UNIVERSIDADE FEDERAL FLUMINENSE

DOUGLAS CUBA DOS SANTOS

CUSTOMER'S QUALITY OF SERVICE PREDICTION MODELS FOR A LARGE FIXED BROADBAND SERVICE PROVIDER

NITERÓI

2023

DOUGLAS CUBA DOS SANTOS

CUSTOMER'S QUALITY OF SERVICE PREDICTION MODELS FOR A LARGE FIXED BROADBAND SERVICE PROVIDER

Dissertation presented to the Computing Graduate Program of the Universidade Federal Fluminense, in partial fulfillment of the requirements for the Degree of Master in Computing. Area:. Computer Science

Advisor:
ANTONIO AUGUSTO DE ARAGÃO ROCHA

Co-advisor:
ADITA KULKARNI

NITERÓI

2023

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

S237c Santos, Douglas Cuba dos
CUSTOMERS QUALITY OF SERVICE PREDICTION MODELS FOR A LARGE FIXED BROADBAND SERVICE PROVIDER / Douglas Cuba dos Santos. - 2023.
68 f.: il.

Orientador: Antonio Augusto de Aragão Rocha.
Coorientador: Adita Kulkarni.
Dissertação (mestrado)-Universidade Federal Fluminense, Instituto de Computação, Niterói, 2023.

1. Rede de computadores. 2. Aprendizado de máquina. 3. Produção intelectual. I. Rocha, Antonio Augusto de Aragão, orientador. II. Kulkarni, Adita, coorientadora. III. Universidade Federal Fluminense. Instituto de Computação.IV. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

DOUGLAS CUBA DOS SANTOS

CUSTOMER'S QUALITY OF SERVICE PREDICTION MODELS FOR A LARGE FIXED BROADBAND SERVICE PROVIDER

Dissertation presented to the Computing Graduate Program of the Universidade Federal Fluminense, in partial fulfillment of the requirements for the Degree of Master in Computing. Area:. Computer Science

Approved in December, 2023.

DISSERTATION COMMITTEE

Prof. ANTONIO AUGUSTO DE ARAGÃO ROCHA - Advisor, UFF

Prof. FLÁVIA COIMBRA DELICATO, UFF

Prof. LISANDRO GRANVILLE, UFRGS

Niterói

2023

Acknowledgements

Firstly, I thank God for all the graces and blessings achieved throughout this journey, being paramount in my life.

To Elen, my wife, for all the love, affection, and dedication given to me in the most difficult moments. Your partnership and sacrifices made in favor of this project were essential for this to be another achievement for us.

To my son Lucas, a special gift from God, for overflowing my heart with love and joy and being the driving force towards a better human being.

To my parents Angélica and Daniel, who from an early age taught me the path of studies and persistence, as well as, of course, all the love, care, and affection received by them.

To my advisor, Professor Antonio Augusto de Aragão Rocha for guiding me in this project. His vision and experience were essential to my motivation as difficulties arose along the way.

I have the same gratitude towards my co-advisor Adita Kulkarni, whose distance did not impede exchanging experiences and advice during this work.

To the Fluminense Federal University and the teaching staff of the Computing Institute where I had the opportunity to gain experience and academic training.

Resumo

Na última década, o número de assinantes de banda larga fixa no Brasil aumentou consistentemente. No entanto, apesar da crescente procura, os clientes enfrentam vários desafios com o serviço de banda larga. Assim, neste trabalho, estabelecemos parceria com uma das maiores empresas provedoras de serviços de banda larga fixa do Brasil, para analisar parâmetros de serviço dos clientes e prever o alcance da taxa de download contratada pelos clientes. Levamos em consideração 7,8 milhões de registros coletados em 31 dias, a partir de 22 de maio de 2021 e 1º de janeiro de 2022, de 5% de todos os clientes. Este trabalho propõe dois modelos ECOC, ECOC-5 e sua versão aprimorada, ECOC-20, um modelo baseado em Error-Correcting Output Codes (ECOC), e propõe o framework Oracle, ambos para prever com precisão a qualidade do serviço, particularmente o download taxa, alcançada pelos clientes usando apenas recursos categóricos relacionados à localização do cliente, plano de Internet e equipamentos. Nossos experimentos demonstram que ambos os modelos ECOC superaram as linhas de base comparativas e o framework Oracle equilibrou os resultados de uma competição entre modelos multiclasse. A empresa pode usar nossas contribuições para aprimorar sua infraestrutura técnica de banda larga fixa.

Palavras-chave: Serviços de banda larga de Internet, previsão da qualidade do cliente, ECOC - Error-Correcting Output Codes, Oracle framework.

Abstract

Over the last decade, the count of fixed broadband subscribers in Brazil has consistently risen. However, despite the growing demand, customers encounter various challenges with the broadband service. Thus, in this work, we established a partnership with one of the largest fixed broadband service providers in Brazil, to analyze customers' service parameters and predict customers' achievement contracted download rate. We take into account 7.8 million logs that were gathered over 31 days starting from May 22, 2021, and January 1, 2022, from 5% of all customers. This work proposes two ECOC models, ECOC-5 and its enhanced version, ECOC-20, a model based on Error-Correcting Output Codes (ECOC), and proposes the Oracle framework, both to accurately predict the quality of service, particularly the download rate, achieved by the customers using only categorical features related to customer location, Internet plan, and equipment. Our experiments demonstrate that both ECOC models outperformed the comparative baselines and the Oracle framework balanced the results of a competition between multiclass models. The company can use our contributions to improve its fixed broadband technical infrastructure.

Keywords: Internet Broadband Services, predicting Customer Quality, ECOC - Error-Correcting Output Codes, Oracle framework.

List of Figures

List of Tables

List of Abbreviations and Acronyms

4G LTE Long Term Evolution 4 Generation

ANATEL Agência Nacional de Telecomunicações

BRAS Broadband remote access server

CDF Cumulative Distribution Function

CRF Conditional Random Fields

DRF Distributed Random Forest

DSL Digital Subscriber Line

DT Decision Tree

ECOC Error-Correcting Output Codes

FF-ANN FeedForward Artificial Neural networks

FTTH Fiber-to-the-Home

GBM Gradient Boosting Machine

GLM Generalized Linear Model

HE-ECOC Hierarchical Ensemble of Error Correcting Output Codes

HTTP Hypertext Transfer Protocol

ISP Internet Service Provider

ITU International Telecommunication Union

KNN K-nearest Neighbors

ML Machine Learning

MSAN multi-service access node

NB Naive Bayes

NMS Network Management System

QoE Quality of Exerience

QoS Quality of Service

RF Random Forest

SVM Support Vector Machine

XGBoost Extreme Gradient Boosting

Summary

1 Introduction

Global Internet usage continued to exhibit significant growth and impact on various aspects of society. As of ITU's (International Telecommunication Union) annual global assessment of digital connectivity (ITU, 2022), 4.9 billion was the estimated number of people using the Internet in 2021. The proliferation of smartphones, social networks, increased accessibility to broadband, and the expansion of digital infrastructure contributed to a significant rise in Internet users worldwide. Fixed-broadband Internet subscriptions continue to grow steadily, at an average annual growth rate averaging 6.7 percent over the last 10 years (ITU, 2022).

In Brazil, the utilization of fixed broadband services has increased over the last decade (ANATEL. . . , s.d.). Until October of 2023, the number of fixed broadband subscribers had surged by 2.1 million, reaching a cumulative total of 47.5 million subscribers (ANATEL. . . , s.d.) Despite its widespread adoption, users encounter challenges with broadband services. According to Anatel, the Agência Nacional de Telecomunicações in Brazil, 25% of the total complaints registered against telecommunication providers in 2022 were related to fixed broadband (ANATEL. . . , s.d.). The most common complaints among them were grouped into 5 categories: billing issues, quality of broadband and repair, user unsubscription, offers, and customer service quality in the specified order (ANATEL, 2022). Ensuring user satisfaction and optimizing resource utilization to address these issues are evolving into significant challenges for telecommunications companies.

To effectively manage the growing need for good quality broadband service while keeping costs to a minimum, it is critical to investigate the QoS (Quality of Service) of customers of large broadband service providers (LI; LU, 2009). This investigation is crucial in discerning the factors that contribute to user satisfaction and pinpointing areas that may require improvement. By understanding the complexities of QoS within large-scale broadband networks, telecommunications companies can implement targeted enhancements, allocate resources more efficiently, and provide a superior and cost-effective service to their clientele. This proactive approach not only ensures customer satisfaction

but also positions companies to meet the evolving challenges of the telecommunications landscape.

Even more, problems directly associated with the customer's perception of quality impact the churn rate of the subscriber base. As churn has a significant impact on the revenue of this industry, more and more companies in the sector have turned to developing machine learning models to predict factors that prevent customer dissatisfaction (BHARAMBE et al., 2023; PRAKASH et al., 2022).

This work partners with a telecommunications company in Brazil that provides fixed broadband services and has more than 0.8 million users and a market share of around 1.7% of the national market segment. As of 2022, Anatel considers this as one of the largest service providers in Brazil, along with Vivo, Claro, Oi, and SkY (ANATEL..., s.d.). As per Anatel's 2022 complaint balance report (ANATEL, 2022), the two most complained about issues for these large companies are billing and quality of fixed broadband.

Therefore, in this work, we bring two proposals that accurately predict the customers' achievement of contracted download rate using categorical attributes related to customers' location, broadband plan, and broadband equipment. We obtained around 7.8 million measurements of fixed broadband QoS parameters for 5% of the company's total customers residing in two states of Brazil. This collected data consists of 31 days starting from May 22nd, 2021 (we refer to as dataset 1) and from January 1, 2022 (we refer to dataset 2 for Rio de Janeiro and dataset 3 for São Paulo state). As the other 95% of customers do not have any kind of measurement, objective data on quality metrics about these customers is not available. This becomes a challenge since these transmission quality metrics data (numerical features) are intrinsically linked to the target chosen for prediction. In this context, as we have categorical attributes in datasets would be possible to predict the customers' achievement of contracted download rate only with categorical features?

The goal is to answer that question by proposing two machine learning models, ECOC-5 and ECOC-20, to predict the quality to be received by customers. Furthermore, we also propose the Oracle framework, an approach to select the best model among the available models intending to improve the overall classification accuracy. This way the company can analyze the cause of a customer's complaint and implement appropriate solutions based on the download rate classification of the customer. Through our proposals, the company can preemptively address customer concerns, mitigating issues before they escalate to the point of formal complaints.

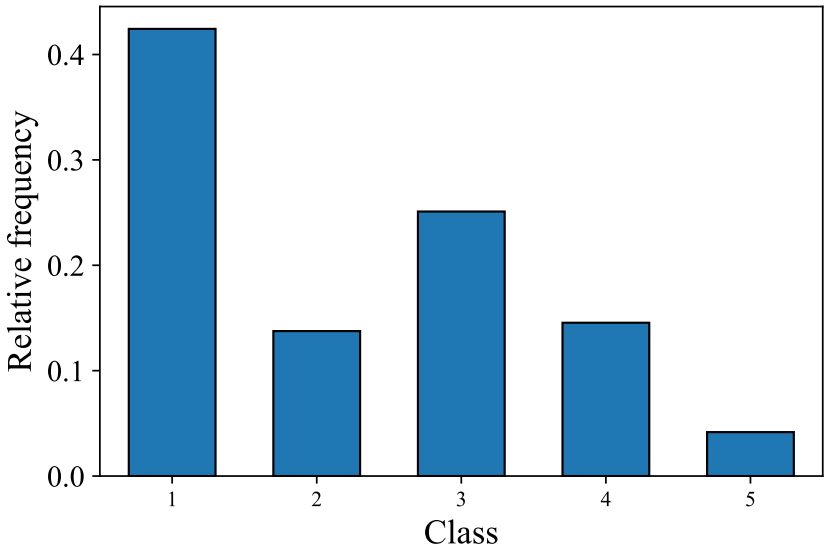
1.1 Problem statement

In this work, we aim to design a framework that accurately predicts the download rate of the customers with respect to fixed broadband. We aim to train models on the data collected from 5% of the company's customers to predict the contracted download rate achieved for the remaining 95% customers for whom measurements are not collected. This will allow the company to accurately identify and handle customer issues.

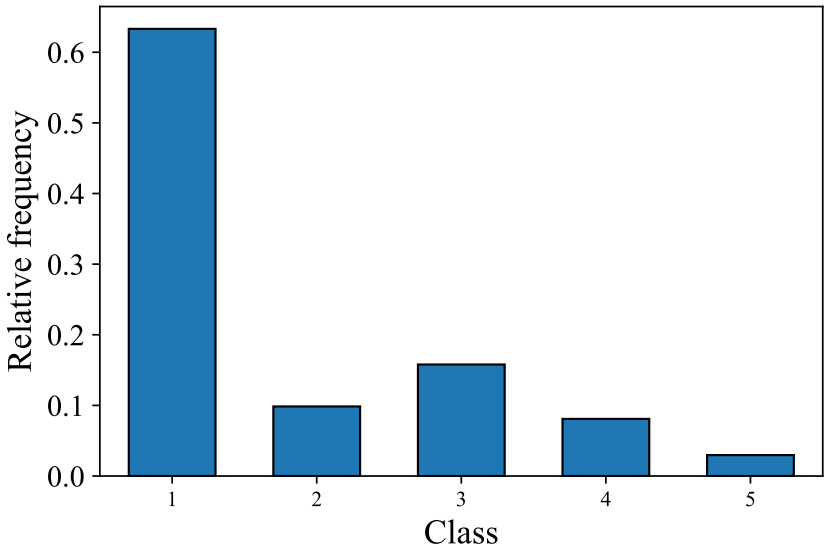
We formulate our problem as a multiclass classification problem where the download rate achieved by customers is predicted using the following customer attributes — *Contracted plan, Configured plan, City, Neighborhood, Zip code,* and *Equipment attributes (BRAS, MSAN, Port Number).* Although our dataset has both numerical features related to the network as well as the aforementioned categorical features related to the customers, we only use the categorical features in our framework since network-related features are unavailable for the remaining 95% customers. Our framework classifies the achievement of contracted download rate into one of the following — Class 1: (80%-150%], Class 2: (60%-80%], Class 3: (40%-60%], Class 4: (20%-40%], Class 5: [0%-20%]. The measurements belonging to Class 1 are considered the best as they are closest to the contracted download rate.

Figure 1 shows the distribution of classes in datasets considering raw measurements. We observe from the figure that although between 45% and 65% measurements fall in Class 1 for all three datasets, the remaining measurements belong to the classes where the download rate achieved is farther from the contracted rate. Thus, it is important for the company to identify the reason behind a customer complaint and take necessary actions to resolve it depending on which class the download rate of a customer belongs to. Our framework will enable the company to proactively deal with customer issues even before the customer files a complaint.

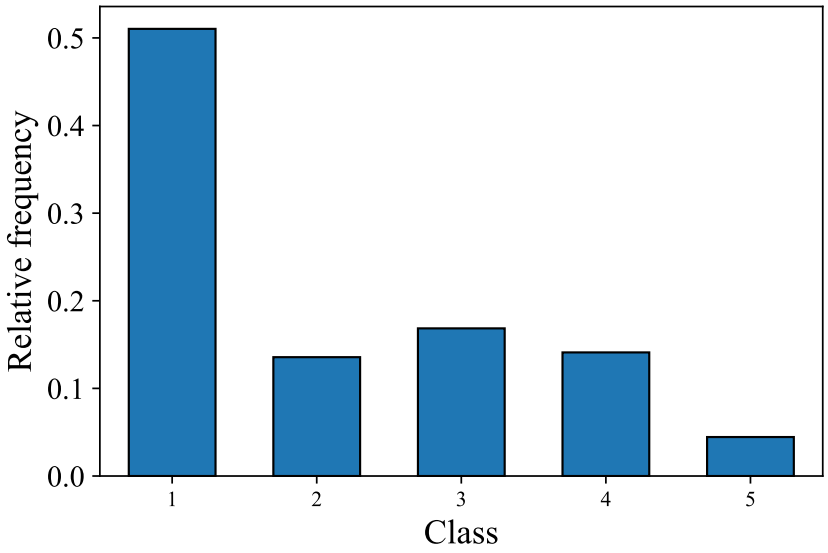
In this dissertation, we propose two ECOC models, ECOC-5 and ECOC-20, to solve the multiclass classification problem. Additionally, we introduced a new framework called Oracle to select which of several machine learning models, including ECOC-5 and ECOC-20, would be best for each dataset.



(a) Dataset 1



(a) Dataset 2



(a) Dataset 3

Figure 1: Histogram of dataset classes

1.2 Objectives and contributions

To achieve the goals we frame the question as a a multiclass classification problem, whereby we categorize client download rates into five distinct groups, which may subsequently be utilized to address consumer concerns. In the first proposal, we use ECOC (Error-Correcting Output Codes) (DIETTERICH; BAKIRI, 1994) to develop an ECOC classifier, named as ECOC-5 model, with two novel approaches which are a Hamming Distance tie-breaking method and the flexibility to select one among a set of algorithms for each base classifier. Further, the second proposal is developing the ECOC-20 model, which was based on the expansion of parameters used in ECOC-5. The results, detailed in Chapter 6, show that both the ECOC models outperformed the comparative baselines in specific datasets as will be explained further.

In an attempt to select the best solution among the available models during the training phase (ECOC models and other ML models), we address it with the third proposal which is the Oracle framework development. We use both ECOC models developed as a new baseline to compare the performance against the Oracle framework. The experiments for the Oracle framework, in Chapter 7 show that we obtain reasonable average accuracy and stay in an intermediate range between both ECOC models.

The main findings from our work are summarized below:

- Our results indicate that both ECOC models beat the average accuracy of the comparative baselines with ECOC-5 winning with 82.88% on dataset 1 and ECOC-20 on datasets 2 and 3 with 93.37% and 93.25% respectively.
- About misclassifications made by our models on the three datasets, we observe that most of them are off by only one or two classes except ECOC-5 in dataset 2 (around 42% of all misclassifications).
- The proposed Oracle framework is feasible and has an average accuracy of 68.13%, 91.87%, and 88.99% on Datasets 1, 2, and 3 respectively.

