International Journal of Computational Science and Engineering Research
ISSN: 3107 -8605 (Online) , http://www.ijcser.com/
Regular Issue , Vol. 2, Issue. 4 (October – December), 2024 , Pages: 44 - 50
Received: 29 May 2024 ; *Accepted:* 28 October 2024 ; *Published:* 30 October 2024
https://doi.org/10.63328/IJCSER-V1RI4P9
Original Paper : Applied Research – Report Research Paper

Bike Sharing : A Forecast using Machine Learning Techniques

M Veeresh Babu [1*], A Tejesh Babu [2], V Sasi Kumar Reddy [3], S. Sameer Ahamed [4], K Udaya Kiran [5]

[1-5] Department of Computer Science and Engineering , Aditya College of Engineering , Madanapalle, India

** Corresponding Author: M Veeresh Babu ; sveeruchandra@gmail.com*

Abstract: Bike sharing systems have emerged as sustainable and convenient transportation alternatives in urban environments. Understanding and predicting bike usage patterns is crucial for optimizing system efficiency and enhancing user experience. In this paper developed a machine learning model to predict bike sharing demand based on various factors, such as weather conditions, time of day, day of the week, and other relevant features. The dataset utilized for this paper includes historical bike sharing data, encompassing information on the number of bikes rented at different time intervals, weather conditions, and temporal attributes. The primary objective is to create a robust predictive model capable of forecasting bike demand accurately. The methodology involves preprocessing and cleaning the dataset, feature engineering to extract relevant information, and employing a machine learning algorithm, such as a regression model or ensemble method. The model will be trained on a subset of the dataset and validated using another portion to ensure generalizability.

Keywords: CSV (Comma Separated Values), Seaborn, NumPy, UML, Kernel.

1. Introduction

A shared micromobility service for short-term bike rentals is the bike-sharing system. The service may be provided for free (for example, at the expense of the city) or at a cost. Worldwide, there are about 2000 different bike-sharing systems, the most of which are located in urban areas. In Amsterdam, 50 bikes were placed around the city unlocked for public use in 1965 by a group known as Provo, which marked the beginning of the city's bike-sharing program. A coin-deposit system was the second-generation bike-sharing arrangement. Using a coin that was returned to them when the bike was returned, users could unlock the bicycles under this system. Customers of bike sharing can use the service whenever they want and are not dependent on certain routes for transit. They can get there without encountering any traffic or parking issues.

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated. A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and 1 utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."

Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. Below are some points that explain why we should use the Random Forest algorithm: ◎ It takes less training time as compared to other algorithms. ◎ It predicts output with high accuracy, even for the large dataset it runs efficiently. ◎ It can also maintain accuracy when a large proportion of data is missing. Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable and a series of other variables (known as independent variables). Although it is a useful technique for identifying correlations between variables in data, regression analysis is not always able to establish causality. In the fields of business, finance, and economics, it has multiple applications. Investment managers, for example, use it to assess assets and comprehend correlations between things like commodity prices and the stocks of companies that trade in those commodities.

2. Literature Survey

Bike share schemes have been around for 45 years and have gone through three generational changes (DeMaio, 2003; DeMaio & Gifford, 2004). On July 28, 1965, Witt Fiesten introduced the first iteration of the bike sharing scheme in Amsterdam, the Netherlands. A standard white-painted bike type was available for general rental use. The program was accessible to all users and could provide continuous usage if it was positioned close to a destination for the next user, but it had malfunctions and was terminated in a matter of days (DeMaio, 2009). In 1991 and 1993, Denmark launched the second iteration of the bike share program (Nielsen, 1993). It started off as a tiny program with 26 different types of bike rental services available at 4 stations, then in 1995 it hosted a larger program named City Bike in Copenhagen. The Copenhagen Bike was created especially for everyday use, with wheels and durable rubber tires covered in advertising plates. At specific points across the city, it might even be given back in exchange for payment and reception of coin deposits. A next-generation bike-sharing program with better tracking capabilities was developed as a result of the program's theft problems, which persisted even after it was established and run by a non-profit organization and station firm (DeMaio 2009).

