International Journal of Advanced Technology in Engineering and Science
Vol. No. 10, Issue No. 04, April 2022
www.ijates.com
ijates
ISSN 2348 - 7550

Analytics: Application of Machine Learning and Pre-Screened Parameters in Predicting Number of Bike Share Riders

Bongs Lainjo

Cybermatic International Montréal QC Canada

Email: bsuru@icloud.com

ABSTRACT

This research was carried out with an aim of identifying the best Machine learning model that predicts the number of bike rentals. In other words, how effective is predictive analytics in modelling bike share parameters? This study employed machine learning classifiers to predict the number of bike share renters. The train-test (80/20%) split procedure was used to estimate the performance of machine learning algorithms used to make new data attribute predictions. The data were analyzed using R, PYTHON and SPSS software. The machine learning classifiers include Random Forest, Decision Trees, Nearest Neighbor and XGBoost. The five most important variables are year, temperature, humidity, seasons and windspeed. The Random Forest Algorithm had an Area Under Curve (AUC) of 97.67%. The classifier identified the most important variables that predict the number of bikes rented are temperature, Year, Windspeed, humidity and Seasons. The decision tree model had an AUC of 87.27%. According to the KNN algorithm, the five most important variables are temperature, year, fall, humidity and windspeed. The KNN algorithm had AUC of 93.07%. The logistics regression model had an AUC of 89.41%. The XGBoost model had an AUC of 97.35. According to the XGBoost algorithm the top five most important predictors are Temperature, year, Seasons, weather situation. Although XGBoost had the best accuracy, based on the Receiver Operating Characteristics (ROC) curve's area under curve, the Random Forest model was selected as the best model because it had the highest AUC. Classifier performance metrics – ROC, precision, recall, F1 etc. - are presented in detail elsewhere in the paper. All the variables ranked by these classifiers were pre-screened and reported in an earlier study. While he limited dataset – 731 observations – was of major concern, the promising outcomes do confirm that reliable, plausible and robust predictions can be made if copious data are made available. It is hence recommended that these findings be interpreted with caution

KEYWORDS: Bike-sharing-programs, Machine Learning Algorithms, Classifiers, Receiver Operating Characteristics, Area Under Curve, predictive analytics.

INTRODUCTION

This paper's objective is the application of machine learning algorithms an using pre-screened parameters to predict the number of bike share riders. The paper seeks to determine the best classification algorithm for predicting the number of bike share riders and use the algorithm to identify the most important predictors of the number of bike share riders. Fuzzy systems, artificial neural networks, and evolutionary computing are examples of computational intelligence that have produced major achievements in modeling, learning, and search and optimization challenges for smart city applications [1]. Machine learning has been used by smart city researchers in a variety of domains, including public space usage and public bus charging station placement [2]. Because the quantity of bicycles is important for the long-term growth of a bike system (BS), this study used quantitative analytics and a machine learning technique to manage bike sharing. Finally, the use of information gleaned from current dock less bicycle operating data to influence the numbering and administration of public bicycles was investigated.

In the data-driven business climate of the mobility sector, precisely predicting client demand is a critical component of success [3]. It is extremely easy to identify global organizations that supply their varied services to clients based on demand forecast findings in the data technology era. Companies have had short-term success because they have accurately forecasted demand based on internal and external factors. Most sectors throughout the globe are undergoing digital transformations based on machine learning and deep learning algorithms [4]. Fast-growing start-ups, in particular, frequently exploit crucial business choices on data and algorithms in tandem with managerial expertise or intuition. The bike is an environmentally beneficial method of transportation that benefits from individuals optimizing its exercise effects especially during the COVID-19 pandemic and in a modern world where environmental concerns are becoming increasingly important. Since it was changed into a "shared" item a few years ago, the bike has received international notice. It is not only for recreational reasons, but also for transportation [5].

When it comes to creating sustainable transportation systems, shared bike programs are seen to respond to climate change and energy challenges in many places across the world. A shared bike system is an important and required tool for encouraging people to use bikes and advancing the implementation of an efficient urban transportation system. Despite its numerous benefits, a growing number of communities and organizations are still hesitant to implement a shared bike program because of the drawbacks of high fixed expenses such as installation and operation [6]. As a result, reliable demand forecasting is essential to keep a bike sharing program running. The bike sharing system was one of the first shared economy models in the transportation business. Capital Bikeshare, which started renting bikes in the Washington, D.C. region, held a data science competition on the Kaggle Competition platform to estimate consumer bike share demand [7]. As a result, data scientists and analysts all around the world have been attempting to forecast demand using various data mining approaches. This study is one of those ttempts to predict number of bike share users applying classifiers and using historical data provided by capital Bikeshare.

Much prior research has emphasized the relevance of model parameters such as the distance between the rental and return locations ([8]; [6]). Within provided datasets, these works mostly focused on engineering feature approaches and statistical modeling. As a result, past research had model constraints since they only employed certain datasets. This research, on the other hand, is focused on investigating a new characteristic from a live data source. Existing research articles have used statistical approaches to estimate bike demand, but few have used machine learning methodologies, which have lately come into the forefront. The goal of this research is to offer a machine learning prediction model that incorporates factors that influence shared bike demand. Furthermore, a predictive model that contains these factors is provided from the perspective of an integrated public transportation system.

Significance of the Research

This paper proposes a prediction model for predicting bike share in integrated public transit systems. Using characteristics that influence demand for shared bikes, a machine learning prediction model is provided. The findings are expected to have policy implications for a city's integrated public transportation system's efficient functioning. Machine learning methods were used in this work to provide good prediction performance of bike sharing demand. This study, on the other hand, is significant since it focuses on investigating a novel characteristic from an Open Data external data source. This new feature exploration will aid mobility organizations in resolving business operations issues like bike supply rearrangement or excess supply cost of bikes or bike station decks, which are frequently caused by a failure to effectively estimate customer bike sharing demand.

Literature

The literature review will cover the bike share programs around the world, bike share models, methodologies used in previous literature and even gaps. The research strategy that yielded the information for this study was discussed in the first section of this review. A search of University student databases yielded a flurry of results, including multiple articles about the Bike Share program. Following initial searches through Academic Search Complete at the library, I discovered further particular subtopics, and the addition of bikeshare-related databases revealed a wide range of relevant research. To find peer-reviewed online resources, the researcher will use government reports, Google Scholar, a literature matrix (created by the researcher to allow quick comparisons among publications to determine scope), and EBSCOhost. The search began with phrases and keywords such as Bike share, regression models, Nearest Neighbor, XGboost, Random Forest and Decision Trees in databases. The usage of these keywords and phrases ensures a complete investigation of all aspects of Bike Share.

Bike Share Programs

Additionally, [9] stated that in the present scenario, China is the leading nation in the world in terms of the growth of public BS along with the private electric bike. The current projections confirm the feasibility of implementing large-scale shared e-bike systems across the nation. Public BS systems can be one of the fastest-growing public transportation modes in the world. The industry is growing at an average rate of 37% every year since the year 2009 [9]. In this context, "The current trajectory of bikeshare adoption, the popularity of e-bikes, and the presence of e-bikeshare pilot projects in other countries all support a future of e-bike sharing in China" [9]. . Given the rapid evolution of transportation in China, it is not well understood how such a system will differ from standard bikeshare and how both types of shared bikes (hereafter ''shared bike" is used to refer to both bikeshare and e-bikeshare) systems can best address the needs of urban China" [9]. Based on the study findings of [10], the growth in bike-sharing is receiving attention, as societies are becoming highly aware of the significance of ‘active non-motorized traffic modes. Since BS is perceived to be a promising transport system, it increases the use of a bicycle, especially in circumstances providing different pick-up, as well as drop-off locations, self-service, etc., thereby making it convenient to users. Furthermore, bike-sharing is found to be offering an efficient solution to the transport system, thus can be perceived as an alternative to other transit systems. Additionally, [11] asserted as a safety measure that it is essential for riders to use a helmet while using a bike. A survey conducted in the year 2013 under ‘New York City’s bikeshare program’ found that about 85.3% of the riders did not wear helmets [12]. This raises an important question – whose liability is it? Even though there are no statistics available on BS riders’ accidents, a well-designed and comprehensive program should include access to helmets. The additional challenge in operating BSPs is motivating riders to use these helmets. If current rider safety behavior and awareness is any indication, there is a strong likelihood that most BS riders will use them. This is a challenge that policy makers need to consider and prioritize in their respective BS programs.