The work not only helps resolve customer complaints but also establishes the groundwork for future service improvements from the telecommunications company, like proactive maintenance scheduling and effective resource allocation that will raise customers' quality of experience even further.

1.3 Dissertation outline

The rest of the text of this dissertation is structured as follows. Chapter 2 presents the existing literature related to the broadband network context, the network quality prediction, and the ECOC applied domains.

The detailed data sources and the respective preprocesses applied to them are presented in Chapter 2.

The necessary ECOC fundamental concepts to better understand the proposals and the different ways to encode and decode information are presented in Chapter 4.

Chapter 5 proposes the ECOC-5 and ECOC-20 models and their evaluations are presented and discussed in Chapter 6.

The motivation and experiments of the Oracle framework are presented in Chapter 7.

Chapter 8 concludes the work, presenting the contributions and pointing out future research directions.

2 Related works

This Chapter presents research related to broadband networks, machine learning approaches to predict the QoS for such networks, and existing research based on the ECOC. We then discuss our previous work on predicting download rate for a fixed broadband service provider and discuss how this paper differs from the existing research.

2.1 Broadband networks

In the past two decades, research related to broadband networks has focused mainly on user performance analysis (SUNDARESAN et al., 2011), user behavior or user experience analysis (BISCHOF; BUSTAMANTE; STANOJEVIC, 2014; MADANAPALLI; GHARAKHIELI; SIVARAMAN, 2019), network traffic analysis (MAIER et al., 2009), congestion management (WONG et al., 2015).

Authors in (SUNDARESAN et al., 2011) study the broadband network performance from home gateways measured over two deployments in the USA. A full suite of measurement tools is used in both deployments to measure throughput, latency, packet loss, and jitter regularly. They analyze performance achieved by users and how various factors ranging from the user's choice of a modem to the ISP's (Internet Service Provider) traffic shaping policies can affect performance.

Bischof *et al.* evaluate the impact of service characteristics (capacity, latency, and loss), their broadband pricing, and user demand by studying three broadband datasets in (BISCHOF; BUSTAMANTE; STANOJEVIC, 2014).

In (MADANAPALLI; GHARAKHIELI; SIVARAMAN, 2019), authors infer Netflix user experience from broadband traffic patterns in real-time by correlating network activity with client playback behavior. They developed, FlixMon, a measurement tool that plays videos, measures their network activity, and stores measured records. 8000 Netflix Streams (750 hours of video) were collected to understand Netflix stream behavior. The

authors used a RandmForest machine learning algorithm to classify the phase of a video streaming playback

The authors describe traffic characteristics by monitoring the network activity for more than 20,000 residential DSL (Digital Subscriber Line) customers from a major European ISP in (MAIER et al., 2009). They analyzed DSL characteristics such as session duration, termination causes, the number of concurrent sessions, and application usage patterns where HTTP (Hypertext Transfer Protocol) carries most of the traffic (>50%).

In (WONG et al., 2015), authors suggest moving congestion management to the network edge to address issues with broadband network congestion during peak hours. For that, they propose a two-level bandwidth allocation solution for broadband network congestion problems. In Level 1, gateways purchase bandwidth on a shared link using QoE (Quality of Exerience) credits. In Level 2, they distribute the bandwidth they have paid for among their devices and apps. Analytically, the authors demonstrate that a credit distribution scheme results in an equitable distribution of bandwidth amongst gateways.

A facet that sets our work apart from these studies is that our work contributes by introducing models designed to predict the download quality metric. As discussed in the next subsection, while many studies exist on this topic, not all share the same focus on developing predictive models for essential metrics.

2.2 Network quality prediction

The task of predicting the quality of fixed broadband networks using machine learning has been extensively studied over the past years (MASLO et al., 2021; WANG et al., 2021; ALIPIO; BURES, 2023; ŻELASKO; PŁAWIAK; KOŁODZIEJ, 2020). Authors in (MASLO et al., 2021) conduct an overview of the QoS and QoE parameter measurement and prediction carried out by Bosnia and Herzegovina's leading telecom operators. Recognizing the value of QoE, operators are gearing up to implement practical measurements to improve service quality and increase profits, emphasizing the adoption of machine algorithms and targeted marketing for enhanced customer retention and acquisition.

In

Authors in (ŻELASKO; PŁAWIAK; KOŁODZIEJ, 2020) propose a ML (Machine Learning) solution, specifically applied to the novel Pay&Require approach for transmission quality assurance in computer networks. The study is limited to 100 samples and 4

tested users, emphasizing the need for future work with a larger user base. The focus is on four classifiers (Nu-SVC, C-SVC, kNN, and Random Forest), and the paper suggests exploring other classifiers and techniques to enhance accuracy and sensitivity. While the current classification accuracy is 87%, deemed quite good, the authors anticipate the possibility of achieving better results with further testing and optimization of classifiers and parameter values.

(ALIPIO; BURES, 2023) successfully developed a proactive model employing ML techniques, such as KNN (K-nearest Neighbors), DT (Decision Tree), RF (Random Forest), NB (Naive Bayes), SVM (Support Vector Machine), and FF-ANN (FeedForward Artificial Neural networks), to detect and identify NMS (Network Management System) parameter degradation in FTTH (Fiber-to-the-Home) networks. The proposed FTTH system flowchart and an alarm system facilitated the identification of specific conditions. KNN exhibited the highest accuracy at 89%, further improved to 89.36% with gradient boosting. The ML models, especially KNN, effectively predicted faults, localized issues, and recommended prescriptive maintenance processes. The study emphasizes the integration of ML techniques for well-monitored and well-maintained networks in IoT-driven sustainable smart homes. It suggests further exploration of ML techniques, including deep neural networks and reinforcement learning, for enhanced accuracy in diagnosing FTTH anomalies. Additionally, the modification of the dataset for balance and exploration of various preprocessing techniques is recommended for providing more relevant information. The study underscores the role of ML in developing FTTH NMSs, offering automation and efficient management solutions.

There has also been a significant amount of work related to predicting QoS in mobile broadband (cellular) and wireless networks. Studies (RAJ et al., 2021; KULKARNI et al., 2019; SAMBA et al., 2017; LEE et al., 2020) aim to predict the channel quality or bandwidth in cellular networks with a goal of improving future QoE. E.g., Raj *et. al.* design discriminative sequence-to-sequence probabilistic graphical models to predict future channel quality variations (RSRP, RSRQ, Download rate, and Upload rate) in 4G LTE (Long Term Evolution 4 Generation) networks based on past channel quality data in (RAJ et al., 2021). Authors in (SAMBA et al., 2017) use an RF model to predict transmission throughput (downlink data rate) considering user context, cellular link quality, and access network performance data. Authors in (ADEKITAN; ABOLADE; SHOBAYO, 2019) design machine learning models to predict the future upload and download Internet traffic based on past traffic using data generated in Covenant University, Nigeria.

Based on the studies surveyed, the uniqueness of our approach lies in our dedication to providing effective and accurate models that improve understanding and prediction of these metric qualities based on the ECOC method.

2.3 ECOC

ECOC is a straightforward yet effective framework for handling multiclass categorization that uses binary classifier embeddings (base classifiers). It was originally introduced by (DIETTERICH; BAKIRI, 1994) and since then a lot of research has been done to expand many of the aspects of its initial structure in various directions (PATEL; POLADI, 2022). ECOC has been applied to solve problems in different domains such as disease diagnosis (ALMUKHTAR, 2023; RUKHSAR, 2022; ÜBEYLI, 2007; LIU; ZENG; NG, 2016), computer networks (XIE et al., 2009; MAJIDIAN et al., 2023; GUO; RAMAMOHANARAO; PARK, 2008), speech recognition (ZHAO; SHU, 2023; XIAO-FENG; XUEYING; JI-KANG, 2010), vision problems (BAGHERI; MONTAZER; ESCALERA, 2012; ESCALERA; PUJOL; RADEVA, 2007a; YE; LIANG; JIAO, 2011), and text classification (ZHANG; YU; TANG, 2017; BALAMURUGAN et al., 2022; LI; VOGEL, 2010). The application of these works is briefly described below, as they are not the focus of this dissertation.

About the disease diagnosis domain, (ALMUKHTAR, 2023) proposed a method using a combination of ECOC and SVM to classify suspicious regions and diagnose lung cancer with an average accuracy of 96.67%. In another cancer classification context, (LIU; ZENG; NG, 2016) proposes a hierarchical ensemble strategy, named HE-ECOC (Hierarchical Ensemble of Error Correcting Output Codes). In this strategy, different feature subsets extracted from a dataset are used as inputs for three data-dependent ECOC algorithms, to produce different ECOC coding matrices. In a similar combination using ECOC and SVM, (ÜBEYLI, 2007) classifies four types of electrocardiogram beats are presented for classification (normal beat, congestive heart failure beat, ventricular tachyarrhythmia beat, atrial fibrillation beat) resulting in high accuracies.In (RUKHSAR, 2022), proposes a multiclass decision tree classier combining with ECOC using ensemble learning applied over electroencephalogram recorded data to classify epileptic seizure.

In the computer networks domain, (XIE et al., 2009) classifies network traffic based on artificial neural network ensemble with ECOC achieving an improvement in the multiclass classification accuracy by 12%-20% on dataset captured on the backbone router of a

campus. (MAJIDIAN et al., 2023), proposes a method to detect denial-of-Service attacks using ECOC and an adaptive neuro-fuzzy Inference System for detection. (GUO; RAMAMOHANARAO; PARK, 2008) decrease the training time and computational resource requirements of CRF (Conditional Random Fields), used to improve web page prediction, by training the ECOC method.

In the speech recognition domain, the combination of gamma classifier and ECOC is applied to classify features and extract emotions in (ZHAO; SHU, 2023). In (XIAOFENG; XUE-YING; JI-KANG, 2010), authors used a combination of SVM a common-used encodings of ECOC to improve classification in speech recognition.

In the computer vision problems, (BAGHERI; MONTAZER; ESCALERA, 2012) used Decision Tree and AdaBoost as the base classifiers in ECOC modeling to predict logo recognition and shape classification. Authors in (ESCALERA; PUJOL; RADEVA, 2007a) present a novel methodology to detect and recognize objects in cluttered scenes by using ECOC with a forest of optimal tree structures on public datasets. In (YE; LIANG; JIAO, 2011), the authors train the base classifiers of ECOC histogram of oriented gradient features and SVM to decide if an image and video frame is pedestrian or not.

About the text classification topic, (ZHANG; YU; TANG, 2017) propose a new partial label learning strategy is studied which refrains from conducting disambiguation using only the ECOC approach. In order to learn text classifiers, the work in (BALAMURUGAN et al., 2022) uses ECOC by comparing two types of dichotomizers to each corresponding monolithic classifier. In (LI; VOGEL, 2010), improved the binary base classifiers in ECOC taking into account the sub-class division distribution in each base classifier to categorize and classify text documents.

Despite the existence of a vast literature, to the best of our knowledge, ECOC has not been used in the domain of telecommunication before. In this work, we designed two ECOC models (referred to as ECOC-5 and ECOC-20) to analyze categorical characteristics of fixed broadband Internet customers and predict their achievement of download rates contracted with a telecom company. Intending to expand studies, this work proposes to add contributions as will be seen in detail later in Chapters 5 and 7.

3 Data

In this work, customer data represents a crucial part of the proposed solution and is the basis on which subsequent analyses and modeling are built. The way data is collected, the types of data involved, and the preprocessing steps undertaken are presented in the next two subsections. The process of data collection by the company is a regulatory obligation and an overview of the data is detailed in Subsection 3.1. The raw data collected is diverse in nature encompassing quantitative measures and categorical features each requiring customized preprocessing strategies discussed in Subsection 3.2.

3.1 Datasets

In this subsection, we provide an overview of our dataset. We obtain broadband customer network measurement data from a large telecommunications company in Brazil. The data used in this work consists of 5% of the total customers of the company residing in two states of Brazil, Rio de Janeiro and São Paulo. A measurement of QoS parameters is performed on the broadband equipment of the customers every hour if they are not using the Internet at that time. Data is collected for a period of 31 days starting from May 22, 2021, for the Rio de Janeiro state (we refer to this as '*Dataset 1*') and for a period of 31 days starting from January 1, 2022, for Rio de Janeiro (we refer to this as '*Dataset 2*') and São Paulo (we refer to this as '*Dataset 3*') states. The total data consists of around 7.8 million measurements of which *Dataset 1* consists of around 4 million, *Dataset 2* consists of around 1.5 million and *Dataset 3* consists of around 2.3 million measurements, respectively.

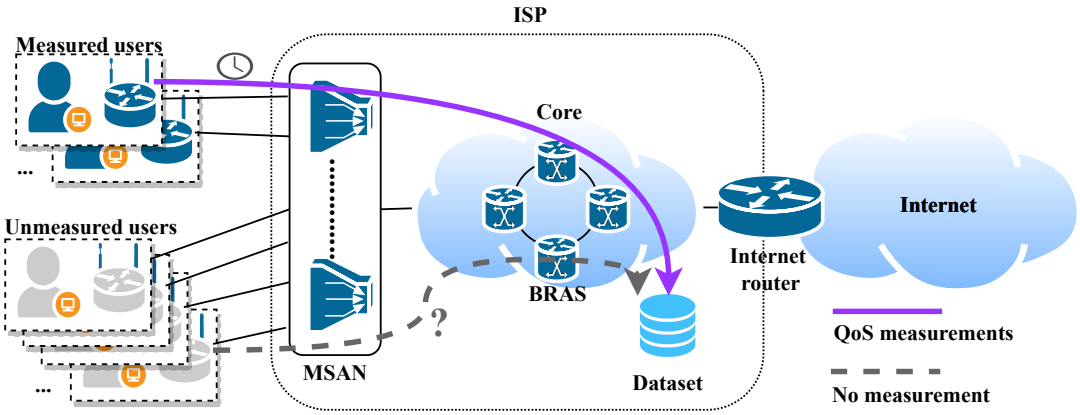


Figure 2: Data Collection Process

Figure 2 shows the data collection process employed by the telecommunications company. We see in the figure that customer equipment is connected to the ISP's MSANs (multi-service access node) which are then connected to the core network consisting of multiple BRAS (Broadband remote access server). This core network is connected to the rest of the Internet via an Internet router. The data is periodically collected from 5% of the total customers (denoted by measured users) and is stored at the ISP. This data collection takes place every hour when the customer's equipment is idle (i.e., the customer is not using the Internet). If the customer is using the Internet at the time of the probe, the measurement does not occur as it would interfere with using the Internet at that moment.

The network measurements comprise 20 quantitative variables and 30 categorical variables (Appendix A). From these, we use the following variables in our work:

- **Device ID** - Each customer uses a unique device that has a unique identifier. Thus, we use this variable as the customer identifier (anonymized). *Dataset 1*, *Dataset 2*, and *Dataset 3* comprise an average of 126,815, 46,020, and 75,065 number of customer identifiers per day, respectively.

- **Contracted plan** - It is the plan that customers purchase from the telecom operator. *Dataset 1* has seventeen different plans that the customers choose from whereas *Dataset 2* and *Dataset 3* have sixteen different plans.

- **Configured plan** - It is the plan that is configured by the telecom operator for the customer. Ideally, this value should be equal to the contracted plan for a particular customer. In some instances, we observe that this value is higher than the contracted plan because of either of the three reasons — i) the company gives a higher plan to

the customer as a trial for a certain period of time, ii) the company gives a customer a higher plan for a few months as compensation when customers complain due to some error from the company's side, iii) there is some glitch on the company's end.

- **Download Rate** - It is the broadband speed actually incurred by the customer.
- **City** - It is the city in which the customer resides. Customers in *Dataset 1*, *Dataset 2*, and *Dataset 3* reside in 11, 6, and 9 different cities of the states, respectively.
- **Neighborhood** - It is the neighborhood in which the customer resides. Customers in *Dataset 1*, *Dataset 2*, and *Dataset 3* reside in 415, 259, and 1,215 different neighborhoods of the states, respectively.
- **Zip code** - It is the zip code where the customer resides. Customers in *Dataset 1*, *Dataset 2*, and *Dataset 3* have 5,631, 2,445, and 5,022 different zip codes, respectively.
- **Equipment attributes** - There are three equipment attributes logged in our dataset — BRAS, MSAN, and Port Number. BRAS accepts client connections, checks credentials, records accounting data via back-end servers, and allows customers network access via that connection. Customers in our dataset connect to one of the five different BRAS. MSAN is a device installed in a telephone switch that connects customers' telephone lines to the core network to provide broadband. *Dataset 1*, *Dataset 2*, and *Dataset 3* have 1,440, 682, and 957 unique MSAN values, respectively. Port Number refers to the MSAN equipment's port.

The Appendix B shows a training sample of these features.

3.2 Preprocessing

We preprocess the data before training our model on it. We undertake the following preprocessing steps.

- The *Download Rate* variable mentioned in Subsection 3.1 provides the instantaneous download rate in bits per second. We convert this value to represent the percentage of download rate achieved with respect to the configured plan.
- We drop all duplicate measurements from the dataset. *Dataset 1* and *Dataset 2* have around 20% duplicate measurements whereas *Dataset 3* has around 18% duplicate measurements.

- We drop the measurements with null values for the *Download Rate* variable. *Dataset 1*, *Dataset 2*, and *Dataset 3* have around 18%, 8%, and 16% null values, respectively.
- In some instances, the value of the download rate goes above 100% because of the reasons mentioned in Subsection 3.1. Thus, we remove the measurements where the download rate is above 150%. We drop 2.12% measurements from *Dataset 1*, 0.5% measurements from *Dataset 2*, and 0.64% measurements from *Dataset 3*.
- Multiple measurements for each customer are recorded in our dataset per day. We aggregate all the measurements for a customer every day by taking the median of the download rate. Thus, after aggregation, each customer has only one measurement per day.

