ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 11 (2024) No. 6, pp. 77-91
DOI:10.33168/JLISS.2024.0605

Examining Data Needs and Predictive Models for Broadband Take-Up Ratio Estimation through CRISP-DM and Literature Review

Wahyu Haris Kusuma Atmaja [1], Haryono [1], Arief Ramadhan [2], Muhammad Zarlis [1]

[1] Computer Science Department, Bina Nusantara Graduate Program - Doctor of Computer Science, Bina Nusantara University, West Jakarta, Indonesia

[2] School of Computing, Telkom University, Bandung, Indonesia

Abstract. This paper aims to examine the data requirements and predictive modeling methodologies for estimating broadband take-up ratio (TUR). The author applies the CRISP-DM approach to define data requirements based on the results of the analysis of the company's internal as well as external factors. Using Kitchenham approach, the author applies the findings of the literature review on machine learning predictive model to the second study area. The CRISP-DM strategy has been demonstrated through the creation of data needs based on business imperatives, data understanding activities that include defining data analysis requirements, identification of attributes for internal and external variables as well as determining resource sources available in a database information system. In addition, as indicated in the system data sources, seven internal factors and nineteen external factors have been identified within the Customer Management System, Survey System, Relations or Micro Clusters. Consequently, neural networks are superior to SVMs and XG Boost when analyzing the data to predict how much uptake there will be. The Neural Network model performed best with an AUC of 79-89, an accuracy level of 75.23-99, an F Measure of 80.97-95.35, an F1 Score of 94.5-98.5, and an R^2 score of 88. SVM, with an AUC of 98.8, accuracy of 92.83-98.35 and F Measure 84.22-95.37 is rated as a prediction model in second place. In contrast, XGBoost had an AUC of 77-99.1, accuracy of 77-98.9, and F1 score of 98.3 as the 3rd. In the case of TUR classification use, XG Boost and SVM were also becoming increasingly common. In terms of AUC, accuracy, F measure and F1 score, XGBoost is more successful than SVM. The AUC of the XG Boost is 99.1, an accuracy range of 77-98.6 and an F1 value of 98.3. Although SVM's AUC is almost the same at 98.8, its F1 score is lower than XG Boost. In conclusion, the research contributes to the identification of relevant data sources and provides an initial framework for the selection of the TUR model methodology. To establish the optimal approaches, additional studies should focus on robust testing and validation of models.

Keywords: FTTH, Internet Broadband, CRISP-DM, broadband adoption, customer acquisition, Machine Learning

1. Introduction

The Pandemic COVID-19, which has been functioning for almost two years, has provided good momentum for economic growth following the pandemic's recovery. During the recovery period, the expansion of broadband internet in non-commercial and commercial areas served as the foundation for community activities that rely on hybrid (online and offline) services. According to the Indonesian Internet User Profile Report published in March 2023 by APJII (the Indonesian Internet Service Provider Association), internet penetration in Indonesia was 78.19% or around 215.63 million users out of a total population of 275.77 million. Over the previous year, internet user growth climbed by 1.17%. 74.34% of the 215.63 million internet users utilize mobile broadband, while 25.66% use fixed broadband (APJII,2023).

Based on these characteristics, mobile and fixed broadband internet access is common and rising year after year. Broadband, also known as the low-cost and simple network transmitter (Abdullah, D. M., & Ameen, S. Y.,2021), is a high-speed communication network that uses a wide bandwidth frequency and can be broadcast in several channels (Merriam-Webster, 2022). The term "Internet broadband" refers to broadband technology that is used to provide Internet services. Customers are provided at their endpoints via optical fiber media. The endpoint is the foundation for designating access; for example, FTTH (fiber to the home) refers to the withdrawal of optical fiber to the customer's endpoint at home. Active optical access networks (OAN) and passive optical networks (PON) can connect centrally to FTTX. Passive access has now surpassed dynamic access in popularity.

Telecommunications operators or internet service provider organizations (ISPs) carry out these responsibilities to actualize the expansion and provision of these services based on the benefits of increasing broadband internet adoption (Briglauer, W. and Gugler, K., 2019). They will design network architecture, build fiber optic networks, provide internet services and content, design consumer endpoints, and activate terminal devices. The capital expenditures budget (CAPEX) supports these initiatives (Spruytte, J., Benhamiche, A., Chardy, M., Verbrugge, S., & Colle, D., 2019). The investment budget must be evaluated regarding its values using cost-benefit analysis (Van den Velden, J., & Sadowski, B. M., 2023) and providing benefits to ISPs (Asgarirad, Mohammadreza; Jahromi, Mansour Nejati, 2020). The investment decision has significant implications for future profits, which depend on the income of clients after deducting their liabilities and operating expenses. The number of users and the monthly membership costs paid to ISP businesses are the two main factors that can influence this decision.

The availability of internet services in a particular region is primarily determined by the percentage take-up rate (TUR). This rate is based on the number of subscribers who have subscribed to the entire capacity of internet connection units that are available to consumers. TUR serves as the basis for cost estimation and decision-making by ISP management in terms of developing and providing internet services.

According to Asgarirad et al (2020), the condition of a place should be known to produce profits based on the optimal use of Capex (Sprytte et al., 2019) and the evaluation of capex utilized (Van de Celden et al., 2023). For at least the first three years, mitigation must be implemented in accordance with the Take-Up Rate criteria. According to the results of the previous Literature Review, only one reference was found regarding predictions of TUR in the European Union based on FTTH Council Europe and EU DAE Dashboard data for 2003-2015, as stated by Briglauer and Gugler (Briglauer & Gugler, 2019), namely less than 30% of basic FTTH broadband services. Customers are reluctant to switch from other mobile or FTTH providers to FTTH broadband services. Briglauer & Gugler's upper bounds for 2SLS and OLS coefficient estimations (Briglauer & Gugler, 2019) were used to calculate the greatest value of hybrid and end-to-end fiber take-up rates observed at the end of our analysis period (TUR of 2015 ~20% for hybrid fiber and TUR of 2015 ~25% for FTTH/B).

