CAPSTONE PROJECT REPORT

(Project Term January-May 2023)

Evaluation of popular places In Indian Tourism using Machine Learning

Submitted by

Parth Chuchra	**Registration Number :11904034**
Samaksh Bansal	**Registration Number :11912543**
Roshan Prakash Urkude	**Registration Number :11913133**
Maurya Aryan Anand	**Registration Number :11911493**
Soumya Siddharth Jena	**Registration Number :11913785**

Project Group Number: CSERG0190

Course Code: CSE 445

Under the Guidance of

Chetna Vaid Kwatra
(Assistant Professor)

School of Computer Science and Engineering

LOVELY
PROFESSIONAL
UNIVERSITY

INDIA S LARGEST UNIVERSITY *
LOVELY PROFESSIONAL UNIVERSITY
Transforming Education, Transforming India

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering (SCSE)

Program : P132::B.Tech. (Computer Science and Engineering)

COURSE CODE : CSE445 **REGULAR/BACKLOG:** Regular **GROUP NUMBER:** CSERG0190

Supervisor Name : Chetna Vaid Kwatra **UID :** 27308 **Designation :**Assistant Professor

Qualification : M.Tech **Research Experience :** 5 Years

SR.NO.	NAME OF STUDENT	Prov. Regd. No.	BATCH	SECTION	CONTACT NUMBER
1	Soumya Siddharth Jena	11913785	2019	K19LC	6370343137
2	Roshan Prakash Urkude	11913133	2019	K19PA	9067984882
3	Samaksh Bansal	11912543	2019	K19AP	9826982062
4	Maurya Aryan Anand	11911493	2019	K19AP	9619814271
5	Parth Chuchra	11904034	2019	K19JC	8433076032

SPECIALIZATION AREA : Networking and Security-II **Supervisor Signature:** ChetnaKwatra

PROPOSED TOPIC : Evaluation of popular places In Indian Tourism using Machine Learning

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	6.20
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	6.60
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	6.80
4	Project Supervision: Project supervisor’s is technically competent to guide students, resolve any issues, and impart necessary skills.	7.80
5	Social Applicability: Project work intends to solve a practical problem.	6.20
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	6.60

PAC Committee Members		
PAC Member (HOD/Chairperson) Name: Dr.Max Bhatia	UID: 16870	Recommended (Y/N): Yes
PAC Member (Allied) Name: Ravishanker	UID: 12412	Recommended (Y/N): Yes
PAC Member 3 Name: Dr. Manjit Kaur	UID: 12438	Recommended (Y/N): Yes

Final Topic Approved by PAC: **Evaluation of popular places In Indian Tourism using Machine Learning**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 13897::Dr. Deepak Prashar **Approval Date:** 16 Mar 2023

DECLARATION

We hereby declare that the project work entitled "Evaluation of popular places In Indian Tourism using Machine Learning" is an authentic record of our own work carried out as requirements of Capstone Project for the award of Bachelor of Technology degree in Computer Science Engineering from Lovely Professional University, Phagwara, under the guidance of Mrs. Chetna Vaid Kwarta during January to May 2023. All the information furnished in this capstone report is based on our own intensive work and is genuine.

Project Group Number: KRGC0255

Name of Student 1: Parth Chuchra
Registration Number: 11904034

Name of Student 2: Samaksh Bansal
Registration Number: 11912543

Name of Student 3: Roshan Prakash Urkude
Registration Number: 11913133

Name of Student 4: Maurya Aryan Anand
Registration Number: 11911493

Name of Student 5: Soumya Siddharth Jenna
Registration Number: 11913785

Parth 10.05.2023

Samaksh Bansal 10.05.2023

Roshan 10.05.2023

Aryan 10.05.2023

Siddharth 10.05.2023

CERTIFICATE

This is to certify that the declaration statement made by this group of students is correct to the best of my knowledge and belief. They have completed this Capstone Project under my guidance and supervision. The present work is the result of their original investigation, effort, and study. No part of the work has ever been submitted for any other degree at any University.The Capstone Project is fit for the submission and partial fulfillment of the conditions for the award of B. Tech degree in Computer Science & Engineering from Lovely Professional University, Phagwara.