In machine learning, grouping data by the median can be a helpful preprocessing step, especially when working with specific types of data distributions or when handling outliers.

The original data consists of around 7.8 million total measurements. After applying the preprocessing steps, we have around 770K total measurements of which *Dataset 1* consists of around 340K measurements, *Dataset 2* consists of around 160K measurements, and *Dataset 3* consists of around 270K measurements.

4 ECOC fundamental concepts

In multiclass classification, the original problem can be broken down into numerous binary classification subproblems using several techniques. Error-Correcting Output Codes (ECOC) (DIETTERICH; BAKIRI, 1994) is one such technique that has been effective and adaptable for several domains.

ECOC is an ensemble multiclass classifier inspired by signal transmission coding theory, where error detection and correction techniques ensure consistency of data sent and received in a communication channel susceptible to sources of interference. In machine learning, interference can be considered when a classification error of the underlying models occurs, after which the encoding and decoding process (described in Subsections 4.1 and 4.2) provides error correction.

ECOC applies the *Divide-and-Conquer* principle by decomposing a multiclass classification problem into several binary classification subproblems (*dichotomies*). Each of these dichotomies is tackled by a base classifier (*dichotomizer*) and the final solution is obtained by decoding the base classifiers' aggregated results as shown in Figure 3. In aggregation, the output of each base classifier is concatenated into a sequence of symbols called "codeword", and checked against the expected codeword for a given class. The class whose codeword matches this output is chosen as the resulting class. If a complete codeword match does not occur, the final predicted codeword consists of correcting the output codeword that matches one of the existing class patterns or considering a closer class pattern according to the decoding method chosen.

The ECOC framework consists of two main phases: an encoding phase, where each class is mapped to a codeword, and a decoding phase, where, given a test sample, the best matching codeword and the corresponding class is selected as the result.

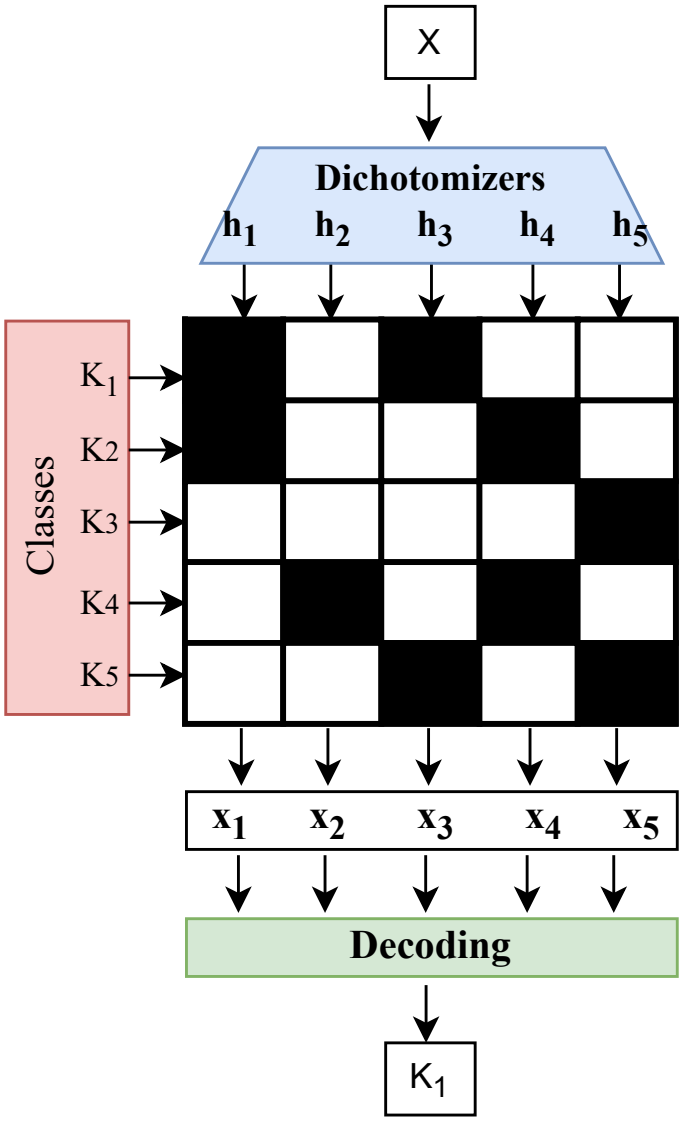
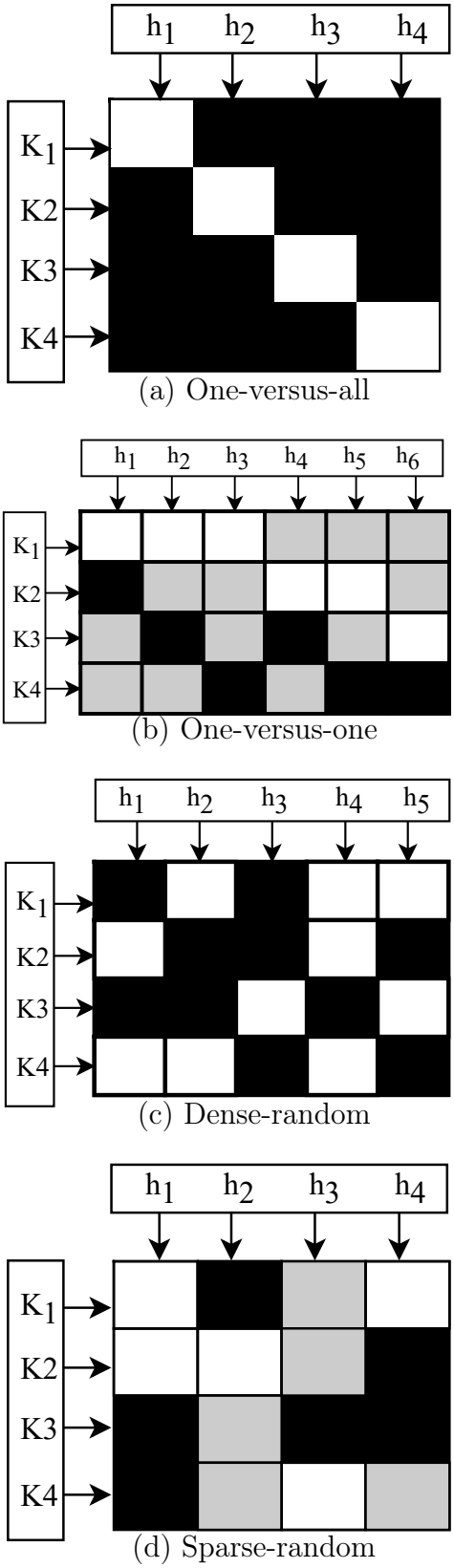


Figure 3: ECOC design example with 5 classes

4.1 Encoding

In this phase, the code matrix plays a key role in mapping classes into codewords. Generally, two coding types are used when designing coding matrices: binary codes and ternary codes. As more detailed further below, Figure 4 shows these two coding types in four different coding matrix designs where binary codes are represented by matrices with black and white colors, and the ternary codes are represented by matrices with black, white, and gray colors. The white regions are coded as 1 (seen as one class by the respective dichotomizer), the dark regions as -1 (considered as the other class), and the gray regions correspond to 0 (classes not considered by the respective dichotomizer in ternary encoding).



(a) One-versus-all

(b) One-versus-one

(c) Dense-random

(d) Sparse-random

Figure 4: ECOC coding designs

- *Binary code*: In binary codes, an ECOC coding matrix $\mathbf{M} \in \{-1,1\}^{k \times n}$ is built, where k denotes the number of classes and n denotes the number of bi-partitions each consisting of a base classifier. The length of a code is the number of columns in the code. The number of rows in the code equals the number of classes in the multiclass learning problem. A codeword is a row in the code and is unambiguously defined to represent a class. As per (DIETTERICH; BAKIRI, 1994), a good error-correcting output code for a multiclass problem should satisfy two properties — i) Row separation: each codeword should be well-separated in Hamming distance from each of the other codewords. ii) Column separation: columns in the matrix should be uncorrelated i.e., they should not be identical or complementary.

- *Ternary code*: Originally, only binary code was used in ECOC encoding phase but (ALLWEIN; SCHAPIRE; SINGER, 2000a) introduced a third value defining a ternary code matrix as $\mathbf{M} \in \{-1,0,+1\}^{k \times n}$, where a class with a 0 is not considered by a base classifier, and +1 and -1 symbols represent binary class instances.

Several problem-dependent and problem-independent code designs based on binary and ternary codes are proposed in existing literature. The most well-known approaches are discussed below.

- *Problem-independent designs*: There are several code designs based on binary and ternary codes and most of them are predefined or “fixed" regardless of the data or the problem domain. The most popular designs are:

  - One-versus-All (NILSSON, 1965): This is the most well-known and simple coding design that produces an identity matrix. For K classes, the model trains K base classifiers and the codewords have length K. This design is shown in Figure 4[One-versus-all].
  - One-versus-One (HASTIE; TIBSHIRANI, 1998): This strategy considers all possible pairs of classes and then produces a codeword of length $\frac{K(K-1)}{2}$. This design is shown in Figure 4[One-versus-one].
  - Dense-random (ALLWEIN; SCHAPIRE; SINGER, 2000b): This strategy generates a high number of random coding matrices of length n using $P(-1) = 1 - P(+1)$, where $P(-1)$ and $P(+1)$ is the probability of the presence of symbols -1 and $+1$, respectively. Then, from a set of generated random matrices, the one that maximizes a decoding measure among all possible rows is selected.

The suggested number of base classifiers in this design is $n = 10 \log k$. This design is shown in Figure 4[Dense-random].

- Sparse-random (PUJOL; RADEVA; VITRIA, 2006): This is similar to the dense-random strategy, but it adds the symbol 0 appearing with a probability of $P(0) = 1 - P(-1) - P(1)$. The suggested number of base classifiers is $n = 15 \log k$. This design is shown in Figure 4[Sparse-random].

- *Problem-dependent designs*: These designs consider the problem domain to select the best binary subproblem representation trying to keep the code length small. Most popular designs are:

  - DECOC (PUJOL; RADEVA; VITRIA, 2006): The Discrimant ECOC employs tree structure to learn partitions of the domain problem. Each internal node of the tree is a column in the coding matrix. The need for base classifiers is $k - 1$.

  - Forest-ECOC (ESCALERA; PUJOL; RADEVA, 2007b): Here, the number of base classifiers is $(k-1).T$, where T is the number of binary tree structures to be embedded.

  - ECOC-ONE (PUJOL; ESCALERA; RADEVA, 2008a): This method is an extension of DECOC and uses a validation subset to train relevant binary subproblems. The design suggests $n = 2.k$ base classifiers.

4.2 Decoding

In ECOC, the main goal of decoding is to fix mistakes that might have happened during the classification procedure. Thus, the decoding phase is the last stage where the model decides what to assign for a particular input's class. During the classification process, there may be instances where the output code does not perfectly match any predefined class. Decoding strategies help handle such ambiguities by assigning the input to the most appropriate class based on predefined rules. The intermediate output codes, that is the concatenation of each base classifier output, are converted into a conclusive classification result.

In this subsection, we present the decoding strategies. The most commonly used binary decoding strategies are Hamming decoding (NILSSON, 1965), Inverse Hamming decoding (TERRY; REZA, 2003), and Euclidean decoding (PUJOL; RADEVA; VITRIA,

2006). The most commonly used ternary decoding strategies are attenuated Euclidean decoding (PUJOL; ESCALERA; RADEVA, 2008b), loss-based decoding (ALLWEIN; SCHAPIRE; SINGER, 2000b), and probabilistic-based decoding (PASSERINI; PONTIL; FRASCONI, 2004).

- Hamming Decoding (HD): In this approach, the hamming distance between two strings of equal length is computed based on the number of positions at which the corresponding symbols are different. It is defined as $HD(x,k_i) = \sum_{j=1}^{n}(1 - sign(x^j.k_i^j))/2$, where x is the test codeword, k_i is the class codeword in a code matrix, and n is the length of the codeword.

- Inverse Hamming Decoding (IHD): (ALLWEIN; SCHAPIRE; SINGER, 2000a) demonstrates that the IHD strategy's practical behavior closely resembles the HD strategy's behavior. This measure is defined as $IHD(x,k_i) = \max(\Delta^{-1}D^T)$, where the proportionality of each class codeword in the test codeword is represented by the values of $\max(\Delta^{-1}D^T)$, where $\Delta(i_1,i_2) = HD(k_{i1},k_{i2})$, and D is the vector of Hamming decoding values of the test codeword x for each of the base codewords k_i

- Euclidean Decoding: This is another well-known decoding technique defined as $ED(x,k_i) = \sqrt{\sum_{j=1}^{n}(x^j - k_i^j)^2}$.

- Attenuated Euclidean Decoding: This approach is a modified version of Euclidean decoding which is defined as $AED(x,k_i) = \sqrt{\sum_{j=1}^{n}|k_i^j|x^j|(x^j - k_i^j)^2}$.

- Loss-based Decoding: In this approach, the total loss on a proposed data sample is the decoding measure given a Loss-function model: $LB(\rho,k_i) = \sum_{j=1}^{n} L(k_i^j.f^j(\rho))$, where $k_i^j.f^j(\rho)$ is the *margin* and L is a loss function that is affected by the binary classifier's nature, where ρ is a test sample, and f is a real-valued function $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

- Probabilistic-based Decoding (PD): To deal with ternary decoding, this strategy relies on the classifier's continuous output. The decoding measure is given by $PD(k_i,x) = -log\left(\prod_{j=1:M(i,j)\neq 0}^{n} P(x^j = M(i,j)|f^j) + K\right)$, where K is a constant factor that collects the probability mass dispersed on the invalid codes, and the probability $P(x^j = M(i,j)|f^j)$ is estimated by means of $P(x^j = k_i^j|f^j) = \frac{1}{1+\mathbf{e}^{k_i^j(\vartheta^j f^j + \omega^j)}}$, where vectors ϑ and ω are obtained by solving an optimization problem.

In ECOC, decoding is essential to completing a classification procedure as it fixes mistakes and resolves ambiguities.

5 Proposed models

In this work, we solve our multiclass classification problem using two different ECOC models for each day in available Datasets. The initial ECOC model is named ECOC-5 and after the result analysis, we decided to enrich it to build a new model with a novel variation, the ECOC-20, to explore how the performance could be. Regarding ECOC design, both models have the same design choices described in the following subsection. Although there are many different designs in the encoding and decoding phases, we chose the traditional approach to build the ECOC models. After that, in Subsection 5.2, all other modeling steps that differentiate ECOC-5 from ECOC-20 are detailed.

5.1 Chosen ECOC design

In ECOC modeling, the code matrix is an essential element. In our design choice, Table 1 shows the relationship between classes and the respective binary code words assigned to them. At the encoding phase, we use a dense-random encoding strategy to encode the five classes into codewords as the matrix shown in Table 1 for both ECOC-5 and ECOC-20 models.

Table 1: ECOC Code Matrix

Download Rate	**Class**	**Classifier**														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
(80%-150%]	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(60%-80%]	2	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
(40%-60%]	3	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0
(20%-40%]	4	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0
[0%-20%]	5	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

Every entry in the matrix denotes the presence or absence of a particular coding bit for a given class and is represented by the binary value (0 or 1). To guarantee that classes

can be distinguished and that mistakes can be fixed during the classification process, the binary codes are meticulously created.

Following (DIETTERICH; BAKIRI, 1994), if there are $3 \leq k \leq 7$classes, there will be at most $2^{k-1} - 1$ usable columns at Exhaustive Codes method. As we have five classes for predictions, we get a 15-column code. Each column in the matrix is a base classifier. Since we have 15 columns, we have 15 base classifiers. In the Exhaustive code method, the building of matrices process consists of filling Row 1 with all ones. Row 2 consists of 2^{k-2} zeros followed by $2^{k-2} - 1$ ones. Row 3 consists of 2^{k-3} zeroes, followed by 2^{k-3} ones, followed by 2^{k-2} zeroes, followed by $2^{k-2} - 1$ ones. In Row i, there are alternating runs of 2^{k-i} zeroes and ones. We applied just the complementary method, where would be 1 we fill with zeroes and vice-versa as a simple fashion style.

As our $k = 5$, the coding matrix produced has inter-row Hamming distance 8 and satisfies the row and the column separation property, that is, the rows are well-separated in Hamming Distance and the columns are not identical or complementary respectively.

The minimum Hamming Distance between any codewords is a measure of the quality of an error-correcting code. If the minimum Hamming distance is d, then it is possible to correct at least $\lfloor \frac{d-1}{2} \rfloor$ single bit errors. Then, we can correct at least 3 bits.

Each row in the matrix corresponds to a class. In our work, Class 1 denotes a download rate between 80 and 150%, Class 2 denotes a download rate between 60 and 80%, Class 3 denotes a download rate between 40 and 60%, Class 4 denotes a download rate between 20 and 40%, and Class 5 denotes a download rate between 0 and 20%.

As we chose the Dense-random encoding strategy that is built on binary codes, every base classifier acts as a binary classifier. E.g., according to the matrix shown in Table 1, Classifier 1 treats the binary classes as follows — if the download rate is above 80%, it predicts 0, else it predicts 1. Classifier 2 classifies as follows — if the download rate is above 80% or if the download rate is between 0% and 20%, it predicts 0, else it predicts 1. Similarly, the rest of the classifiers perform binary classification following Table 1. Since we are performing binary classification at this step, the classes are unbalanced. We apply oversampling to balance the classes and use the RandomOverSampler Python library (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017).