Bike Sharing Classification Models

Different machine learning classifiers can be used to predict the number of bike share customers. Classification is a supervised learning approach that involves training a classifier on a set of samples that already have a class label for it to categorize an unknown sample. Hundreds of classifiers exist in the field of machine learning to address real-world categorization challenges [14]. Random forest classifier, K-nearest neighbor classifier, logistic regression, support vector machine, and artificial neural network are the five most often used classification techniques [13]. The most extensively used supervised machine learning algorithm is the random forest classifier (RFC). It's a strong tool that typically produces decent results without requiring constant tweaking of the parameters. The decision tree is the fundamental unit of random forests. A random forest is a classifier that consists of numerous decision trees, and the output category is decided by the mode of the individual tree output categories (Breiman, 2001). N trees will have N classification results for an input sample. The random forest considers all of the categorization voting results and outputs the category with the most votes. It provides several benefits, including the ability to accommodate thousands of input variables without deleting any of them, as well as estimations of which variables are most significant in the categorization. K-nearest neighbor (KNN) is a classification approach that measures the distance between multiple feature values. The test object is a vector formed of attribute values and an unknown category label, given a training set B and a test object m. The distance (or similarity) between m and each training object must be calculated by the algorithm. The list of closest neighbors can be determined in this manner. Then, in the nearest neighbor, assign m to the category with the most instances. The benefit is that it is simple to comprehend, and good results may be achieved without a lot of tweaking. The downside is that the prediction speed is poor, and a dataset with many features cannot be processed. It is susceptible to data inconsistency. Furthermore, the output's interpretability is poor [4]. Logistic regression (LR) is a linear classifier that involves establishing a regression formula on the classification boundary line depending on the data to be classified. This approach has a low calculation cost and is simple to learn and apply. The fitted parameters clearly show how each element affects the outcome. And most of the time is spent on training, with classification occurring quickly once the training is done. However, it is simple to underfit, and classification accuracy is low. The fundamental reason for this is because LR is a linear fitting algorithm, yet many objects do not fulfill linearity in practice [15]. The support vector machine (SVM) converts the nonlinear separable issue in the original sample space into a linear separable problem in the feature space, allowing for the discovery of the best hyperplane for classification. Then, using the hyperplane as a guide, categorize the set. SVM can produce accurate predictions on data outside of the training set and has a low generalization error rate, minimal processing cost, and simple-to-understand outputs, but it is too sensitive to parameter changes and kernel function parameters [16]). The artificial neural network (ANN) is a data processing system that mimics the structure and function of the neural network in the brain. The ANN algorithm is made up of a series of continuous input/output units, each of which has a weight assigned to it. The proper class label for the sample to be learnt can be anticipated in the learning stage by modifying the weights of the neural network. High classification accuracy and powerful distributed parallel processing capabilities are two of the ANN algorithm's benefits (Grossi & Buscema, 2007). For datasets containing a huge quantity of noisy data, artificial neural networks offer high robustness and fault tolerance; nevertheless, the learning process cannot be viewed, and the output results are difficult to comprehend, reducing the findings' reliability and acceptance. It also necessitates a significant number of variables, such as network structure, weight initialization, and thresholds.

Different types of bike-sharing models can be also used for predicting flows in every station. Contextually, [17], robust linear regression models are one of such models that help in predicting flows. The developed environment variables used in the model are often identified within a buffer zone (300 meters) in every bike-sharing station. Thus, to predict the bike-sharing flow, linear regression can be used in the busiest time of a weekday. The integration of a robust linear regression method can be helpful in improving rider needs specifically and program optimization in general. Furthermore, the arrival and exit flow in terms of hourly level can be integrated into the regression model at each station. On the other hand, [18] highlighted another bike-sharing model system dynamics simulation. In this context, the simulation method helps in the modeling of factors along with operations, processes, and policies to be considered in the dock less bike-sharing programs operations. It further helps in assessing effective sustainable strategies, which focus on enhancing the overall system performance. Maintaining an adequate balance between expenditure and revenue is an important factor especially in the sustainable development of dock less bike-sharing programs in a specific area. Thus, both revenue and expenditure of the dock less bike-sharing programs need to be fully considered in the respective system. Additionally, the system model performs different simulations and further evaluates the dynamic behavior of the respective system.

The study findings of [19] further stated that the main objective of a dynamic simulation model is that it helps "to minimize the vehicles repositioning costs for bike-sharing operators, aiming at a high-level users satisfaction, and if it increases with the probability to find an available bike or a free docking point in any station at any time. The proposed model considers the dynamic variation of the demand". Based on the study findings of [20], docking stations along with bikes are passive agents, while the amount and location are perceived as inputs to the simulation. Thus, it indicates that the simulation is dependent on the higher-level model for establishing the optimality in the respective inputs. Furthermore, “users and repositioning trucks are the active agents who take decisions, which result in an efficient flow of bikes between stations” [20]. For instance, the case study of Barcelona´s Bicing can be taken into consideration, wherein a 24-hour simulation was performed with the inputs from a real case study. Bicing is a bike-sharing system in Barcelona, which was selected as a benchmark. Additionally, “the open data policy of the Barcelona council, including Bicing's data, was decisive in such selection. The Bicing open data portal includes the real-time occupancy of every station with a one-minute update frequency. These data allow assessing some aspects of the performance of the simulator” [20].

Methodologies Used in Previous Literature

During the fourth industrial revolution, multinational corporations and start-ups experimented with the sharing economy idea, attempting to better fulfill consumer demand by incorporating demand forecast findings into their operations. Companies must enhance their prediction model to better estimate client demand in a more precise manner to survive in today's harsh competition. [21] investigated a novel feature for bike sharing demand prediction models, which enhanced the RMSLE (Root Mean Squared Logarithmic Error) score. The RMSLE score results increased by adding this new feature to the number of daily car accidents recorded in the Washington, DC region to the random forest, XGBoost, and LightGBM models.

Machine learning techniques were used by [22]to estimate the availability of bikes at San Francisco Bay Area Bike Share stations. As univariate regression methods, Random Forest (RF) and Least-Squares Boosting (LSBoost) were utilized, while as a multivariate regression technique, Partial Least-Squares Regression (PLSR) was used. The number of available bikes at each station was modeled using univariate models. PLSR was used to decrease the number of prediction models necessary and to represent the geographical correlation between network stations. The results clearly suggest that univariate models predict errors more accurately than multivariate models. The results of the multivariate model, on the other hand, are plausible for networks with many spatially associated stations. Station neighbors and the forecast horizon time are also important factors, according to the findings. 15 minutes was the most efficient forecast horizon period for minimizing prediction error.

Citizens have benefited from the bike-sharing program, which has functioned as a useful addition to public transportation. Each docking station has a defined space to keep bikes for docked bike-sharing service, and the station may be empty or saturated at various times. Bike-sharing companies often move bikes between stations by driving trucks based on their previous experiences, which might result in a waste of human resources. Accessing this service is inefficient for operators and cumbersome for users. As a result, both operators and riders benefit from accurately forecasting the quantity of available bikes in the stations. [2] focused primarily on short-term docking station utilization predictions in Suzhou, China. With one-month historical data, two new and extremely efficient models, LSTM and GRU, are used to forecast the short-term available number of bikes in docking stations. As a comparison, Random Forest is utilized as a baseline. Both RNNs (LSTM and GRU) and Random Forest may achieve good performance with tolerable error and comparative accuracies, according to the results. In terms of training time, random forest is preferable, while

LSTM with complicated structures can predict better in the long run. The highest discrepancy between real data and anticipated value is just one or two motorcycles, indicating that the created models are basically suitable for implementation.

Because the number of bicycles is essential to the long-term success of dock less PBS, this study used OFO bike operation data in Shenzhen to test the implementation of a machine learning method to quantity management. To identify the bicycle gathering area, [23] employed two clustering methods, and the available bike number and coefficient of available bike number variation were evaluated in each type of cycling gathering area. Second, using 25 impact variables, five classification algorithms were evaluated in their accuracy in classifying the kind of bicycle collecting locations. Finally, the use of information accessed from current dock less bicycle operating data to influence the numbering and administration of public bicycles was investigated. According to the findings, there were 492 OFO bicycle collecting places that were classified as high inefficient, normal inefficient, highly efficient, and normal efficient. Around 110,000 bicycles with minimal utilization were gathered in the high inefficient and normal inefficient zones. The categorization algorithm's accuracy will be impacted when more types of bicycles collecting areas are added. In five classification methods, the random forest classification had the greatest performance in detecting bicycle gathering area kinds, with an accuracy of more than 75%. In four different types of bicycles gathering spaces, there were notable variances in the features of 25 impact elements. It is possible to estimate area type using these criteria to maximize the number of bicycles available, save operating costs, and enhance usage efficiency. Using a machine learning technique, this research aids operators and the government in understanding the features of dock less PBS and contributes to the system's long-term sustainability.

In recent years, bike sharing systems have seen remarkable expansion and scholarly interest. The key drivers to this growth were environmental awareness, technological advancements, and the desire for socially fair transportation choices. However, as these systems continue to expand, businesses are confronted with the perpetual need to rebalance them to satisfy rising demand. As a result, managing organizations are constantly looking for the best methods for predicting flow. [24] investigated three machine learning techniques, focusing on the overlooked topic of multiple seasonality in time-series models. The goal of the study was to look at the link between bike sharing and the weather, as well as the people who utilize it. The four various strategies are then constructed and assessed to select the best-performing algorithm and to recommend other research topics in traditional time series models.

While the advantages of shared bicycle use in terms of greater mobility, accessibility, and urban environmental quality are well established, the effects of increased bicycling on traffic safety need to be evaluated and managed further. Helmet use and behaviors of bikeshare users and other cyclists are contrasted based on observational studies in one of the nation's most extensive and successful bikeshare programs (Honolulu, Hawaii). In 25 different places throughout the city, 5431 bicycles, mopeds, motorbikes, and other two-wheeled vehicles were spotted. To examine the links between helmet wear, bicyclist characteristics, roadway, traffic violations, location, and environmental factors, [25] employed two logistic regression models. Bikeshare users, visitors, ladies, and those carrying earbuds are less likely than other categories to wear helmets.

Bicyclists who ride during rush hour and on weekdays, as well as those who ride on conventional road lanes, are more likely to wear helmets. Bikeshare riders are also more prone to break the law than other bicyclists. In addition to raising awareness of the traffic safety issues connected with the growing popularity of biking and bikeshare, the report includes recommendations for enforcement, education, engineering, and risk management. There is a pressing need to boost helmet wear and overall cyclist safety.