The Bike around program, which was introduced in 1996 at the University of Portsmouth in England, was the third iteration of the bike-sharing scheme. Through a variety of technical advancements, including on-board computers, smart card and smartphone access, electronic locking and communication systems, and more, the bike share system became intelligent (Demaio, 2009). This is due to the fact that since the invention of electronic cards, University of Portsmouth students have been renting bikes (Black et al., 1998).

In Lyon, France, JCDecaus installed 1,500 bikes as part of the third generation of the program, which has evolved gradually. One to two new bike sharing programs are launched year (Optimising Bike Sharing in Europe Cities, 2009). Out of all the third-generation bike sharing schemes, this one was the biggest. Bikes were utilized by 15,000 members 6.5 times every day. France's capital, Paris, also took advantage of the opportunity to highlight the initiative (Henley, 2005). After two years, Paris introduced its own bike-sharing scheme, which saw an increase in the number of bikes in the city and suburbs from roughly 7,000 to 23,600. This widespread adoption has significantly altered how bike-sharing systems operate globally.

Since 2007, the total demand for programs that provide bike sharing has been rising. In 2017, there were an estimated 639 bike-sharing programs with around 643,000 bikes spread across 53 countries in almost every part of the globe. Since 2010, there have been more than 88 million shared bike rides in the United States. Over 28 million rides were taken in a year in 2016, which is the same number of journeys taken on the Amtrak system annually. A survey released in May stated that 85% of all bike trips in the United States were made possible by the five largest bike share programs: Citi Bike in New York, Capital Bike share in metro Washington, D.C., Citi Bike in Miami, Divvy in Chicago, and Hub way in metro Boston.

Out of fifteen American cities with the highest traffic congestion, Washington, D.C. ranked second with an annual per capita traffic jam hour loss of 155. Each driver's annual cost of congestion was $2,161, while the city as a whole paid $4.6 billion in costs associated with traffic congestion. There were numerous recorded car accidents in Washington, D.C. as a result of the severe traffic congestion. Since 2008, the overall number of collisions has been steadily rising (Department of Transportation, 2018).

According to the study mentioned above, the demand for bike shares has been rising steadily on a global scale, and the primary incentive for signing up for a program is to get around more quickly and easily. This suggests that the need for alternate modes of transportation arose as a result of outside factors like daily traffic accidents or traffic congestion. In order to improve the forecast results for bike share demand, this study addresses predicting bike share demand using both external and given data. Depending on the features of the program's operational location, the bike share demand forecasting research contains a variety of important impacting aspects. Furthermore, different researchers use predictive models in different ways. We provide an overview and a presentation of the current bike share demand projection study in this part.

Two approaches were used by Alhusseini (2014) to forecast the demand for bike sharing. The first method considered bike demand as a numerical attribute and predicted the demand using a support vector machine (SVM) algorithm. In Alhusseini's second method, bike share demand was treated as a categorical attribute with five class labels, and SoftMax regression and SVM algorithms were employed. In order to forecast the time-to-time bike share demand, Du et al.

(2014) used hourly data from a dataset on bike sharing demand. For demand estimation, Du employed the generalized boosted model and a random forest. The principal component revision (PCR), support vector revision (SVR), conditional reference tree (Ctree), and generalized linear models with explicit net regulation (GLMNet) methods were also used to forecast demand.

Additionally, Kim et al. (2018) suggested a study that used a graph convolutional neural network to forecast bike sharing demand. Wang (2016) used neural networks, decision trees, random forests, and multiple linear regression to forecast the hourly demand for CitiBikes in New York. Wang's research concentrated on using the random forest ensemble approach to increase accuracy. By changing the dependent variable, enhancing the quality of each independent attribute, and adding new ones, he was able to generate several trees. Wang (2016) developed the dataset for himself, in contrast to prior studies. In order to prepare his statistics, he had to gather information from three different sources: historical weather data, rider information, and official New York holiday data.

