





Article

# On Predicting Ticket Reopening for Improving Customer Service in 5G Fiber Optic Networks

Lorenzo Ricciardi Celsi <sup>1,2,\*</sup> , Andrea Caliciotti <sup>2,3</sup> , Matteo D'Onorio <sup>4</sup> , Eugenio Scocchi <sup>5</sup> ,  
Nour Alhuda Sulieman <sup>6</sup> and Massimo Villari <sup>6</sup>

- <sup>1</sup> ELIS Innovation Hub, Via Sandro Sandri 81, 00159 Roma, Italy
  - <sup>2</sup> Department of Computer, Control, and Management Engineering Antonio Ruberti, Sapienza Università di Roma, Via Ariosto 25, 00185 Roma, Italy; andrea.caliciotti@enel.com
  - <sup>3</sup> Enel Green Power S.p.A., Viale Regina Margherita 125, 00198 Roma, Italy
  - <sup>4</sup> DIAEE, Sapienza Università di Roma, Corso Vittorio Emanuele II 244, 00186 Roma, Italy; matteo.donorio@uniroma1.it
  - <sup>5</sup> ERG S.p.A., Via Bissolati 76, 00187 Roma, Italy; escocchi@erg.eu
  - <sup>6</sup> Dipartimento di Scienze Matematiche e Informatiche, Scienze Fisiche e Scienze della Terra, Università di Messina, Piazza Pugliatti 1, 98122 Messina, Italy; nosulieman@unime.it (N.A.S.); mvillari@unime.it (M.V.)
- \* Correspondence: l.ricciardicelsi@elis.org

**Abstract:** The paper proposes a data-driven strategy for predicting technical ticket reopening in the context of customer service for telecommunications companies providing 5G fiber optic networks. Namely, the main aim is to ensure that, between end user and service provider, the Service Level Agreement in terms of perceived Quality of Service is satisfied. The activity has been carried out within the framework of an extensive joint research initiative focused on Next Generation Networks between ELIS Innovation Hub and a major network service provider in Italy over the years 2018–2021. The authors make a detailed comparison among the performance of different approaches to classification—ranging from decision trees to Artificial Neural Networks and Support Vector Machines—and claim that a Bayesian network classifier is the most accurate at predicting whether a monitored ticket will be reopened or not. Moreover, the authors propose an approach to dimensionality reduction that proves to be successful at increasing the computational efficiency, namely by reducing the size of the relevant training dataset by two orders of magnitude with respect to the original dataset. Numerical simulations end the paper, proving that the proposed approach can be a very useful tool for service providers in order to identify the customers that are most at risk of reopening a ticket due to an unsolved technical issue.

**Keywords:** 5G fiber optic networks; data-driven service assurance; next generation networks; predictive analytics



**Citation:** Ricciardi Celsi, L.; Caliciotti, A.; D'Onorio, M.; Scocchi, E.; Sulieman, N.A.; Villari, M. On Predicting Ticket Reopening for Improving Customer Service in 5G Fiber Optic Networks. *Future Internet* **2021**, *13*, 259. <https://doi.org/10.3390/fi13100259>

Academic Editor: Michael Mackay

Received: 17 September 2021

Accepted: 4 October 2021

Published: 9 October 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Effective customer care and data-driven service assurance have become a vital need for telecommunication companies, especially as regards the progression towards automating the management of technical tickets: in addition, this enables a thorough and objective evaluation of the performance of the service assurance functions, based on generated reports and prescribed Key Performance Indicators (KPIs), which implies increased productivity, improved quality of service and, in some cases, even personalized satisfaction of the end user.

More in detail, designing a successful data-driven service assurance system requires:

- in-house design and development in order to offer a solution that is at the same time cheap and adaptable to all service needs;
- tight integration with the customer portal in order to allow the access to all relevant online services with the same login credentials;

- support to all activities that are characteristic of an incident management process, such as opening, classifying, assigning, solving, closing and archiving an IT incident;
- enabling the exchange of files and comments regarding an IT incident, as this makes the served customers satisfied by allowing them to interact with IT agents in order to clarify the problem or understand the solution;
- provision of email notifications that allow for keeping all involved parties informed on any ticket update;
- automatic classification and ticket rerouting to the relevant IT agents, which is expected to significantly minimize resolution time of each ticket and consequently enhance user satisfaction;
- generation of reports to measure the KPIs that are chosen to evaluate the success of the customer care service.

For detailed references, the advantages and challenges of using data-driven service assurance in an organization have already been explored in the Information Technology Infrastructure Library (ITIL) framework of best practices for delivering IT services (see [1–8]). Moreover, the positive influence of ticketing services on the implementation of the incident management process is stressed in [1–10]. Indeed, being able to automate several key activities such as identification, prioritization, assignment, diagnosis and closure of technical tickets plays a relevant role in enhancing the said process. In addition, data-driven service assurance (see [4–6]) facilitates the measurement and improvement of several important KPIs for the IT business processes, such as the percentage of incidents detected and solved in the first attempt, the mean number of incidents that occurred per day and the average lifetime of an incident. In particular, in [8], it is clearly stated that automatic incident classification proves to be extremely effective at minimizing ticket resolution time. Providing an automated solution to the challenge of ticket classification is a relevant emerging task also according to [10].

Some related additional features that are in part beyond the scope of this work but testify the interest of the scientific and industrial community are the multimedia chat service architecture introduced in [11], as well as rule-based reasoning for fault diagnosis and visual dashboards for helping desk tickets monitoring, which is illustrated in [12,13]. In addition, a probabilistic framework for IT ticket annotation and search based on natural language processing is introduced in [14–17]. In [15], a predictive model based on Support Vector Machines (SVMs) and K-Nearest Neighbours (KNN) is discussed with the aim to automate incident categorization with the specific help of ticket description and other relevant ticket attributes. In a similar way, in [18], the dispatch of a ticket to the correct resolution group is successfully automated by means of a tool that combines SVMs and discriminative term-based classification techniques. Alternatively, Multinomial Naive Bayes (MNB) and Softmax Regression Neural Network (SNN) are used in [19] for text classification purposes aimed at categorizing user tickets. Finally, several methods to detect duplicate tickets/bugs are proposed in [20–22].

In particular, this paper, with respect to the emerging need for preventing customers from issuing a request for technical ticket reopening on customer service platforms of telecommunication companies, provides the following contributions:

- the identification of relevant correlations between the reopening of a technical customer service ticket, on the one hand, and Quality of Service (QoS) parameters of the 5G fiber optic networks, on the other hand, based on the actual use the customer is currently making of the fixed network itself;
- based on such correlations, the design of a data-driven model capable of predicting whether a customer will call the assistance service once again even though his/her technical ticket has already been closed.

Incidentally, reopened tickets are to be considered as those tickets that were formerly solved and have been reopened [23].