Gaps

Scholars have produced impressive results when it comes to determining the indicators of the bicycle system and the elements that influence riding. Regression models are commonly used in research approaches. Except for a few research, most of the knowledge acquired comes from the dock bike sharing [26]. Because there may be other models better than the regression models, the paper therefore explores the predictive ability of other machine learning classifiers. Because research on Machine Learning algorithms has grown at an exponential rate in recent years, it is desirable to implement a variety of modifications on the fundamental Machine Learning algorithm structure. Further study is needed to determine the best model structures, sequence length, and time interval for improved prediction. Limited availability of standardized BSP parameters is a challenge that needs to be explored and addressed. Such a strategy will facilitate model comparison, research duplication and consensus among different BSP organizations and regions.

Contribution to the Research Field

It is possible to predict the number of renters using the criteria – pre-screened parameters - identified in this study to maximize the number of available bicycles, lower operating costs, and enhance usage efficiency. Using a machine learning technique, this research aids operatives and the government in understanding the features of bike sharing and adds to the system's long-term sustainability. Bike sharing operators may use Machine Learning algorithms to properly estimate demand for bikes in real time. They may assess all the docking stations within easy walking distance and convey the number of bikes accessible to each consumer at any given moment. So that there is no gap between a growing demand and the ability to supply that need, personnel are dispatched to regions where bikes are required. The study will also help researchers to identify the machine learning classifier with the best accuracy and the least implementation time in such scenarios.

The Study Settings

Capital Share

This study will be based on a US bike-sharing provider Capital Bikeshare company (CBC). A mountain bike guided tour and rental facilities are part of the business. The program is jointly owned and sponsored by the District of Columbia and Arlington County, VA and operated by Alta Bicycle Share, Inc. Its coverage includes both regions (see figure 1)

CBC provides rides for people of all skill levels. They will range in difficulty from easy family rides to intense fast-paced expert rides. These tours might take anywhere from a few hours to a week to complete. The company offers a package deal that includes bikes, transportation to and from the trailhead, lunches, and a personal tour guide to show them around the trails and provide information about the area. It also offers a complete bike rental fleet. The bikes range in price from a low-cost city cruiser to a full-fledged mountain bike with full suspension. Customers are not required to be on a tour to hire bicycles from the company. The company has 4500 bicycles and 500 stations.

Due to the ongoing Corona pandemic, the company's revenues have recently dropped significantly. In the current market environment, the company is struggling to stay afloat. As a result, it has decided to develop a thoughtful business plan in order to increase revenue as soon as the current lockdown ends and the economy returns to a healthy position. CBC hopes to have a better understanding of people's demand for shared bikes after the current Covid-19-related complications end across the country. They planned this to position themselves to meet people's demands whenever the situation improves all around, to differentiate themselves from other service providers, and to profit handsomely. They want to know what factors influence demand for these shared bikes in the United States.

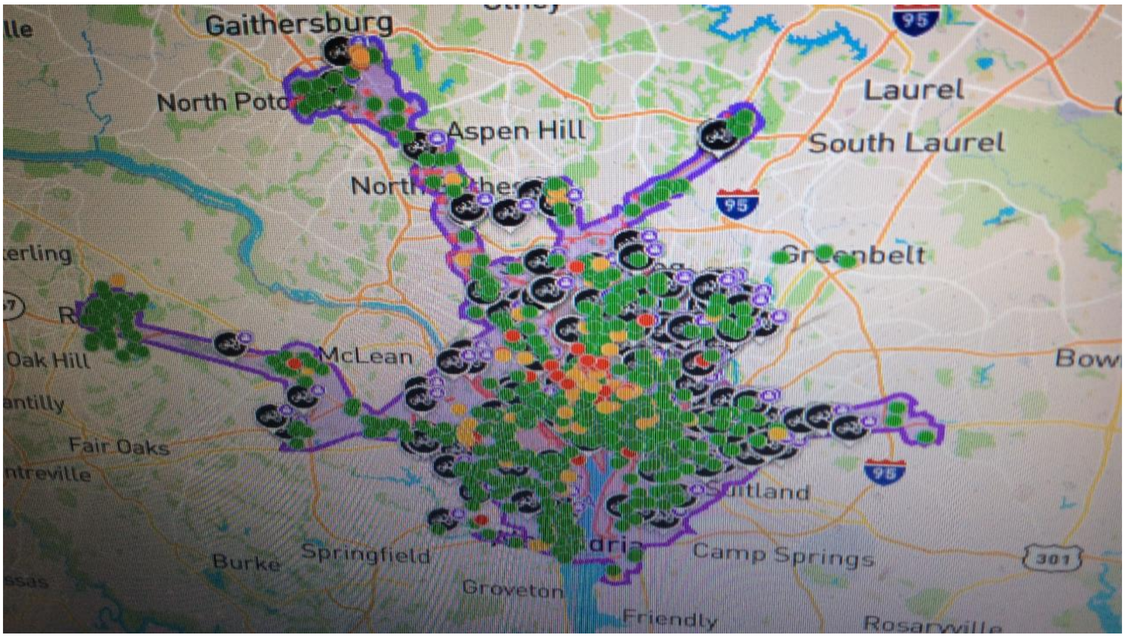


Fig 1: **Capital Bikeshare (CB) Coverage: (Source: CB website)**

Global Distribution of Bike Sharing Programs

A global graphic view of BSPs highlights the availability and access of these programs in different countries. The wealthy countries as expected continue to be trail blazers in making these programs available to its citizens. China remains the country with the highest number of BSPs Given the overall distribution, there is every reason to believe that with the environmental awareness and need for economic optimization, these programs will continue to increase and improve. Real time data collection and instant data analyses contribute significantly in making these programs efficient, sustainable and reliable. The links below are based on the latest (December 2021) data available and should be helpful in increasing readers' knowledge of the BSP dynamics in different parts of the world.

US Bike Share Distribution ; Europe Bike Share Map; Asia Bike Share Locations; Canada Bike Share Sites

South America Bike Share Points; Africa only Bike Share Organization

Research Limitations

In general, many scientific research studies are prone to different types of limitations. This one is no exception. The inexhaustive list of this study's shortcomings includes:

- Study design being influenced by existing data sets.
- Lack of a sampling protocol.
- Data set being enhanced by a third party.
- Inability to verify inconsistent data points – wrong data entry, recording errors etc. - : example following up outlier data.
- Inadequate data cleaning
- Limited number of cases especially when divided into TRAINIG (80 %) and TEST (20 %) Datasets
- Model Replication only valid within the applied parameters

Dataset

A bike-sharing system refers to a service that makes bikes accessible for shared use to individuals for a fee or free on a short-term basis. Such systems let users rent a bike from a "dock," which is frequently computer-controlled and where the user enters payment information, and the system unlocks the bike. After that, the bike can be returned to another dock in the same system. The original database has N=731 observations. These have been reduced to 703 cases after adjusting for deleted cases identified as outliers. Table 1 shows the data dictionary that provides details of the data set attributes used in the study. The data used are free and publicly provided by BSC and Hadi Fanaee-T Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto: original data provider and data compiler respectively. As indicated earlier, one limitation of the database has been the inability to include a comprehensive data cleaning strategy during this process. Data collection for the study was conducted between 2011 and 2012 inclusive. In this dataset, there were no missing values.

Table 1: Study Database Dictionary (Source: Author)

VARIABLE	DESCRIPTION	VALUE LABELS
Independent		

variables		
season	Seasons of Year	1=Spring, 2=Summer, 3=Fall, 4=Winter
month	1=Jan, 2=Feb etc.	
holiday	Day Holiday or not	
weekday	Day of the Week	
Working day	Weekend or Holiday =0, Otherwise = 1	
Weather sit	1=Clear, Few Clouds, Partly Cloudy, Mist	
	2=Mist+Cloud, Mist +Broken Clouds, Mist+Few Clouds	
	3=Light Snow, Light Rain+Scattered Clouds	
	4=Heavy Rain+Ice Pellets+Thunderstorm+Mist, Snow+Fog	
temp	The temperature in Celsius. Normalized by division by 41	
atemp	Temperature Feel Division by 50	
hum	Humidity Division by 100	
windspeed	Division by 67	
Dependent Variable		
cnt	Total number of registered & unregistered bikers: renters	

The dataset was split into ‘Train’ and ‘Test’ datasets in the ratio of 80% to 20%.

Methodology

The research is motivated by my interest in learning more about the factors that influence demand for these shared bikes using Machine Learning classifiers. In this study Machine learning classifiers namely, Random Forest, Decision Trees, Nearest Neighbor and XGBoost were employed to determine the most important predictors of the number of bike share renters, the classifier that was most accurate was the best. The data was analyzed using SPSS V25, R and PYTHON. The dependent variable ‘cnt’ was recoded as ‘1’ for the number of renters greater than 4500 and ‘0’ for those less than or equal to 4500 (which is the approximate average from the dataset). The train-test (80/20%) split procedure was used to estimate the performance of machine learning algorithms when they were used to make predictions on data. It's a quick and simple technique that allows you to evaluate the performance of several machine learning algorithms for your predictive modeling challenge. Although the process is straightforward to use and comprehend, there are occasions when it should not be utilized, such as when you have a tiny dataset or when further setup is necessary, such as when it is used for classification and the dataset is not balanced.

Preliminary Analysis

Data visualization

To begin, we looked at the distribution of the response variable Total Bike Rentals (cnt). According to the histogram in figure 1, the total number of rental bikes appears to follow a relatively normal distribution. The distribution's mean and variance are the same, and when the mean grows larger, the distribution approaches a normal distribution.