Godavarthy et al. (2017) used data from the Great Rides Bike Share program in Fargo, North Dakota to study the operational aspect, travel behavior, and travel mode shift. They discovered that the program's implementation in Fargo increased the amount of time that residents and students of North Dakota State University (NDSU) rode bicycles. Additionally, Liu et al. (2019) used multi-time step models and conventional long short-term memory (LSTM) models to forecast the demand for bike shares.

Pan et al. (2018) demonstrated improved performance by employing a recursive neural net (RNN) model to estimate bike sharing demand. In order to forecast client demand, Li (2019) gathered data on bike share in Chicago and used a Gaussian mixture model (GMM). A station-centric model was presented by Zeng et al. (2016) to account for global factors

An auto-regressive moving-average (ARMA) model was employed by Kaltenbrunner et al. (2010) to predict the quantity of bikes and docks at each station. Yoon et al. (2012) predicted the resources available at each station using a modified autoregressive integrated moving average (ARIMA) model. They took temporal and spatial interaction into account in their experiments.

In order to statistically forecast the demand for Seoul bikes, Lim and Chung (2019) created a time series analysis model. Specifically, the Holt-Winters technique, which was employed to gauge the demand for electricity was adjusted and applied to forecast demand. Sensitivity research on the impact of parameter variations on the actual demand prediction was also carried out.

Chen et al. (2016) used a weighted correlation network model to forecast surplus demand for renting public bicycles. A hierarchical prediction model was put forth by Li et al. (2015) to forecast the check-in and check-out of every station cluster. Using a bipartite clustering technique, they initially grouped stations based on their geographic positions and transition patterns. They then used a gradient boosting regression tree model to forecast traffic for the entire city. A multi-similarity-based inference model predicted the rental proportion between clusters.

Additionally, Min et al. (2017) suggested an analysis methodology that might use visualization to determine the significance of the data. The location of stops was determined based on utilization rates, and the time, day, and month-specific bike use trends were examined. examination of the trip route revealed the pattern of use between stops, and examination of the destination ratio at each stop revealed the reason for use. The development path of Daejeon's public bike rental systems was given based on these facts.

Based on Goyang-si, Korea public bike data, Kim et al. (2012) investigated the impact of weather on demand for public bikes. Temperature variables indicated that bike use decreased if temperatures dropped noticeably or increased above 23°C. Bicycle utilization is negatively impacted by precipitation; for every 10 cm increase in rainfall, bicycle use decreases by approximately 60%. The quantity of clouds in the sky also has an adverse effect on bike usage. While there are some variations in the values when compared to studies conducted abroad, these results demonstrate a similar tendency. With the exception of elementary school kids, Lee et al. (2011) developed a bike demand estimation model whose features included the number of passenger cars and pupils.

Vogel et al. (2014) investigated activity patterns and carried out a case study on the strategy development of bike sharing systems using data mining. In order to determine the ideal rebalancing strategy for Seoul's public bike sharing networks, Lee & Son (2019) estimated the number of bikes using an integer programming technique. In order to address the bike sharing demand rebalancing issue, Liu et al. (2016) created a Meteorology Similarity Weighted K-Nearest Neighbor (MSWK) regressor to forecast station pick-up demand based on extensive historical trip records. Lu (2016) introduced mathematical programming techniques that produce the best possible daily distribution of bikes to a bike-sharing system's stations. To explain time-dependent bike movements within the system, a time-space network was built. Based on the time-space network, a bike fleet allocation model was then designed, taking into account fixed fleet size and average historical demand. Lu's approach attempted to address demand imbalance in bike-sharing systems, where flow from one station to another is rarely equal to the flow in the other direction.

The following Table 1 provides an overview of the aforementioned bike share prediction studies: Numerous prior research has examined various attributes and factors related to the prediction of bike share. Only the characteristics of each variable, such as their numerical or categorical nature, were examined in these earlier investigations. This study, on the other hand, highlights the significance of every feature for every algorithm and centers on a fresh feature investigation from an external data source.