The proposed approach may prove particularly useful in the domain of 5G enabling technologies. Indeed, even with the advent of 5G, optical fiber is the most suitable means

for wireless backhaul networks. Indeed, even in networks where this is not the case, the wireless backhaul actually has to be connected into a fiber backhaul. For this reason, fiber technology is increasingly being preferred for the so-called fronthaul, especially when it comes to connecting the dense mesh of 5G small cells. There are several benefits, such as increased speeds matched with lower attenuation, significant immunity with respect to electromagnetic interference, relatively small size, and practically unlimited potential in terms of bandwidth. Hence, customer service in order to address any technical issue relative to the Quality of Service (QoS) perceived in 5G fiber optic networks has a critical role, especially with the advent of the emerging Fixed Wireless Access (FWA) paradigm [24].

The paper contribution lies in the fact that the effectiveness of the proposed approach is evaluated on the customer complaints that arise in conditions of intensive usage of the fixed network of a major Italian network operator. Indeed, according to IBM analyses in [25] and to [26], by automating up to 85% of the customer service process thanks to the usage of predictive tools such as the one presented in this work, an increase in efficiency up to 90% can be obtained, together with a reduction in the operating costs between 25% and 30%.

Similar machine learning methods have already been used in order to solve resource allocation problems in order to improve the perceived QoS in [27–29]. In more detail, Pietrabissa et al. have already focused on distributed load balancing in Software Defined Networking relying on Lyapunov-based decision-making algorithms in [27], on optimal buffer allocation for guaranteeing QoS in multimedia Internet broadcasting for mobile networks in [28], and on predictor-based control design for improving Quality of Experience in delay-sensitive Future Internet frameworks in [29]. In contrast with the cited works, this paper specifically focuses on the prediction of the ticket reopening phenomena that characterize fixed networks and therefore also 5G fiber optic networks. To this aim, we exploited several machine learning approaches and compared the obtained performance results.

The paper is organized as follows: Section 2 introduces the so-called Analytical Base Table (ABT), namely the dataset to be given as an input to the predictive model for training and test purposes, as well as the data collection and preparation activity that has been carried out to assemble said ABT. Section 3 presents the performance achieved by different machine learning based classifiers, comparing the results obtained on the original dataset against those obtained on the reduced dataset. Section 4 discusses a further dimensionality reduction effort before introducing the Bayesian network classifier whose performance is the best in class. Concluding remarks end the paper.

## 2. Data Collection, Data Preparation and Analytical Base Table

With the aim of effectively predicting ticket reopening relative to the QoS perceived in 5G fiber optic networks—namely, according to the emerging FWA paradigm—the methodological approach inspiring this work can be structured as follows.

- A first *data collection* and *data preparation* effort were made, by collecting, aggregating, cleaning and preparing the relevant data for the subsequent processing phase, resulting in a homogeneous ABT to be fed to the predictive analytics/classification engine that was subsequently designed and developed. This first activity was carried out in the framework of an extensive joint research initiative on Next Generation Networks between ELIS Innovation Hub and Vodafone over the years 2018–2021.
- Then, the *KPI definition* was necessary, that is, the identification of the relevant KPIs that allows for quantifying the benefits that are ultimately yielded by the adoption of the proposed predictive analytics engine. In this respect, we chose to evaluate the performance of the proposed predictive analytics engine in terms of accuracy, Gini coefficient, Youden index, and Area Under the ROC Curve (AUC).
- The design and training of the *predictive analytics engine* followed, together with the evaluation of the performance obtained on a suitable test dataset.

In more detail, the relevant input data are collected from two heterogeneous data sources,

1. the former related to the Virtual Unbundled Local Access (VULA) technology—VULA offers a means to any licensed network operator to effectively join the ultrabroadband network infrastructure of the backbone provider by virtually accessing the last mile only;
2. the latter related to the Sub-Loop Unbundling (SLU) technology—SLU is alternative to VULA, yet it joins the network infrastructure of the backbone provider only at the copper section, thus exhibiting lower performance than VULA.

The *original* dataset  $X$  is therefore represented by an  $m \times n \times t$  matrix aggregating the inputs from both VULA and SLU technologies. After being suitably cleaned, it reports  $n = 307$  network QoS parameters collected over a period of  $t = 30$  days of intensive usage for a group of  $m = 600,000$  users.

The  $i$ -th row of  $X$  (for  $i = 1, \dots, m$ ),

$$x_i = (x_{u,v}^i)_{\substack{u=1,\dots,n \\ v=1,\dots,t}} \quad (1)$$

associated with a user  $i$  that the customer service has already closed a ticket for (since the aim of the paper is to predict ticket *reopening*), accounts for the values of the  $n$  network QoS parameters perceived by user  $i$  on day  $v$ . Among the  $m$  users, in the considered period, 25% reopened a ticket due to a technical issue that is still unsolved even though a ticket in that respect had already been closed. In particular, data collection was carried out with a specific data matching criterion: namely, in the case of a reopened ticket, only the last data point—i.e., the last values of the  $n$  features—that is, the temporally closest to ticket reopening for the considered user is collected and stored in the dataset  $X$ , whereas, for all other tickets—that have already been closed and have not been reopened yet—all data in the time interval between ticket closing and the last possible sampling instant are collected and stored in the dataset  $X$ .

In order to obtain a smaller dataset, with reduced dimensionality ( $p < n = 307$ ) but with very similar information content, we first performed feature reduction on the original dataset  $X$  by removing all features with very low variance, i.e., proving to be redundant when it comes to estimating the probability of ticket reopening.