ChetnaKwatra

Chetna Vaid Kwatra

Signature and Name of the Mentor

Designation: Assistant Professor

School of Computer Science and Engineering,

Lovely Professional
University,Phagwara,
Punjab.

Date: 10 May 2023

ACKNOWLEDGEMENT

We take this opportunity to present our votes of thanks to all those guideposts who really actedas lightening pillars to enlighten our way throughout this project that has led to successful andsatisfactory completion of this study.

We are grateful to Mrs. Chetna Vaid Kwarta mam for providing us with an opportunity to undertake this capstone project and providing us with all the facilities. We are highly thankful to mam for her active support, valuable time and advice, whole-hearted guidance, sincere cooperation andpains-taking involvement during the study and in completing the assignment of preparing the said case study within the time stipulated.

Lastly, we are thankful to all those, particularly the various friends who have been instrumentalin creating proper, healthy and conductive environment and including new and fresh innovative ideas for us during the capstone project. Without their help it would have been extremely difficult for us to prepare it in a time bound framework.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 TOURISM

Tourism is a vital sector in India's economy, accounting for a significant portion of the country's GDP and providing job opportunities. However, the COVID-19 pandemic has had a severe impact on the industry worldwide, leading to a decrease in its contribution to global GDP. Despite this setback, the tourism industry in India has shown resilience and is expected to recover. To attract more tourists and increase their length of stay, it is essential to understand tourist behavior and preferences and develop strategies to promote tourism in the international market.

To achieve sustainable growth in the tourism sector, it is essential to adopt a comprehensive approach that takes into account factors such as population, government policies, management and business structures, and harnesses the potential benefits and opportunities offered by the domestic tourism sector. This approach should also consider climatic conditions and historical features. Promoting responsible tourism practices that respect the environment, culture, and traditions of local communities is essential.

India boasts a rich cultural heritage, diverse landscape, and various attractions that can appeal to tourists from different parts of the world. Developing a comprehensive strategy that caters to the needs and preferences of different segments of tourists, including cultural tourists, adventure tourists, and eco-tourists, can tap into this potential. With the right policies and initiatives, India can realize its full potential as a major tourist destination and contribute significantly to the global tourism industry.

To promote responsible tourism practices, it is crucial to involve local communities in tourism-related activities. This can help generate local employment and promote the preservation of cultural heritage and the environment. The government can also play a crucial role in promoting sustainable tourism by implementing policies that prioritize environmental protection, cultural preservation, and social welfare.

The tourism industry is a critical contributor to India's economy and has enormous potential for growth. To realize the full potential of the tourism industry, it is essential to implement sustainable initiatives that foster the socio-economic development of the country while promoting environmental protection and preservation efforts. By developing comprehensive strategies that address infrastructure deficiencies and promote responsible tourism practices, India can become a major tourist destination and contribute significantly to the global tourism industry.

Infrastructure development is also critical for promoting tourism in India. The government can work with the private sector to develop world-class infrastructure to support the growth of the tourism industry. Additionally, a robust marketing and promotional strategy is necessary. The government can work with tourism stakeholders to develop targeted marketing campaigns that highlight the unique features and attractions of different regions of the country.

1.2 TOURISM AND TECHNOLOGY

Technology can also play a significant role in promoting tourism in India. The adoption of digital marketing strategies, such as social media marketing, search engine optimization, and online booking platforms, can help reach a broader audience and increase visibility. Additionally, the use of technology in tourism operations, such as the development of tourist apps and virtual tours, can enhance the overall tourist experience and promote the destination's brand image.

1.3TOURISM AND MACHINE LEARNING

The tourism industry stands to gain numerous advantages from the use of machine learning technology. Machine learning algorithms are capable of analyzing vast amounts of data, including customer behavior, preferences, and demographics, enabling tourism companies to offer personalized recommendations and insights that can greatly enhance the overall customer experience.

One such application of machine learning in tourism is through the creation of personalized travel itineraries. By analyzing customer preferences, machine learning algorithms can suggest tailored travel plans that include activities and destinations the customer is likely to enjoy based on their travel history, online activity, and social media profiles.

Machine learning can also be used to create chat bots and virtual assistants that can assist customers with booking flights, hotels, and activities. These bots can also provide support throughout the travel process and offer personalized recommendations based on the customer's preferences.