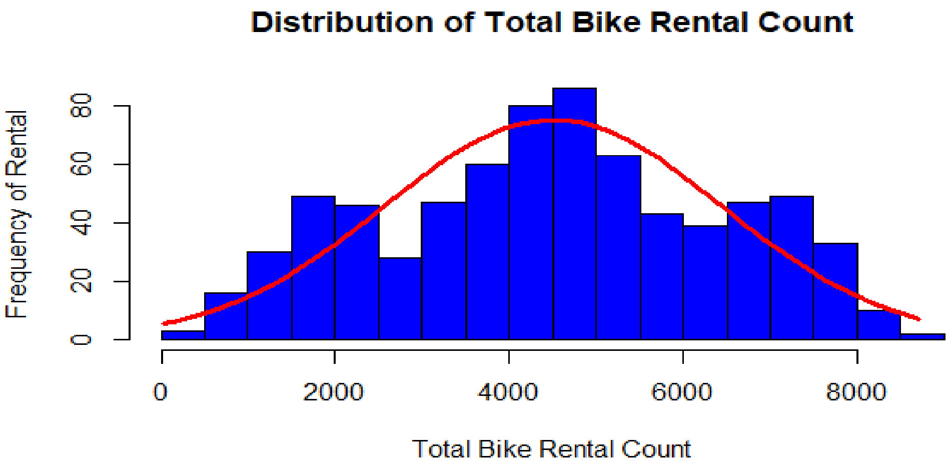


Fig 2: Histogram showing the count of Bike Rentals

To begin, we looked at the distribution of the response variable Total Bike Rentals (cnt). According to the histogram in figure 1, the total number of rental bikes appears to follow a relatively normal distribution. The distribution's mean and variance are the same, and when the mean grows larger, the distribution approaches a normal distribution.

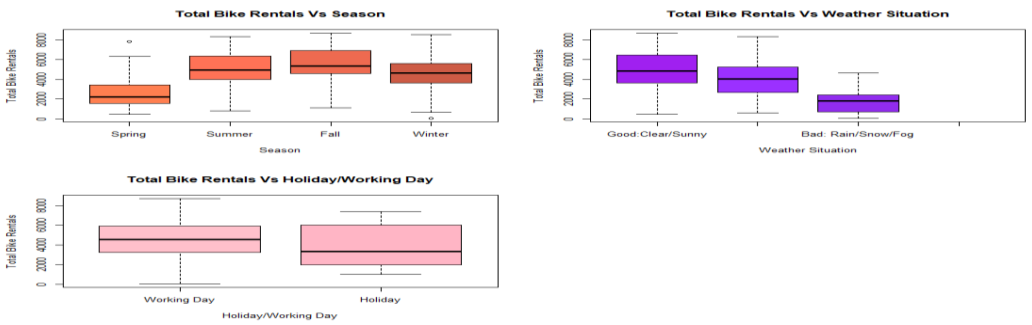


Figure 3: Boxplots for select variables (Bike Rentals by Seasons, Weather situation and Whether it is a Holiday or Working day)

The boxplots in figure 3 depicts the association between the variable Total Bike Rentals(cnt) and the season. During the summer and fall, the average number of bike rentals is at its peak. The graph depicts the link between the variable Total Bike Rentals(cnt) and the holiday. We can see that the average number of bike rentals is larger on weekdays than on weekends. The graph depicts the association between the variable Total Bike Rentals(cnt) and the weather. When the weather is terrible, there is a distinct downward tendency in bike rentals. The graph depicts the association between the variable Total Bike Rentals(cnt) and the year. During the two-year period, we can see that the overall trend has risen. Moreover, during the summer and fall seasons of each year, there are a large number of bike rentals.

Table 1 Descriptive statistics

		temp	atemp	hum	windspeed	casual	registered	cnt
Mean		.50	.47	.63	.19	848.18	3656.17	4504.35
Median		.50	.49	.63	.18	713.00	3662.00	4548.00
Minimum		.06	.08	.00	.02	2.00	20.00	22.00
Maximum		.86	.84	.97	.51	3410.00	6946.00	8714.00
Percentiles	25	.34	.34	.52	.13	315.00	2493.00	3141.00
	50	.50	.49	.63	.18	713.00	3662.00	4548.00
	75	.66	.61	.73	.23	1097.00	4790.00	5976.00

Machine Learning Results

1. Introduction

We claimed that no single learning algorithm can consistently outperform others across all data sets. As a result, we used an empirical approach to determine the accuracy of the candidate algorithms for the problem and then choose the one with the best accuracy. In machine learning, a classifier is an algorithm that sorts or categorizes data into one or more of a set of "classes" automatically. A classifier is the algorithm - the principles that robots use to categorize data. The product of your classifier's machine learning, on the other hand, is a classification model. Historical data sets are used to train the model, and the model is then used to classify your data. Both supervised and unsupervised classifiers are available. Unsupervised machine learning classifiers are fed just unlabeled datasets, which they sort into categories based on data structures, pattern recognition, and anomalies. Training datasets is applied to supervised and semi-supervised classifiers, which teach them how to classify data into specified categories. Classification is a type of supervised learning helps to segregate large amounts of data into discrete values. Classification has numerous uses in a variety of fields, including medical diagnosis, credit approval, and target marketing.

2. Types of Classifiers

The classifiers that were used in this study include: Random Forest, Decision Trees, Nearest Neighbor, Logistics regression and XGBoost.

Random Forest

A random forest is a machine learning technique for solving classification and regression problems. It makes use of ensemble learning, which is a technique for solving complex problems by combining multiple classifiers. Many decision trees make up a random forest algorithm. Bagging or bootstrap aggregation are used to train the 'forest' formed by the random forest method. Bagging is a meta-algorithm that increases the accuracy of machine learning methods by grouping them together. The random forest method determines the outcome based on decision tree forecasts. It forecasts by averaging or averaging the output of various trees. The precision of the result improves as the number of trees grows. A random forest method overcomes the drawbacks of a decision tree algorithm. It reduces dataset overfitting and improves precision.

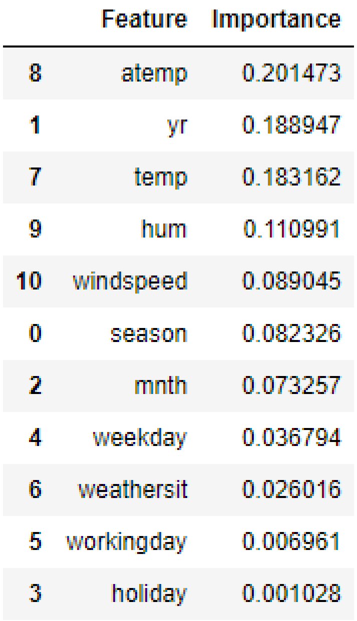
A random forest algorithm’s building component are decision trees. A decision tree is a decision-making tool with a tree-like structure. A basic understanding of decision trees will aid our understanding of random forest algorithms. There are three parts to a decision tree: decision nodes, leaf nodes, and a root node. A decision tree method separates a training dataset into branches, each of which is further divided into branches. This pattern repeats until a leaf node is reached. There is no way to separate the leaf node any farther. The attributes utilized to forecast the outcome are represented by the nodes in the decision tree. The leaves are connected to the decision nodes.

The fundamental distinction between the decision tree and the random forest algorithms is that the latter randomly establishes root nodes and segregates nodes. The bagging method is used by the random forest to generate the required forecast. Rather than using a single sample of data, bagging includes using many samples (training data). A training dataset is a collection of observations and attributes used to make predictions. Depending on the training data provided to the random forest algorithm, the decision trees produce varied results. These outputs will be ranked, and the one with the best score will be chosen as the final product. Random forest classification uses an ensemble methodology to achieve the desired result. Various decision trees are trained using the training data. This dataset contains observations and features that will be chosen at random when nodes are split [27]. Various decision trees are used in a rain forest system. There are three types of nodes in a decision tree: decision nodes, leaf nodes, and the root node. Each tree's leaf node represents the final output produced by that decision tree.

The other duty that a random forest algorithm does is regression. The principle of simple regression is followed by a random forest regression. In the random forest model, the values of dependent (features) and independent variables are passed. In terms of data extrapolation, random forest regression isn't ideal [28]. Unlike linear regression, which uses present data points to estimate values outside of the observation range, nonlinear regression employs existing observations to estimate values outside of the observation range. This explains why most random forest applications are related to classification. When the data is sparse, random forest does not generate good results. In this situation, the bootstrapped sample and the subset of features will result in an invariant space. This will result in ineffective divides, which will have an impact on the outcome.

In our analysis, the Random Forest algorithm had a prediction accuracy of 92%.