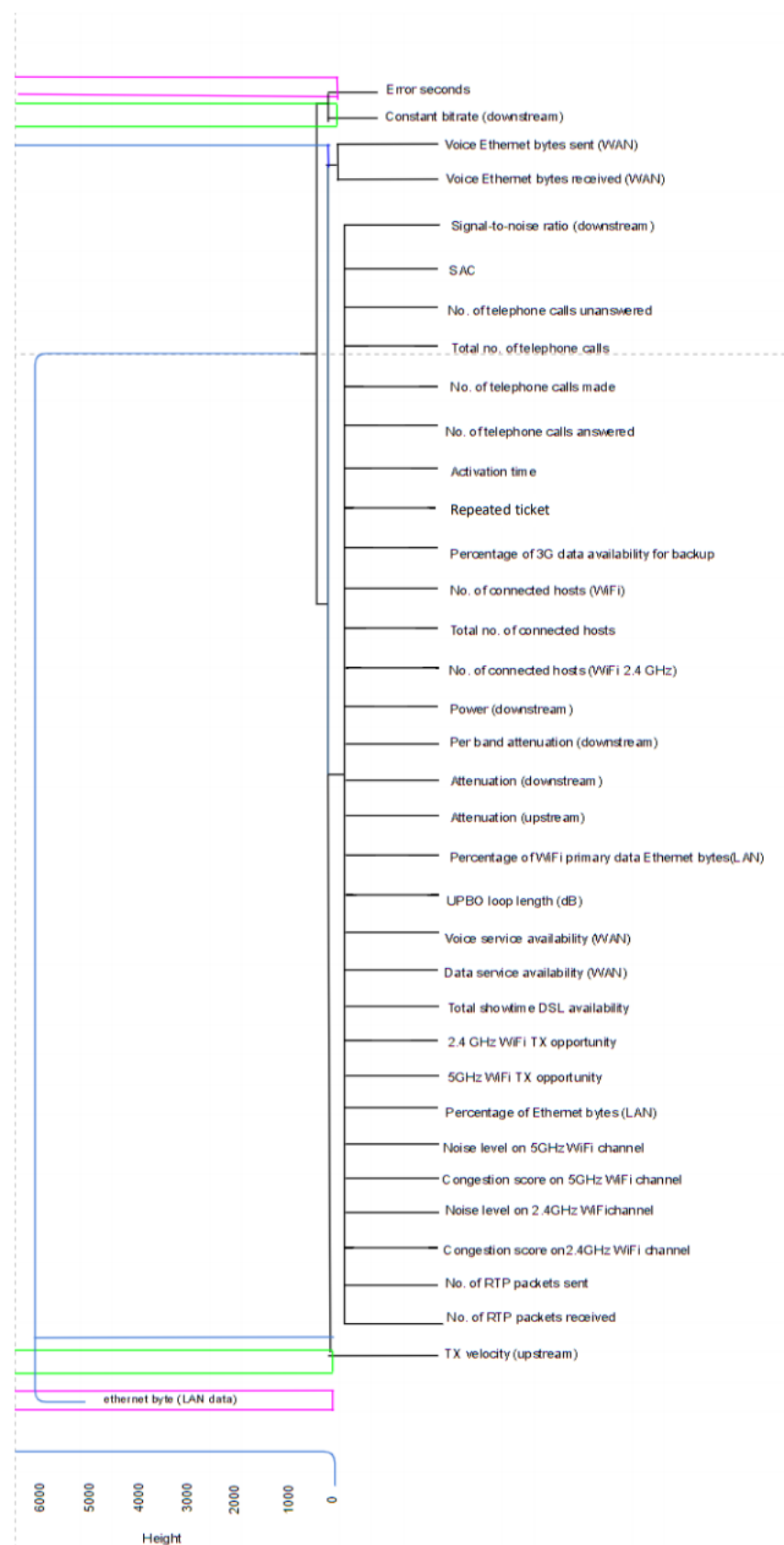
Then, we carried out a linear correlation analysis among the features which, however, did not yield any relevant results; this is why we chose to resort to *hierarchical clustering* in order to shed light on any existing nonlinear correlations.

The result of hierarchical clustering, performed on the portion of  $X$  accounting for the VULA and SLU inputs alternatively, is represented by the *dendrograms* shown in Figures 1 and 2, which represent the resulting hierarchies of clusters, by reporting:

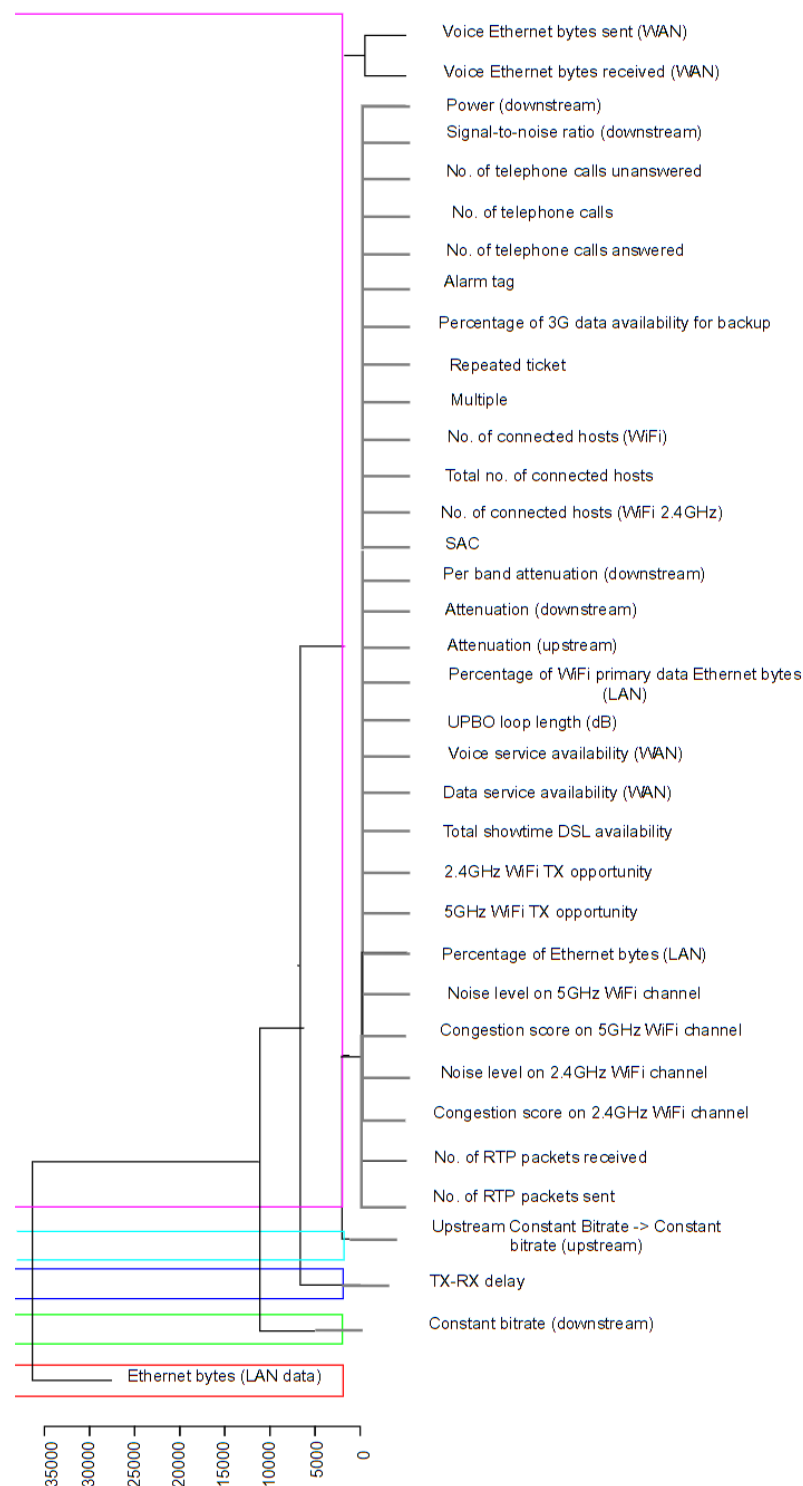
- along the  $x$ -axis the logical distance between clusters according to the Manhattan metric;
- along the  $y$ -axis the hierarchical level of aggregation (denoted with 'Height' in Figures 1 and 2) measured in integer positive values.

In both dendrograms, important correlations emerged—which had not been as evident from linear correlation analysis at all—between ticket reopening (accounted for by the 'Repeated ticket' variable) and some network parameters. The most relevant evidence is related to the correlations of the Repeated ticket variable with the following:

- Percentage of 3G data availability for backup;
- Alarm tag;
- Number of connected hosts (WiFi);
- Total number of connected hosts.



**Figure 1.** Dendrogram showing the results of hierarchical clustering on the VULA portion of the original dataset X.



**Figure 2.** Dendrogram showing the result of hierarchical clustering on the SLU portion of the original dataset X.

As regards the last two, it is reasonable that the higher the number of connected devices, the higher the performance degradation.