3. Theory / Calculation

Evaluation Measures for Predictive Analytics for Software Defect Detection:

The various software defect prediction metrics, including true positive (TP), true negative (TN), false positive (FP), and false negative (FN), will be covered in this section. In software, TP represents the number of occurrences of defective software that are accurately classified as such, while TN represents the number of instances of clean software that are correctly labeled as such. The numbers FP and FN represent the number of software instances that are incorrectly identified as faulty and clean, respectively, and the number of software instances that are incorrectly classed as defective. Classification accuracy, sometimes known as the correct classification rate, is one of the main straightforward measures used to assess the effectiveness of predictive models. It is used to gauge how closely the cases that have been successfully classified relate to the overall number of cases.

The number of cases accurately categorized as faulty (TP) divided by the total number of cases classified as defective (TP + FP) yields another metric known as accuracy. Furthermore, recall quantifies the proportion of accurately identified defective cases (TP) to the overall number of faulty cases (TP+FN). F-score metrics are widely used in the literature, and they are a harmonic mean of precision and recall.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)\ \text{Precision} = TP/(TP+FP)$$

$$\text{Recall} = TP/(TP+FN)$$

$$\text{F-Score} = (2\ \text{Precision Recall})/\ (\text{Precision} + \text{Recall})$$

4. Experimental Method/ Procedure

The proposed system for the "Bike Sharing - A forecast using machine learning techniques" entails the integration of advanced machine learning techniques to enhance the forecasting capabilities of the existing bike sharing system. Leveraging historical usage data, weather forecasts, time of day, and other relevant factors, the system will employ machine learning algorithms such as regression analysis, decision tree algorithm, and random forest algorithm to generate more accurate predictions of bike demand at various locations and times.

By doing so, the system aims to optimize bike distribution, anticipate fluctuations in demand due to factors like weather and events, and ultimately improve the availability and accessibility of bikes for users. Overall, the integration of machine learning into the bike sharing system holds the potential to enhance operational efficiency, user satisfaction, and overall service quality.

Advantages:

- Enhanced Predictive Accuracy
- Flexibility and Adaptability
- Feature Engineering
- Real-time Insights
- Customization and Optimization

Algorithms :

This is the most basic and simple algorithm you might have ever seen this tool is regression, regression is very commonly used when it comes to machine learning. Nonlinear regression in Machine Learning can be done with the help of decision tree regression. The main function of the decision tree regression algorithm is to split the dataset into smaller sets. The subsets of the dataset are created to plot the value of any data point that connects to the problem statement. The splitting of the data set by this algorithm results in a decision tree that has decision and leaf nodes. ML experts prefer this model in cases where there is not enough change in the data set.

```
Step-1: START
Step-2: Importing the libraries
Step-3: Importing the dataset
Step-4: Splitting the dataset into the Training set and Test
        set
Step-5: Training the Decision Tree Regression model on
        the training set
Step-6: Predicting the Results
Step-7: Comparing the Real Values with Predicted Values
Step-8: Visualising the Decision Tree Regression Results
Step-9: END
```

Random Forest and Decision Tree: The most effective and widely used technique for prediction and categorization is the decision tree. A decision tree is a tree structure that resembles a flowchart, with each internal node signifying an attribute test, each branch representing a test result, and each leaf node (terminal node) holding a class label. Building the Decision Tree: By dividing the source set into subsets according to an attribute value test, a tree can be "taught." This procedure, known as recursive partitioning, is carried out iteratively on every derived subset.