In the decoding phase, we take the predictions from the 15 classifiers, which is a 15-bit sequence of 0's and 1's. We apply Hamming distance on this 15-bit sequence and each of the five codewords in the matrix. The final class predicted by the model is the one where

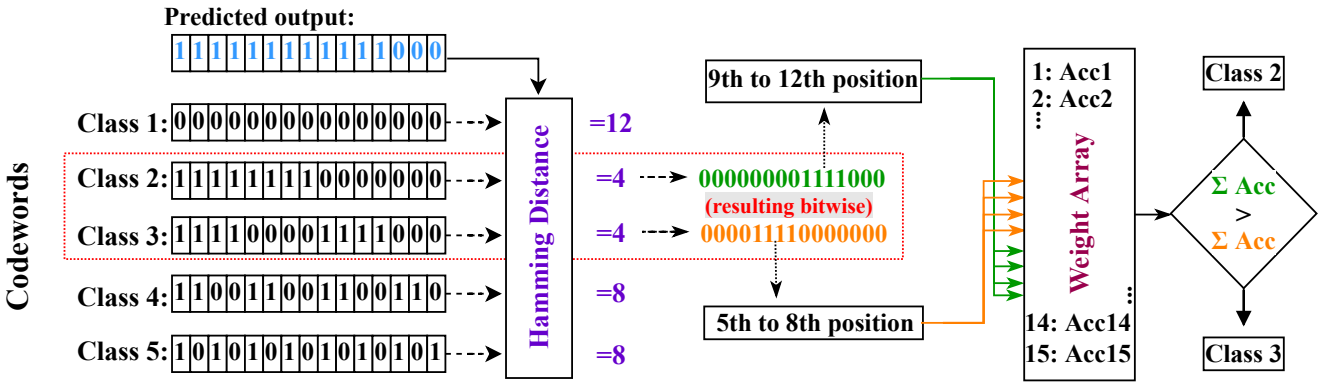


Figure 5: Tie breaking mechanism

the distance between the 15-bit sequence and the codeword is the lowest.

If there is a tie for the hamming distance for two or more classes, we break the tie as described using an example in Figure 5. As seen in the example, the hamming distance between the 15-bit predicted output and the codewords for Class 2 and Class 3 results in a tie. In such a case, we perform an exclusive OR operation between the codeword and the predicted 15-bit sequence to obtain which bit positions differ. We observe that for Class 2, they differ from bits 9 to 12, and for Class 3, they differ from bits 5 to 8. We next obtain the accuracy array, which consists of the training accuracy for each of the 15 classifiers. We sum the accuracy for classifiers 9 to 12 for Class 2, and we sum the accuracy for classifiers 5 to 8 for Class 3. The final class predicted by the model is the one with a higher value of this computed sum. To the best of our knowledge, considering the vast literature about the ECOC, the described tie-breaking method is a novel contribution.

5.2 Modeling steps

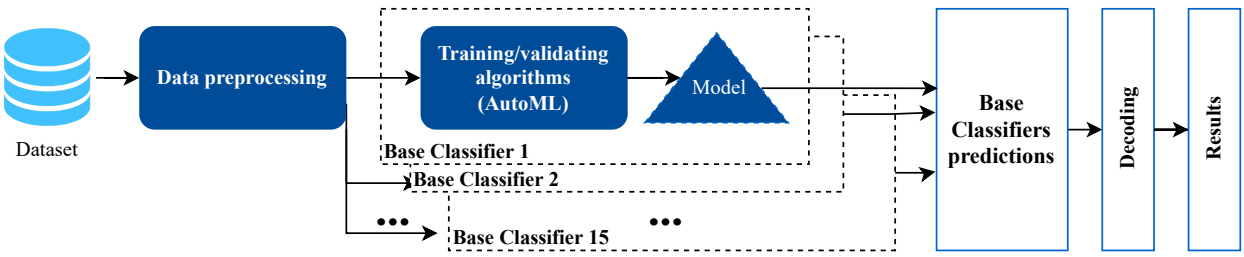


Figure 6: ECOC modeling

Figure 6 represents the ECOC modeling for both ECOC-5 and ECOC-20 in this work. Each dotted line region represents, for each base classifier, the modeling steps described in

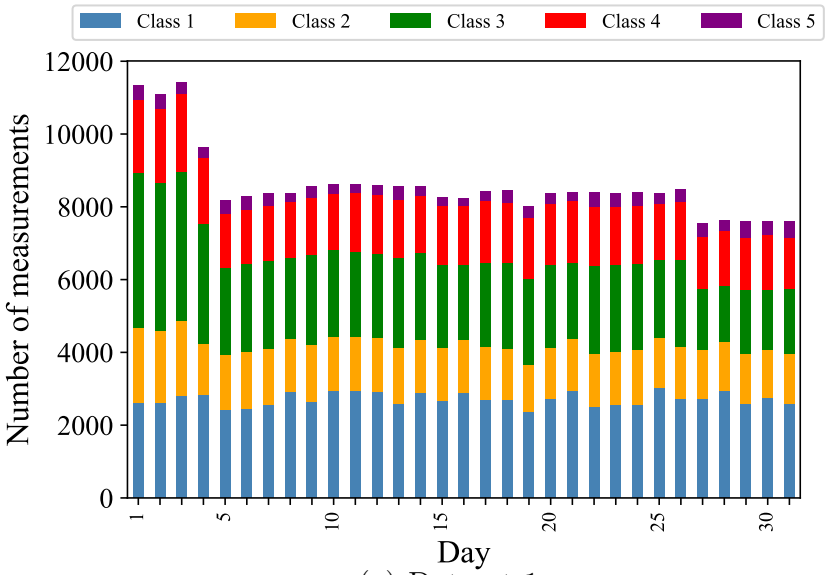
the subsections below (except for dataset splitting). After data preprocessing on datasets, the modeling steps are performed culminating in a trained base classifier (blue triangle). Each base classifier predicts a test sample and all outputs are concatenated to be decoded according to the chosen decode strategy. At the end of the process, the ECOC model gives a predicted output.

The ECOC-5 and ECOC-20 models act as distinct classifiers. Although they perform the same modeling steps, they only differ in one parameter in Subsection 5.2.4.

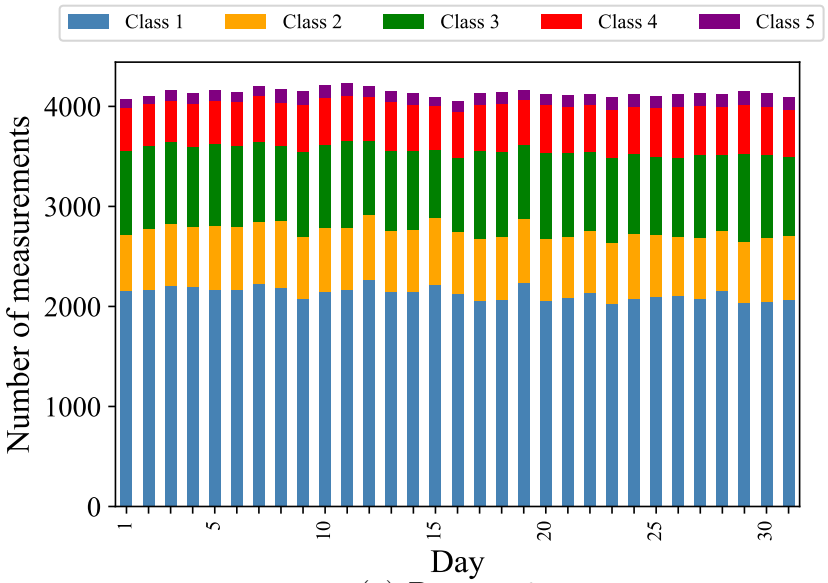
5.2.1 Dataset split

The creation of models with good generalization to unknown data is the main objective of machine learning. A key technique for achieving this goal is dataset splitting, which simulates the model's performance on fresh, real-world instances. This method remains a fundamental and effective approach to tackling machine learning problems. Authors in (RAYKAR; SAHA, 2015) emphasized that the primary goal of dividing datasets into distinct sets is to prevent bias resulting from repeatedly using the test or validation sets. The test set is used to evaluate the model's performance on data that was not seen during training after it has been trained on the training set. This makes it easier to determine if the model is picking up on data patterns or training set memorization. The test set is used as a last assessment to determine the model's expected performance on entirely new data.

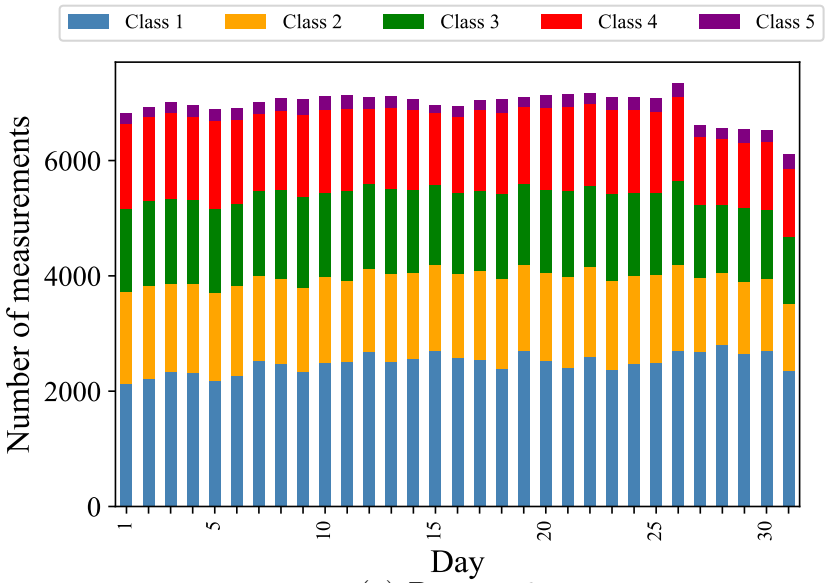
At this modeling step, the data is divided into 80% for training and 20% for testing. Figure 7 shows, in stacked bars, the total number of training measurements per day in all Datasets that belong to each of the five classes. We see from Figure 7[Dataset 1], 7[Dataset2] and 7[Dataset 3] that the highest number of measurements in the training set belongs to Class 1 and Class 3 whereas Class 5 has the lowest number of measurements. Class 2 and Class 4 have almost the same distribution. All data sets have very few samples representative of class 5, the worst in terms of achieving the contracted download rate. The average measurements per day are 8000, 4000, and 6000 for Datasets 1,2 and 3 respectively. Dataset 1 has the highest number of measurements per day. Different from other datasets, the number of user measurements in the first 4 days has a different amount than the other days with class 3 being the main contributor to this phenomenon. Dataset 2 presents a proportion of class 2 samples that are different from the other Datasets.



(a) Dataset 1



(a) Dataset 2



(a) Dataset 3

Figure 7: Number of training measurements for each class per day

5.2.2 Categorical features transformation

When supplying categorical data to machine learning models, it is imperative to convert it into numerical data for multiple reasons. Many machine learning algorithms are designed to work with numerical input (SREE et al., 2021). Categorical data with numerical representations can improve machine learning model performance. More significant insights are frequently derived by algorithms from numerical features, enabling improved generalization and prediction accuracy. Often used methods for converting categorical data into numerical format are one-hot encoding and label encoding. While one-hot encoding generates binary columns for each category, indicating its presence or absence, label encoding gives each category a distinct numerical label.

In the datasets, the selected features mentioned in Subsection III-C are categorical. They must be converted to numeric to be subsequently submitted to the classifiers. To perform that, we use CatBoost (DOROGUSH; ERSHOV; GULIN, 2018), an open-source library for gradient boosting on decision trees with categorical features support. As a result, each categorical feature value or feature combination value is assigned a numerical feature.

5.2.3 Standardization

To guarantee equitable and efficient learning across diverse features and to improve the performance and convergence of different algorithms, standardization is an essential preprocessing step in machine learning.

Some of the categorical features have high cardinality e.g. "neighborhood" and "zipcode". After transformation to numerical in the previous step, these differences in the ranges of numerical features can cause trouble for many machine learning models. To mitigate that, we use Scikit-learn StandardScaler (PEDREGOSA et al., 2011), an open-source machine learning library to standards.

5.2.4 Base classifiers

Base classifiers play a key role in ECOC modeling. For each day, we train the 15 base classifiers separately and we use 5-fold cross-validation to validate each of them. In this step, we bring a contribution. To the best of our knowledge, the other works choose the same classifier algorithm to be used in all base classifiers. For both models we built,

ECOC-5 and ECOC-20, the choice of classifier algorithm is flexible for every base classifier. We use H2O's Automatic Machine Learning (AutoML) (LEDELL; POIRIER, 2020), an open-source machine learning tool, to build our base classifiers. The set of available algorithms in version 3.44.0.2 are DRF (Distributed Random Forest), XGBoost (Extreme Gradient Boosting), GBM (Gradient Boosting Machine), and GLM (Generalized Linear Model).

From here we differentiate the proposed models we build by the number of algorithms and their variations used to build the base classifiers. Each base classifier is represented by the best algorithm selected for that purpose. The selection is based on the higher algorithm's ACC (Accuracy) metric which is the default H2O AutoML metric to sort the leader model. For ECOC-5 model, we use H2O AutoMl to train five GLM algorithm variations for each base classifier specifically.

At this point, a question appeared to us. What would happen if we increased the number of algorithm variations in this step? Would the results be intuitively better than the initial version of ECOC-5? Hereupon, for the ECOC-20, we train twenty variations of all H2O AutoML available algorithms by tunning random grid-search hyperparameters. The 5-fold cross-validation technique is performed for every algorithm variation.

5.2.5 Controlling randomness

The reproducibility of experiments is ensured by controlling randomness. An experiment can be repeated exactly by setting a seed for random number generation. This makes it easier to compare various models or algorithms and lets others confirm your findings.

Some parts of the modeling are inherently random like the cross-validation splitter, random oversampling, and H2O AutoML estimators. We control the randomness of these objects by setting the "random_state" or "seed" parameter for reproducible results across executions.

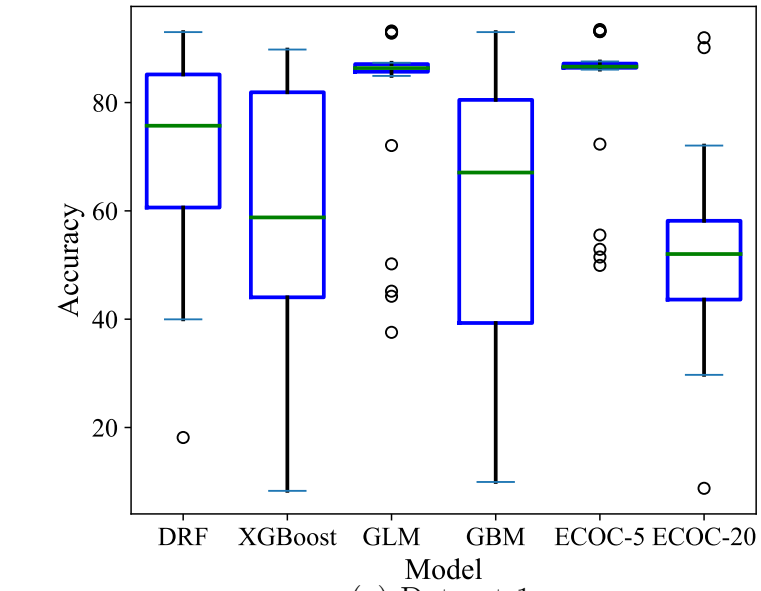
6 ECOC models evaluation

In this Chapter, we present the experimental results that demonstrate the superior performance of our models (ECOC-5 and ECOC-20). The main metric used in our evaluation is accuracy. We begin with discussing the accuracy of our models, then discuss the level of deviation of the model between the actual class and the predicted class, and finally discuss the qualitative results. We compare the performance of our models with four baselines—Distributed Random Forest (DRF), Extreme Gradient Boosting (XGBoost), Gradient Boosting Model (GBM), and Generalized Linear Model with regularization (GLM), which are a part of the H2O AutoML tool and run independently outside our ECOC framework. We present accuracy, classification error, and qualitative results for all three datasets in the following subsections.

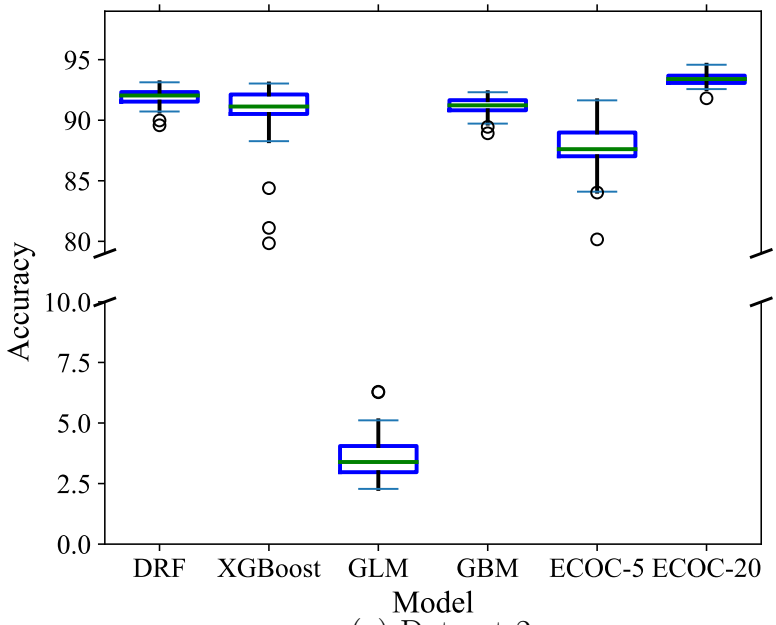
6.1 Accuracy

We obtain classification results by running our model separately on the test set of each day, thus generating results for a total of 31 days for each dataset. Figures 8[Dataset1], 8[Dataset 2] and 8[Dataset 3] show the boxplot denoting the accuracy distribution of our ECOC models and the comparative baselines for Datasets 1, 2, and 3, respectively. A boxplot, sometimes referred to as a box-and-whisker plot, is a type of graphic that shows the distribution and central tendency of a dataset visually. It shows important statistical metrics and draws attention to any outliers. When it comes to comparing the distribution of various groups or datasets, the boxplot is especially helpful.