A specific remark has to be made relative to the tight correlation of the Repeated ticket variable with the Alarm tag. Indeed, the Alarm tag parameter is a boolean variable which is valued 1 if the so-called customer-premises equipment parameter hits the alarm and 0 otherwise according to two alarm-triggering rules:



- if in almost 4 out of 8 measurements, the SAC parameter goes below threshold;
- if in 3 consecutive measurements, the SAC parameter goes below the threshold.

The SAC variable, which specifically accounts for attenuation-related QoS, can have values in the range from 1 to 6 (where 1 accounts for ‘bad’ and 6 for ‘good’) and is obtained by comparing other two parameters, namely signal-to-noise ratio (downstream) and attenuation (downstream), respectively.

The correlation between Repeated ticket and Alarm tag can be considered as *primary* since it is the tightest one. Therefore, in addition to the primary correlations highlighted by the dendrograms, starting from the tight correlation between Repeated ticket and Alarm tag, other *secondary* correlations also emerge between the Repeated ticket variable, on the one hand, and SAC, signal-to-noise ratio (downstream) and attenuation (downstream), on the other hand. In particular, by observing the relationship among these variables, we can notice the following high degrees of correlation:

- −47% correlation between Alarm tag and SAC for both VULA and SLU technologies;
- +65% correlation between SAC and signal-to-noise ratio for the SLU technology and +68% correlation between the same variables for the VULA technology;
- −55% correlation between SAC and attenuation (downstream) for the SLU technology and −45% between the same variables for the VULA technology.

It is also possible to evaluate the similarity degree between the two dendrograms by computing the so-called *entanglement* parameter—ranging from 0, which accounts for no entanglement—to 1—which accounts for full entanglement. A low entanglement score implies a good alignment degree between the QoS performance of the two technologies with respect to the time period and set of users observed. In the considered case, the entanglement parameter is evaluated to be equal to 0.175, thus allowing us to consider as ABT a reduced dataset  $X^{red}$  extracted from the original one  $X$ . In particular, the Repeated ticket variable occupies very similar positions in both dendrograms. In light of this, it is reasonable to adopt the same predictive algorithm to predict ticket reopening for both VULA and SLU technologies.

As a result, the following lessons can be learned from the data preparation activity carried out in this section:

- no relevant linear correlations emerge between the Repeated ticket variable and QoS parameters;
- the variables exhibiting primary nonlinear correlations with the Repeated ticket variable are the percentage of 3G data availability for backup and the Alarm tag;
- the variables exhibiting secondary nonlinear correlations with the Repeated ticket variable are SAC, signal-to-noise ratio, and attenuation in downstream;
- the datasets related to the VULA and SLU technologies are very much correlated with each other due to the 0.175 entanglement coefficient computed between the two dendrograms in Figures 1 and 2.

Hence, the *reduced* dataset  $X_{red}$  we are going to resort to hereinafter can be regarded as a  $p \times n \times t$  matrix reporting the following  $p = 15$  network QoS parameters for the same group of  $m$  users at time  $t$ : (i) SAC, (ii) signal-to-noise ratio (downstream), (iii) attenuation (downstream), (iv) constant bitrate (downstream), (v) downstream maximum rate, (vi) Ethernet bytes (LAN data), (vii) percentage of Ethernet bytes (LAN), (viii) percentage of WiFi primary data Ethernet bytes (LAN), (ix) attenuation (upstream), (x) current bitrate (upstream), (xi) upstream maximum rate, (xii) power (upstream), (xiii) UPBO (upstream power back-off) loop length (dB), (xiv) ticket close code, and (xv) repeated ticket.

The ticket close code has not been mentioned so far but was already also present in the original dataset  $X$ : it testifies that the network operator has acknowledged the causes that are attributable to the variations in the QoS parameters for which the user is complaining. The ticket close code variable may take one of the following values: (a) activity on the OCA protocol, (b) existence of a known problem that is pending for resolution, (c) line/device problem needing for device reboot, (d) line/device problem

needing for device reset, (e) unexploited link/GNP (Geographic Number Portability), (f) LNI-minimum bitrate, (g) maximum obtainable performance, (h) no trouble found, (i) OLO2OLO problem—OLO2OLO is the Italian platform allowing the migration of access lines between different network operators insisting on the same network infrastructure—, (j) performance degradation resulting from monitoring, (k) macroproblem due to the network infrastructure, (l) known network outage, (m) network outage detected as a result of monitoring, (n) access degradation of the network infrastructure resulting from monitoring, (o) access network degradation relative to the backbone provider, (p) access network degradation relative to the network operator, and (q) wrong assignment of the network service to a user. For the sake of completeness, in Table 1, we show the relative frequency of each ticket close code in the considered dataset.

**Table 1.** Relative frequencies of the values exhibited by the ticket close code variable.

Ticket Close Code	Relative Frequency
Known network outage	42.99%
Network outage—monitoring	22.35%
Access network degradation—backbone provider	13.14%
Network infrastructure macroproblem	12.48%
No trouble found	4.10%
Line/device problem—reset device	1.74%
LNI-minimum bitrate	1.71%
Known problem pending for resolution	0.44%
OLO2OLO problem	0.41%
Activity on the OCA protocol	0.30%
Access network degradation—network operator	0.10%
Maximum obtainable performance	0.10%
Network infrastructure access degradation—monitoring	0.04%
Performance degradation—monitoring	0.04%
Unexploited link/GNP	0.03%
Line/device problem—reboot device	0.02%
Wrong assignment	0.02%

We now present our predictive model. In more detail, we are going to address the following classification problem: *is a user at risk of reopening a ticket that was previously closed even though the related technical issue (in terms of perceived QoS) was still not solved?*

### 3. Classification for Predicting Ticket Reopening

In this section, we show the performance achieved by different machine learning based classifiers trained both on the original dataset  $X$  and on the reduced dataset  $X_{red}$  in order to predict if a monitored ticket will be reopened or not. This allows us to evaluate the effectiveness of the dimensionality reduction activity discussed in the previous section.

#### 3.1. Different Approaches to Classification

Machine learning is a branch of artificial intelligence based on the idea that systems can learn from data and make reasonable decisions with minimal human intervention. In contrast with many statistical modeling approaches, which generally value inference over prediction, the focus of machine learning is predictive accuracy (see [30]). High predictive accuracy is usually achieved by training complex predictive models, often involving advanced numerical optimization routines, on a very large number of training examples.

According to the survey provided in [31], in this paper, the following supervised classification techniques are considered: decision tree, random forest, boosting, logistic regression, Artificial Neural Network (ANN) and SVM.

The chosen architecture for the decision tree based classifier is inspired by [32]. The random forest based approach to classification follows [33]. The setup for boosting resembles [34], whereas the logistic regression one is inspired by [35].



Instead, in the case of the ANN, we consider a two-layer fully connected network. For the hidden layer, we resort to ReLU nonlinearity, whereas, for the output layer, we have a Softmax loss function. The size of the neural network for the input and output layers is dependent on the input dataset ( $X$  and  $X_{red}$ , alternatively) and classes respectively, while the hidden layer is arbitrarily set.