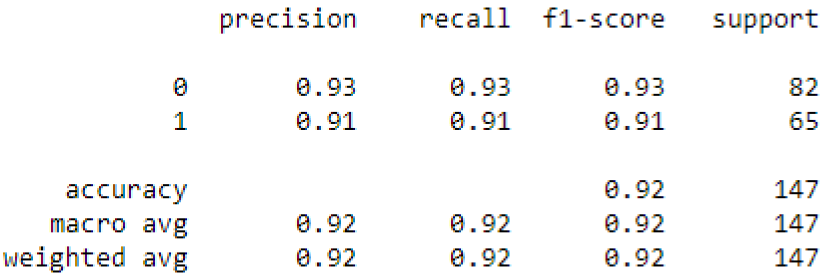
Variable Importance



	Feature	Importance
8	atemp	0.201473
1	yr	0.188947
7	temp	0.183162
9	hum	0.110991
10	windspeed	0.089045
0	season	0.082326
2	mnth	0.073257
4	weekday	0.036794
6	weathersit	0.026016
5	workingday	0.006961
3	holiday	0.001028

Fig. 4 Ranked Variables by RF Algorithm (Random Forest)

According to the Random Forest model (shown in Figure 4), the five most important variables are year, temperature, humidity, seasons and windspeed. The algorithm was then tested for negative values. The sum of the negative values is zero implying that the algorithm is not predicting any negative values.



```
              precision    recall  f1-score   support

           0       0.93      0.93      0.93        82
           1       0.91      0.91      0.91        65

    accuracy                           0.92       147
   macro avg       0.92      0.92      0.92       147
weighted avg       0.92      0.92      0.92       147
```

Fig5: RF Accuracy Measures (Random Forest)

The Figure 5 shows the values for precision, recall and f1-score. The precision and recall score suggest that the Random Forest correctly predicts number of bike rentals greater than 4500 correctly 91% of the time and those equal to or less than 4500 correctly 93% of the time. The macro average suggests that the average of the scores is 92%. The support column suggests that the number of samples that are true for those equal to or less than 4500 is 82, the true samples for those greater than 4500 is 65.

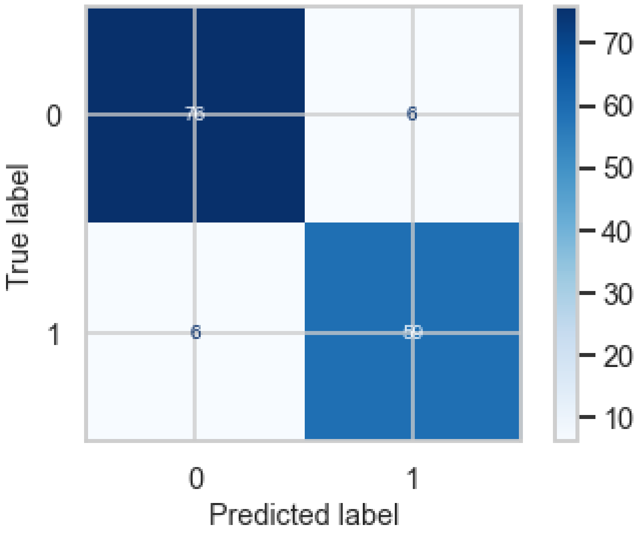
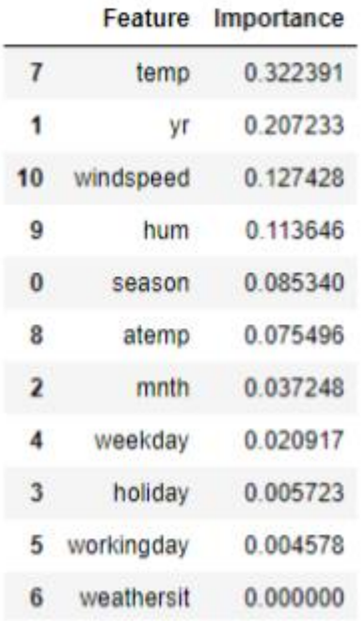


Fig. 6 RF Confusion Matrix (Random Forest)

Figure 6 shows the confusion matrix, the confusion matrix suggests that out of 82 true samples predicted 76 observations of bike renters were correctly predicted as being less than or equal to 4500. Out of 65 True observations, 59 observations were correctly predicted as having more than 4500 renters.

Decision Trees

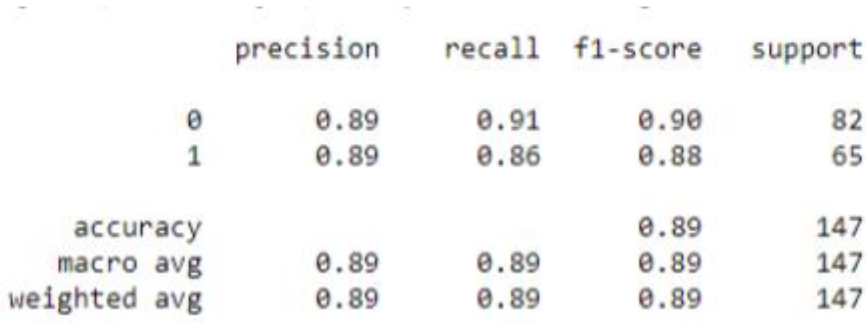
Decision Tree Analysis is a general-purpose predictive modeling tool with a wide range of applications. In general, decision trees are built using an algorithm that finds different ways to split a data set based on various conditions. It's one of the most popular and practical supervised learning methodologies [29]. Decision Trees are a non-parametric supervised learning technique that can be used for regression and classification. The objective is to generate a model that predicts the values of targeted variables using simple decision rules inferred. A decision tree is a tree-like graph with nodes indicating the point at which we select an attribute, edges representing the responses to the query, and leaves represent the actual output. With a simple linear decision surface, they are employed in non-linear decision making [30]. The examples are classified using decision trees by sorting them along the tree from the root to a leaf node, with the leaf node supplying the classification to the example. Each node in the tree represents a test case for an attribute, with each edge descending from that node representing one of the test case's possible solutions. This is a cyclical procedure that occurs for each subtree rooted at the new nodes.



	Feature	Importance
7	temp	0.322391
1	yr	0.207233
10	windspeed	0.127428
9	hum	0.113646
0	season	0.085340
8	atemp	0.075496
2	mnth	0.037248
4	weekday	0.020917
3	holiday	0.005723
5	workingday	0.004578
6	weathersit	0.000000

Fig 7: Ranked Variable (Decision Tree)

The Figure 7 shows a decision tree that shows that the most important variables that predict the number of bikes rented are temperature, Year, Windspeed, humidity and Seasons. The algorithm was then tested for negative values.



```
              precision    recall  f1-score   support

           0       0.89      0.91      0.90        82
           1       0.89      0.86      0.88        65

    accuracy                           0.89       147
   macro avg       0.89      0.89      0.89       147
weighted avg       0.89      0.89      0.89       147
```

Fig 8: RF Accuracy Measures (Decision Tree)

The Figure 8 shows the values for precision, recall and f1-score. The precision and recall score suggest that the Decision Tree correctly predicts bike rentals greater than 4500 correctly 89% and 86% of the time respectively and those equal to or less than 4500 correctly 89% and 91% of the time respectively. The macro average suggests that the average of the scores is 89%. The support column suggests that the number of samples that are true for those equal to or less than 4500 is 82, the true samples for those greater than 4500 is 65.

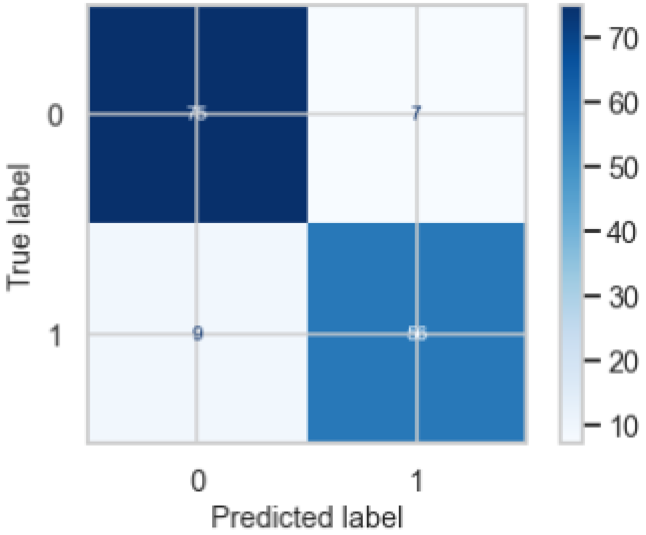


Fig. 9: (Confusion Matrix -Decision Tree)

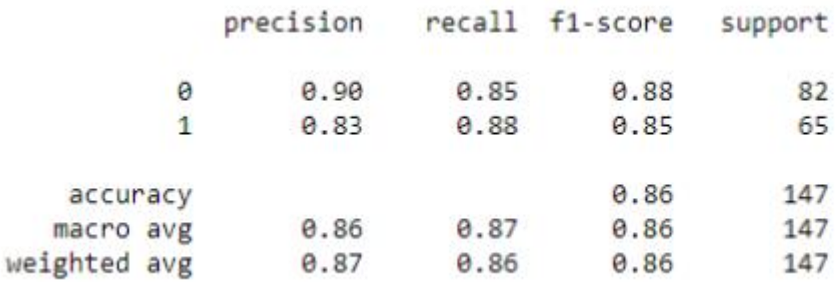
Figure 9 shows the confusion matrix, the confusion matrix suggests that out of 82 true samples predicted 75 observations of bike renters were correctly predicted as being less than or equal to 4500. Out of 65 True observations, 56 observations were correctly predicted as having more than 4500 renters.

Nearest Neighbor and Deep Learning

Because of its simplicity, ease of implementation, and efficacy, KNN (k-nearest neighbor) is a widely used classification technique. It is one of the top 10 data mining algorithms and has a wide range of applications. KNN has a few flaws that impair its categorization accuracy. It has a lot of memory requirements and a lot of time complexity [31]. The value of k must be chosen carefully for KNN to work. In real-world data sets, certain classes have more data points than others. In most circumstances, if k is a fixed, user-defined value, the result will be biased towards the majority class. Dynamic KNN is another good method for learning the best k value during training period (DKNN). It is based on the leave-one-out cross-validation method, which is a hybrid of eager and lazy learning [32].