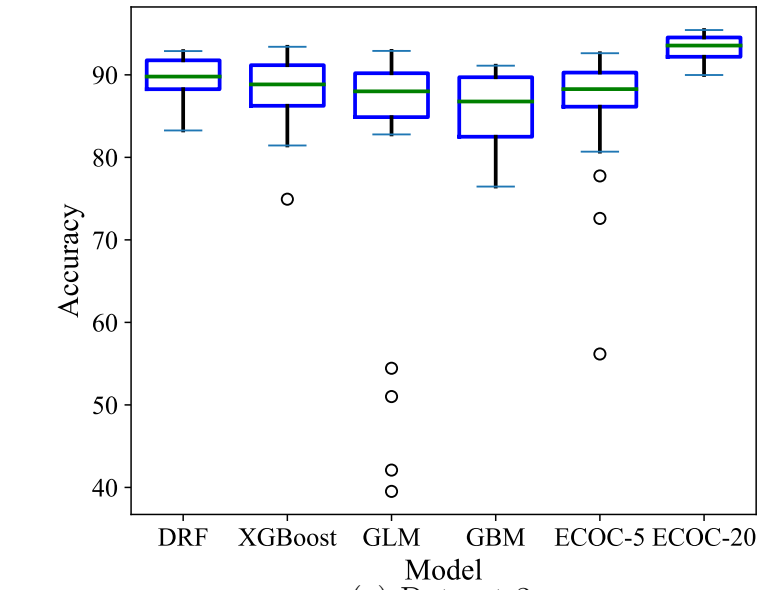
We observe from Figure 8 that for all the datasets at least one of our ECOC models outperforms the baselines. Figure 8[Dataset1] indicates that ECOC-5 and GLM models are more stable than XGBoost and GBM as the boxplot shows a less dispersion of accuracy whereas XGBoost and GBM have a wide range of accuracy values among the 31 days. We see from Figures 8[Dataset2] and 8[Dataset3] that accuracy for all days in all models is closer in range.



(a) Dataset 1



(a) Dataset 2



(a) Dataset 3

Figure 8: Boxplot of accuracy for each model

Table 2 summarizes the mean accuracy for the entire period in each dataset. The comparative baseline algorithms are from the second to fifth columns. We observe from the table that ECOC-5 has the best accuracy for Dataset 1 and ECOC-20 shows has the best accuracy for Dataset 2 and Dataset 3. In Dataset 2, the GLM algorithm has a significant drop in performance representing the worst result while it has the competitive values for Datasets 1 and 3. In Dataset 3, all models have competitive results.

Dataset	DRF	XGBoost	GLM	GBM	ECOC-5	ECOC-20
1	70.04%	59.87%	81.49%	60.74%	82.88%	51.55%
2	91.87%	90.45%	3.65%	91.08%	87.74%	93.37%
3	89.51%	88.13%	82.82%	85.64%	86.59%	93.25%

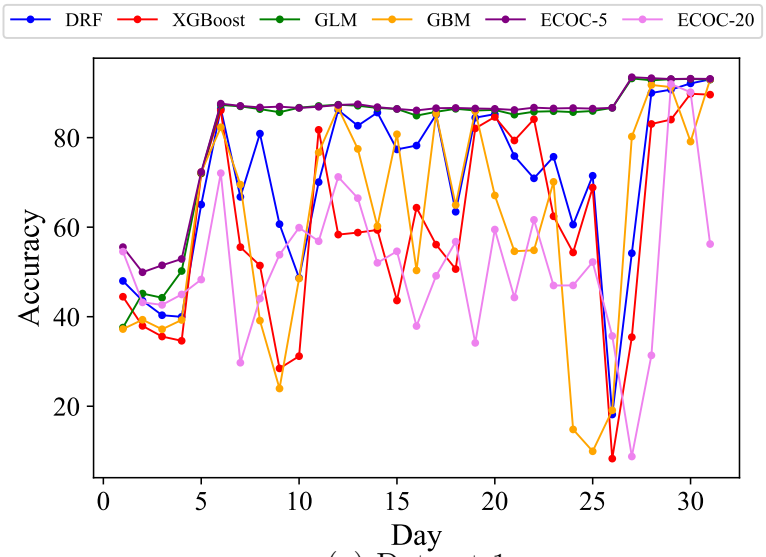
Table 2: Mean accuracy for all datasets.

Since the performance of our ECOC models is very close to the performance of the top baseline models, we further investigate our results by comparing the accuracy achieved per day by all models in Figure 9 where each model is represented by a colored line.

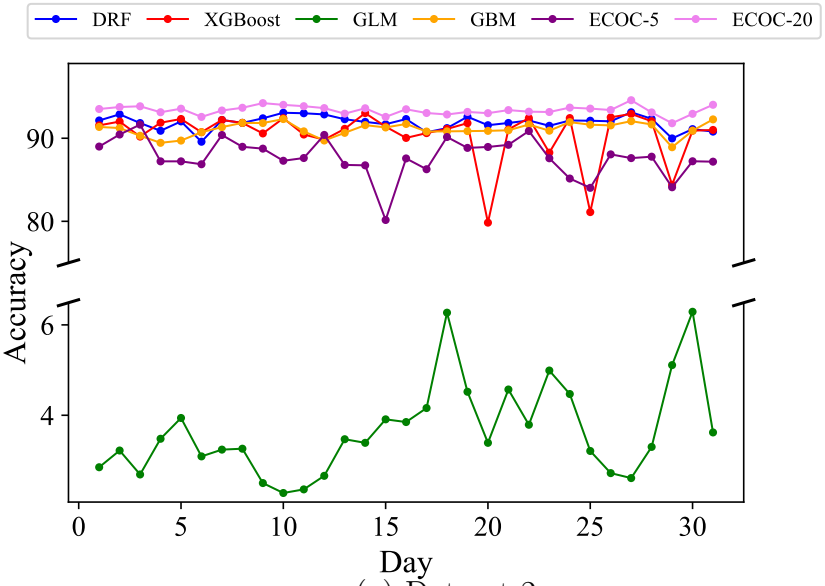
We observe in Figure 9[Dataset 1] that, although the ECOC-5 model (violet line) and GLM (green line) show close performance for Dataset 1, there are some days where our model beats the baseline significantly (Days 1 to 4). For the rest of the days, DRF, XGBoost, GLM, GBM, and ECOC-20 vary the accuracy range with a huge drop between days 6 to 9 and 25 to 28. In these drops, ECOC and GLM present steadfast values in their lines after day 5.

In Figure 9[Dataset 2], the bad performance of the GLM algorithm is very noticeable while the other models compete in the high range of accuracy. Among these, except for GLM, ECOC-5 has the lower accuracy values per day and ECOC-20 has the superiority in the majority of days.

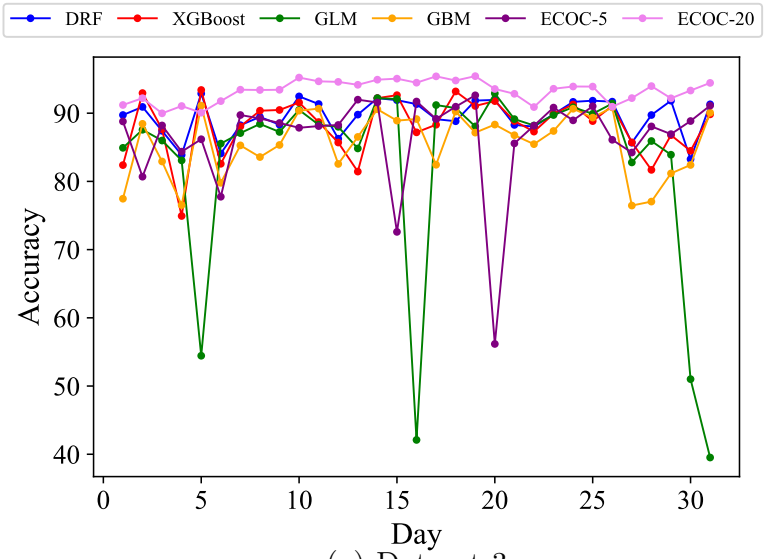
In Figure 9[Dataset 3], GLM has significant drops on days 5, 16, 30, and 31 and ECOC-20 again is on the top. Except for GLM in the mentioned days, all algorithms are competitors in high accuracy values.



(a) Dataset 1



(a) Dataset 2



(a) Dataset 3

Figure 9: Accuracy per day for each model

Continuing the analysis of the results, Figure 10 has a better visualization of the results by ranking the position for the daily accuracy for each model. The higher the accuracy, the better the model is positioned.

We observe from Figure 10[Dataset 1] that ECOC-5 does consistently better than the comparative baselines followed by GLM. The rest of the models compete through 3rd to 6th positions during the days.

For all days in Dataset 2, ECOC-20 is the leader, as well as GLM, is the worst. The rest of the models compete in the intermediate position zone.

We see from Figures 10[Dataset 2] and 10[Dataset 3] that ECOC-20 beats the baselines on all days for Datasets 2 and 3, respectively, except two days for Dataset 3 although does consistently better than the baselines for the rest of the days.

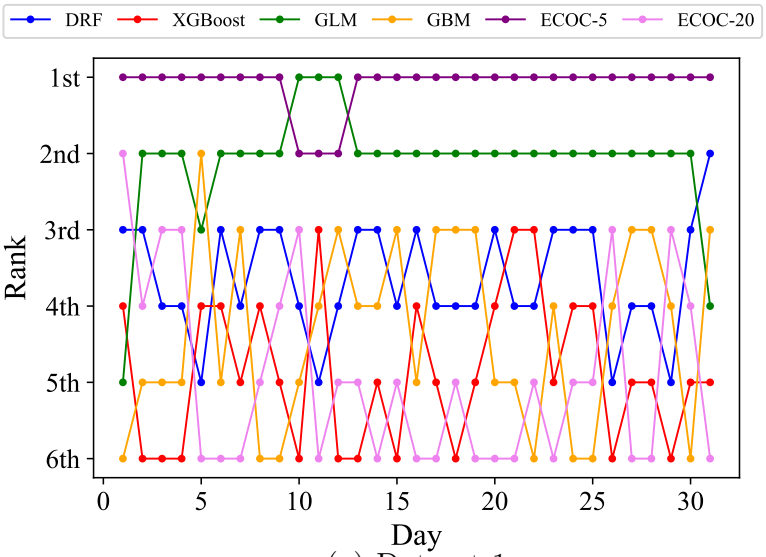
6.2 Classification Error

We next discuss how far off our best models predict from the actual class when a miss occurs. From Table 2, a misclassification rate is obtained by the complementary values of accuracy. Then, Dataset 1 has a misclassification rate of around 17% on average for the ECOC-5 model, around 6.6% and 6.75% on average for the ECOC-20 model for Datasets 2 and 3, respectively. Figure 11 shows the distance between the mispredicted class and the actual class for both ECOC models for all three datasets. As we have 5 classes, the distance between any two different classes can be up to 4 classes.

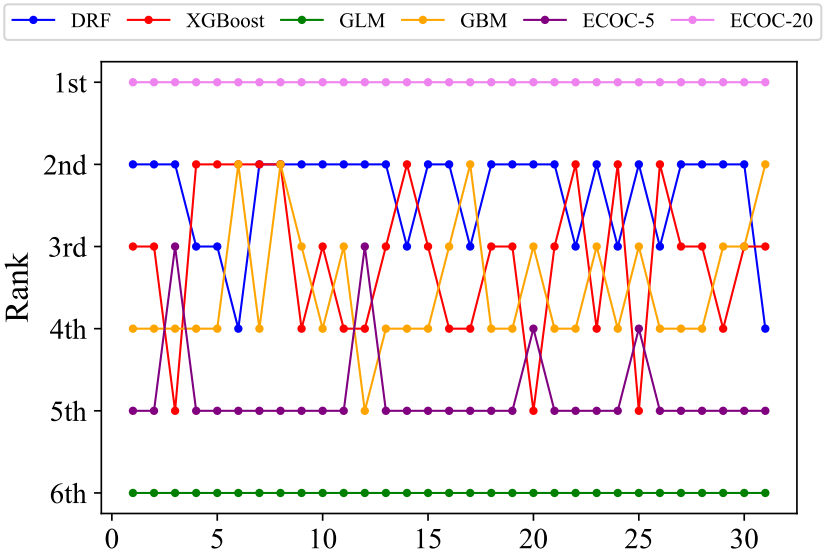
Figure 11[Dataset 1] denotes that out of around 17% of total misclassifications by ECOC-5 for Dataset 1, 58.5% of these classification errors are by one or two classes. Similarly, Figures 11[Dataset 2] and 11[Dataset 3] denote that out of 6.6% and 6.75% of the total misclassifications by ECOC-20 for Dataset 2 and Dataset 3, respectively, 74% of these classification errors are by one or two classes for both datasets.

The ideal shape of the bars would be as skewed to the left as possible for each ECOC model denoting that, when the model misses, the miss is from less distance as possible from the actual class. In general, ECOC-20 has descending bars from the left which is a good indicator of a missing context. On the other hand, ECOC-5 has a kind of hill shape on the bars for Dataset 1 and 2 but it has a similar shape to ECOC-20 in Dataset 3.

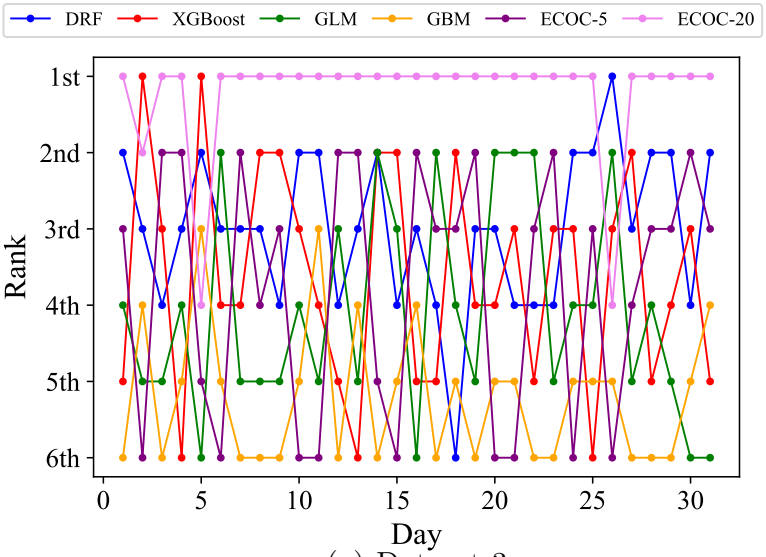
Thus, our models perform quite reasonably for the measurements they originally predicted incorrectly.



(a) Dataset 1



(a) Dataset 2



(a) Dataset 3

Figure 10: Accuracy per day for each model

6.3 Qualitative Results

Finally, this subsection investigates the qualitative results of ECOC-5 and ECOC-20 models for the datasets. We discuss the qualitative results for only the best-performing models i.e., ECOC-5 for Dataset 1 and ECOC-20 for Datasets 2 and 3.

We obtain the predicted download rate for each test measurement and compute the absolute difference between the value and the closest bound of the actual class that it classifies into. Then, we get the CDF (Cumulative Distribution Function) for each of the five classes for both our models for all datasets which are shown in Figures 12 and 13.

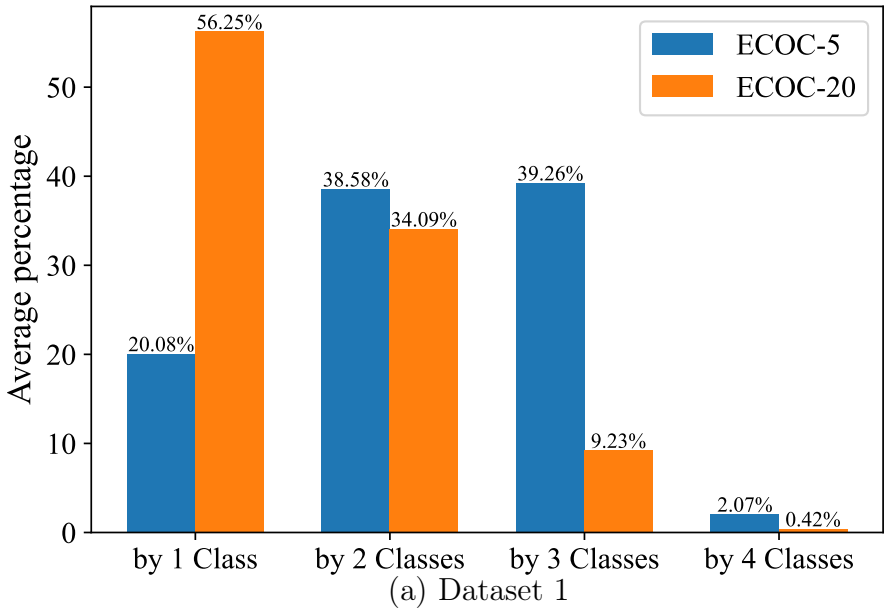
It is possible to observe from Figure 12[Dataset1] that the ECOC-5 model gives the best results for Dataset 1 for measurements belonging to Class 1 (blue line), whereas it shows the worst performance for Class 5 (violet line)measurements.

Similarly, we observe from Figures 13[Dataset 2] and 13[Dataset 3] that the ECOC-20 model shows the best results for measurements belonging to Class 1, whereas it shows the worst performance for Class 5 measurements. A possible explanation for how our models perform badly for Class 5 data can be because, as seen in Figure 7, we have a low number of measurements in the training data belonging to Class 5. However, bad performance for Class 5 measurements does not affect the overall accuracy of our model much because there are fewer Class 5 measurements in our test data.