Finally, the SVM classifier follows the classical approach from [36].

### 3.2. Numerical Simulations and Results

In this subsection, we compare the different classification approaches (described in Section 3.1) in order to predict ticket reopening via supervised learning. According to Section 2, we consider both datasets: the first one ( $X$ ) in the original form and the second one in the reduced form  $X_{red}$  (through feature selection analysis).

Given a month (i.e., 30 days) of data collected according to the format discussed in Section 2, we reordered the dataset by picking six groups of four days as training sets ( $k \in \{1, 6, 11, 16, 21, 26\}$ ), denoting them, with a slight abuse of notation, with  $X_{training}[k]$  in the case of the original dataset and with  $X_{training}^{red}[k]$  in the case of the reduced dataset, that is,

$$X_{training}[k] := \begin{bmatrix} x_{u,(v=k)}^i \\ x_{u,(v=k+1)}^i \\ x_{u,(v=k+2)}^i \\ x_{u,(v=k+3)}^i \end{bmatrix}, \quad \forall u, \quad k \in \{1, 6, 11, 16, 21, 26\}, \quad i = 1, \dots, m, \quad (2)$$

and analogously for  $X_{training}^{red}[k]$ .

We then considered the dataset portion relative to each of the remaining six days of the considered month ( $q = k + 4$ ) as a one-day subset of the ABT providing a suitable test set, denoted with

$$X_{test}[q] := \begin{bmatrix} x_{u,(v=q)}^i \end{bmatrix}, \quad \forall u, \quad k \in \{1, 6, 11, 16, 21, 26\}, \quad i = 1, \dots, m, \quad (3)$$

in the case of the original dataset, and with  $X_{test}^{red}[q]$  defined analogously, in the case of the reduced dataset.

In order to ensure the statistical robustness of the learned models, we proceeded in the following way. We first trained each classification algorithm on  $X_{training}[k]$  and  $X_{training}^{red}[k]$  alternatively, in order to obtain the learned models for each iteration  $k$  (training phase). Then, we tested each model learned at iteration  $k$  on the one-day test set  $X_{test}[q]$  in the case of the original dataset, and on the one-day test set  $X_{test}^{red}[q]$  in the case of the reduced dataset (test phase).

Eventually, we measured the KPIs listed below for each couple  $(k, q)$  of training and test sets and we reported in Tables 2 and 3 the average KPI values over all  $(k, q)$  couples.

For both data preparation and supervised learning algorithms, all codes are written in R. All the simulation runs were performed on a dual-core Intel Core i7-7500U 2.70GHz (up to 3.50 GHz) processor equipped with 16 GB RAM and running Ubuntu 18.04.

Numerical results are provided in terms of *Accuracy*, *Gini coefficient*, *Youden index* and *AUC*.

- *Accuracy* [37]: the accuracy measure tells how well a machine learner, which learned the hypothesis  $h$  as the approximation of the target classification function  $V$ , performs in terms of classifying a novel unseen example correctly. The true error of hypothesis  $h$  is the probability that it will misclassify a randomly drawn example  $x$ , that is,

$$\text{error}(h) = \Pr[V(x) \neq h(x)]. \quad (4)$$

With this in mind, accuracy has the following definition:

$$Accuracy = \frac{\text{number of correct predictions}}{\text{total number of predictions}}. \quad (5)$$

For binary classification, as in the considered case, accuracy can also be calculated in terms of positives and negatives as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (6)$$

where  $TP$ ,  $TN$ ,  $FP$  and  $FN$  are the number of true positives, true negatives, false positives and false negatives, respectively.

- *Gini coefficient* [37]: It provides a measure of the degree or probability of the target variable being wrongly classified. It has values between  $-1$  and  $1$ . The closer it is to  $1$  the better.
- *Youden index* [37]: it is a goodness-of-fit index that represents the maximum separation between the model Receiver Operating Characteristic (ROC) curve and the baseline ROC curve.
- *AUC* [37]: an ROC curve shows the probability distribution just for the events, the positive class. It compares the events predicted correctly (true positives) against the events predicted incorrectly (false positives). The higher the sensitivity (true positive rate close to  $1$ ) and the lower the false negative rate (specificity close to  $1$ ), the better the model. The AUC gives us the area under the ROC curve: the greater the area, the higher the index.

**Table 2.** Numerical results for the dataset  $X$ .

Classifier Trained on the Original Dataset $X$	Accuracy	Gini Coefficient	Youden Index	AUC
Decision Tree	69%	0	0.18	0.5
Random Forest	75%	0.86	0.80	0.93
Gradient Boosting	71%	0.79	0.79	0.89
Logistic Regression	70%	0	0.18	0.5
ANN	66%	0.05	0.06	0.53
Support Vector Machine	73%	0.79	0.79	0.89

**Table 3.** Numerical results for the dataset  $X_{red}$ .

Classifier Trained on the Original Dataset $X_{red}$	Accuracy	Gini Coefficient	Youden Index	AUC
Decision Tree	79%	0.1	0.20	0.55
Random Forest	80%	0.9	0.83	0.95
Gradient Boosting	79%	0.8	0.84	0.90
Logistic Regression	81%	0.12	0.25	0.56
ANN	67%	$-0.12$	0.06	0.46
Support Vector Machine	69%	0.78	0.78	0.89

The best one in the first case is the random forest algorithm. In the second case, the most accurate is the logistic regression algorithm, but the best performing one in general remains the random forest algorithm.

#### 4. A Bayesian Network Classifier Trained on a Further Reduced Dataset

We now propose another data-driven classification model, namely based on a Bayesian network, aimed at improving the performance already obtained on the dataset  $X_{red}$  by means of a further dimensionality reduction, namely resorting to the further reduced dataset  $X'_{red}$ .

#### 4.1. Bayesian Network Classifier

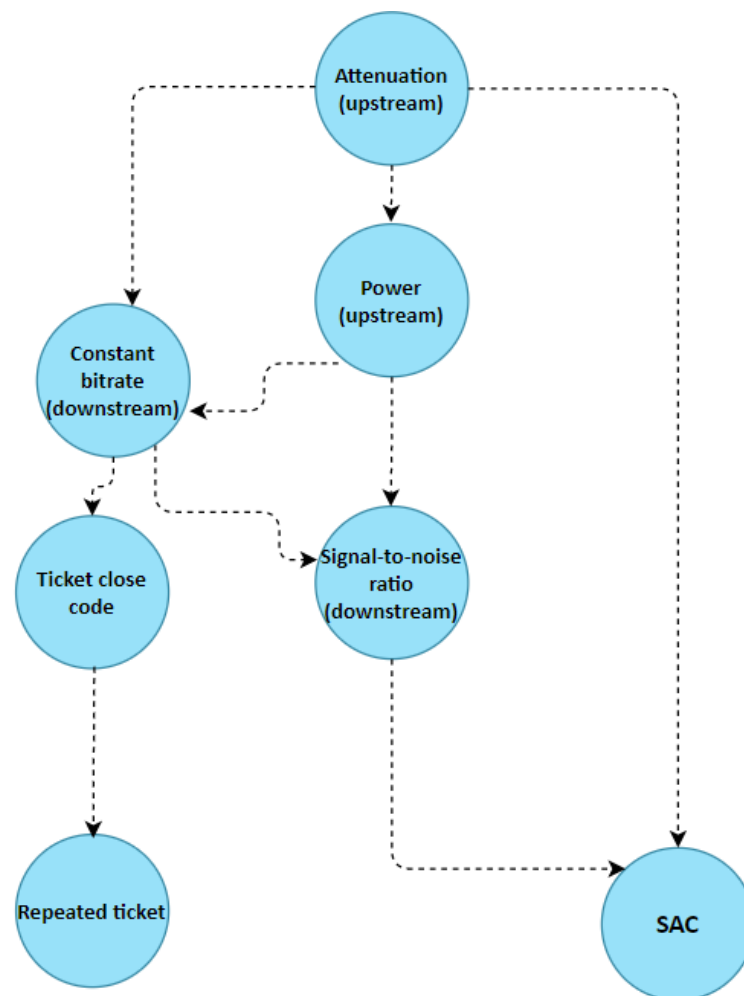
Based on the reduced dataset, we trained a classifier resorting to a Bayesian network.