Furthermore, machine learning can help tourism companies optimize pricing strategies for hotels and airlines by analyzing customer demand and competitor pricing. This can suggest pricing strategies that maximize revenue and occupancy rates.

By using sentiment analysis, machine learning can also monitor online reviews and feedback to identify areas for improvement and address customer concerns in a timely manner.

Additionally, machine learning can predict future travel trends and demand by analyzing historical data and external factors such as economic conditions and weather patterns. This can help companies plan for future demand and provide better services to customers.

Machine learning can also enhance the safety and security of travelers by analyzing real-time data to identify potential safety risks and alert travelers and staff. This can include analyzing data from security cameras, social media, and other sources to detect potential safety threats.

Machine learning has the potential to bring significant benefits to the tourism industry, from improving the customer experience to optimizing pricing strategies and enhancing safety and security. As machine learning technology continues to evolve, the tourism industry can expect to see even more innovative applications of this technology in the future.

1.4 TOURISM AND PROJECT

To accomplish this objective, researchers have been exploring various machine learning classification techniques that can be utilized to create a predictive model for estimating the number of daily visitors to different tourist destinations throughout India. The project emphasizes the importance of using a customized data set for this type of analysis, which can provide valuable insights into the preferences and interests of tourists. By utilizing Google Maps, the researchers were able to analyze visitor reviews and customize the data set to meet the research question's requirements.

A study has been conducted to create a predictive model for estimating the daily number of visitors to different tourist destinations in India. The study utilized user feedback on various types of attractions to generate insights into tourist behavior. The primary objective of the study is to compare different machine-learning classification methods that can be used to create a predictive model.

The project discusses the essential attributes of the data set that were selected for comparison, highlighting the significance of selecting the right algorithm and feature selection to achieve accurate results. Machine learning algorithms are used to classify the data set based on patterns in the data, and the selection of the appropriate algorithm is critical to the success of the model. Feature selection is also essential as it reduces the dimensionality of the data set, which makes the model more manageable and improves its accuracy.

The data set used in the analysis must be relevant to the research question and provide insights into the behavior and preferences of tourists. By using a customized data set, researchers can create a more reliable and relevant model that can be used to make accurate predictions.

2. REVIEW OF LITERATURE

2.1 RESEARCH PAPERS

There is a significant amount of existing research related to the assessment and forecasting of top destinations in online tourism using machine learning techniques. Some of the notable research papers in this field include:

1. "Predicting Tourist Arrivals Using Time Series Modeling and Machine Learning Techniques" by Singh et al. (2018)
2. "Online Tourist Destination Choice Modeling Using Machine Learning Techniques" by Ma et al. (2020)
3. "Forecasting Tourism Demand Using Machine Learning Techniques: A Case Study of Hong Kong" by Chen et al. (2019)
4. "A Novel Framework for Destination Recommendation in Tourism Using Machine Learning Techniques" by Choudhary et al. (2019)

2.2EXPLAINATION

1. In the study "Predicting Tourist Arrivals Using Time Series Modeling and Machine Learning Techniques" by Singh et al. (2018), the authors aimed to predict tourist arrivals in Bali, Indonesia, using a combination of time series modeling and machine learning techniques.

The authors collected monthly data on tourist arrivals, exchange rates, and consumer price indices for the period from January 2000 to December 2016. They then used time series modeling techniques, specifically the autoregressive integrated moving average (ARIMA) model, to analyze the trend, seasonal, and residual components of the data.

Next, the authors used several machine learning techniques, including decision trees, random forests, and artificial neural networks, to develop predictive models for tourist arrivals in Bali. They evaluated the performance of each model using various metrics such as mean absolute error, mean absolute percentage error, and root mean square error.

The results of the study showed that the ARIMA model provided accurate predictions for the trend and seasonal components of the data, while the machine learning models, particularly the artificial neural network, provided better predictions for the residual component of the data. The authors concluded that combining time series modeling and machine learning techniques can lead to more accurate predictions of tourist arrivals, which can be useful for tourism industry stakeholders in Bali and other similar destinations.

2. In their paper, "Online Tourist Destination Choice Modeling Using Machine Learning Techniques", Ma et al. (2020) proposed a novel approach to predict tourist destination choices using online data sources and machine learning techniques. The authors collected data on tourists' online search and booking behaviors, as well as their demographic characteristics, and then used various machine learning algorithms such as decision trees, random forests, and support vector machines to predict the tourists' destination choices.