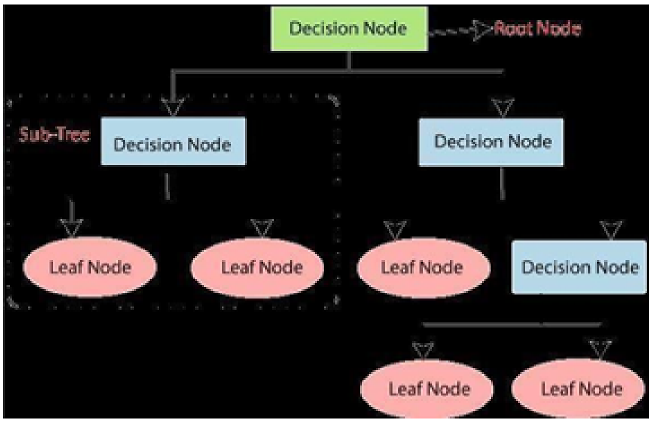


Figure. 1 Working of Decision Tree

The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of a decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high- dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

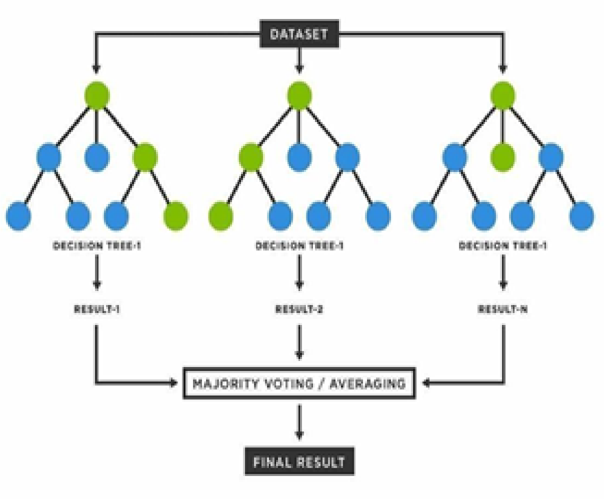


Figure. 2 Working of Random Forest

Among the supervised learning methods is the well-known machine learning algorithm Random Forest. It can be applied to ML issues involving both classification and regression. Its foundation is the idea of ensemble learning, which is the process of merging several classifiers to solve a challenging issue and enhance the model's functionality. H decision trees on different dataset subsets and averages the results to increase the dataset's predicted accuracy." Rather than depending on a single decision tree, the random forest forecasts the outcome based on the majority vote of projections from each tree.

5. Flow Chart

The system architecture gives an overview of the working of the system. The working of this system is described as follows: Downloading the dataset is collecting the data which contains weather details. Attributes selection process of rainfall. After identifying the available data resources, they are further selected, cleaned, made into desired form. Different classification techniques stated will be applied or preprocessed data to predict the accuracy of rainfall. Accuracy measures compares the accuracy of different classifiers.

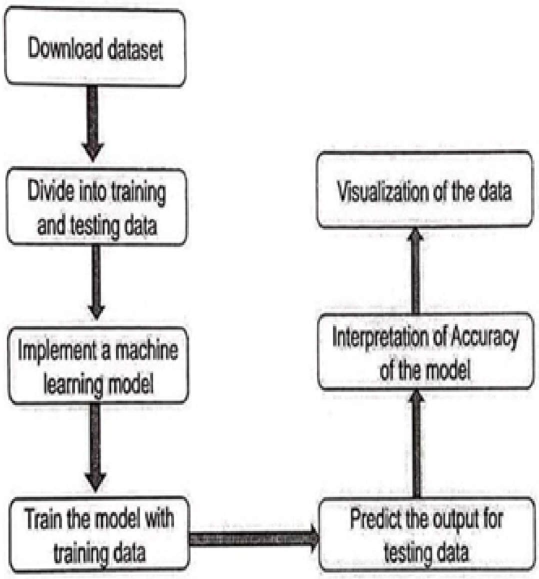
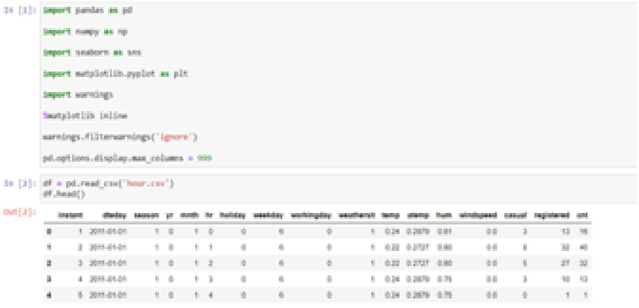