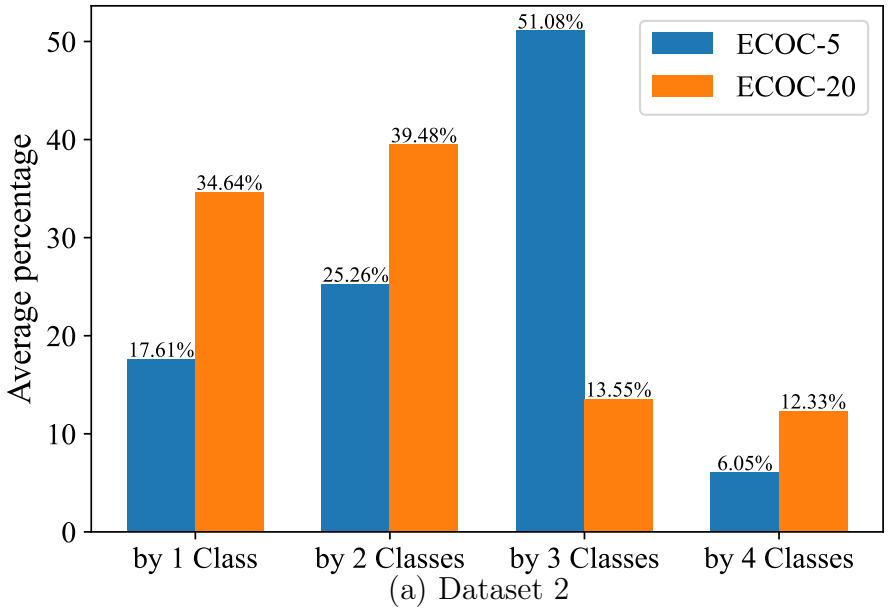
6.4 ECOC models discussion

Since ECOC was designed, it has brought many practical advantages (PATEL; POLADI, 2022) as it is commonly used in many areas, is flexible to use with any learning algorithm, is not prone to overfitting, has the possibility to try variants, improves classification accuracy, can reduce the bias and variance produced by the learning algorithm, and has a low computational cost.

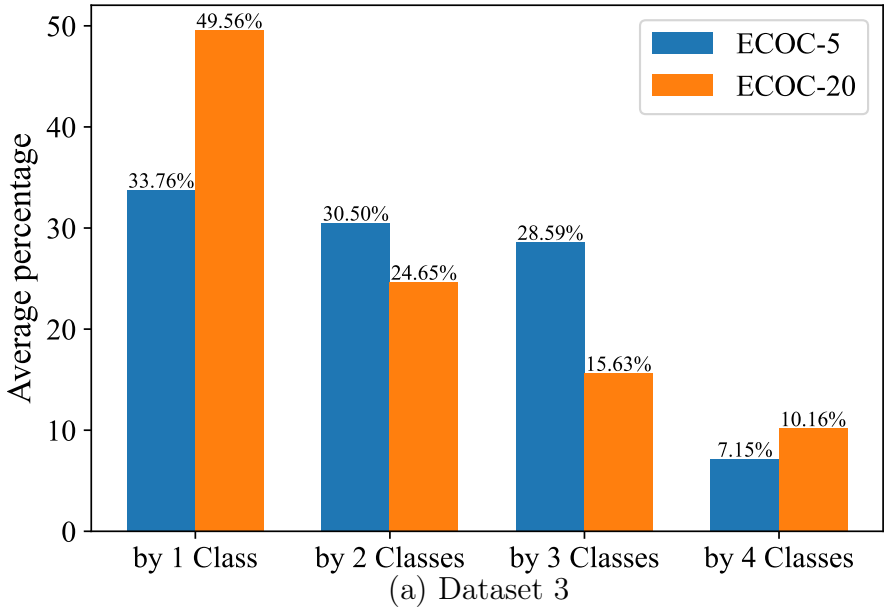
It also has some limitations, for example, it is only successful if the errors in the various bit positions are relatively uncorrelated so that there are few simultaneous errors across various bit positions at the same time. If there are simultaneous errors that exceed the error correction limit, the ECOC will not be able to correct them. ECOC is not effective if each individual codeword is not well separated from other codewords by a large Hamming distance. Depending on how the code matrix is constructed, another difficulty is how some binary classifiers can divide the class space into non-contiguous patterns, presenting



(a) Dataset 1

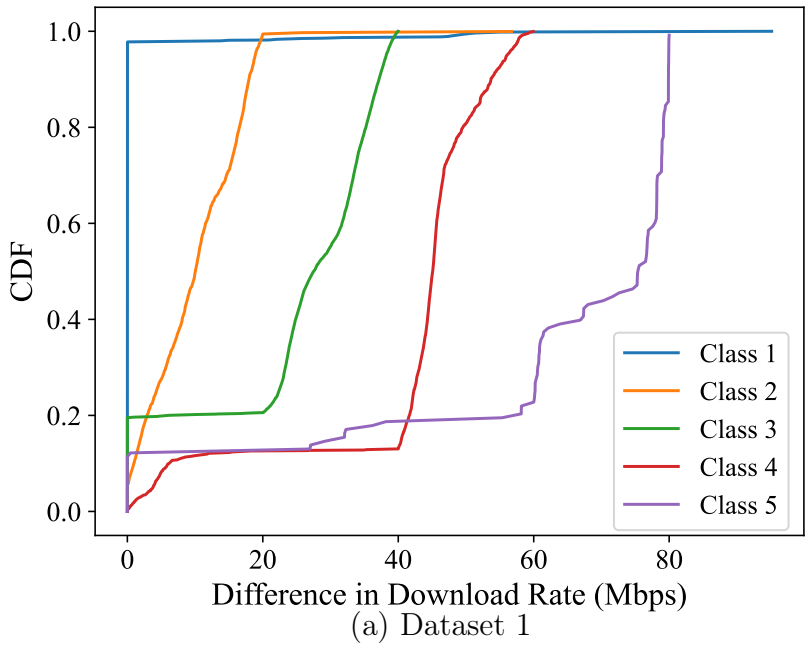


(a) Dataset 2

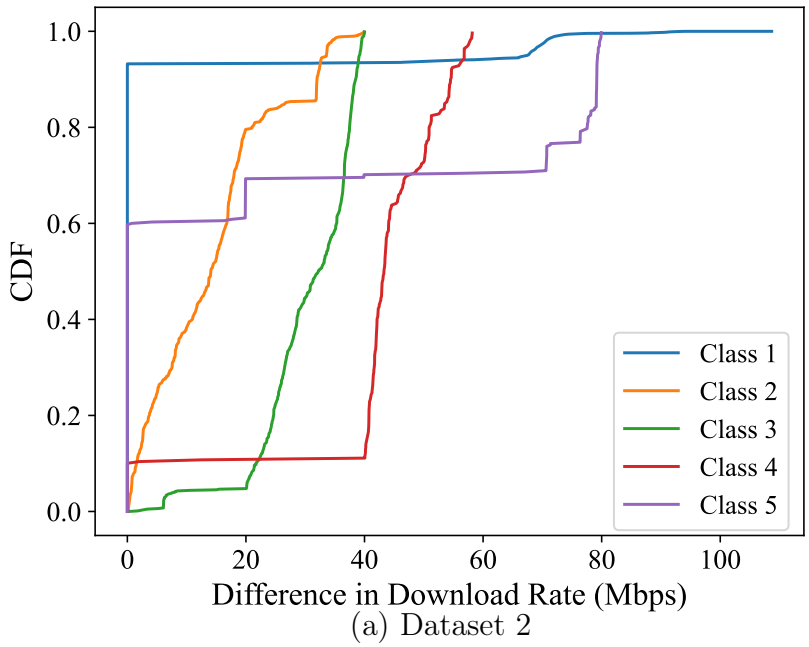


(a) Dataset 3

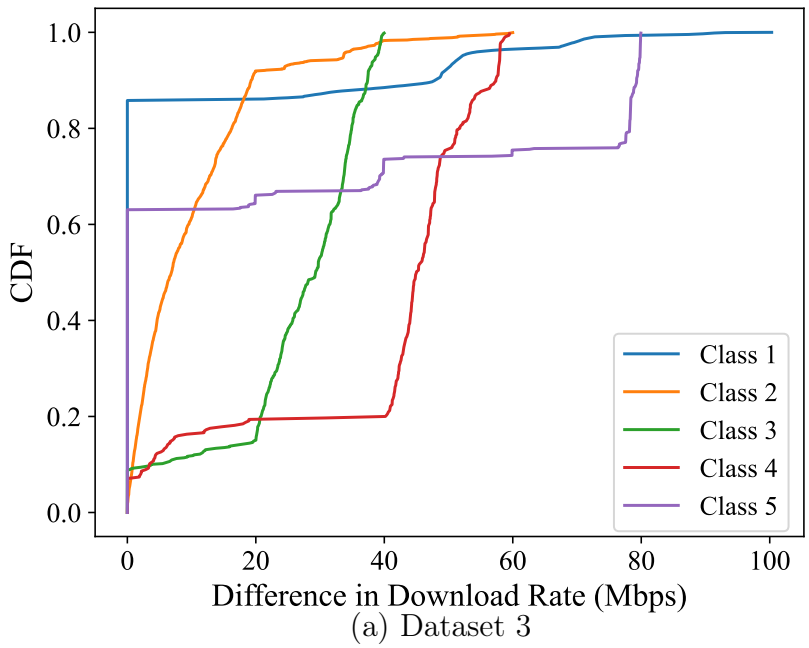
Figure 11: Classification error of our models: level of deviation from the actual class



(a) Dataset 1

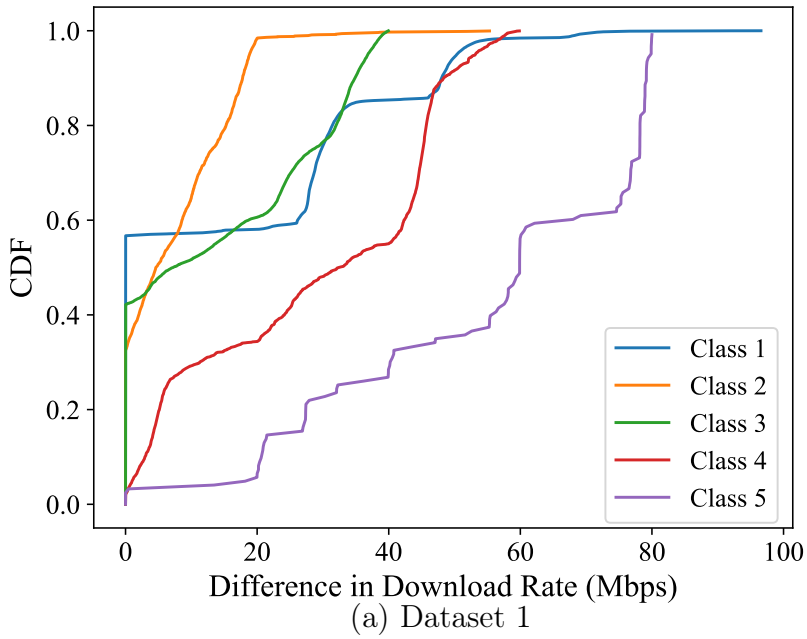


(a) Dataset 2

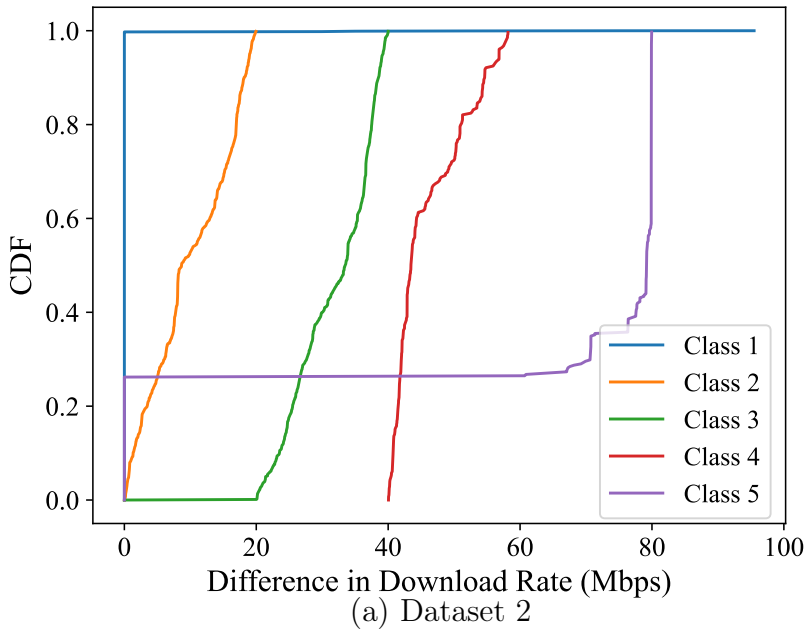


(a) Dataset 3

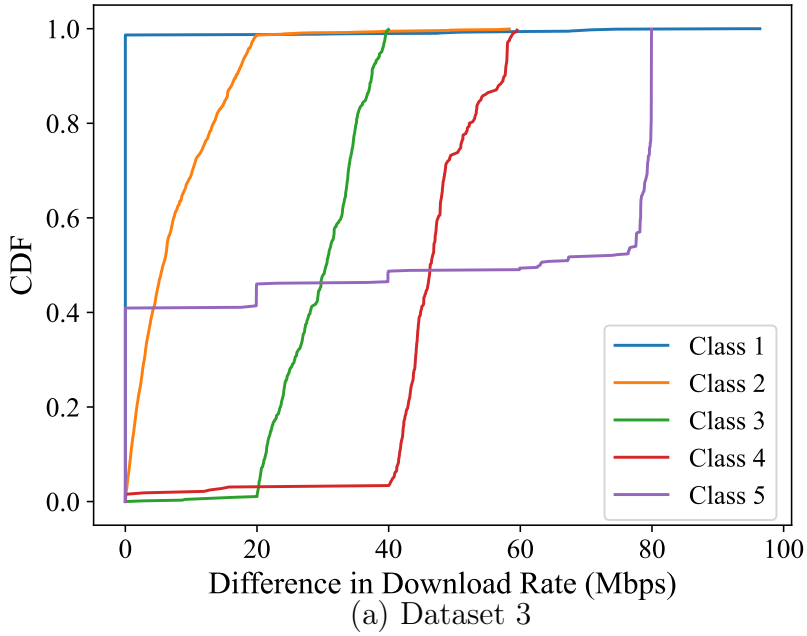
Figure 12: ECOC-5: CDF of Difference in Download Rate (Mbps)



(a) Dataset 1



(a) Dataset 2



(a) Dataset 3

Figure 13: ECOC-20: CDF of Difference in Download Rate (Mbps)

a special challenge to base classifiers, for example, according to column 6 of the Table 1.

There are many ways to build an ECOC classifier as described in Chapter 4. The described methods can derive multiple combinations between encoding and decoding designs. We decided to go with the original design proposal although there are many opportunities to explore other design combinations and other machine learning algorithms, for example, unsupervised algorithms to be used in base classifiers.

In the decoding phase, the tie-breaking process can play an important role as only 1 bit can lead to the correct or incorrect codeword in error correction. Also, the are many ways to improve and explore this step.

The results indicate that our ECOC models are superior compared to the baselines, especially the ECOC-5 for Dataset 1 and the ECOC-20 for Datasets 2 and 3. Comparing both ECOC models to each other, although ECOC-20 has lower classification errors as seen in Figure 11 and has higher average accuracy on Datasets 2 and 3, it does not show the best performance for Dataset 1 consistently.

Despite increasing the number of submodels in each base classifier in the ECOC models from five to twenty, the results for Dataset 1 are counter-intuitive. It was expected that ECOC-20 would win in all datasets. Comparing Dataset 1 against Datasets 2 and 3, we see from Figure 7 that Dataset 1 has almost twice as many measurements on average over the days as compared to the remaining two datasets and the proportions of its classes are not similar to those in Datasets 2 and 3. These differences could be a possible explanation for our results even though classification for Dataset 1 is challenging for all tested models as seen by the large variance in Figure 8[Dataset1].

Although the measurements in all datasets represent 5% of total customers by the telecommunication company, it is unclear if this data is a well-representative sample. We assume this data to be a proper sample as a premise in this work.

Based on our experimental results, although the ECOC models appear to be the best choice to accurately predict the contracted download rate by customers, the best models are selected based on the test accuracy, and ECOC-5 and ECOC-20 achieve the best results for the problem in different scenarios. We address this drawback in the next subsection where we propose our Oracle framework. The proposal is to select the best-performing model by looking at the training accuracy of the models and have the option of selecting the best models used from all the comparative baselines as well as ECOC models.

7 Oracle framework

As mentioned at the end of the previous subsection, our ECOC models give good results for different Datasets. In a way to select the best model, the Oracle framework is proposed. The main idea is to compare the training accuracy of the models to determine which one would perform the best in the test phase. Additionally, users will be able to choose the top models from both ECOC models and all of the comparative baselines. The next subsections discuss the details of this proposal and its experimental evaluation organized in subsections 7.1 and 7.2 respectively.

7.1 Proposed framework

The proposed framework consists of selecting not only the ECOC-5 and ECOC-20 models but also the models used as a comparative baseline - DRF, XGBoost, GLM, and GBM - with the best performance in the training phase and then applying it in the testing phase. The key idea is to capture the best of proposed ECOC models for different scenarios.

Figure 14 describes the flow of the Oracle framework used in this work. It follows some classical steps to address a machine learning problem. The datasets are divided into 80% for training and 20% for testing (grey boxes). During the training phase, the cross-validation technique is applied in the portion intended for training, with 5 folds (dark blue box), for all models (orange boxes).

At the modeling selection phase (the light blue region), the accuracy of each of the 5 folds is computed for each model (light orange boxes) and then the average accuracy score is calculated (training evaluation box). The model with the best training average accuracy score then is selected (yellow box) to apply to the test portion of the dataset. At the end, the final accuracy of the chosen model is computed (green box).