As a result, this study will delve deeper into the prediction method or TUR calculation derived from the analysis of micro variables implementing the Internet Service Provider, to produce a more realistic computation operationally. Simultaneously, data is acquired from Customer Relationship Management (CRM) about the elements that determine the amount of TUR in a cluster using standard approaches to achieve better results. Therefore, this study focuses also on identifying the factors that influence TUR, with the first research question identifying the factors within the data in the enterprise system. As a second research topic, we are exploring various methods for predicting TUR.

This paper is divided into four sections. The first section provides context and an introduction to FTTH internet broadband. The second section comprises the previous study conducted, followed by the third section, which describes the research method. Next, the result and discussion will be explained in the fourth section. The final component, the fifth section, gives conclusions and prospects for further research.

2. Literature Review

According to Cambini and Jiang (Cambini and Jiang, 2009), broadband, also known as the Next Generation Network, is a high-speed communication network that uses frequency bandwidth (band) or wide band, which is divided into numerous channels or streams. It is referred to as "broadband internet" when it is implemented over the internet. High-speed internet connections are typically used to send large amounts of data. The Federal Communication Commission (FCC) (FCC, 2022) states that broadband internet speeds can vary depending on the technology, media, and service sought. The media used to share broadband services is classified as either wired or wireless. According to Milanovic (Milanovic, 2014) and Pradeep (Pradeep et al., 2017), fiber optic cables are connected from the optical line terminal (OLT) on the operator's equipment, which acts as a symmetrical divider of broadband services, to the optical network terminal (ONT) at the customer's location. The connection is set up for FTTX broadband internet service. According to Briglauer and Gugler (Briglauer and Gugler, 2019), internet service providers or telecommunications operators (ISP) operate as the developers and connectors of broadband internet services in an area.

The telecommunications provider has the responsibility to carry out internet broadband activities such as infrastructure deployment, sales, operations & maintenance, and compete with other operators. To ensure the performance of the telecommunications provider, Sharma & Misra (Sharma et al., 2017) suggest examining two criteria. The first factor is the company's profile and services, while the second is the overall location profile. To classify the factors that impact on the take-up rate, we divide them into internal company and external location profiles. The percentage of the take-up rate in a location is still a conditional/dependent factor. According to the articles, the internal company profile has 18 factors, with affordable prices being the most important one. Other internal factors include demand forecast, activation time, product bandwidth, broadband quality, availability of marketers, expected revenue, churn area & individual churn prediction, subscription method, ease of obtaining, product value, service quality, facilities, low latency, premium services, coverage, and package product.

Twenty-one external factors can affect the performance of telecommunication providers. These factors include the competitors in the market, population density, income per area, demographics, skills of the population, willingness to pay for services, geographical location, income levels, education levels, the number of personal computers and gadgets in use, the authority of local municipality leaders, the utility of building businesses, the bandwidth of competitors, aerial location, the age of the population, English literacy rates, the number of teenagers below 18 years, and new housing areas. Therefore, internal and external factors are compared with the sources of company data and can be processed later.

According to Schneir and Xiong's research article on TUR percentage prediction in (Limbach et al. 2016), the take-up rate for the first three years ranged from 25% to 70%. The churn and take-up rates are equal or remain constant for the remaining investment term. The take-up and churn rates are frequently believed to be the same from the fourth to the tenth year of an investment. This is because certain turnover causes are like those that influence take-up rates. Various methods for measuring churn rate are evaluated and used to calculate the take-up rate.

3. Research Methodology

To address the two questions stated in the preceding part, a methodology is required as a research reference.

RQ1 addresses two questions: what factors drive TUR within and outside the organization, and how data attributes within the system can be used to collect data that supports these aspects. Next, RQ2 deals with different data science techniques to predict TUR. Answer these questions using the methods described below.

3.1 Systematic Literature Review

The following protocol using Kitchenham's systematic literature protocol (Kitchenham, B., & Charters, S.M., 2021), which conducted a thorough literature review to determine the TUR variables. Kitchenham's process consists of three steps: planning, conducting, and reporting the review. When conducting a review activity, investigate the background and identify the items to focus on.

The first step in planning activities is to determine the research issue, which is followed by setting selection criteria, selection techniques, and assessing the quality of articles, as well as extracting data from the literature review. Next, a report is created to share the findings of the literature review. This is represented in Figure 1 below.

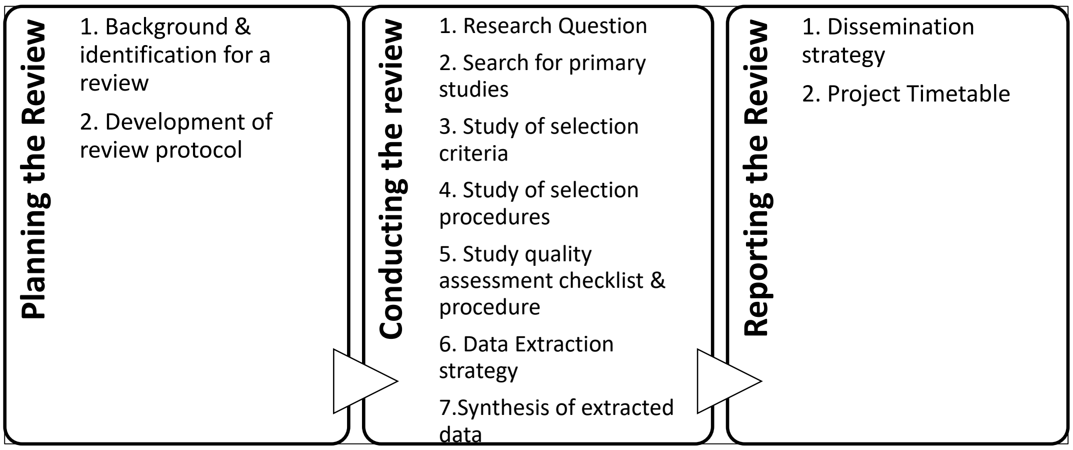

Fig.1: Systematic Literature Review by Kitchenham

Once the factors are found, factor matching is performed using CRISP-DM against the data attributes available in the system.