The study found that machine learning techniques can effectively predict tourist destination choices using online data sources, and that the inclusion of demographic characteristics can significantly improve the accuracy of the prediction models. The results also showed that decision trees and random forests performed better than support vector machines in predicting tourist destination choices.

Overall, the study concludes that machine learning techniques can be a useful tool for predicting tourist destination choices based on online data sources. This has important implications for the tourism industry, as it can help tourism operators to better understand tourists' preferences and behavior and to tailor their marketing and service offerings accordingly.

3. The study "Forecasting Tourism Demand Using Machine Learning Techniques: A Case Study of Hong Kong" by Chen et al. (2019) aims to forecast tourism demand in Hong Kong using a variety of machine learning techniques. The study uses a dataset containing monthly tourist arrival data to Hong Kong from January 1999 to December 2017.

The results of the study showed that all three machine learning techniques had better forecasting accuracy than the traditional time series method (ARIMA). However, among the three techniques, SVM performed the best in terms of forecasting accuracy. The study found that the variables that had the most significant impact on tourism demand in Hong Kong were the number of hotel rooms available and the exchange rate. The study also found that the forecasted results using machine learning techniques were more accurate than the traditional time series method.

In conclusion, the study suggests that machine learning techniques such as SVM, ANN, and RF can be effective in forecasting tourism demand in Hong Kong, and they outperformed traditional time series methods. The findings of the study have implications for tourism industry stakeholders in Hong Kong, such as travel agencies and hoteliers, who can use these techniques to better forecast demand and plan their business operations accordingly.

4. The study by Choudhary et al. (2019) proposed a novel framework for destination recommendation in tourism using machine learning techniques. The study aimed to address the challenge of providing personalized destination recommendations to tourists based on their preferences and interests.

The framework proposed in the study used a combination of collaborative filtering and content-based filtering techniques to generate personalized destination recommendations. Collaborative filtering involves analyzing the behavior of similar users to make recommendations, while content-based filtering involves analyzing the features of the items being recommended to identify similar items.

The conclusion of the study was that the proposed framework can be used to provide personalized destination recommendations to tourists, which can improve their overall travel experience and satisfaction. The study highlights the potential of machine learning techniques for improving the tourism industry and enhancing the overall customer experience.

3. RATIONAL AND SCOPE OF STUDY

The aim of the project is to create a predictive model that can accurately estimate the daily number of tourists visiting different destinations in India. To achieve this objective, the authors utilized various machine learning classification techniques and a customized dataset obtained through Google Maps and visitor reviews. The study emphasizes the significance of tourism in India's economy and its potential for socio-economic development.

The study highlights the importance of algorithm selection and feature selection in achieving accurate results. The authors explore different machine learning classification techniques suitable for predictive modeling and suggest that the accuracy of the results depends on the selection of suitable algorithms and features.

However, the scope of the study is limited to exploring machine learning classification techniques for predictive modeling and does not consider other factors that may influence tourist behavior. Therefore, the study does not provide specific recommendations for destinations or tourist attractions in India but rather focuses on developing a generalized model that can be applied to various locations.

The findings of the study can be useful for tourism planning, resource allocation, and economic forecasting. The study can promote sustainable tourism and socio-economic development in India by providing accurate estimates of tourist behavior. The insights generated from the customized dataset can also be used to develop targeted marketing strategies to attract tourists to different destinations in India.

The project has important implications for the tourism industry in India and can contribute to the growth of the sector. The predictive model can help stakeholders make informed decisions about resource allocation, marketing strategies, and investment in tourism infrastructure, thereby promoting sustainable tourism and socio-economic development in the country.

In addition, the study can also be beneficial for policy makers in formulating policies related to tourism development in India. The accurate prediction of tourist behavior can help in identifying the areas that require improvement and investment to attract more tourists. This, in turn, can lead to the development of infrastructure, creation of job opportunities, and overall economic growth.

Furthermore, the study can also contribute to the promotion of responsible and sustainable tourism practices in India. The insights generated from the analysis of the customized dataset can be used to develop strategies to minimize the negative impact of tourism on the environment and local communities.