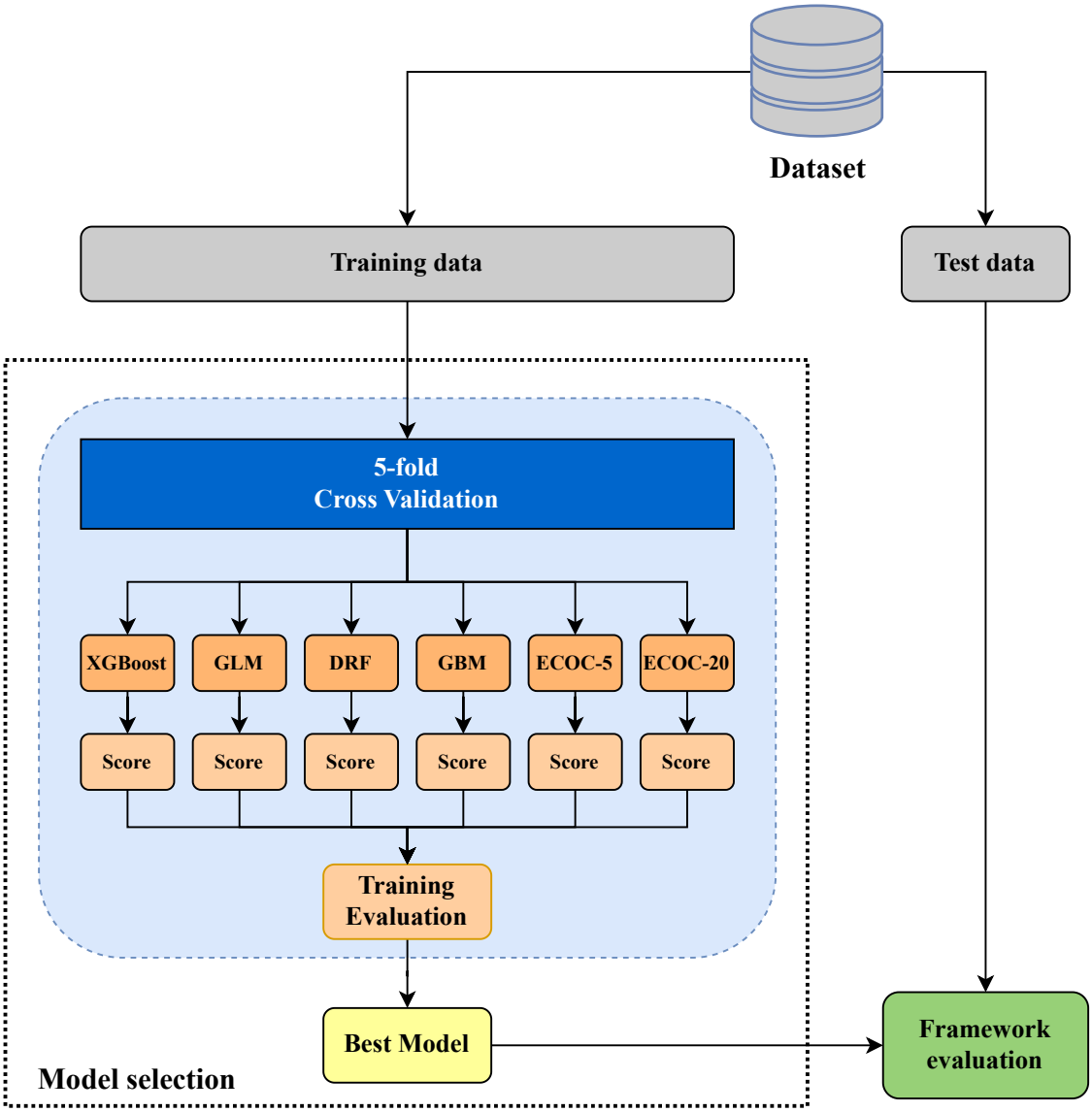


Figure 14: The Oracle framework

7.2 Oracle framework evaluation

For the Oracle framework evaluation, we elevate ECOC-5 and ECOC-20 results as a new comparative baseline, and Figure 15 is the boxplot of the accuracy distribution for all Datasets.

In Dataset 1, Oracle framework values from the first quartile to the third quartile are above 60% accuracy. on the other hand, 75% of ECOC-20 values are below 60% accuracy. ECOC-5 has a compressed shape that denotes very low dispersion denoting that ECOC-5 has good generalization and is the best choice for that Dataset.

In Dataset 2, 100% of the Oracle framework values are above 90% accuracy staying in an intermediate position compared with ECOC models. ECOC-20 is the leader as it has low dispersion and high accuracy values.

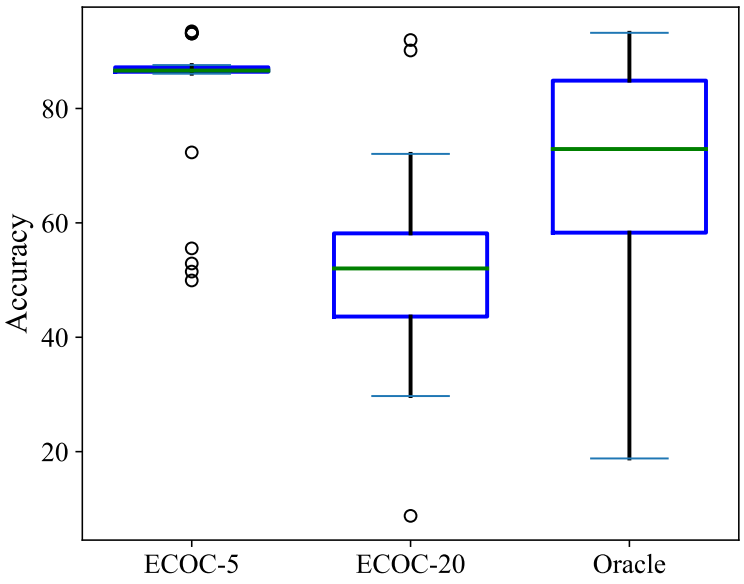
The results in Dataset 3 are similar to Dataset 2 but ECOC-5 is better and has less dispersion than its version in Dataset 2. Moreover, the Oracle framework and ECOC-5 results are pretty similar. Again, ECOC-20 is the leader with the highest accuracy values and very low dispersion.

Although Oracle framework shows a not better distribution in some cases, for example, in Dataset 1 is worse than ECOC-5, and in Datasets 2 and 3 is not better than ECOC-20, its distribution shows that Oracle framework has satisfactory results, especially in Datasets 2 and 3. The average accuracy is 68.13%, 91.87%, and 88.99% on Datasets 1, 2, and 3 respectively.

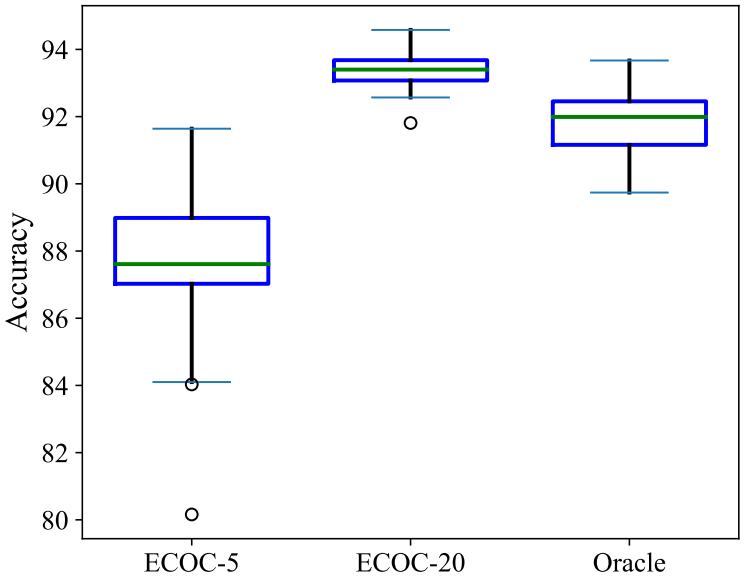
Figure 16 shows the ranking of the models' accuracy in the training phase for each day. As mentioned at the beginning of this chapter, for each day 6 models are trained and their scores are computed. Then, the model with the highest score is selected to be applied in the test phase.

In a general way, the rank results can be summarized since DRF, XGBoost, and GBM always share the top 3 positions for all Datasets, some higher, others lower, and vice-versa. Similarly, GLM, ECOC05, and ECOC-20 always share the bottom positions. Although, it is counter-intuitive that none of the ECOC models are selected to compose the Oracle framework a possible explanation for this behavior is the high training score of those models suggesting maybe an overfitting behavior that could prevent the ECOC models from being selected.

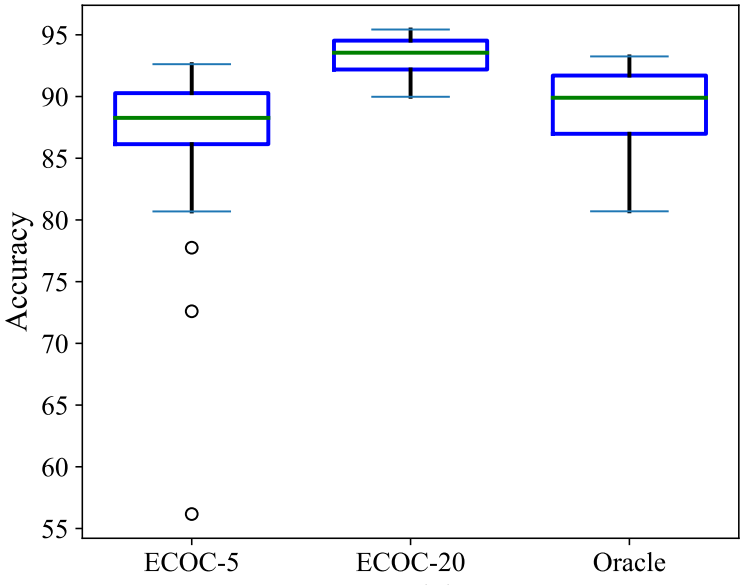
The results indicate that the Oracle framework is not the best nor is not the worst



(a) Dataset 1



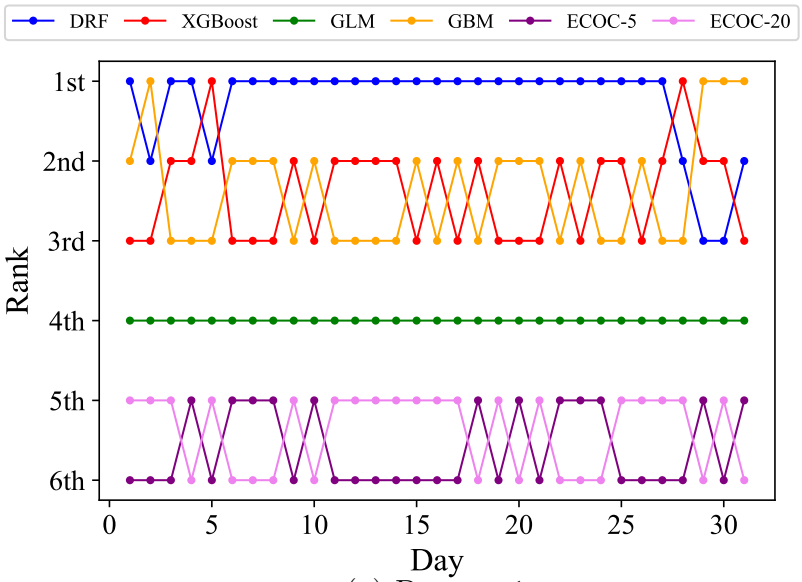
(a) Dataset 2



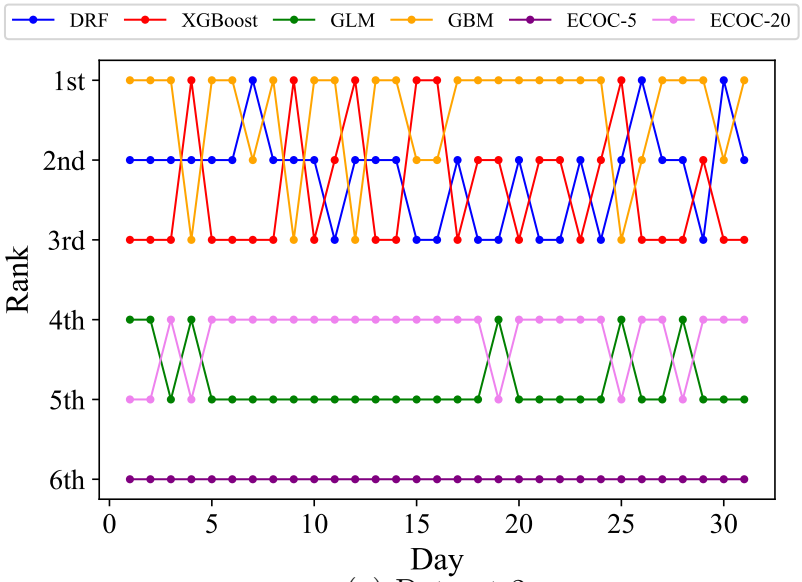
(a) Dataset 3

Figure 15: Boxplot: Accuracy

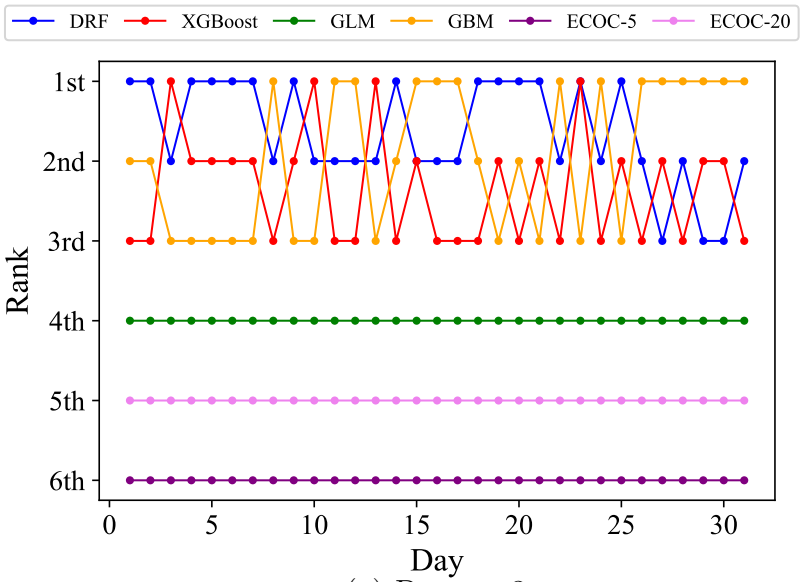
among the ECOC baselines. Although it is not always a winner, it is reasonably well positioned regarding the others. As the average accuracy of the framework is established at intermediate values between ECOC-5 and ECOC-20, it would be reasonable to say that the Oracle framework can be used to obtain more stable results between the two models.



(a) Dataset 1



(a) Dataset 2



(a) Dataset 3

Figure 16: Oracle framework model selection rank

8 Final considerations

In this Chapter, we discuss the conclusions, the future research directions, and the scientific contributions. This work investigates the problem of predicting the quality of broadband users in a large telecommunications company using only categorical features rather than probe network measurements.

8.1 Conclusion

The work answers the proposed question by suggesting two ECOC models, ECOC-5 and ECOC-20, and the Oracle framework. In ECOC-20, we changed two aspects regarding ECOC-5 which is the raised number of different variations to be experimented with for each base classifier from five to twenty, and made flexible the number of machine learning algorithms on these variations from only one (GLM) to four (GLM, GBM, XGBoost, and DRF). Similarly to the ECOC-5, the ECOC-20 classifies the download rate achieved by customers into five different classes using features related to customer location, Internet plan, and broadband device. Our results demonstrate that, in general, our ECOC models are superior to the baseline algorithms, with ECOC-5 winning on dataset 1 and ECOC-20 on datasets 2 and 3. Our experiments showed that the ECOC-20 achieved an average accuracy of around 51.55%, 93.37%, and 93.25% on datasets 1,2, and 3 respectively.

In the Oracle framework, we structured a way to select the winner model among the GLM, GBM, XGboost, DRF, ECOC-5, and ECOC-20 models in the training phase to be applied in the test phase. The results of the Oracle framework show that the task of finding the best model has reasonable accuracy among the best models in the testing phase with average accuracy of 68.13%, 91.87%, and 88.99% on Datasets 1, 2, and 3 respectively.

Those proposals, ECOC-5, ECOC-20, and the Oracle framework are applied to the data obtained from 5% of the total customers of the telecommunications company consisting of 31 days starting from May 22nd, 2021 in Rio de Janeiro state (dataset 1), and

from 1st January 2022 in Rio de Janeiro and São Paulo state (datasets 2 and 3).

The contributions can be used by the telecommunications company to improve its quality of service, maintain user satisfaction, and retain existing customers.

8.2 Contributions

The results of the research carried out are presented in an article published at The International Wireless Communications & Mobile Computing Conference - IWCMC 2023 and is currently in the process of being submitted to a scientific journal.

1. CUBA, Douglas; KULKARNI, Adita; ROCHA, Antonio A de A. Predicting Customer Quality of Service for a Large Fixed Broadband Service Provider. In: IEEE. 2023 International Wireless Communications and Mobile Computing (IWCMC 2023). P. 669–674. Available at: https://doi.org/10.1109/IWCMC58020.2023.10183169

2. In process of Journal submission.

In the first paper, we investigated the same problem of predicting the download rate of broadband users in a large telecommunications company. We designed the ECOC-5 based on the Error-Correcting Output Codes approach and H2O's Automatic Machine Learning that is trained on the data obtained from 5% of the total customers of the company in May and June, 2021. Our framework classified the download rate achieved by customers into five different classes using features related to customer location, Internet plan, and broadband device. The experiments showed that our framework achieved an accuracy of around 83% on average.