3.2 CRISP-DM Process

CRISP-DM (Cross Industry Standard Process for Data Mining) was developed by data mining practitioners in 1999. It is a de facto standard approach for delivering data mining (Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N. & Flach, P., 2019) and a data mining-specific process model (Schröer, C., Kruse, F., & Gómez, J. M., 2021).

According to this point of view, research for TUR prediction, which is classified as a data mining activity, is ideal for putting this method into action. CRISP-DM contains six stages for doing data mining research, as indicated in Table 1: business understanding, data understanding, data preparation, modeling, model assessment, and model deployment. Table 1 below explains the six process flows in practice.

Table 1: Explanation Of CRISP-DM Process Flow

No.	Process Flow	Definition
1.	Business Understanding	This part establishes the problem definition, study objectives, data mining type, requirements, and success criteria for data mining to be achieved
2.	Data Understanding	Data collection from data sources, data description, analysis initiation, data quality knowledge, and data evaluation to accomplish the study objectives utilizing data mining techniques.
3.	Data Preparation	In this part, we'll go through how to create and prepare data sets for mining methods, how to convert raw data sets into final data sets through transformation & scaling, defining inclusion & exclusion attributes, cleansing, and how to evaluate the modeling tools.
4.	Modeling	It includes options for defining a data mining method (model) depending on the parameters specified, building test cases, building a model, and calibrating or setting the parameters of the model
5.	Model Evaluation	The results are then reviewed and calibrated for fit with the business objectives during the model evaluation.
6.	Model Deployment	This last stage involves deploying and re-assessing input on study objectives with users via numerous evaluation scenarios.

4. Results and Discussion

The study is intended to address two research objectives: how can data on the prediction of TUR using internal and external internet connection factors be acquired, as referred to in section 1 at the beginning of this discussion? This question shall be used for the TUR predictions or classification of internet connections, and what method approach shall be used in calculating their forecast? The suitability of each method for TUR prediction computations is examined in this topic.

4.1 Result of RQ1

The following study results in Sections 1 and 2 will be used to investigate question 2 using CRISP-DM method as it is important to understand business knowledge. The data understanding part will be explained accordingly.

To answer RQ1, relevant data is needed to indicate the internal and external factors influencing TUR in a cluster. Two activities are required in CRISP-DM to answer RQ1: business understanding and data understanding. In the business understanding part, the following tasks are carried out: determining business problems, determining analytical activities designed to solve business challenges, and determining data requirements. Section 2 states the internal and external variables of a telecommunication business that are successful in providing broadband internet services. CRISP-DM can categorize the company's internal and external elements as factors that influence internet connectivity and whether it affects TUR or not. The analytic activity utilized to complete the business knowledge is the classification of the availability/unavailability of the influence on the internet connection. Based on the definition of business understanding and analytic activity, data requirements are then determined to represent these elements through the availability of data in the firm.

The created data framework will be used as a dataset framework containing internal and external TUR parameters that can be quantitatively examined to find the primary deciding factors of TUR. The collection incorporates information derived from CRM databases, assets, and a preliminary survey of the FTTH cluster. The features and content of the data will be compared to factors derived from earlier studies. If applicable, the traits and data will be included. If they aren't present, secondary data will be sought. This model will be used as a reference for establishing TUR influencing variables and forecasts.

Table 2 shows the identification of business understanding, analytic efforts, and data requirements on internal and external factors influencing TUR at the firm.

Table 2: Identification Business Understanding, Analytic Efforts, and Data Requirements

Business Problems	Analytic Purpose	Attribute	Description	Data Source Avail.
Determine Internal Factors of TUR	classify the cluster's availability and unavailability of broadband internet connectivity elements	1. Affordable Price	Price of Internet Broadband	C: CRM 2019-2022
		3. Activation Time	Time to Instal / Activate Services	C
		4. Marketer	Name of the salesperson/sales partner in charge of customer acquisition	C
		5. Broadband Quality	Broadband SLA Quality	C
		6. Product Bandwidth	Broadband Speed	C
		7. Package Product	Name Package Product	C
		2. Demand Forecast	Percentage Forecast of Demand in an area per month	R: Rismon 2022
Determine External Factors of TUR	classify the cluster's availability and unavailability of broadband internet connectivity elements	5. Skill	Customer's Skill	R
		6. Willingness to Buy/Pay	Percentage of willingness to pay in an area	R
		8. Income	Customer Income Range	R
		9. Education Level	Customer Education Level	R
		10. PC Population	No. of PC in an area	R
		11. Gadget Population	No. of Mobile Phones / Tablets in an area	R
		12. Municipality Leader Authority / Decision Maker	Is it difficult to have a permit from the Municipality Leader?	R
		16. Population Age (Median)	Median of Customer Age in an area	R
		17. English Literature	Customer English Literacy	R
		19. New Housing Area Development	Is it a new housing area	R
		20. Perception/Citizen Perception	Classification of perception	R
		21. Influencer Role	Role of Salesperson/Partner	R
		1. Number of Competitors (New and Existing) / FTTH Availability	No. of competitors in an area	M: Micro 2019-2022
		2. Density (m2)	The density of the population in an area	M
		3. GDP Area/ Income per Area (Median)	Average GDP (revenue) of population in an area	M
		4. Demographic	No. of Population and Gender	M
		7. Geolocation	Classification of geo-area	M
		13. Building / Developer/Business Utilities / (Trade Analysis)	Is it a special Building?	M
		14. Competitor Bandwidth (min.)	Broadband bandwidth competitor	M
		15. Aerial Location	Does it have an aerial pole?	M
		18. Teenagers < 18 years	Is the Customer an 18-year-old teen?	M

In a previous study, researchers identified 18 factors that influence TUR (Take-Up Rate). These factors were mainly internal and were used to determine the attributes in CRM (Customer Relationship Management), Asset Management, and Marketing Surveys (Rismon) that affect TUR. For instance, the Monthly Price attribute with float data type in CRM corresponds to the Affordable Price attribute, while the Activation Actual Date attribute with date time data type and Product Bandwidth attribute is represented by data in the CRM database with the bandwidth column. However, if there is no descriptive data attribute, such as talents, in the database, it will not be used as an influencing factor.