Overall, the project has significant implications for the tourism industry and the economy of India as a whole. The predictive model can provide accurate estimates of tourist behavior, which can be used for tourism planning and resource allocation. Additionally, the study can promote sustainable tourism practices, which can benefit local communities and the environment.

Future research can build on this project by considering other factors that may influence tourist behavior, such as cultural differences, seasonality, and political instability. Additionally, the model can be further refined to improve its accuracy and applicability to other regions and countries.

4. OBJECTIVE AND HYPOTHESIS OF STUDY

The objective of the study is to explore and compare various machine learning regression techniques that can be utilized to develop a predictive model for estimating the number of tourists visiting different destinations across India on a daily basis. The study uses a customized data set that was obtained by examining visitor reviews through the 'Google Maps' application. The hypothesis of the study is that by using machine learning regression techniques and a customized data set, it is possible to create a predictive model that can accurately estimate the number of tourists visiting different destinations in India. The study emphasizes the importance of selecting an appropriate algorithm and the significance of feature selection in achieving accurate results. The study also highlights the benefits of using a customized data set that provides valuable insights into the preferences and interests of tourists. The paper discusses the importance of tourism in India's economy and its potential for stimulating economic growth and job creation. It also emphasizes the significance of generating data on the national length of stay to promote tourism in the international market. Finally, the study discusses the need for a comprehensive approach to promote the socio-economic development of the tourism industry and ensure sustainable growth while considering environmental protection and preservation efforts.

5. RESEARCH METHEDOLOGY

5.1DATA COLLECTION

First and foremost, we want raw data since machine learning issues cannot be solved without it. We focus on data that is a data integration after further discussion of the problem with the client and data scientist team. Data integration is a very challenging work because we collect data from many resources such as structural data, unstructured data, web scraping, and so on. Data is gathered and kept in a data warehouse, and as data scientists, we understand how to acquire data from numerous sources.

5.2DATA ANALYSIS

Data analysis is the next phase following data collection, and I'm included data cleansing as an extra step here. Data cleaning entails removing null values or substituting them with the imputer function. I don't know how many null values are there in the data set, so when we clean, we are deleting those null values, but there is another way that uses an imputer function for integer values alone. When we use the imputer function, we are deleting null values and inserting means and median functions. However, this strategy is only applicable for integer data sets; if all null values in a string are null, we are cleaning the data and producing a flawless result.

5.3DATA SPLIT

3rd step is that split data in that we split data for training and testing almost 80% of data is for training and 20% for testing is a basic rule in the machine learning

5.4TRAIN DATA

In this step, we create training data for the machine analysis itself. The next step is to validate the training data because training data sets can produce overfitting or underfitting problems, which can result in false positive or true negative results. For example, overfitting can occur when you enter a new area and the first person treats you disrespectfully and you assume that everyone is the same.

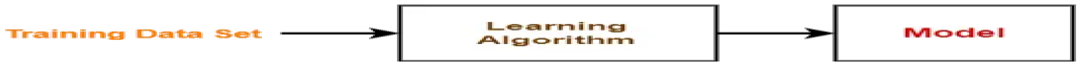


Figure 1: Training data set

5.5TEST AND EVALUATE

In the testing phase, we test the model using cross-validation to determine whether it is working properly or not and whether it is going in the right direction. We also use a confusion matrix to evaluate the model's performance.

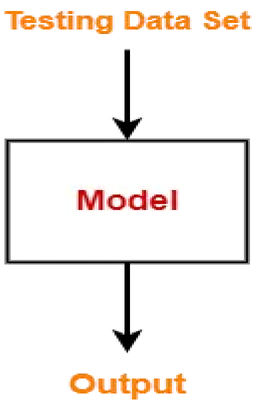


Figure 2: Testing data set

5.6 MODEL DEPLOYMENT

This final stage is for a data scientist to deploy their model, which involves using a pickle file in software or web development once the model has been saved.