The second paper presents the development of the ECOC-20, by expanding our previous ECOC-5 model, and the Oracle framework in the same context but they are trained in two more datasets, with 31 days starting from 1st January 2022 for Rio de Janeiro and São Paulo states. The ECOC-20 has the same modeling steps as ECOC-5 except for increasing the number of models to be considered for each base classifier from five to twenty models and increasing the set of available machine learning algorithms to be considered from only GLM to DRF, XGBoost, GLM, and GBM. Our results demonstrate that our ECOC-20 model is superior to the comparative baseline algorithms, being the best model on datasets 2 and 3. Our experiments showed that the ECOC-20 achieved an average accuracy of around 93.37%, and 93.25% on datasets 2 and 3 respectively.

Furthermore, the Oracle framework is also proposed and has reasonable results with an average accuracy of 68.13%, 91.87%, and 88.99% on Datasets 1, 2, and 3 respectively.

In summary, the main contributions are:

- The development of two ECOC models, ECOC-5 and ECOC-20, to accurately predict the achievement of customers' download rate. In each ECOC model, we flexibly model the choice of machine learning algorithm. According to our best understanding, this flexibility has not been explored in the literature.
- In the literature, this is the first time that ECOC modeling has been used in the telecommunications domain, specifically in predicting broadband Internet metrics.
- Considering the vast ECOC literature, we introduce a new tie-breaking method that takes into account the best training metric of each base classifier for value vector composition so that there is a final tie-breaker based on the position of the tied bits.
- We built the Oracle framework as a way of selecting the best classification algorithm from the presented algorithms.

8.3 Future works

The ECOC model has a rich structure that allows the exploration of multiple variations in its design. This work uses the classical approach in ECOC modeling. In the future, can be prospected some different design combinations in the encoding and decoding phase. For example, expanding our ECOC model to use Ternary code and use *problem-dependent* approaches, and for the base classifiers can experiment with unsupervised learning techniques. The tie-breaking process for the same Hamming Distance values can be improved by taking into account other aspects like history for hits by the base classifier, how good a contribution certain base classifiers have in the code matrix, and even a random fashion to decide to untie. We also plan to leverage the temporal aspect of our data and model the multiclass classification as a time series problem, where we aim to use a sliding window approach to predict customer quality based on data available for the past n time steps. We have an intuition that this approach will aid in increasing the performance of our model.

In the Oracle framework, we can add extra decision parameters to be considered in the decision of the winner model in the training phase like other machine learning evaluation metrics and time to run the task.

In the current work, we focus on predicting only one metric — the download rate achieved by customers. Other metrics such as jitter, packet loss, and latency are also important for several applications. For example, gaming requires that customers do not incur high latency; it is important to minimize jitter and packet loss for efficient video streaming and video calling. Thus, in the future, we plan to extend our framework to accurately predict other transmission quality metrics like jitter, packet loss, and latency incurred by customers. Our work focuses on predicting customer quality for a fixed broadband Internet network. This can be extended in the future for other types of networks such as mobile broadband and client-server networks. Other domains can benefit from using the ECOC modeling as its intuition is a *divide-and-conquer* the problem in multiple binary sub-problems. For example, these sub-problems can be executed separately in different *Internet of Things* (IoT) hardware and then a central entity can group the symbols and decode the codeword to the properly expected label.

REFERENCES

ADEKITAN, Aderibigbe Israel; ABOLADE, Jeremiah; SHOBAYO, Olamilekan. Data mining approach for predicting the daily Internet data traffic of a smart university. **Journal of Big Data**, Springer, v. 6, n. 1, p. 1–23, 2019.

ALIPIO, Melchizedek; BURES, Miroslav. Intelligent Network Maintenance Modeling for Fixed Broadband Networks in Sustainable Smart Homes. **IEEE Internet of Things Journal**, IEEE, 2023.

ALLWEIN, Erin L; SCHAPIRE, Robert E; SINGER, Yoram. Reducing multiclass to binary: A unifying approach for margin classifiers. **Journal of machine learning research**, v. 1, Dec, p. 113–141, 2000.

_____._____. **Journal of machine learning research**, v. 1, Dec, p. 113–141, 2000.

ALMUKHTAR, Firas H. Lung cancer diagnosis through CT images using principal component analysis (PCA) and error correcting output codes (ECOC). **Journal of Control and Decision**, Taylor & Francis, p. 1–11, 2023.

ANATEL. **Panorama - Complaints 2022**. [S. l.], 2022.

ANATEL Complaints. [S. l.: s. n.]. https://www.gov.br/anatel/pt-br/consumidor/compare-as-prestadoras/reclamacoes-na-anatel.

ANATEL Fixed Broadband panel. [S. l.: s. n.]. https://informacoes.anatel.gov.br/paineis/acessos/banda-larga-fixa.

ANATEL History. [S. l.: s. n.]. https://informacoes.anatel.gov.br/paineis/acessos/historico.

BAGHERI, Mohammad ali; MONTAZER, Gholam Ali; ESCALERA, Sergio. Error correcting output codes for multiclass classification: application to two image vision problems. In: IEEE. THE 16th CSI international symposium on artificial intelligence and signal processing (AISP 2012). [S. l.: s. n.], 2012. P. 508–513.

BALAMURUGAN, V et al. Multi-label Text Categorization using Error-correcting Output Coding with Weighted Probability. **International Journal of Engineering**, Materials e Energy Research Center, v. 35, n. 8, p. 1516–1523, 2022.

BHARAMBE, Yashraj et al. Churn Prediction in Telecommunication Industry. In: IEEE. 2023 International Conference for Advancement in Technology (ICONAT). [S. l.: s. n.], 2023. P. 1–5.

BISCHOF, Zachary S; BUSTAMANTE, Fabián E; STANOJEVIC, Rade. Need, want, can afford: Broadband markets and the behavior of users. In: PROCEEDINGS of the 2014 Conference on Internet Measurement Conference. [S. l.: s. n.], 2014. P. 73–86.

DIETTERICH, Thomas G; BAKIRI, Ghulum. Solving multiclass learning problems via error-correcting output codes. **Journal of artificial intelligence research**, v. 2, p. 263–286, 1994.

DOROGUSH, Anna Veronika; ERSHOV, Vasily; GULIN, Andrey. CatBoost: gradient boosting with categorical features support. **arXiv preprint arXiv:1810.11363**, 2018.

ESCALERA, Sergio; PUJOL, Oriol; RADEVA, Petia. Boosted Landmarks of Contextual Descriptors and Forest-ECOC: A novel framework to detect and classify objects in cluttered scenes. **Pattern Recognition Letters**, Elsevier, v. 28, n. 13, p. 1759–1768, 2007.

_____._____. **Pattern Recognition Letters**, Elsevier, v. 28, n. 13, p. 1759–1768, 2007.

GUO, Yong Zhen; RAMAMOHANARAO, Kotagiri; PARK, Laurence AF. Error correcting output coding-based conditional random fields for web page prediction. In: IEEE. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. [S. l.: s. n.], 2008. v. 1, p. 743–746.

HASTIE, T.; TIBSHIRANI, R. Classification by Pairwise Grouping. **Proc. Conf. Neural Information Processing Systems**, v. 26, p. 451–471, 1998.

ITU. **Measuring digital development: Facts and Figures 2022**. [S. l.], 2022.

KULKARNI, Adita et al. Deepchannel: Wireless channel quality prediction using deep learning. **IEEE Transactions on Vehicular Technology**, IEEE, v. 69, n. 1, p. 443–456, 2019.

LEDELL, Erin; POIRIER, Sebastien. H2O AutoML: Scalable Automatic Machine Learning. **7th ICML Workshop on Automated Machine Learning (AutoML)**, jul. 2020. Disponível em: <https://www.automl.org/wp-content/uploads/2020/07/AutoML%7B%5C_%7D2020%7B%5C_%7Dpaper%7B%5C_%7D61.pdf>.

LEE, Jinsung et al. PERCEIVE: deep learning-based cellular uplink prediction using real-time scheduling patterns. In: PROCEEDINGS of the 18th International Conference on Mobile Systems, Applications, and Services. [S. l.: s. n.], 2020. P. 377–390.

LEMAÎTRE, Guillaume; NOGUEIRA, Fernando; ARIDAS, Christos K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. **Journal of Machine Learning Research**, v. 18, n. 17, p. 1–5, 2017. Disponível em: <http://jmlr.org/papers/v18/16-365>.

LI, Baoli; VOGEL, Carl. Improving multiclass text classification with error-correcting output coding and sub-class partitions. In: SPRINGER. CANADIAN Conference on Artificial Intelligence. [S. l.: s. n.], 2010. P. 4–15.

LI, Zhen-Jun; LU, Yun-Ting. A network QoS evaluation method based on customer satisfaction indices. In: IEEE. 2009 International Conference on Machine Learning and Cybernetics. [S. l.: s. n.], 2009. v. 3, p. 1328–1332.

LIU, Kun-Hong; ZENG, Zhi-Hao; NG, Vincent To Yee. A hierarchical ensemble of ECOC for cancer classification based on multi-class microarray data. **Information Sciences**, Elsevier, v. 349, p. 102–118, 2016.

MADANAPALLI, Sharat Chandra; GHARAKHIELI, Hassan Habibi; SIVARAMAN, Vijay. Inferring netflix user experience from broadband network measurement. In: IEEE. 2019 Network Traffic Measurement and Analysis Conference (TMA). [S. l.: s. n.], 2019. P. 41–48.

MAIER, Gregor et al. On dominant characteristics of residential broadband internet traffic. In: PROCEEDINGS of the 9th ACM SIGCOMM Conference on Internet Measurement. [S. l.: s. n.], 2009. P. 90–102.

MAJIDIAN, Zohre et al. An intrusion detection method to detect denial of service attacks using error-correcting output codes and adaptive neuro-fuzzy inference. **Computers and Electrical Engineering**, Elsevier, v. 106, p. 108600, 2023.

MASLO, Anis et al. Machine Learning and Quality of Customer Experience in Leading Telecom Providers of Bosnia and Herzegovina. In: IEEE. 2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). [S. l.: s. n.], 2021. P. 241–245.

NILSSON, NJ. Learning machines McGraw-Hill. **New York**, v. 19652, 1965.

PASSERINI, Andrea; PONTIL, Massimiliano; FRASCONI, Paolo. New results on error correcting output codes of kernel machines. **IEEE transactions on neural networks**, IEEE, v. 15, n. 1, p. 45–54, 2004.

PATEL, Rinkal K; POLADI, Irfan. A STUDY PAPER ON ERROR CORRECTING OUTPUT CODE BUILD ON MULTICLASS CLASSIFICATION. **International Journal of Engineering Applied Sciences and Technology**, v. 7, n. 10, p. 124–128, 2022.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PRAKASH, U et al. A Survey on Artificial Intelligence in Telecommunication for Churn Prediction. In: IEEE. 2022 6th International Conference on Electronics, Communication and Aerospace Technology. [S. l.: s. n.], 2022. P. 1261–1265.

PUJOL, Oriol; ESCALERA, Sergio; RADEVA, Petia. An incremental node embedding technique for error correcting output codes. **Pattern Recognition**, Elsevier, v. 41, n. 2, p. 713–725, 2008.

_____._____. **Pattern Recognition**, Elsevier, v. 41, n. 2, p. 713–725, 2008.

PUJOL, Oriol; RADEVA, Petia; VITRIA, Jordi. Discriminant ECOC: A heuristic method for application dependent design of error correcting output codes. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 28, n. 6, p. 1007–1012, 2006.

RAJ, Raushan et al. Wireless Channel Quality Prediction using Sparse Gaussian Conditional Random Fields. In: 2021 IEEE 18th Annual Consumer Communications Networking Conference (CCNC). [S. l.: s. n.], 2021. P. 1–6. DOI: 10.1109/CCNC49032.2021.9369651.

RAYKAR, Vikas C; SAHA, Amrita. Data split strategiesfor evolving predictive models. In: SPRINGER. MACHINE Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15. [S. l.: s. n.], 2015. P. 3–19.

RUKHSAR, Salim. Discrimination of multi-class EEG signal in phase space of variability for epileptic seizure detection using error correcting output code (ECOC). **International Journal of Information Technology**, Springer, v. 14, n. 2, p. 965–977, 2022.

SAMBA, Alassane et al. Instantaneous throughput prediction in cellular networks: Which information is needed? In: IEEE. 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM). [S. l.: s. n.], 2017. P. 624–627.

SREE, KPNV Satya et al. Optimized conversion of categorical and numerical features in machine learning models. In: IEEE. 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC). [S. l.: s. n.], 2021. P. 294–299.

SUNDARESAN, Srikanth et al. Broadband internet performance: a view from the gateway. **ACM SIGCOMM computer communication review**, ACM New York, NY, USA, v. 41, n. 4, p. 134–145, 2011.

TERRY, Windeatt; REZA, Ghaderi. Coding and decoding for multi-class learning problems [J]. **Information Fusion**, v. 43, n. 4, p. 11–21, 2003.

ÜBEYLI, Elif Derya. ECG beats classification using multiclass support vector machines with error correcting output codes. **Digital Signal Processing**, Elsevier, v. 17, n. 3, p. 675–684, 2007.

WANG, Jinling et al. Novelty Prediction in Broadband Line Multi-variate Time Series Using a Deep Long Short-Term Memory Network. In: IEEE. 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET). [S. l.: s. n.], 2021. P. 1–6.

WONG, Felix Ming Fai et al. Improving user QoE for residential broadband: Adaptive traffic management at the network edge. In: IEEE. 2015 IEEE 23rd International Symposium on Quality of Service (IWQoS). [S. l.: s. n.], 2015. P. 105–114.

XIAO-FENG, Liu; XUE-YING, Zhang; JI-KANG, Duan. Speech recognition based on support vector machine and error correcting output codes. In: IEEE. 2010 First International Conference on Pervasive Computing, Signal Processing and Applications. [S. l.: s. n.], 2010. P. 336–339.

XIE, Xiao et al. Network traffic classification based on error-correcting output codes and nn ensemble. In: IEEE. 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery. [S. l.: s. n.], 2009. v. 3, p. 475–479.

YE, Qixiang; LIANG, Jixiang; JIAO, Jianbin. Pedestrian detection in video images via error correcting output code classification of manifold subclasses. **IEEE Transactions on Intelligent Transportation Systems**, IEEE, v. 13, n. 1, p. 193–202, 2011.

ŻELASKO, Dariusz; PŁAWIAK, Paweł; KOŁODZIEJ, Joanna. Machine learning techniques for transmission parameters classification in multi-agent managed network. In: IEEE. 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID). [S. l.: s. n.], 2020. P. 699–707.

ZHANG, Min-Ling; YU, Fei; TANG, Cai-Zhi. Disambiguation-free partial label learning. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 29, n. 10, p. 2155–2167, 2017.

ZHAO, Yunhao; SHU, Xiaoqing. Speech emotion analysis using convolutional neural network (CNN) and gamma classifier-based error correcting output codes (ECOC). **Scientific Reports**, Nature Publishing Group UK London, v. 13, n. 1, p. 20398, 2023.

APPENDIX A – Dataset sample

Columns	**Sample**
DIADATPER	44285
DEVICEID	000AC2-FHTT23A7D3D0
TIMESTAMP	44285,87477
ISP	tim
TYPE	scm
MANUFACTURER	FiberHome
CPE_MODEL_RAWDATA	AN5506-04-FAT
SOFTWARE_VERSION	RP2662
IP_ADDRESS	179,54,157,181
TEST_POINT	186,231,1,202
DOWNLOAD_STATE	Completed
DOWNLOAD_FILESIZE_TEST	321090210
DOWNLOAD_RATE	73280367,23
UPLOAD_TEST	Completed
UPLOAD_FILESIZE_TEST	47185920
UPLOAD_RATE	61389218,04
UDP_STATE	Completed
UDP_ESTIMATED_TRAFFIC	6400
UDP_LATENCY	84,79995
UDP_JITTER	43,34978
UDP_PACKET_LOSS_PERC	0
TOTAL_TRAFFIC_SENT	49271278
TOTAL_TRAFFIC_RECV	328476608
TOTAL_TRAFFIC	377747886
TEST_ORIGIN	APP

SOURCE	Auto
MAC_ADDRESS	
CPE_MODEL_REGMAN	Fiberhome AN5506-04 GPON TR-181
FIRMWARE_REGMAN	RP2662
PLANO_ATUAL	TIM Live A 150 Mega Plus-GGO
PLANO_CNT	150MB/40MB
BAIRRO	ANAPOLIS CITY
CIDADE	ANAPOLIS
UF	GO
CEP	75094170
PLANO_CFG	150MB/40MB
DESTLVNAS	RMAGANS-ANS012-01
MSAN	GOGNA_MSAN0047
PORTA	1/9:011
QTD_DIST_IP	1
QTD_DIST_PLN	1
QTD_DIST_MSAN	1
QTD_DIST_MSAN_PORTA	1
QTD_DIST_NAS	1
QTEMBYUPL	1098,711634
QTEMBYDNL	8203,760454
QTEEMBYTOT	9302,472088
QTEEMBYTOT_ALL	9302,472088
STATUS_CONTRATO	a
CLASSIF	FTTH

APPENDIX B – Categorical sample

	Sample 1	**Sample 2**	**Sample 3**
Device ID	000AC2-FHTT23A4CB98	000AC2-FHTT23A4CBB8	000AC2-FHTT23A4CE00
Contracted Plan	100MB/40MB	150MB/60MB	150MB/40MB
Configured Plan	100MB/40MB	150MB/60MB	300MB/150MB
City	DUQUE DE CAXIAS	RIO DE JANEIRO	RIO DE JANEIRO
Neighborhood	PARQUE LAFAIETE	VILA ISABEL	VILA ISABEL
Zipcode	25025104.0	20551070.0	20540365.0
BRAS	RMAGSMI-MRT01-03	RMAGRJO-RJO03-02	RMAGRJO-RJO03-03
MSAN	RJRJO_MSAN1467	RJRJO_MSAN1433	RJRJO_MSAN1428
Port	2/10:030	3/9:031	4/8:020
Download Rate	113.557716	110.160914	73.068992