CRM is a tool used to manage client data, transactions, and service bundles. It contains various information, including Monthly Pricing, Actual Activation Date, Sales, File Uploaded, Bandwidth, Product, and No-Initial Acquirer. On the other hand, Rismon is a survey application used to collect cluster data performance and potential customer information, including Potential Revenue, Education Prediction, No Device Connected, Location Permit, Age Prediction, newly installed Houses, Citizen

Opinion, and Sales Name. The GIS micro cluster is then used to link these features to spatial data, which includes prospective revenue, education prediction, and age in Table 4 with information on the Micro cluster data source.

Table 2 presents a list of seven internal and twenty-one external factors that can affect the internet connection and data source identification from enterprise systems. After that, Table 3 can be used to determine the data requirements of the enterprise system in terms of the appropriate data structures and sources. Table 3 includes seven internal data requirements for Customer Relationship Management (CRM) and Survey Systems (Rismon), such as data type, column name, and table/file name. However, due to company data privacy, the file name has been abbreviated.

Table 3: Identification Of Internal Factor Data Requirements

Attribute	Source of Data	Data Type	Column Name	File/ Table Name
1. Afford-able Price	CRM	float	Monthly_Price	S*_ AL
3. Activation Time	CRM	date time	ActivationActualDate	S*_ AL
4. Marketer	CRM	text	Sales	S*_ AL
5. Broad-band Quality	CRM	text	File upload	I*_Date
6. Product Bandwidth	CRM	text	Bandwidth	S*_AL
7. Package Product	CRM	text	Product	*_AL
2. Demand Forecast	Survey (Rismon)	float	No_Initial_Acquirer	F*_S*_S*

The following table, Table 4, lists twenty-one external data requirements in CRM and Survey Systems (Rismon), which include data type, column name, and table/file name.

Table. 4: Identification Of External Factor Data Requirements

Attribute	Source of Data	Data Type	Column Name	File/ Table Name
5. Skill	Not Available	Not Available	Not Available	Not Available
6. Willing-ness to Buy/Pay	Survey (Rismon)	text	internet_budget_allocation	F*_S*_S*
8. Income	Micro-cluster	float	Potential_Revenue	R*_S*
9. Education Level	Micro-cluster	Text	Education_Prediction	B* D*
10. PC Population	Survey (Rismon)	float	no_device_connected	F*_S*_S*
11. Gadget Population	Survey (Rismon)	float	no_device_connected	F*_S*_S*
12. Municipality Leader Authority / Decision Maker	Survey (Rismon)	Text	Location_Permit	F*_S*_S*
16. Population Age (Median)	Micro-cluster	Text	Age_Prediction	B* D*
17. English Literature	Micro-cluster	Text	Education_Prediction	B* D*
19. New Housing Area Development	Micro-cluster	Boolean	New_Installed_Pred	P*_A*
20. Perception/Citizen Perception	Survey (Rismon)	Text	Citizen_Price_Opinion	F*_S*_S*
21. Influen-cer Role	Survey (Rismon)	text	Sales	S*_AL
1. Number of Competitors (New and Existing) / FTTH Availability	Micro-cluster	float	Competitor_No_Classificat ion	M*
2. Density (m2)	Micro-cluster	float	Persil_Classification	M*
3. GDP Area/Income per Area (Median)	Micro-cluster			M*
4. Demographic	Micro-cluster	float	Density_Class	M*
7. Geolocation	Micro-cluster	float	Usage_Permit	M*
13. Building / Developer/Business Utilities	Micro-cluster	float	Usage_Permit	M*
14. Competitor Bandwidth (min.)	Micro-cluster	Not Available	Not Available	Not Available
15. Aerial Location	Micro-cluster	coordinates	Poles_Code	M*
18. Teenagers < 18 years	Micro-cluster	float	Age	M*

Most of the data in Table 4 comes from the Micro cluster system, except for two parameters that are not available: competitor bandwidth and customer skill. The results of data acquisition identification in CRISP-DM are shown in Tables 3 and 4. The data will then be extracted based on the columns and tables described in the system data sources accessible from January 2020 to August 2021.

Result of RQ2

To address this issue, question 2 will be completed through a literature review. The outcome of the procedure is displayed in Figure 2 along with the index search service, which enables auditing and replication [24]. We employ certain queries to carry out automated searches on Google Scholar, Science Direct, and Pro Quest, which are restricted to FTTH (fiber-to-the-home) and include research questions with one to three keywords that don't contain the " " symbol, as well as specific keywords that do. The advanced query effectively retrieves articles that measure take-up or take-rate factors, predict take-up or take-rate research methods, or identify take-up rate research issues, and retrieves articles on the topic of take-up FTTH or take-up rate FTTH., as illustrated in Figure 2.