A Bayesian network is a probabilistic graphical model that, by representing a set of variables and their conditional dependencies via a directed acyclic graph  $\mathcal{G}$ , allows for predicting the likelihood that one of several possible known causes is the contributing factor behind the occurrence of a specific event. In the considered case, the aim is that of predicting if a combination of network QoS parameters belongs to the discrete class variable Repeated ticket.

In more detail, we learned a Naive Bayes network structure  $\mathcal{G}$  as in Figure 3, revolving around the following input variables, which therefore compose the new reduced dataset  $X'_{red}$  as a  $p' \times n$  matrix, with  $p' = 7$ :

- Repeated ticket;
- Ticket close code;
- Signal-to-noise ratio (downstream);
- SAC;
- Constant bitrate (downstream);
- Power (upstream);
- Attenuation (upstream).

We set the Upstream Attenuation node as root node and the SAC and Repeated ticket nodes as leaf nodes.



**Figure 3.** Naive Bayes network structure  $\mathcal{G}$  behind the chosen Bayesian network classifier.

The variables' alarm tag and percentage of 3G data availability for backup have been preliminarily excluded from the ABT because, from a statistical viewpoint, they were not suitable for the algorithm procedure that is behind training a Bayesian network classifier.

In addition, before training the classifier, we chose to perform *discretization*—referred to as the process of grouping values into intervals in order to limit the number of possible states—on the input data according to the type and distribution of each variable, in order to optimize the performance in the creation of the Bayesian network graph. The following two methods were applied onto the continuous variables of the ABT: namely, quantile (subdivision by frequency) and uniform (subdivision into a suitable number of groups of the same size) discretization.

Quantile discretization was performed onto the constant bitrate (downstream), signal-to-noise ratio (downstream), attenuation (upstream) and power (upstream) variables, by grouping the values of each variable into four same size bins, split based on percentiles.

Uniform discretization, instead, was performed onto the ticket close code and SAC, grouping the values of each variable into four same-width discrete bins depending on the span of possible values for each considered variable.

The Repeated ticket variable was discretized into two disjoint bins, of which 15% are repeated tickets and the rest are non-repeated.

Table 4 shows the characteristics of the Bayesian network in detail. The value of the Pearson correlation coefficient (denoted with 'Strength' in the table) indicates the existing degree of correlation between the variables considered: the closer this value is to 1, the greater the correlation between the variables. On the other hand, the 'Direction' column indicates the degree of reliability of the links that introduce a hierarchy between the variables: in this case, too, the closer the value is to 1, the more the direction of the link accounting for the existing relationship between the considered variables is reliable.

**Table 4.** Pearson correlation coefficient between the variables in  $X'_{red}$  and degree of reliability of the links characterizing the learned Naive Bayes network structure  $\mathcal{G}$ .

From	To	Strength	Direction
Signal-to-noise ratio (downstream)	SAC	1.0	0.93
Power (upstream)	Signal-to-noise ratio (downstream)	1.0	0.9
Attenuation (upstream)	SAC	1.0	1.0
Attenuation (upstream)	Power (upstream)	1.0	0.7
Power (upstream)	Signal-to-noise ratio (downstream)	1.0	0.6
Attenuation (upstream)	Constant bitrate (downstream)	1.0	1.0
Power (upstream)	Constant bitrate (downstream)	0.87	0.85
Constant bitrate (downstream)	Ticket close code	0.93	0.9
Ticket close code	Repeated ticket	0.93	0.95

It is clear from Figure 3 and Table 4 that the correlation coefficient with the variables closest to the Repeated ticket variable is always greater than 0.93, which implies that the proposed tree structure can be considered as highly reliable for our classification purpose.

The combinations of conditional probabilities calculated by the Bayesian network as a result of discretization generate a number of scenarios to which it is possible to associate the probability of occurrence of the event of ticket reopening. For the considered reduced dataset  $X'_{red}$ , more than one thousand different simulation scenarios were generated. Namely, the 13 intervals shown in Table 5 are combined with the 17 possible causes identified within the ticket close code variable. Constant bitrate in downstream is measured in bits per second. Attenuation in upstream is measured in dB and power in upstream is measured in dBmV.

**Table 5.** Relevant distributions of the variables of  $X'_{red}$  into suitable bins as a result of discretization.

Variable	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
Constant bitrate (downstream) [bits per sec]	<36,200	[36,200; 54,000]	[54,000; 64,200]	[64,200; 87,100]	>87,100
Attenuation (upstream) [dB]	<10.2	[10.2; 16.1]	[16.1; 21.1]	>21.1	-
Power (upstream)	<-19.2	[-19.2; -8.6]	[-8.6; -0.7]	>0.7	-

**Remark 1.** The number of relevant scenarios may vary depending on the discretization type and the number of tickets associated with the different combinations. This has proven to be the best choice, given the characteristics of the considered dataset.

#### 4.2. Model Performance Evaluation and Discussion on the Results

For the purpose of evaluating the model performance, we adopted the same approach as discussed at the beginning of Section 3.2. However, in order to further test the robustness of the Bayesian network classifier, we also carried out the experiment of creating 1000 random pairs of training/test sets according to the 70/30 rule, i.e., 70% of the  $X'_{red}$  dataset was used for training purposes and the rest for testing. The measured KPIs were very similar, thus testifying the effectiveness of the Bayesian network classifier as well as its robustness. The average values of the KPIs measured throughout these tests are reported below in Table 6 in order to compare them with the results of the different classifiers trained in Section 3.

The validation process was completed by comparing the accuracy measure achieved by the Bayesian network classifier against the performance of the other classifiers.

From Table 6, it is clear that the Bayesian Network classifier trained on  $X'_{red}$  outperforms the classifiers introduced in Section 3.