Download dataset
Divide into training and testing data
Implement a machine learning model
Train the model with training data
Predict the output for testing data
Interpretation of Accuracy of the model
Visualization of the data

Figure. 3 System Architecture

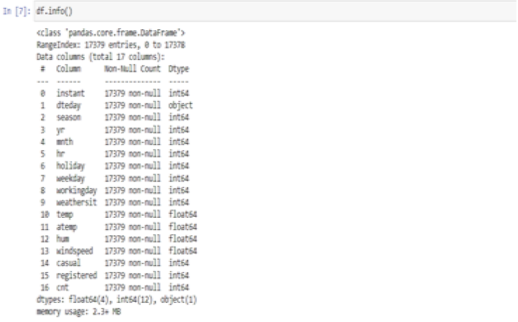
6. Results and Discussion

In the above code import pandas, NumPy, seaborn and matplotlib.pyplot libraries and ignore any warnings. Then read csv file into the notebook and display the data.



```
In [1]: import pandas as pd

        import numpy as np

        import seaborn as sns

        import matplotlib.pyplot as plt

        import warnings

        %matplotlib inline

        warnings.filterwarnings('ignore')

        pd.options.display.max_columns = 999

In [2]: df = pd.read_csv("hour.csv")
        df.head()

Out[2]:
   instant      dteday  season  yr  mnth  hr  holiday  weekday  workingday  weathersit  temp   atemp   hum  windspeed  casual  registered  cnt
0        1  2011-01-01       1   0     1   0        0        6           0           1  0.24  0.2879  0.81        0.0       3          13   16
1        2  2011-01-01       1   0     1   1        0        6           0           1  0.22  0.2727  0.80        0.0       8          32   40
2        3  2011-01-01       1   0     1   2        0        6           0           1  0.22  0.2727  0.80        0.0       5          27   32
3        4  2011-01-01       1   0     1   3        0        6           0           1  0.24  0.2879  0.75        0.0       3          10   13
4        5  2011-01-01       1   0     1   4        0        6           0           1  0.24  0.2879  0.75        0.0       0           1    1
```

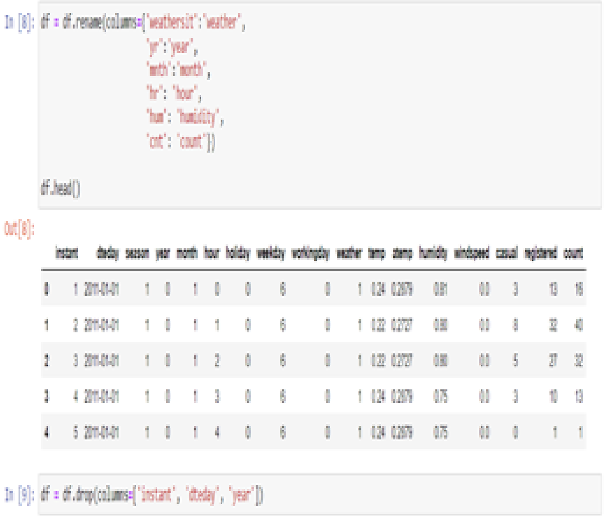
Figure. 4 Import Libraries



```
In [7]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17379 entries, 0 to 17378
Data columns (total 17 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   instant     17379 non-null  int64
 1   dteday      17379 non-null  object
 2   season      17379 non-null  int64
 3   yr          17379 non-null  int64
 4   mnth        17379 non-null  int64
 5   hr          17379 non-null  int64
 6   holiday     17379 non-null  int64
 7   weekday     17379 non-null  int64
 8   workingday  17379 non-null  int64
 9   weathersit  17379 non-null  int64
 10  temp        17379 non-null  float64
 11  atemp       17379 non-null  float64
 12  hum         17379 non-null  float64
 13  windspeed   17379 non-null  float64
 14  casual      17379 non-null  int64
 15  registered  17379 non-null  int64
 16  cnt         17379 non-null  int64
dtypes: float64(4), int64(12), object(1)
memory usage: 2.3+ MB
```