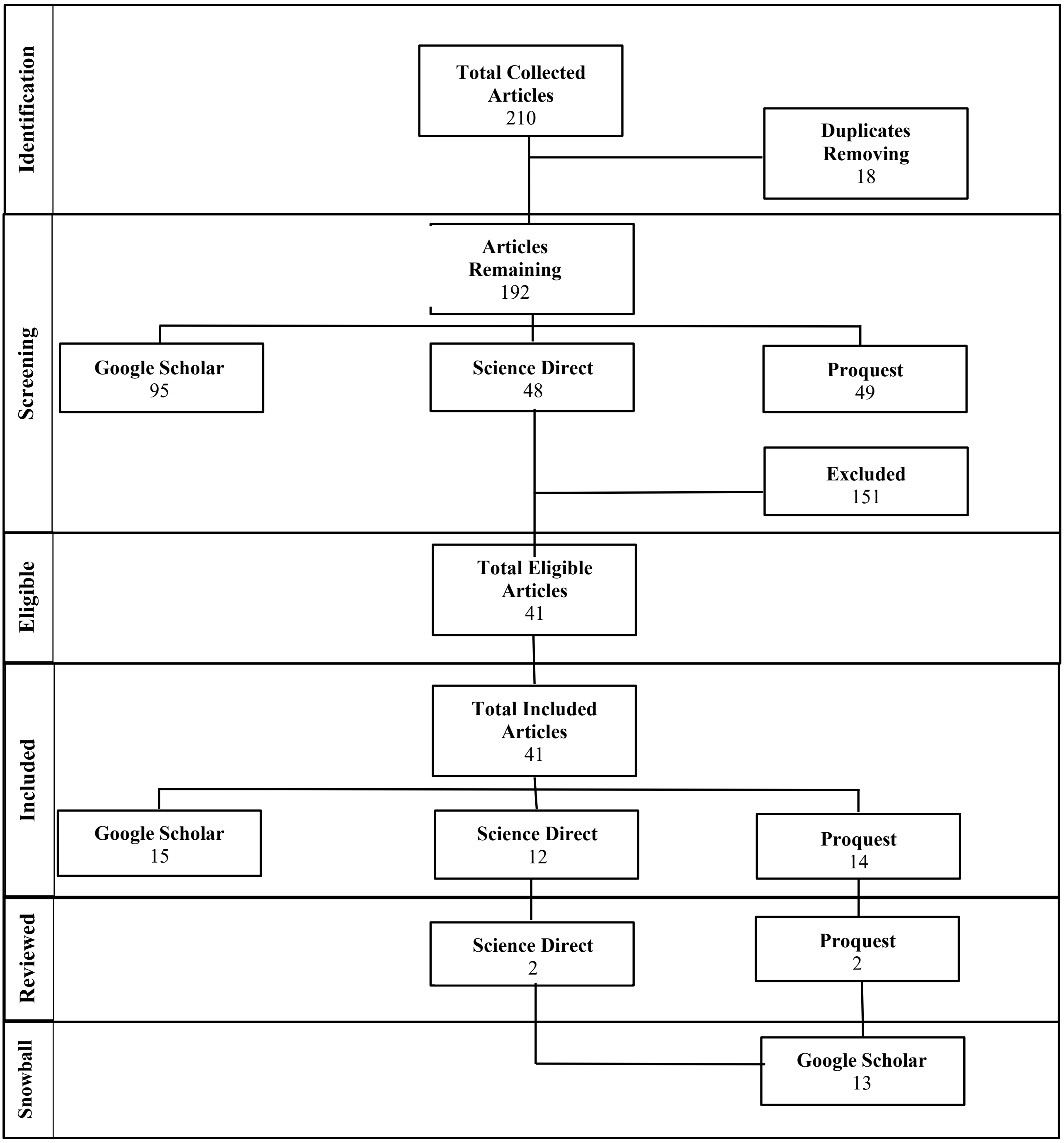

Fig. 2: Systematic Literature Review Diagram

After discovering suitable papers, a screening process was carried out using the precise inclusion criteria listed below.

1. Papers or articles must be written in English.
2. Papers or articles must be complete and inseparable literature.

3. Complete papers and articles must be available for download.
4. Papers or articles must have been published in scientific journals, presented at conferences, or included in proceedings between 2003-2022.
5. Papers or articles must discuss FTTH or FTTH take-up rates, preferably in the telecommunications industry.

Next, the following set of criteria for exclusion has been established:

1. Incomplete or unfinished literature will not be considered.
2. Articles or papers that cannot be downloaded in their entirety will be excluded.
3. Papers or articles published in books or grey literature will not be considered.
4. Any paper or article that does not mention FTTH or the FTTH take-up rate (preferably in the telecommunications industry) will be excluded from consideration.

Only articles or papers that meet these criteria will be considered for review based on the exclusion and inclusion results. There were four articles found within the criteria. The first one is research conducted by Iwashita et al. (2017). He suggests that TUR predictions can be obtained by monitoring people's movements in and out of an area. Typically, a social and communicative person is selected to convince their peers to use a particular broadband provider's internet services. On the other hand, Bermolen and Rossi in Yu et. al (2020) propose the use of support vector regression algorithms to calculate moving averages and their regressions for forecasting TUR. This methodology employs classification and regression. Mycek, M., and Zotkiewicz (2019) suggest that predictions can be obtained from optimal trajectories for Internet broadband FTTH cables. Additionally, Limbach et al. (2016) demonstrated that their TUR predictions at year three are constant or comparable between TUR and churn rates. This implies that the estimated TUR in year three is equal to the churn rate. According to Ahn et al. (2020), churn is defined as the tendency of clients to stop doing business with a company within a specified period or contract. The number of discontinuations is considered so that the company does not obtain excessively low TUR estimates.

We discovered four papers based on the findings of the previous articles. To supplement the results of the churn strategy, which is equivalent to TUR, we used the snowballing process. Snowballing is a technique for increasing the number of references to SLR for prediction algorithms that are believed to be comparable enough to be directly compared with methods employed for FTTH. This technique involves searching the literature references of the four publications in the title, abstract, and keywords. Then articles with appropriate content are once again screened using inclusion and exclusion criteria. Finally, we found 13 papers that met these requirements on the churn calculating methodology, equal to TUR, in our search for articles using the snowballing mechanism (Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. 2020). Table 5 summarizes the information.