The similarity or difference between the training and test instances is measured by KNN using standard Euclidean distance. It takes into account the equal participation of all of the instance's qualities, whether or not they are significant. As a result, when there are a high number of irrelevant qualities, the distance function's value becomes erroneous, which is referred to as the Curse of Dimensionality [32]. To solve this problem, assign varying degrees of priority to each attribute and weight each attribute differently when computing distance between two instances.

The final value used for the model was k = 10. The algorithm was then tested for negative values. The sum of the negative values is zero implying that the algorithm is not predicting any negative values.



```
              precision    recall  f1-score   support

           0       0.90      0.85      0.88        82
           1       0.83      0.88      0.85        65

    accuracy                           0.86       147
   macro avg       0.86      0.87      0.86       147
weighted avg       0.87      0.86      0.86       147
```

Fig 10: Accuracy measures for Nearest Neighbor

The Figure 9 shows the values for precision, recall and f1-score. The precision and recall score suggest that the Decision Tree correctly predicts bike rentals greater than 4500 correctly 83% and 87% of the time respectively and those equal to or less than 4500 correctly 90% and 85% of the time respectively. The macro average suggests that the average of the scores is 87%. The support column suggests that the number of samples that are true for those equal to or less than 4500 is 82, the true samples for those greater than 4500 is 65.

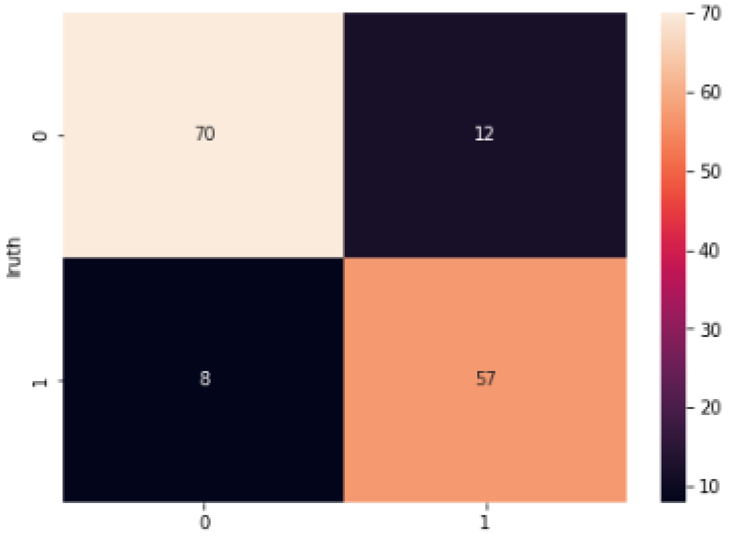
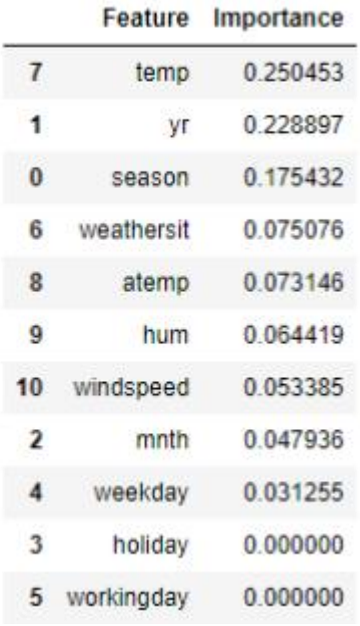


Fig. 11: Confusion Matrix-Nearest Neighbor

Figure 11 shows the confusion matrix, the confusion matrix suggests that out of 82 true samples predicted 70 observations of bike renters were correctly predicted as being less than or equal to 4500. Out of 65 True observations, 57 observations were correctly predicted as having more than 4500 renters.

XGBoost

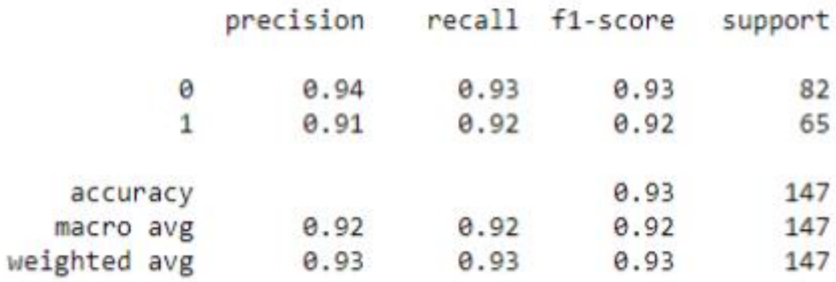
The XGBoost classifier is a machine learning technique that may be used to classify both structured and tabular data. XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees. XGBoost is a gradient boost technique with high gradients. As a result, it's a large Machine Learning algorithm with a lot of moving pieces. XGBoost is capable of handling huge, complex datasets. XGBoost is a strategy for ensemble modeling. XGBoost is a method of ensemble learning. It may not always be enough to rely on the outcomes of a single machine learning model. Ensemble learning is a method for combining the predictive abilities of numerous learners in a systematic way. The end result is a single model that combines the outputs of numerous models. The foundation learners, or models that make up the ensemble, could be from the same learning algorithm or from distinct learning algorithms. The most extensively used ensemble learning models are bagging, boosting, stack generalization, and expert mixtures. Bagging and boosting, on the other hand, are two highly appreciated ensemble learners. Though these two strategies can be applied to a variety of statistical models, decision trees have been the most popular. In this study, XGBoost was selected as the best model.



	Feature	Importance
7	temp	0.250453
1	yr	0.228897
0	season	0.175432
6	weathersit	0.075076
8	atemp	0.073146
9	hum	0.064419
10	windspeed	0.053385
2	mnth	0.047936
4	weekday	0.031255
3	holiday	0.000000
5	workingday	0.000000

Fig. 12: Ranked Variable (XGBoost)

Figure 12 shows that the most important predictors according to the XGBoost algorithm are; Temperature, year, Seasons, weather situation, weather situation.



```
              precision    recall  f1-score   support

           0       0.94      0.93      0.93        82
           1       0.91      0.92      0.92        65

    accuracy                           0.93       147
   macro avg       0.92      0.92      0.92       147
weighted avg       0.93      0.93      0.93       147
```

Fig. 13: Accuracy measures for XGBoost

The Figure 13 shows the values for precision, recall and f1-score. The precision and recall score suggest that the Decision Tree correctly predicts bike rentals greater than 4500 correctly 91% and 92% of the time respectively and those equal to or less than 4500 correctly 94% and 93% of the time respectively. The macro average suggests that the average of the scores is 87%. The support column suggests that the number of samples that are true for those equal to or less than 4500 is 82, the true samples for those greater than 4500 is 65.

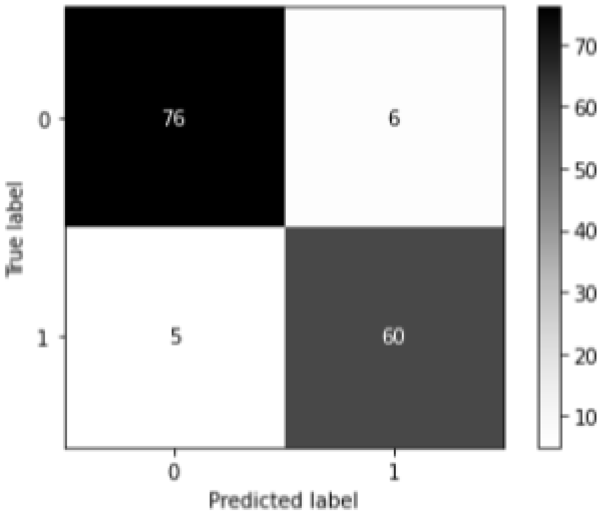
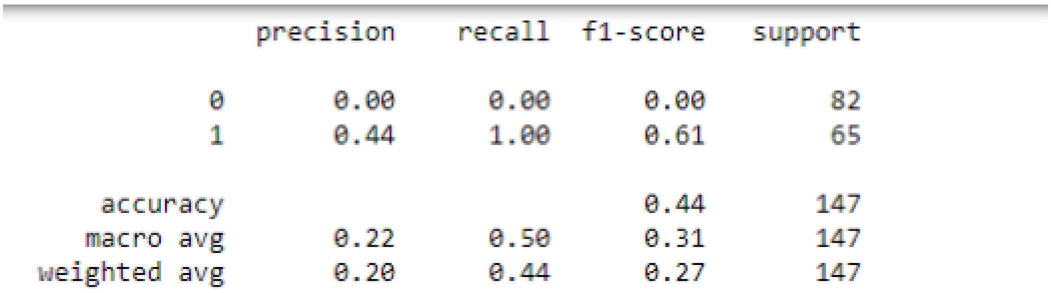


Fig. 14: Confusion Matrix for XGBoost

Figure 14 shows the confusion matrix, the confusion matrix suggests that out of 82 true samples predicted 76 observations of bike renters were correctly predicted as being less than or equal to 4500. Out of 65 True observations, 60 observations were correctly predicted as having more than 4500 renters.