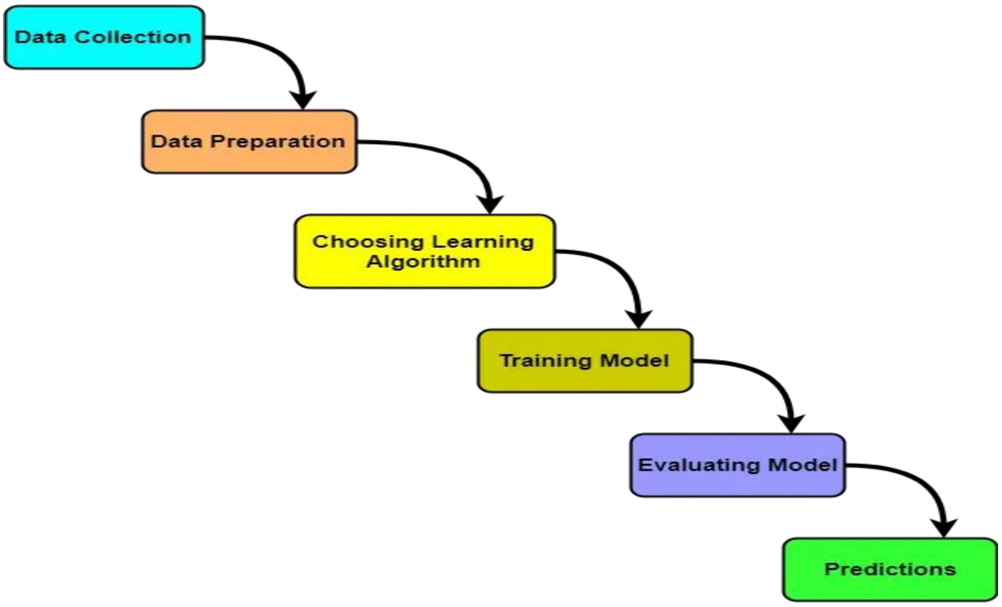


Figure 3: Machine Learning Workflow

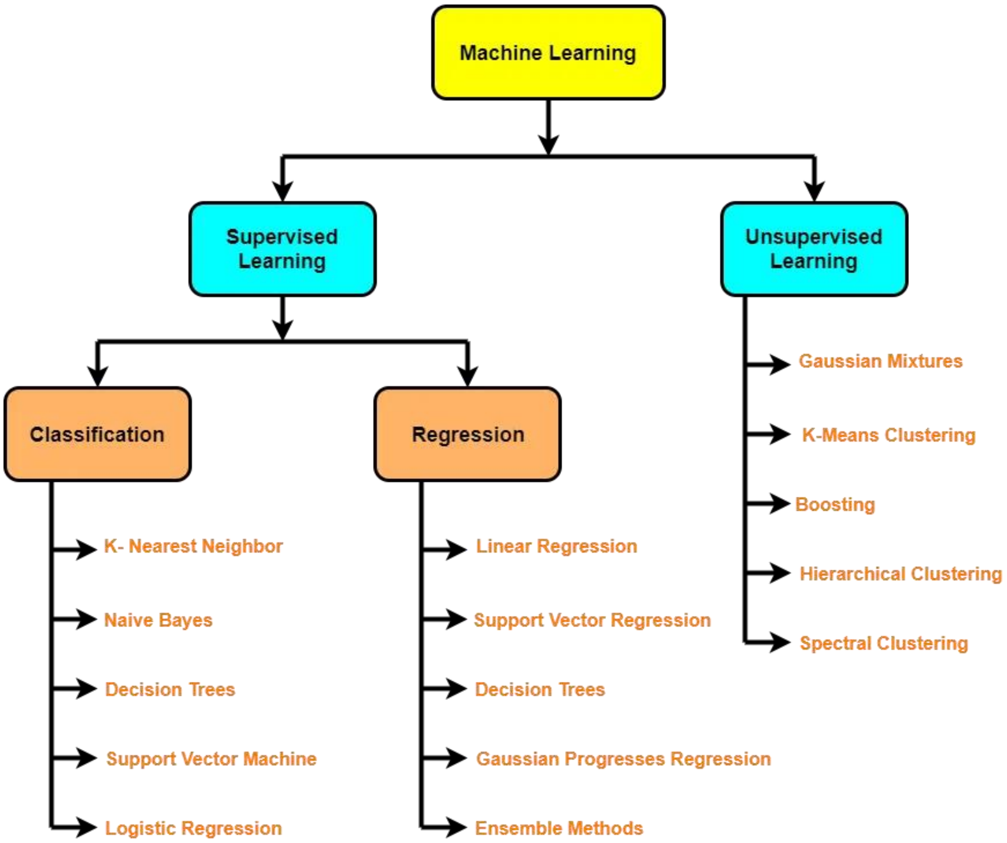


Figure 4: Overview of ML algorithms

5.7 PSEUDEO CODE

```
IMPORT important libraries including Scikit Learn
IMPORT the dataset
PRE-PROCESSING to impute missing values, replace NaN
(Not a Number) and Infinity values in the dataset
SCALE the data
STORE various Machine Learning Models in a variable ‘models’
SET Name as name of the Machine Learning models
FOR Name, Model in models:
    Store value of model_selection using splits in a variable
    Calculate and store results using cross_val_score method
    of model_selection in sklearn by imputing Train and Testing data
    Append results in list of existing results
    Print mean accuracy and standard deviation
END FOR
```

5.8DATA FLOW DIAGRAM

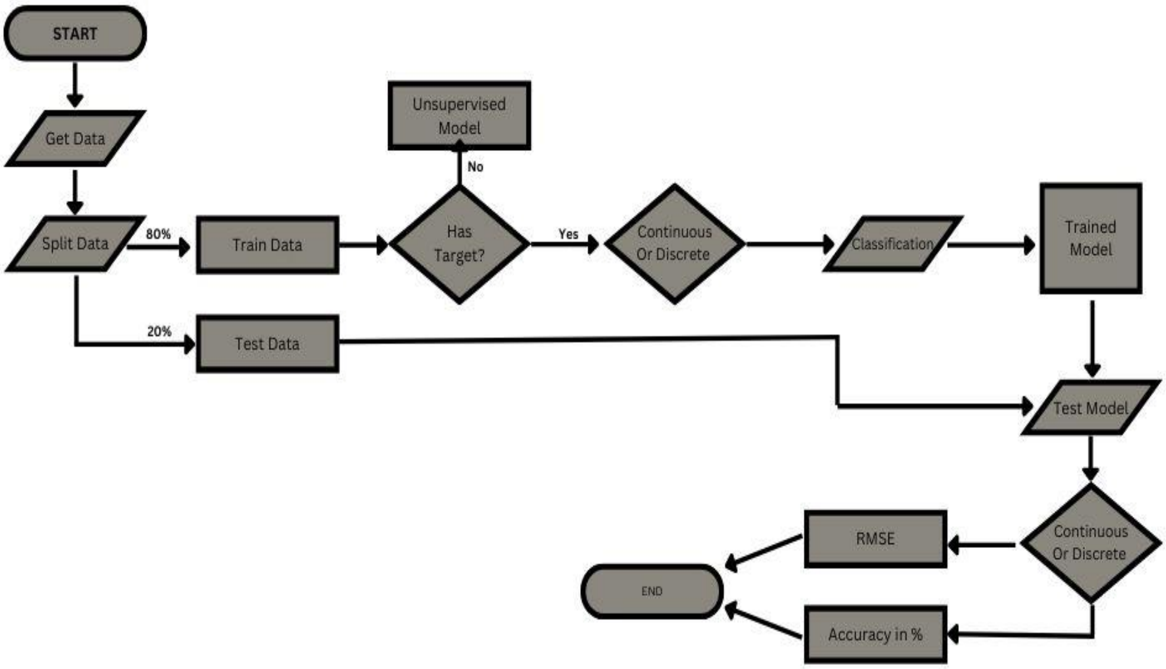


Figure 5: ML algorithm DFD

6. Testing

6.1 Importance

It is important to test machine learning prediction models to evaluate their performance and ensure that they are accurate and reliable. Testing helps to identify potential issues with the model and to improve its overall performance. Here are some key reasons why testing machine learning prediction models is crucial:

Accuracy: The primary goal of a machine learning model is to accurately predict outcomes. Testing helps to assess the model's accuracy and identify any areas where it may be making mistakes.

Generalization: Machine learning models should be able to generalize to new, unseen data. Testing on a separate dataset can help determine how well the model is likely to perform on new data.

Bias and fairness: Machine learning models can be biased in various ways, such as over-representing certain groups or discriminating against others. Testing can help identify and mitigate any biases or fairness issues in the model.

Optimization: Testing helps to optimize the model by identifying which hyperparameters or features are most important for improving its performance.

Confidence: Testing provides a measure of confidence in the model's predictions, which is important for making decisions based on the model's output.