Table 5: RQ2 Answers for Methodology of Calculating TUR Results

No	Writer	Year	Prediction Methodology	Type of Methodology
1.	Bermolen & Rossi in Yu et. al (2020)	2009	Support Vector Regression	Supervised
2.	Iwashita et.al (2017)	2015	The Moving Average Algorithm of Influencer employs a graph model. Multiple regression is used in locations where telco infrastructure has not been installed.	Generic Supervised
3.	Limbach et.al (2016)	2016	Equal to Churn Methodology	Generic
4.	Mycek & Zotkiewicz (2019)	2019	Optimal Trajectory using MIP Optimalization	Generic
5.	Hung et.al (2006)	2006	Neural Network, Decision Tree, K-Means	Supervised & Unsupervised
6.	Huang et.al (2009)	2007	Logistic Regression, Decision Tree, Multi-layer perceptron Neural Network, SVM	Supervised & Unsupervised
7.	Cheng et.al (2009)	2009	K-Means	Supervised
8.	Vafeiadis et.al (2015)	2015	SVM Polynomial with Ada Boost	Supervised
9.	Maw et. al (2021)	2021	Random Forest	Supervised
10.	Khan et al. (2019)	2019	Artificial Neural Network	Supervised

No	Writer	Year	Prediction Methodology	Type of Methodology
11.	Tariq et.al (2022)	2022	Artificial Neural Network	Supervised
12.	Guo & Huang (2021)	2021	XGBoost	Supervised
13.	Dhini & Fauzan (2021)	2021	Random Forest & XG Boost	Supervised
14.	Zhau et.al (2021)	2021	Logistic Regression	Supervised
15.	Mustafa et.al (2021)	2021	CART, SVM, K-Nearest Neighborhood	Supervised
16.	Wael et.al (2022)	2022	Deep-BP-Neural Network	Unsupervised
17.	Leogrande et.al (2022)	2022	Artificial Neural Network	Unsupervised

In Table 5, there are three main types of TUR calculations - generic algorithms, supervised machine learning, and unsupervised machine learning. By analyzing the papers listed in Table 5, we can determine the most effective approaches for each prediction, classification, or clustering model. Table 8 provides a comparison of model evaluations, which we can use as a basis for selecting the best approach. We consider various criteria, such as researcher recommendations and evaluation metrics like accuracy and precision. Table 6 categorizes the approach based on the prediction, classification, or clustering model, which we can use as a guide for selecting the best method to predict TUR.

Table 6: Modeling Section Methodological Choice Identification Results

No.	Author	Prediction Method	Clustering Method	Classification Method
1.	Hung et.al (2006)	Neural Network*		Decision Tree
2.	Huang et.al (2009)	Neural Network SVM		Logistic Regression
3.	Bermolen & Rossi in Yu et. al (2020)	SVR		SVM*
4.	Cheng et.al (2009)		K-Means	
5.	Iwashita et.al (2017)	Weighing Average of Influencer with Multi-regression		
6.	Vafeiadis et.al (2015)	SVM		SVM Polynomial*
7.	Maw et. al (2021)			Random Forest*, Decision Tree, Light GBM, Gradient Boosting, Logistic Regression, & XG Boost
8.	Khan et al. (2019)	Neural Network*		
9.	Tariq et.al (2022)	Neural Network*		
10.	Mycek & Zotkiewiecz (2019)	Optimal Trajectory		
11.	Guo & Huang (2021)			XGBoost*
12.	Dhini & Fauzan (2021)			XGBoost Random Forest
13.	Zhau et.al (2021)	Logistic Regression		
14.	Mustafa et.al (2021)	SVR		SVM* K-Nearest Neighborhood*
15.	Leogrande (2022)	Neural Network*		
16.	Wael et.al (2022)	Neural Network*		

* Stated as best method by the researcher

Table 6 shows three models in the use case for TUR prediction: neural networks, SVR(SVM), and regression (Logistic and Multi regression). There are six models available for TUR classification application cases: Decision Tree, Logistic Regression, SVM, Random Forest, Gradient Boosting (GB, XG Boost), and K-Nearest Neighborhood. Then there's K-Means for the clustering use case.

Then, once the methodologies are established, we proceed to develop the model assessment using the CRISP-DM approach. This evaluation model is further separated into methodological choices depending on their use for predicting, classifying, or clustering, as well as their combinations, if necessary. The evaluation model is shown in Table 7.

Table 7: Modeling Evaluation Section Methodological Choices Identification Results*

No.	Prediction Method	Clustering Method	Classification Method
1.	Neural Network*		
2.	Neural Network*		Decision Tree*
	Neural Network*		Logistic Regression
3.			Random Forest*

No.	Prediction Method	Clustering Method	Classification Method
			XGBoost*
4.	SVR		SVM*
5.		K-Means*	
6.	SVR		K-Nearest Neighborhood*
7.			Logistic Regression
8.	Optimal Trajectory		
9.	Equal to the Churn Method		
10.	Weighted Moving Average of Influencer Multiregression		
11.	SVM		SVM Polynomial*
12.	Logistic Regression		Decision Tree, Light GBM, Gradient Boosting, & XG Boost

*Best Method populated from Tab. 6

Table 7 presents that three studies have employed neural networks for TUR prediction, two for Support Vector Machine and Regression, and two for Regression models (Logistic and Multi Regression). Among the different models, the Neural Network model produced the best results, as seen in Table 8, with an AUC of 79-89, an accuracy level of 75.23-99, an F-Measure of 80.97-95.35, an F1 Score of 94.5-98.5, and an R2 of 88. SVM and XGBoost were used three times in the TUR classification use case. SVM comes second as a prediction model, with an AUC of 98.8, accuracy of 92.83-98.35, and F-Measure of 84.22-95.37. XGBoost, on the other hand, has an AUC of 77-99.1, an accuracy of 77-98.9, and an F1 Score of 98.3. Decision trees, logistic regression, and random forest were all used twice. K-Nearest Neighbor (KNN) was mentioned only once. Logistic Regression and KNN are the two models that outperform the other three, with AUCs of 98.7, accuracy of 98.4-98.6, and F1 Scores ranging from 97.7-98.5.