Logistics Regression

Logistic Regression is a Machine Learning method that is used to solve classification issues. It is a predictive analytic approach that is based on the probability notion. In this case, the dependent variable in this case has two categories; ‘1’-count greater than 4500 and ‘0’-count less than or equal to 4500.



```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        82
           1       0.44      1.00      0.61        65

    accuracy                           0.44       147
   macro avg       0.22      0.50      0.31       147
weighted avg       0.20      0.44      0.27       147
```

Fig. 15: Accuracy Measures for Logistics Regression

The Figure 15 shows the values for precision, recall and f1-score. The precision and recall score suggest that the logistics regression correctly predicts bike rentals greater than 4500 correctly 44% and 100% of the time respectively and those equal to or less than 4500 correctly 0% and 0% of the time respectively. The macro average suggests that the average of the scores is 87%. The support column suggests that the number of samples that are true for those equal to or less than 4500 is 82, the true samples for those greater than 4500 is 65.

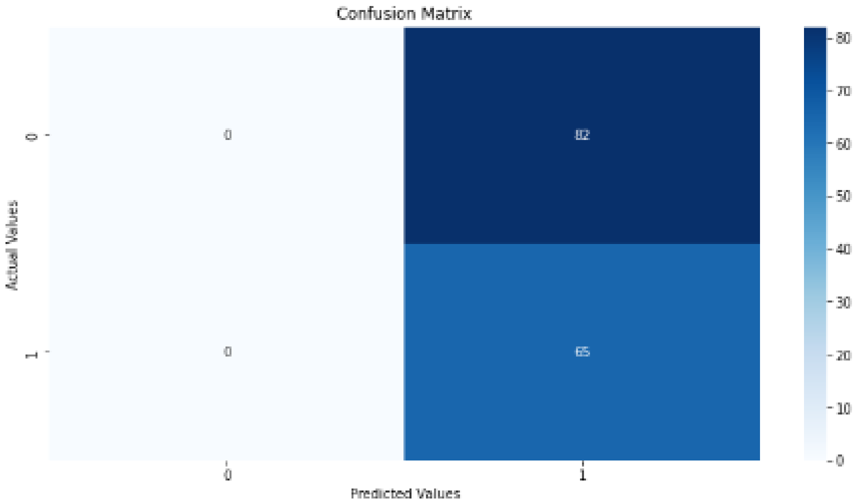


Fig 16: Confusion Matrix- for Logistics Regression

Figure 16 shows the confusion matrix, the confusion matrix suggests that zero observations of bike renters were correctly predicted as being less than or equal to 4500. 65 observations were correctly predicted as having more than 4500 renters.

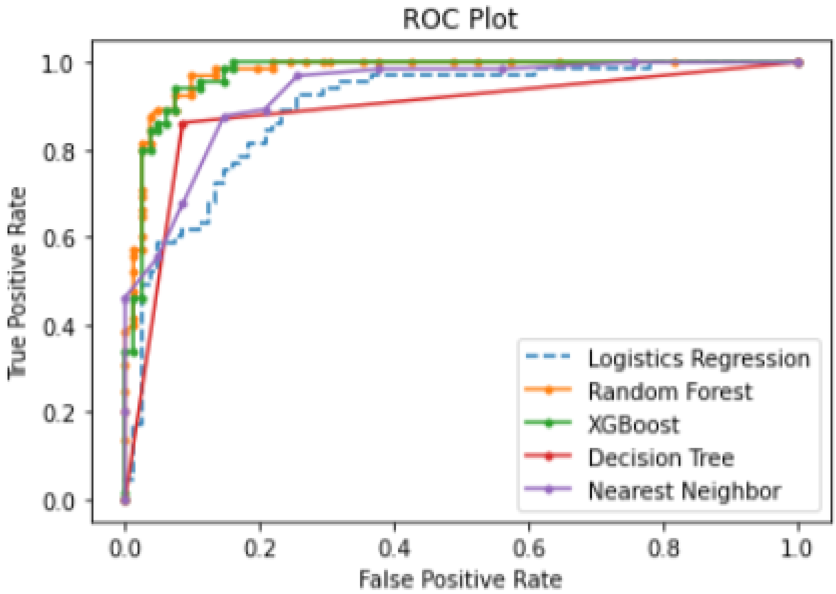


Fig 17: RECIEVER OPERATING CHARACTERISTICS CURVE

Figure 17 shows the ROC curve for the Random Forest algorithm, the ROC suggests that the XGBoost model has an area under curve (AUC) of 97.35%. The ROC suggests that the Decision Tree model has AUC of 87.27%. The ROC suggests that the Random Forest model has an accuracy of 97.67%. The logistics regression model ha AUC of 89.41%. The Nearest Neighbor had AUC value of 93.07%.

Discussion

This study employed machine learning classifiers to predict the number of bike share renters. The machine learning classifiers include Random Forest, Decision Trees, Nearest Neighbor and XGBoost. the five most important variables are year, temperature, humidity, seasons and windspeed. The Random Forest Algorithm had an AUC of 97.67%. The decision shows that the most important variables that predict the number of bikes rented are temperature, Year, Windspeed, humidity and Seasons. The decision tree model has AUC of 87.27%. According to the KNN algorithm, the five most important variables are temperature, year, fall, humidity and windspeed. The KNN algorithm had AUC of 93.07%. The logistics regression model had an AUC of 89.41%. The XGBoost model had an AUC of 97.35. According to the XGBoost algorithm the top five most important predictors are Temperature, year, Seasons, weather situation, weather situation.

Although XGBoost had the best accuracy as consistent with (Kim, Park, Shin and Oh, 2021), based on the Receiver Operating Characteristics curve's area under curve, the Random Forest model was selected as the best model because it had the highest area under curve. The most important variables are atemp (Normalized Temperature in degrees Celsius (20.15%), year (18.89%), temp (Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=8,t_max=+39 (18.32%), humidity (11.10%), Windspeed (8.91%), season(8.23%), month (7.33%), whether it was a weekday or not (3.68%) and the weather situation (2.6%). The other variables had an importance of less than 1%.

Because outliers are inevitable in practice, the model may fail to predict them, this is part of the study's limitations. With correct hyper parameter tuning, there may be improved machine learning or deep learning models that can produce better outcomes. Geographic data such as longitude and latitude data from station locations and bike rental decks could be used in future studies. The distances between those location points can help you make sense of your data. Predicting future mobility demand for other modes of transportation besides bike shares, such as electric motorcycles, is also a possibility. It's difficult to extrapolate the findings from our study to other bicycle-sharing programs around the world because we used data from Washington, DC. As a result, more cities' data must be analyzed in order to make thorough findings and create models that can reflect differences between cities. Also necessary is a model presentation that reflects the changing character of public transportation systems in metropolitan areas following COVID 19.

Conclusion

Bike-sharing systems have become popular in recent years all around the world. Although this trend has resulted in many studies on public cycling systems, there have been few previous studies on the factors influencing public bicycle travel behaviour. A bike-sharing system is a service in which individuals can borrow bikes for a fee or free for a limited period. Many bike share programs allow users to borrow a bike from a system, which is usually computer-controlled. The user enters payment information, and the system unlocks the bike. After that, the bike can be returned to a system-wide dock. The study's goal is to figure out how much demand there is for shared bikes across the country based on compelling parameter estimates. Rental firms arrange this to position themselves to meet people's requirements whenever the situation improves overall, allowing them to stand out from other service providers and earn handsomely. My focus is to apply pre-screened parameters in predicting number of bike share users - demand. How well those variables accurately characterize the bike's requirements. The service provider organization has amassed a vast dataset on daily bike requests across the market based on some parameters which can reliably be applied in predicting potential demand.

To achieve this objective, study employed machine learning classifiers to predict the number of bike share renters. The machine learning classifiers include Random Forest, Decision Trees, Nearest Neighbor and XGBoost. the five most important variables are year, temperature, humidity, seasons and windspeed. The Random Forest Algorithm had an AUC of 97.67%. The decision shows that the most important variables that predict the number of bikes rented are temperature, Year, Windspeed, humidity and Seasons. The decision tree model has AUC of 87.27%. According to the KNN algorithm, the five most important variables are temperature, year, fall, humidity and windspeed. The KNN algorithm had AUC of 93.07%. The logistics regression model had an AUC of 89.41%. The XGBoost model had an AUC of 97.35. According to the XGBoost algorithm the top five most important predictors are Temperature, year, Seasons, weather situation, weather situation. The linear regression analysis carried out determined that the variables that significantly predict the total number of registered & unregistered bikers: renters (cnt) include Temperature, humidity, wind speed, the month of September, Spring, Fall, all weekdays except weekday one and weather situation 3 (Light Snow, Light Rain with Scattered Clouds). Comprehending the temporal features of bike-sharing usage may aid service providers and policymakers in improving bike-sharing services.

Bike sharing programs (BSPs) continue to evolve and expand at a rapid pace. Many countries have implemented various BSP concepts and techniques since the 1960s. There are a variety of versions available, ranging from dock less to electronic real-time monitoring systems. Recreation, errands, work, and other activities may all be done with these BSP. And all signs point to the introduction of more complex and inventive rider-friendly technology in the future. The goal of this article is to apply pre-screened variables established by various operators and streamline them using analytics to discover the most appealing ones. There is a lack of standardization and a single criterion on what is required and what is not, given the contents of existing data sets. There appear to be elements in common among BSP organizations: weather, duration, season, temperature etc. This article is based on historical data provided by a single operator in the Washington, District of Columbia, United States. Several variables were tested, including categorical and continuous data types. Eight of the 18 were deemed acceptable and contributed significantly to the development of usable and reliable predictive model. Bike-sharing systems have grown in popularity around the world in recent years. Even though this trend has resulted in a slew of studies on public bicycle systems, there have been few studies on predicting the factors that influence public bicycle travel behavior. Bike-sharing is a computer-controlled system that allows people to rent bikes for a price or for free for a limited time.