**Table 6.** Numerical results for the second dataset.

Trained Classifier	Accuracy	Gini Coefficient	Youden Index	AUC
Decision Tree trained on $X_{red}$	79%	0.1	0.2	0.55
Random Forest trained on $X_{red}$	80%	0.9	0.83	0.95
Gradient Boosting trained on $X_{red}$	79%	0.8	0.84	0.90
Logistic Regression trained on $X_{red}$	81%	0.12	0.25	0.56
ANN trained on $X_{red}$	67%	-0.12	0.06	0.46
Support Vector Machine trained on $X_{red}$	69%	0.78	0.78	0.89
Bayesian Network classifier trained on $X'_{red}$	96%	0.90	0.79	0.95

Among the trained classifiers, according to the accuracy and AUC measures, the Bayesian Network classifier proves to be the most effective at minimizing the error (8).

From the results obtained, we also infer the combinations of features in  $X'_{red}$  that are most probably the reason for ticket reopening; namely, they are the *events* listed below:

- (1) Ticket close code (d) AND Constant bitrate (downstream) in bin 2 AND Power (upstream) in bin 4 AND Attenuation (upstream) in bin 3;
- (2) Ticket close code (k) AND Constant bitrate (downstream) in bin 1 AND Power (upstream) in bin 1 AND Attenuation (downstream) in bin 3;
- (3) Ticket close code (i) AND Constant bitrate (downstream) in bin 4 AND Power (upstream) in bin 3 AND Attenuation (upstream) in bin 3;
- (4) Ticket close code (i) AND Constant bitrate (downstream) in bin 3 AND Power (upstream) in bin 3 AND Attenuation (upstream) in bin 4;
- (5) Ticket close code (i) AND Constant bitrate (downstream) in bin 2 AND Power (upstream) in bin 2 AND Attenuation (upstream) in bin 3;
- (6) Ticket close code (i) AND Constant bitrate (downstream) in bin 3 AND Power (upstream) in bin 2 AND Attenuation (upstream) in bin 1;
- (7) Ticket close code (p) AND Constant bitrate (downstream) in bin 2 AND Power (upstream) in bin 3 AND Attenuation (upstream) in bin 3;

- (8) Ticket close code (p) AND Constant bitrate (downstream) in bin 3 and Power (upstream) in bin 4 AND Attenuation (upstream) in bin 4;
- (9) Ticket close code (p) AND Constant bitrate (downstream) in bin 4 AND Power (upstream) in bin 1 AND Attenuation (upstream) in bin 2.

Table 7 reports them with the corresponding number of occurrences of such combinations of QoS parameters.

**Table 7.** Combinations of features in  $X'_{red}$  that are most probably the reason for ticket reopening according to the predictions of the Bayesian Network classifier.

Event	Number of Occurrences in $X'_{red}$	Percentage of Events Causing a Repeated Ticket
Event (1)	7	42.86%
Event (2)	29	34.48%
Event (3)	17	29.41%
Event (4)	107	20.56%
Event (5)	102	17.65%
Event (6)	102	12.75%
Event (7)	372	12.37%
Event (8)	447	10.96%
Event (9)	504	10.12%

In general, as can be seen from Table 7, the trained classifier provides the customer service of a network operator with a reliable tool for effectively monitoring customer tickets that, despite being already closed, are at risk of being reopened due to unsolved technical issues related to the perceived QoS.

The complexity of the Bayesian Network classifier, namely the most successful one, is linear in the number of training examples and in the number of features characterizing each training example. Instead, almost all other methods exhibit increased runtime complexity: more precisely, the Decision Tree, Random Forest and Gradient Boosting approaches are such that their complexity is logarithmic in the number of training examples, whereas the complexity of the SVM approach is quadratic in the number of training examples. Only the ANN and the Logistic Regression techniques have comparable computational complexity with respect to the Bayesian Network classifier, but with lower predictive performance (as shown in Table 6).

## 5. Conclusions

The paper proposes a data-driven approach based on machine learning for predicting technical ticket reopening in customer service platforms of telecommunications companies providing 5G fiber optic networks, namely with respect to ensuring that, between end user and service provider, the Service Level Agreement in terms of perceived Quality of Service is satisfied.

The activity was carried out within the framework of an extensive joint research initiative on Next Generation Networks between ELIS Innovation Hub and a major network service provider in Italy over the years 2018–2021.

The authors compare the performance of different approaches to classification—ranging from decision trees to Artificial Neural Networks and Support Vector Machines—and establish that a Bayesian network classifier is the most accurate at predicting whether a monitored ticket will be reopened or not.

In addition, the authors propose a suitable dimensionality reduction strategy that proves to be successful at increasing the computational efficiency by reducing the size of the relevant training dataset by two orders of magnitude with respect to the original dataset.

Numerical simulations show the effectiveness of the proposed approach, proving it can be a very useful tool for service providers in order to identify the customers that are most at risk of reopening a ticket due to an unsolved technical issue.



As future work, the authors look forward to testing the proposed method on Quality of Service datasets coming from additional sources and/or related to other 5G networks, as well as to testing the same method on even larger datasets in orders to further assess its scalability properties.

**Author Contributions:** Investigation, M.V.; Methodology, L.R.C., A.C., M.D., E.S. and M.V.; Software, A.C., M.D. and E.S.; Writing—original draft, L.R.C. and N.A.S.; Writing—review & editing, L.R.C. and N.A.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by ELIS Innovation Hub within a collaboration with Vodafone, grant number Joint Research Project within the framework of the Mindset Revolution Semester.

**Data Availability Statement:** Not Applicable, the study does not report any data.

**Conflicts of Interest:** The work presented in this paper was carried out while Caliciotti and Scocchi were with ELIS Innovation Hub and does not reflect the results of any activity carried out at Enel Green Power S.p.A. and ERG S.p.A., to which these two authors are currently affiliated, respectively.

