16 display the curve in the learning model based on the SVM classifier by plotting the training score compared to the cross-validation score for the Normal, Uniform and Poisson distributions, respectively. The plots show that the optimization process of training score depends on the Gamma parameter.
Time duration (in seconds) of the learning process is depicted in Table 4

**Table 4. Time-simulation complexity**

| Classifiers | Solver type | Scaling-sensitive | Class limits |
|-------------|-------------|------------------|--------------|
| SVM         | 0.01s       | 0.02s            | 0.02s        |
| LDA         | 0.02s       | 0.04s            | 0.03s        |
| NB          | 0.02s       | 0.03s            | 0.3s         |
| XGBoost     | 0.06s       | 0.1s             | 0.07s        |
| KNN         | 0.02s       | 0.03s            | 0.02s        |
| LR          | 0.03s       | 0.02s            | 0.02s        |

The classification scores of the different techniques are presented in table 5. The used standards include Precision, Recall and F1-score.

**Table 5. Classification results of the generated datasets fitting respectively Normal, Uniform and Poisson distributions**

| Distribution | Classifier | Classes | Precision | Recall | F1-score |
|--------------|------------|---------|-----------|--------|----------|
| Normal       | SVM        | 0       | 0.86      | 1.00   | 0.92     |
|              |            | 1       | 1.00      | 0.88   | 0.94     |
|              | XGBoost    | 0       | 0.91      | 0.98   | 0.94     |
|              |            | 1       | 0.98      | 0.93   | 0.96     |
|              | NB         | 0       | 0.85      | 0.76   | 0.81     |
|              |            | 1       | 0.86      | 0.92   | 0.89     |
|              | LR         | 0       | 0.86      | 0.88   | 0.87     |
|              |            | 1       | 0.91      | 0.90   | 0.90     |
|              | LDA        | 0       | 0.80      | 0.95   | 0.87     |
|              |            | 1       | 0.96      | 0.83   | 0.89     |
|              | KNN        | 0       | 0.80      | 0.74   | 0.77     |
Table 6 summarizes the accuracy scores of the supervised machine learning techniques used in the study conducted in this paper.

**Table 6.** Achieved accuracy scores of the applied classifiers fitting respectively Normal, Poisson and Uniform distributions

| Classifiers | Accuracy (Normal) | Accuracy (Poisson) | Accuracy (Uniform) |
|-------------|-------------------|--------------------|-------------------|
| SVM         | 0.93              | 0.95               | 0.83              |
| XGBoost     | 0.95              | 0.96               | 1                 |
| NB          | 0.86              | 0.89               | 0.71              |
| LR          | 0.89              | 0.81               | 0.73              |
| LDA         | 0.88              | 0.89               | 0.74              |
| KNN         | 0.83              | 0.81               | 0.77              |

Table 7 illustrates the confusion matrix of the best two classifiers, namely XGBoost and SVM. Note that Class-0 denotes the normal measures, while Class-1 denotes the abnormal ones.

**Table 7.** Confusion matrix of the optimal two classifiers, namely SVM and XGBoost

| Distribution | SVM | XGBoost |
|--------------|-----|---------|
| Class-0      |     |         |
| Class-1      |     |         |
3.5. Discussion

A closer look at the results indicates that the current research appears to validate the view that SVM and XGBoost return very good scores for all distributions-based datasets. The XGBoost technique, known to be one of the best performing classifiers, scored almost perfect. Dealing with the SVM technique often requires the optimization of Gamma and C parameters. From a time-complexity standpoint, the time execution belonging to [0.01,0.02] seconds provides confirmatory evidence this technique ensures very accurate outcomes especially when dealing with a decision-support system that requires short response times. The XGBoost’s seamless overall performance is due to the time-consuming use of decision trees, which last within [0.06,0.1] seconds.

During this learning process, a minimum threshold of 80% has been established, and only models having calculated accuracy scores above this threshold are subsequently accepted. The day-to-day use of IOT devices often requires a combination of new approaches towards data classification and analysis to meet given decision-making requirements. Moreover, many optimization approaches depend entirely on the data collection process. Deciding on the right classification technique requires a comparative analysis of different available approaches. Ant yet, this was very concisely established within this research paper.

As illustrated in table 3, our study considered several parameters that must be considered when choosing a classifier for intelligent systems. Basically, classification is about predicted labels and regression is a matter of predicting a quantity. The selection of a learning technique must satisfy these two analytical grounds while ensuring accurate and time-efficient results. In the era of big data, intelligent systems operate with sensitive data processing on a large scale. This implies analytical approaches that can ensure these tasks. Thus, both SVM/SVR and XGBoost validate the required assessments dealing with data of small, medium, and large sizes. The consensus view seems to state that these two techniques offer:

- The ability to use either classification solvers or regression solvers, or both.
- Handling scale-sensitive and non-scale-sensitive data aspects

IOT systems are applied across multiple areas handling different data distributions. We list several possible use cases in accordance with the distributions mentioned in this paper.