Figure. 5 Data frame Information

Now we see the information about data frame.



```
In [8]: df = df.rename(columns={'weathersit':'weather',
                               'yr':'year',
                               'mnth':'month',
                               'hr': 'hour',
                               'hum': 'humidity',
                               'cnt': 'count'})

        df.head()

Out[8]:
```

	instant	dteday	season	year	month	hour	holiday	weekday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
0	1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0	3	13	16
1	2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0	8	32	40
2	3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0	5	27	32
3	4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0	3	10	13
4	5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0	0	1	1

```
In [9]: df = df.drop(columns=['instant', 'dteday', 'year'])
```

Figure. 6 Rename the columns

Next, rename the columns in meaningful order such as weathersit as weather, yr as year, mnth as month, hr as hour, hum as humidity and cnt as count.

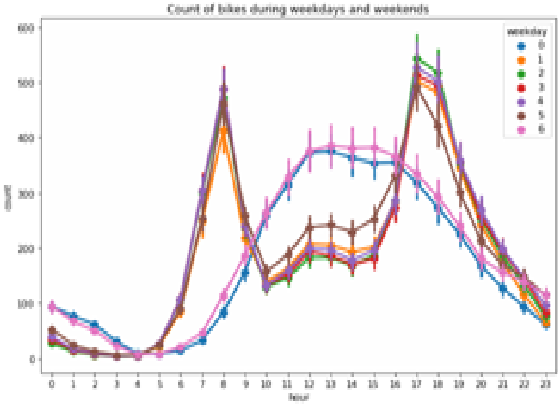


Figure. 3 Graphical Representation (count of bikes during weekdays and weekends)

Now drop the columns instant, dteday and year because it is unwanted data.

In this phase, described the columns season, month, hour, holiday, weekday, working day and weather.

The above graph describes count of bikes during weekdays and weekends .

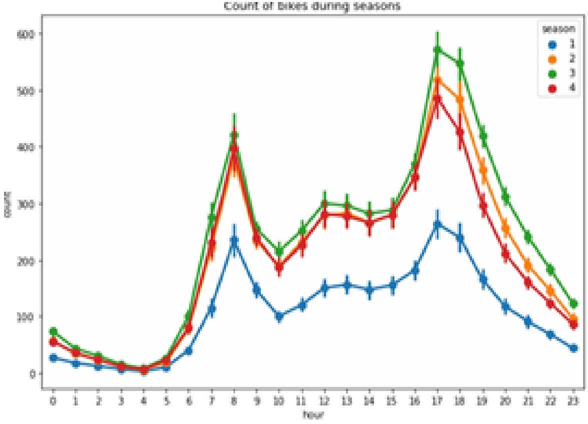


Figure. 8 Graphical Representation (Count of bikes during seasons)

The above graph represents the count of bikes during seasons

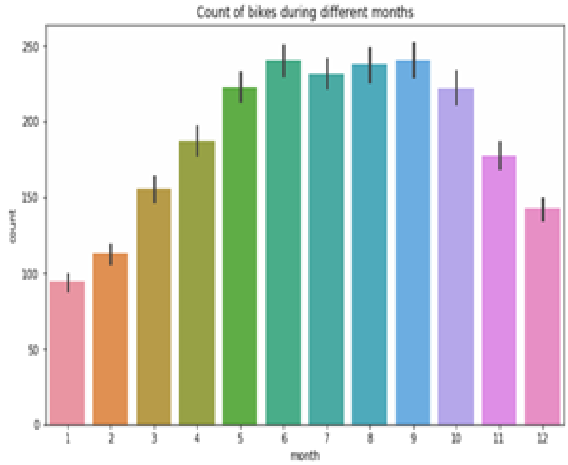


Figure. 9 Bar Graph(Count of bikes during different months)

The above graph represents the count of bikes during different months.