## References

1. Cater-Steel, A.; Valverde, R.; Shrestha, A.; Toleman, M. Decision support systems for IT service management. *Int. J. Inf. Decis. Sci.* **2016**, *8*, 284–304.
2. Keller, A.; Midboe, T. Implementing a service desk: A practitioner’s perspective. In Proceedings of the 2010 IEEE Network Operations and Management Symposium—NOMS 2010, Osaka, Japan, 19–23 April 2010; pp. 685–696.
3. Cusick, J.J.; Ma, G. Creating an ITIL inspired Incident Management approach: Roots, response, and results. In Proceedings of the 2010 IEEE/IFIP Network Operations and Management Symposium Workshops, Osaka, Japan, 19–23 April 2010; pp. 142–148.
4. Tang, X.; Todo, Y. A Study of Service Desk Setup in Implementing IT Service Management in Enterprises. *Technol. Investig.* **2013**, 190–196. [\[CrossRef\]](#)
5. Tanovic, A.; Mastorakis, N. Advantage of Using Service Desk Management Systems in Real Organizations. *Int. J. Econ. Manag. Syst.* **2016**, *1*, 81–86.
6. Tanovic, A.; Butkovic, A.; Orucevic, F.; Mastorakis, N.E. Advantages of the implementation of Service Desk based on ITIL framework in telecommunication industry. In *Recent Researches in Electrical Engineering*; 2014; pp. 165–179. Available online: <http://www.wseas.us/e-library/conferences/2014/Lisbon/ELEL/ELEL-20.pdf> (accessed on 3 October 2021).
7. Harcenko, M.; Dorogovs, P.; Romanovs, A. IT Service Desk Implementation Solutions. *J. Riga Tech. Univ.* **2010**, *44*, 68–73. [\[CrossRef\]](#)
8. Jäntti, M. *Examining Challenges in IT Service Desk System and Processes: A Case Study*; 2012. Available online: <https://www.semanticscholar.org/paper/Examining-Challenges-in-IT-Service-Desk-System-and-J%C3%A4ntti/a0e6be989f664445e5a83535d5dd6b1427dc5e2a> (accessed on 3 October 2021).
9. Serbest, S.; Goksen, Y.; Dogan, O.; Tokdemir, A. Design and implementation of help desk system on the effective focus of information system. *Procedia Econ. Financ.* **2015**, *33*, 461–467. [\[CrossRef\]](#)
10. Robles, V.D. Resolving discourse at technical-support helpdesks. *IEEE Trans. Prof. Commun.* **2018**, *61*, 275–294. [\[CrossRef\]](#)
11. Shae, Z.; Bergstrom, T.; Pinhanes, C.; Podlaseck, M. Multimedia Chat for Helpdesks: A Practical SOA Architecture. In Proceedings of the IEEE Congress on Services—Part I, Honolulu, HI, USA, 6–11 July 2008; pp. 75–77.
12. Zhu, X.; Zhou, W.; Li, T. A visual tool for ticket monitoring and management. In Proceedings of the 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, China, 24–26 November 2017; pp. 1–7.
13. Ilieva, R.; Anguelov, K.; Gashurova, D. Monitoring and optimization of e-Services in IT Service Desk Systems. In Proceedings of the 19th International Symposium on Electrical Apparatus and Technologies (SIELA), Bourgas, Bulgaria, 29 May–1 June 2016; pp. 1–4.
14. Zhou, Y.; Tong, Y.; Gu, R.; Gall, H. Combining text mining and data mining for bug report classification. *J. Softw. Evol. Process.* **2016**, *28*, 150–176. [\[CrossRef\]](#)
15. Silva, S.; Pereira, R.; Ribeiro, R. Machine learning in incident categorization automation. In Proceedings of the 13th Iberian Conference on Information Systems and Technologies (CISTI), Cáceres, Spain, 13–16 June 2018; pp. 1–6.
16. Agarwal, S.; Sindhgatta, R.; Sengupta, B. SmartDispatch: Enabling Efficient Ticket Dispatch in an IT Service Environment. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 12–16 August 2012; pp. 1393–1401.
17. Wongsakthawom, R.; Limpiyakorn, Y. Development of IT Helpdesk with Microservices. In Proceedings of the 8th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 15–17 June 2018; pp. 31–34.
18. Aggarwal, V.; Agarwal, S.; Dasgupta, G.B.; Sridhara, G.; Vijay, E.V. ReAct: A System for Recommending Actions for Rapid Resolution of IT Service Incidents. In Proceedings of the IEEE International Conference on Services Computing (SCC), San Francisco, CA, USA, 27 June–2 July 2016; pp. 1–8.

19. Son, G.; Hazlewood, V.; Peterson, G.D. On Automating XSEDE User Ticket Classification. In Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment, Atlanta, GA, USA, 13 July 2014; pp. 1–7.
20. Wang, D.; Li, T.; Zhu, S.; Gong, Y. iHelp: An Intelligent Online Helpdesk System. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2010**, *41*, 173–182. [[CrossRef](#)] [[PubMed](#)]
21. Wang, X.; Zhang, L.; Xie, T.; Anvik, J.; Sun, J. An Approach to Detecting Duplicate Bug Reports Using Natural Language and Execution Information. In Proceedings of the 30th International Conference on Software Engineering, Virtual, 11–13 October 2021; pp. 461–470.
22. Rajeshwari, M.R.; Kavitha, K.S. Reduction of duplicate bugs and classification of bugs using a text mining and contingency approach. *Int. J. Appl. Eng. Res.* **2017**, *12*, 2840–2847.
23. Available online: <https://freshdesk.com/customer-service-training/customer-service-metric> (accessed on 1 September 2021).
24. Skubic, B.; Fiorani, M.; Tombaz, S.; Furuskar, A.; Martensson, J.; Monti, P. Optical transport solutions for 5G fixed wireless access. *IEEE/OSA J. Opt. Commun. Netw.* **2017**, *9*, 10–18. [[CrossRef](#)]
25. IBM Case Studies Corporate Landing Page, AI for Customer Service. Available online: <https://www.ibm.com/case-studies/> (accessed on 4 October 2021).
26. Daugherty, P.R.; Wilson, H.J. *Human+Machine: Reimagining Work in the Age of AI*; Harvard Business Review Press: Brighton, MA, USA, 2018.
27. Pietrabissa, A.; Celsi, L.R.; Cimorelli, F.; Suraci, V.; Priscoli, F.D.; Di Giorgio, A.; Giuseppi, A.; Monaco, S. Lyapunov-Based Design of a Distributed Wardrop Load-Balancing Algorithm With Application to Software-Defined Networking. *IEEE Trans. Control Syst. Technol.* **2018**, *27*, 1924–1936. [[CrossRef](#)]
28. Caliciotti, A.; Celsi, L.R. On optimal buffer allocation for guaranteeing quality of service in multimedia internet broadcasting for mobile networks. *Int. J. Control Autom. Syst.* **2020**, *18*, 3043–3050. [[CrossRef](#)]
29. Celsi, L.R.; Bonghi, R.; Monaco, S.; Normand-Cyrot, D. Normand-Cyrot. On the Exact Steering of Finite Sampled Nonlinear Dynamics with Input Delays. *IFAC-PapersOnLine* **2015**, *48*, 674–679. [[CrossRef](#)]
30. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
31. Narayanan, U.; Unnikrishnan, A.; Paul, V.; Joseph, S. A survey on various supervised classification algorithms. In Proceedings of the International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 1–2 August 2017; pp. 2118–2124.
32. Ben-Gal, I.; Trister, C. Parallel Construction of Decision Trees with Consistently Non Increasing Expected Number of Tests. *Appl. Stoch. Models Bus. Ind.* **2015**, *31*, 64–78. [[CrossRef](#)]
33. Prinzie, A.; van den Poel, D. Random Forests for multiclass classification: Random MultiNomial Logit. *Expert Syst. Appl.* **2008**, *34*, 1721–1732. [[CrossRef](#)]
34. Elias, C.; Calapez, A.; Almeida, S.; Chessman, B.; Simoes, N.; Feio, M. Predicting reference conditions for river bioassessment by incorporating boosted trees in the environmental filters method. *Ecol. Indic.* **2016**, *69*, 239–251. [[CrossRef](#)]
35. Al-Hawari, F.; Barham, H. A machine learning based help desk system for IT service management. *J. King Saud Univ.—Comput. Inf. Sci.* **2019**, *33*, 702–718. [[CrossRef](#)]
36. Cristianini, N.; Ricci, E. Support Vector Machines. In *Encyclopedia of Algorithms*; Kao, M.Y., Ed.; Springer: Boston, MA, USA, 2008.
37. Powers, D. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.