XG Boost, SVM, Decision Tree, Random Forest, and KNN are the top classification algorithms. When compared to SVM, the model that has the highest AUC, accuracy, F Measure, and F1 Score is XG Boost. As seen in Table 8, XG Boost has an AUC value of 99.1, an accuracy of 77-98.6, and an F1 Score of 98.3. Although SVM's AUC is almost identical at 98.8, its F1 Score is lower than that of XG Boost. Furthermore, if TUR prediction and classification are to be performed on the same dataset, then a pair of models can be developed, namely Neural Network with Decision Tree & Logistic Regression. SVM can be used for forecasting demand and categorizing data, working in conjunction with KNN & Logistic Regression. Finally, the Logistic Regression model can be combined with Decision Trees, Gradient Boosting, & XG Boosting.

Table 8: Model Evaluation Identification Results from Selected Methods

Methods	Author	Evaluation								
		Std. Deviation	Std Error Mean	AUC	RMSE	Accuracy	F-Measure	F1 Score	Precision	R^2
Neural Network	Cheng et.al (2009)					75.23				
Multi-Layer Perceptron	Huang et.al (2009)			84		97.7	97.92			
Back Propagation ANN	Hung et.al (2006) Vafeiadis et.al (2015) Fujo et.al (2022)	7.416 x 10^{-3}	2.141 x 10^{-2}	88		95,05 88	80,97			
ANN	Khan et al. (2019) Leogrande et.al (2022)			79		99				88
CNN	Tariq et.al (2022)					94.5		94.5	95	
PNN	Leogrande et.al (2022)									22
Decision Tree	Hung et.al (2006) Huang et.al (2009) Cheng et.al (2009) Maw et. al (2021) Vafeiadis et.al (2015)	1.146 x 10^{-2}	3.309 x 10^{-3}	89		92.45, 95.09, 83, 89.9	83.87, 92.87			
Tree Ensemble	Leogrande et.al (2022)									76

Methods	Author	Evaluation								
		Std. Deviation	Std Error Mean	AUC	RMSE	Accuracy	F-Measure	F1 Score	Precision	R^2
SVM	Huang et.al (2009) Guo & Huang (2021) Mustafa et.al (2021)			90, 98.8		98.8, 98		97.9		
SVM RBF	Vafeiadis et.al (2015)					96.05, 92.83	84.22, 94.88			
SVM POLY	Vafeiadis et.al (2015)					96.85, 93.48, 98.35	84.57, 95.37			
SVR	Bermolen & Rossi in Yu et. al (2020)				5.8 - 7	97.98				
Moving Average	Bermolen & Rossi in Yu et. al (2020) Iwashita et.al (2017)				6.7 - 7.4					99
Naive Bayes	Cheng et.al (2009) Mustafa et.al (2021)					86.40, 45				
Random Forest (RF)	Maw et. al (2021) Dhini & Fauzan (2021) Leogrande et.al (2022)			95		89				57
Opt RF	Dhini & Fauzan (2021)					98.21				
LGBM	Maw et. al (2021)			76		76				
GB	Maw et. al (2021)			74		74				
XG Boost	Maw et. al (2021) Guo & Huang (2021) Dhini & Fauzan (2021)			77, 99.1		77, 98.9, 98.16		98.3		
OPT XG Boost	Dhini & Fauzan (2021)					98.82				
Logistic Regression (LR)	Maw et. al (2021) Guo & Huang (2021) Dhini & Fauzan (2021) Zhau et.al (2021) Mustafa et.al [23]			66, 98.7, 87.1		66, 98.6, 87.75, 91, 45		98,5	37	
Opt LR	Dhini & Fauzan (2021)					95.42				
Simple Regression Tree	Leogrande et.al (2022)									62
Linear Regression	Leogrande et.al (2022)									80
Polynomial Regression	Leogrande et.al (2022)									83
KNN	Guo & Huang (2021) Mustafa et.al (2021)			98.7		98.4, 97		97.7		
Linear Discriminant Analysis	Mustafa et.al (2021)					45				
CART	Mustafa et.al (2021)					98				
GBT	Leogrande et.al (2022)									68

5. Conclusion

This study conducted a structured literature review and analysis using CRISP-DM to examine methods for data acquisition and predictive modeling for take-up ratio (TUR) in broadband services. The findings highlight useful data fields across customer databases and common modeling techniques such as neural networks and SVMs. The study offers three models for TUR prediction - SVR, regression (Logistic and regression) - and six models for TUR classification: SVM, XG Boost, Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbor. K-Means is another option for the clustering use case. However, it is important to note that the literature assessment, which included 41 publications and 13 snowball publications, needs a more careful evaluation of sources. The justification for selecting the 'best' TUR prediction models based on model evaluation parameters such as AUC, Accuracy, F-Measure, F-1 Score, R2, Precision, and RMSE will need to be strengthened in the future with specific use cases and model evaluation. While providing an initial data and modeling framework, further research should empirically validate the recommendations for various prediction settings, and additional comparison testing would explain optimal methodologies.

Acknowledgement

The author thanks all Research Methodology Lecturers for supporting and sharing experiences on the subjects.

References

Abdullah, D. M., & Ameen, S. Y. (2021). Enhanced mobile broadband (EMBB): A review. Journal of Information Technology and Informatics, 1(1), 13-19.

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A survey on churn analysis in various business domains. IEEE Access, 8, 220816-220839.

APJII. (2023). Indonesia Internet Profile. https://survei.apjii.or.id/survei/2023

Asgarirad, Mohammadreza; Jahromi, Mansour Nejati. A Taxonomy-based Comparison of FTTH Network Implementation Costs.Majlesi Journal of Electrical Engineering; Isfahan Vol. 14, Iss. 2, (2020): 71-80.