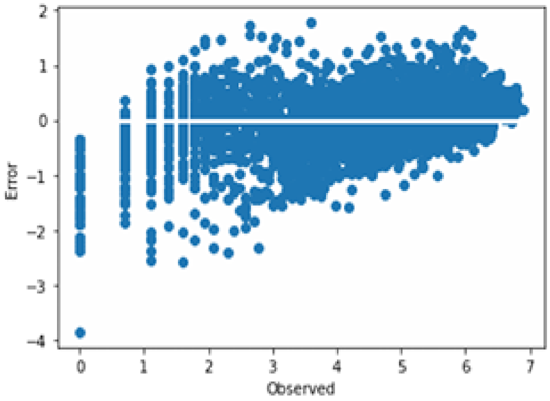
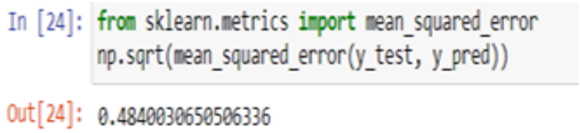


Figure. 10 Scatter Plot

The above diagram shows the scatter plot of the data.
One kind of data visualization that shows the relationship between two continuous variables is the scatter plot. Finding patterns, trends, and correlations in data is one of its main uses.



```
In [24]: from sklearn.metrics import mean_squared_error
         np.sqrt(mean_squared_error(y_test, y_pred))

Out[24]: 0.4840030650506336
```

Figure. 11 Accurate demand prediction

7. Conclusion and Future Scope

The benefits of bike sharing schemes include transport flexibility, reductions to vehicle emissions, health benefits, reduced congestion and fuel consumption, and financial savings for individuals. Through the utilization of a publicly accessible program, people can share bicycles with others and use them "as-needed," eliminating the expenses and liabilities that come with ownership. By doing thus, these programs enable both visitors and locals to benefit from riding bicycles who might not have otherwise done so. By establishing a clear visual cue that bicycles are legitimately allowed on public streets, bike sharing programs can also serve as a catalyst for an increase in bicycle usage. Furthermore, additional studies show that the number of people riding grew in the cities that adopted bike sharing programs, pointing out that these outcomes are a function of both the availability of bike sharing programs and improvements to cycling infrastructure.

Some of them suggesting that the introduction of bike sharing systems can cause cycling to be seen as a safe and normal mode of transport, in contexts where it’s not common.

- Bicycle sharing demand forecast neural networks, which are highly complex yet give excellent accuracy and feature selection automation, should be taken into consideration.
- Even if they can be a little complicated, neural network can significantly increase demand for bike sharing demand classification.
- Bike manufacturing organizations can use them to improve bike sharing demand.
- Government agencies can decide policies and regulations.
- They can help bike organizations create a more appropriate curriculum.

References

[1]. https://www.kaggle.com/code/fatmakursun/bike sharing feature engineering/comments.

[2]. https://www.geeksforgeeks.org/decision-tree.

[3]. https://www.geeksforgeeks.org/random-forest-regression-in-python.

[4]. Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR).

[5]. Decision Trees”, International Journal of Computer Applications (0975 – 8887), Volume 36– No.11, December 2011

[6]. Kaviani, Pouria & Dhotre, Sunita. (2017). Short Survey on Naive Bayes Algorithm. International Journal of Advance Research in Computer Science and Management.

[7]. Kaushik, Manju & Mathur, Bhawana. (2014). Comparative Study of K-Means and Hierarchical Clustering Techniques. International journal of Software and Hardware Research in Engineering. 2. 93- 98.

[8]. Baker, Ryan SJD, Yacef K (2009) The state of educational data mining in 2009: A review and future visions. JEDM 1: 3-16.