- **Uniform distribution:**
  - Fault tolerance (Nodes are no longer sending authentic sensed data due to environmental issues)
  - Security (Nodes communications through a network, i.e. Attacks)
  - Medical (i.e. Storks, Heart Attack, Falling, Gesture)
  - Smart Grids (i.e. Electrical Cyber Attacks, Billing Errors …)

- **Normal distribution:**
  - Medical (Blood Pressure is known to fit this distribution) [8]
  - Business (Stock returns)
  - Anthropology (Heights and weights of a specific population)

- **Poisson distribution**
  - Web-based platforms (i.e. Content delivery network - CDN)

| Normal          | Class-0 | Class-1 | Class-0 | Class-1 |
|-----------------|---------|---------|---------|---------|
| Class-0         | 42      | 0       | 41      | 1       |
| Class-1         | 7       | 51      | 4       | 54      |

| Poisson         | Class-0 | Class-1 | Class-0 | Class-1 |
|-----------------|---------|---------|---------|---------|
| Class-0         | 19      | 5       | 23      | 1       |
| Class-1         | 0       | 76      | 3       | 73      |

| Uniform        | Class-0 | Class-1 | Class-0 | Class-1 |
|----------------|---------|---------|---------|---------|
| Class-0        | 39      | 0       | 39      | 0       |
| Class-1        | 17      | 44      | 0       | 61      |
4. Conclusion

In this paper, several supervised learning techniques were applied to classify decision-making related data. Because of the sensitivity of studying the internal nature of data-driven behaviour, three probability distributions were used to evaluate possible relationships between a data distribution and the most appropriate learning technique with the aim of selecting a classifier that could be useful for decision-making processes. The results yielded by this study provides convincing evidence that two of the classifiers returned strong results for all datasets and their associated distributions, namely SVM and XGBoost. The evaluation of the selected learning models was based on the classical standards of measurement, while also highlighting the execution time required throughout the end-to-end application. Part of the research perspective associated to this work would be to propose new approaches for optimizing the objective functions and internal parameters of these classifiers to better match them to a given data distribution.

References

[1] N., Hashemi, S., & Hamzeh, A. (2011). A precise statistical approach for concept change detection in unlabeled data streams. *Computers & Mathematics with Applications, 62*(4), 1655-1669.

[2] Lip, G. Y., Nieuwlaat, R., Pisters, R., Lane, D. A., & Crijns, H. J. (2010). Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest, 137*(2), 263-272.

[3] O'Mahony, C., Jichi, F., Pavlou, M., Monserrat, L., Anastasakis, A., Rapezzi, C., ... & Omar, R. Z. (2014). A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (HCM risk-SCD). *European heart journal, 35*(30), 2010-2020.

[4] Ghamari, M., Janko, B., Sherratt, R. S., Harwin, W., Piechocki, R., & Soltanpur, C. (2016). A survey on wireless body area networks for ehealthcare systems in residential environments. *Sensors, 16*(6), 831.

[5] Wang, H., Choi, H. S., Agouline, N., Deen, M. J., & Hong, J. W. K. (2010, October). Information-based sensor tasking wireless body area networks in U-health systems. In *2010 International Conference on Network and Service Management* (pp. 517-522). IEEE.

[6] Maciuca, A., Popescu, D., Strutu, M., & Stamatescu, G. (2013, September). Wireless sensor network based on multilevel femtocells for home monitoring. In *2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)* (Vol. 1, pp. 499-503). IEEE.

[7] Nielsen, J. J., Madueño, G. C., Pratas, N. K., Sørensen, R. B., Stefanovic, C., & Popovski, P. (2015). What can wireless cellular technologies do about the upcoming smart metering traffic?. *IEEE Communications Magazine, 53*(9), 41-47.

[8] Cheffira, R., & Rakrak, S. (2019). Enhancing Decision Making through Combined Classification Techniques and Probabilistic Data Analysis for Ubiquitous Healthcare Anomaly Monitoring. *Applied Sciences, 9*(22), 4802.

[9] Xu, X., Zhao, F., Wang, W., & Tian, H. (2016). Adaptive sampling-based information collection for wireless body area networks. *Sensors, 16*(9), 1385.

[10] Mohamed, M., Joseph, W., Vermeeren, G., Tanghe, E., & Cheffena, M. (2019). Characterization of dynamic wireless body area network channels during walking. *EURASIP Journal on Wireless Communications and Networking, 2019*(1), 104.

[11] Lee, M., Gatton, T. M., & Lee, K. K. (2010). A monitoring and advisory system for diabetes patient management using a rule-based method and KNN. *Sensors, 10*(4), 3934-3953.

[12] Bodhe, M. K. D., Sawant, R. R., & Kazi, M. A. (2014). A proposed mobile based health care
system for patient diagnosis using android OS. *International Journal of Computer Science and Mobile Computing, 3*(5), 422-427.

[13] Laya, A., Alonso, L., Chatzimisios, P., & Alonso-Zarate, J. (2015, June). Massive access in the random access channel of LTE for M2M communications: An energy perspective. In *2015 IEEE International Conference on Communication Workshop (ICCW)* (pp. 1452-1457).

[14] Kamio, T., Van, T., & Masumune, K. (2017). Use of machine-learning approaches to predict clinical deterioration in critically ill patients: a systematic review. *International Journal of Medical Research & Health Sciences, 6*(6), 1-7.