Briglauer, W. and Gugler, K. (2019). Go for Gigabit? First Evidence on Economic Benefits of High-speed Broadband Technologies in Europe. JCMS: Journal of Common Market Studies, 57(5), pp.1071-1090.

Cambini, C., & Jiang, Y. (2009). Broadband investment and regulation: A literature review. Telecommunications Policy, 33(10-11), 559-574

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. Expert systems with applications, 36(3), 4176-4184.

Dhini, A., & Fauzan, M. (2021). Predicting Customer Churn using Ensemble Learning: Case Study of a Fixed Broadband Company. International Journal of Technology, 12(5), 1030-1037.

D. P. Reed. (2020). “Examining the prospects for Gigabit Broadband: Lessons learned from Google Fiber,” Telecomm. Policy, vol. 44, no. 5, p. 101957.

Federal Communications Commission. Types of broadband connections, from https://www.fcc.gov/general/types-broadband-connections, accessed on 09-Dec-2022.

F. Queder. (2020). “Competitive effects of cable networks on FTTx deployment in Europe,” Telecomm. Policy, vol. 44, no. 10, p. 102027.

Guo, Y., & Huang, L. (2021). Customer Churn Prediction in the Broadband Service on Machine Learning. In LISS 2020 (pp. 811-821). Springer, Singapore.

Huang, B. Q., Kechadi, M. T., & Buckley, B. (2009). Customer churn prediction for broadband internet services. In International Conference on Data Warehousing and Knowledge Discovery (pp. 229-243). Springer, Berlin, Heidelberg.

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. Expert Systems with Applications, 31(3), 515-524.

H. Galperin, T. V. Le, and K. Wyatt. (2021). "Who gets access to fast broadband? Evidence from Los Angeles County," Gov. Inf. Q., vol. 38, no. 3, p. 101594.

Iwashita, M., Inoue, A., Kurosawa, T., & Nishimatu, K. (2017). Microarea Selection Method for Broadband Infrastructure Installation Based on Service Diffusion Process. International Journal on Advances in Telecommunications Volume 10, Number 1 & 2, 2017.

Kitchenham, B., & Charters, S. M. (2021). Guidelines for performing Systematic Literature Reviews in Software Engineering Guidelines for performing Systematic Literature Reviews in Software Engineering EBSE Technical Report EBSE-2007-01 Software Engineering Group School of Computer Science and Ma. no. October, 2007.

Khan, Y., Shafiq, S., Naeem, A., Ahmed, S., Safwan, N., & Hussain, S. (2019). Customers churn prediction using artificial neural networks (ANN) in telecom industry. International journal of advanced computer science and applications, 10(9).

Leogrande, A., Magaletti, N., Cosoli, G., & Massaro, A. (2022). Fixed Broadband Take-Up in Europe. Available at SSRN 4034298.

Limbach, F., Kuebel, H., & Zarnekow, R. (2016). Improving rural broadband deployment with synergistic effects between multiple fixed infrastructures. Australasian Journal of Information Systems, 20.

Maw, M., Haw, S. C., & Ho, C. K. (2021). Utilizing data sampling techniques on algorithmic fairness for customer churn prediction with data imbalance problems. F1000Research, 10.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N. & Flach, P. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048-3061.

Merriam-Webster. (2022). Broadband Definition. https://www.merriam-webster.com/dictionary/broadband

Milanovic, S. (2014). Case study for a GPON deployment in the enterprise environment. J. Netw., vol. 9, no. 1.

Mustafa, N., Ling, L. S., & Razak, S. F. A. (2021). Customer churn prediction for telecommunication industry: A Malaysian Case Study. F1000Research, 10.

Mycek, M. and Żotkiewicz, M. (2019). On optimal trajectory for the multi-period evolution of FTTx networks. Telecommunication Systems, 71(3), pp.353-376.

Paiva, D. M. B., Freire, A. P., & de Mattos Fortes, R. P. (2021). Accessibility and software engineering processes: A systematic literature review. Journal of Systems and Software, 171, 110819.

Pradeep, M., Pavithra, B., Pooja, R., Parameswari, S., Pandi, M. (2017). A Survey of FTTH Elements Based on Broadband Access Network. Asian Journal of Applied Science and Technology (AJAST), vol. 1, no. 7, pp. 54–59.

R. Sharma and R. Mishra. (2017), "Investigating the role of intermediaries in adoption of public access outlets for delivery of e-Government services in developing countries: An empirical study," Gov. Inf. Q., vol. 34, no. 4, pp. 658–679.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526-534.

Spruytte, J., Benhamiche, A., Chardy, M., Verbrugge, S., & Colle, D. (2019). Modeling the relationship between network operators and venue owners in public Wi-Fi deployment using non-cooperative game theory. EURASIP Journal on Wireless Communications and Networking, 2019, 1-16.

Tariq, M. U., Babar, M., Poulin, M., & Khattak, A. S. (2022). Distributed model for customer churn prediction using convolutional neural network. Journal of Modelling in Management, 17(3), 853-863.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9.

Van den Velden, J., & Sadowski, B. M. (2023). Creating public value with municipal Wi-Fi networks: a bottom-up methodology. Digital Policy, Regulation and Governance, 25(2), 77-103.

Wael Fujo, S., Subramanian, S., & Ahmad Khder, M. (2022). Customer Churn Prediction in Telecommunication Industry Using Deep Learning. Information Sciences Letters, 11(1), 24.

Yu, Y., Liu, S., Wang, L., Teng, F., & Li, S. (2020). Traffic prediction model based on improved quantum particle swarm algorithm in wireless network. In IOP Conference Series: Materials Science and Engineering (Vol. 768, No. 6, p. 062110). IOP Publishing.

Zhao, M., Zeng, Q., Chang, M., Tong, Q., & Su, J. (2021). A prediction model of customer churn considering customer value: an empirical research of telecom industry in China. Discrete Dynamics in Nature and Society, 2021.