ACKNOWLEDGMENTS

The author acknowledges approval provided by CapitalBike and Hadi Fanaee-T : both granted access to the database used in this study. The latter was specifically for refining and structuring the data set for instant use.

AUTHOR CONTRIBUTIONS

The study design, analyses and narrative were conducted by the author.

FUNDING

Personal funds

CONFLICT OF INTEREST: The author declares no conflict of interest

References

[1]V. Albuquerque, M. Sales Dias and F. Bacao, "Machine Learning Approaches to Bike-Sharing Systems: A Systematic Literature Review", ISPRS International Journal of Geo-Information, vol. 10, no. 2, p. 62, 2021. Available: 10.3390/ijgi10020062 [Accessed 1 January 2022].

[2]B. Wang and I. Kim, "Short-term prediction for bike-sharing service using machine learning", Transportation Research Procedia, vol. 34, pp. 171-178, 2018. Available: 10.1016/j.trpro.2018.11.029 [Accessed 1 January 2022].

[3]S. Sohrabi, R. Paleti, L. Balan and M. Cetin, "Real-time prediction of public bike sharing system demand using generalized extreme value count model", Transportation Research Part A: Policy and Practice, vol. 133, pp. 325-336, 2020. Available: 10.1016/j.tra.2020.02.001 [Accessed 1 January 2022].

[4]Y. Wang, D. Zhang, Y. Liu, B. Dai and L. Lee, "Enhancing transportation systems via deep learning: A survey", Transportation Research Part C: Emerging Technologies, vol. 99, pp. 144-163, 2019. Available: 10.1016/j.trc.2018.12.004 [Accessed 1 January 2022].

[5]T. Gu, I. Kim and G. Currie, "Measuring immediate impacts of a new mass transit system on an existing bike-share system in China", Transportation Research Part A: Policy and Practice, vol. 124, pp. 20-39, 2019. Available: 10.1016/j.tra.2019.03.003 [Accessed 1 January 2022].

[6]X. Ma, Y. Ji, M. Yang, Y. Jin and X. Tan, "Understanding bikeshare mode as a feeder to metro by isolating metro-bikeshare transfers from smart card data", Transport Policy, vol. 71, pp. 57-69, 2018. Available: 10.1016/j.tranpol.2018.07.008 [Accessed 1 January 2022].

[7]H. Fanaee-T and J. Gama, 2014. [Online]. Available: https://www.researchgate.net/publication/259382357_Bike-Sharing_Dataset. [Accessed: 01- Jan- 2022].

[8]Z. Zou, H. Younes, S. Erdoğan and J. Wu, "Exploratory Analysis of Real-Time E-Scooter Trip Data in Washington, D.C.", Transportation Research Record: Journal of the Transportation Research Board, vol. 2674, no. 8, pp. 285-299, 2020. Available: 10.1177/0361198120919760 [Accessed 1 January 2022].

[9]A. Campbell, C. Cherry, M. Ryerson and X. Yang, "Factors influencing the choice of shared bicycles and shared electric bikes in Beijing", Transportation Research Part C: Emerging Technologies, vol. 67, pp. 399-414, 2016. Available: 10.1016/j.trc.2016.03.004.

[10]Y. Guo, J. Zhou, Y. Wu and Z. Li, "Identifying the factors affecting bike-sharing usage and degree of satisfaction in Ningbo, China", PLOS ONE, vol. 12, no. 9, p. e0185100, 2017. Available: 10.1371/journal.pone.0185100 [Accessed 1 January 2022].

[11]K. Kim, J. Ghimire, P. Pant and E. Yamashita, "Bikeshare and safety: Risk assessment and management", Transportation Research Interdisciplinary Perspectives, vol. 9, p. 100276, 2021. Available: 10.1016/j.trip.2020.100276 [Accessed 1 January 2022].

[12]C. Basch, E. Zagnit, S. Rajan, D. Ethan and C. Basch, "Helmet Use Among Cyclists in New York City", Journal of Community Health, vol. 39, no. 5, pp. 956-958, 2014. Available: 10.1007/s10900-014-9836-8 [Accessed 1 January 2022].

[13]Q. Zhou, C. Wong and X. Su, "Machine Learning Approach to Quantity Management for Long-Term Sustainable Development of Dockless Public Bike: Case of Shenzhen in China", Journal of Advanced Transportation, vol. 2020, pp. 1-13, 2020. Available: 10.1155/2020/8847752 [Accessed 1 January 2022].

[14]M. Fernandez-Delgado, E. Cernadas, S. Barro and D. Amorim, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?", Journal of Machine Learning Research, 2014. [Accessed 1 January 2022].

[15]C. Peng and T. So, "Logistic Regression Analysis and Reporting: A Primer", Understanding Statistics, vol. 1, no. 1, pp. 31-70, 2002. Available: 10.1207/s15328031us0101_04 [Accessed 1 January 2022].

[16]V. Albuquerque, M. Sales Dias and F. Bacao, "Machine Learning Approaches to Bike-Sharing Systems: A Systematic Literature Review", ISPRS International Journal of Geo-Information, vol. 10, no. 2, p. 62, 2021. Available: 10.3390/ijgi10020062 [Accessed 1 January 2022].

[17]T. Tran, N. Ovtracht and B. d’Arcier, "Modeling Bike Sharing System using Built Environment Factors", Procedia CIRP, vol. 30, pp. 293-298, 2015. Available: 10.1016/j.procir.2015.02.156 [Accessed 1 January 2022].

[18]T. Yang, Y. Li and S. Zhou, "System Dynamics Modeling of Dockless Bike-Sharing Program Operations: A Case Study of Mobike in Beijing, China", Sustainability, vol. 11, no. 6, p. 1601, 2019. Available: 10.3390/su11061601 [Accessed 1 January 2022].

[19]L. Caggiani and M. Ottomanelli, "A Dynamic Simulation based Model for Optimal Fleet Repositioning in Bike-sharing Systems", Procedia - Social and Behavioral Sciences, vol. 87, pp. 203-210, 2013. Available: 10.1016/j.sbspro.2013.10.604 [Accessed 1 January 2022].

[20]F. Soriguera, V. Casado and E. Jiménez, "A simulation model for public bike-sharing systems", Transportation Research Procedia, vol. 33, pp. 139-146, 2018. Available: 10.1016/j.trpro.2018.10.086 [Accessed 1 January 2022].

[21]K. Kim, J. Ghimire, P. Pant and E. Yamashita, "Bikeshare and safety: Risk assessment and management", Transportation Research Interdisciplinary Perspectives, vol. 9, p. 100276, 2021. Available: 10.1016/j.trip.2020.100276 [Accessed 1 January 2022].

[22]H. Ashqar, M. Elhenawy, M. Almannaa, A. Ghanem, H. Rakha and L. House, "Modeling bike availability in a bike-sharing system using machine learning", 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017. Available: 10.1109/mtits.2017.8005700 [Accessed 1 January 2022].

[23]Q. Zhou, C. Wong and X. Su, "Machine Learning Approach to Quantity Management for Long-Term Sustainable Development of Dockless Public Bike: Case of Shenzhen in China", Journal of Advanced Transportation, vol. 2020, pp. 1-13, 2020. Available: 10.1155/2020/8847752 [Accessed 1 January 2022].

[24]Analysis and Prediction of Bike Sharing Traffic Flow., Technical University of Munich Chair of Transportation Systems Engineering, 2020. [Accessed 1 January 2022].

[25]K. Kim, J. Ghimire, P. Pant and E. Yamashita, "Bikeshare and safety: Risk assessment and management", Transportation Research Interdisciplinary Perspectives, vol. 9, p. 100276, 2021. Available: 10.1016/j.trip.2020.100276 [Accessed 1 January 2022].

[26]S. Sun and M. Ertz, "Contribution of bike-sharing to urban resource conservation: The case of free-floating bike-sharing", Journal of Cleaner Production, vol. 280, p. 124416, 2021. Available: 10.1016/j.jclepro.2020.124416 [Accessed 1 January 2022].

[27]A. Sarica, A. Cerasa and A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review", Frontiers in Aging Neuroscience, vol. 9, 2017. Available: 10.3389/fnagi.2017.00329 [Accessed 1 January 2022].

[28]Q. Ren, H. Cheng and H. Han, "Research on machine learning framework based on random forest algorithm", AIP Conference Proceedings, 2017. Available: 10.1063/1.4977376 [Accessed 1 January 2022].

[29]H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms", International Journal of Computer Sciences and Engineering, vol. 6, no. 10, pp. 74-78, 2018. Available: 10.26438/ijcse/v6i10.7478 [Accessed 1 January 2022].

[30]Z. Zhang, Z. Zhao and D. Yeom, "Decision Tree Algorithm-Based Model and Computer Simulation for Evaluating the Effectiveness of Physical Education in Universities", Complexity, vol. 2020, pp. 1-11, 2020. Available: 10.1155/2020/8868793 [Accessed 1 January 2022].

[31]S. Taneja, C. Gupta, S. Aggarwal and V. Jindal, "MFZ-KNN &#x2014; A modified fuzzy based K nearest neighbor algorithm", 2015 International Conference on Cognitive Computing and Information Processing(CCIP), 2015. Available: 10.1109/ccip.2015.7100689 [Accessed 1 January 2022].

[32]M. Gupta, "Dynamic k-NN with Attribute Weighting for Automatic Web Page Classification (Dk-NNwAW)", Birla Institute of Technology and Science Pilani-Dubai, 2012. [Accessed 1 January 2022].