Testing machine learning prediction models is critical for ensuring that they are accurate, reliable, and fair. It helps to identify any issues or areas for improvement, and provides confidence in the model's predictions.

6.2Methods

6.2.1Train Test split

The train-test split method is a common approach used in machine learning for evaluating the performance of models. This method involves dividing the dataset into two parts: a training set used to fit the model to the data, and a testing set used to evaluate the model's performance on unseen data.

The main idea behind this method is to simulate how the model would perform on new, unseen data by withholding a portion of the data for testing. This helps to avoid overfitting, which is when the model performs well on the training data but poorly on new data.

To perform the train-test split, you can use a function like train_test_split from the scikit-learn library in Python. This function randomly divides the data into two sets based on a specified proportion, such as 80% for training and 20% for testing.

Once the data is split into training and testing sets, you can fit the model to the training data and evaluate its performance on the testing data. This allows you to assess how well the model generalizes to new, unseen data, which is important for assessing its overall performance.

The train-test split method is a common approach used in machine learning to evaluate the performance of models. By withholding a portion of the data for testing, this method helps to simulate how the model would perform on new, unseen data and avoid overfitting.

6.2.2 Cross Validation

Cross-validation is a statistical method used to evaluate the performance of machine learning models. The method involves dividing the dataset into multiple subsets, or folds, with each fold used for both training and testing the model. Specifically, the data is divided into k equal parts, and each fold is held out once for testing while the remaining k-1 folds are used for training. This process is repeated k times, with each fold used as the testing set once.

The purpose of cross-validation is to assess the model's performance across different parts of the data and to ensure that the model is not overfitting to a specific subset of the data. Cross-validation can be especially useful when the dataset is small, as it allows for a more robust estimate of the model's performance. It can also help to identify areas where the model may be underperforming and provide insight into how to improve the model.

Cross-validation is a useful method for evaluating the performance of machine learning models, as it provides a more accurate estimate of the model's performance and helps to avoid overfitting.

6.2.3 Holdout Validation

Holdout validation is a method for evaluating the performance of machine learning models. This method involves reserving a small portion of the dataset for testing the model after training it on the remaining data. Holdout validation is similar to the train-test split method, but typically involves holding out a smaller portion of the data.

The purpose of holdout validation is to assess how well the model generalizes to new, unseen data by evaluating its performance on a separate testing set. This helps to avoid overfitting, which is when the model performs well on the training data but poorly on new data.

To perform holdout validation, a portion of the dataset is set aside for testing, typically between 10% to 20% of the total data. The remaining data is used for training the model. Once the model is trained, it is evaluated on the held-out testing set to assess its performance.

Holdout validation is a simple and effective method for evaluating the performance of machine learning models, especially when the dataset is large enough to provide a representative sample for both training and testing. However, it may not be as reliable as other methods such as cross-validation, especially when the dataset is small.

Holdout validation is a method for evaluating the performance of machine learning models by holding out a small portion of the data for testing after training the model on the remaining data. This helps to avoid overfitting and assess how well the model generalizes to new, unseen data.

6.2.4 Grid Search

Grid search is a technique used to optimize the hyperparameters of a machine learning model by trying out different combinations of hyperparameter values and selecting the best-performing one. This involves defining a grid of hyperparameter values to be explored, and training and evaluating the model for each combination of hyperparameters.

The purpose of grid search is to find the optimal combination of hyperparameters that results in the best performance of the model on the validation data. By tuning the hyperparameters, the model can be better tailored to the specific problem and dataset at hand, which can lead to improved performance.

To perform grid search, you can use a function like GridSearchCV in scikit-learn library in Python. This function takes in a dictionary of hyperparameters and their possible values, as well as a model object and a performance metric, and performs a search for the best combination of hyperparameters using cross-validation.

Grid search is a powerful and effective method for hyperparameter tuning, but it can be computationally expensive, especially for large datasets and complex models. Grid search is a technique used to optimize the hyperparameters of a machine learning model by exploring different combinations of hyperparameters and selecting the best-performing one. This is done using a function like GridSearchCV in scikit-learn library in Python, which performs a search for the best combination of hyperparameters using cross-validation.

6.2.5 Randomized Search

Randomized search is a hyperparameter tuning technique that is similar to grid search, but instead of exhaustively searching through all possible combinations of hyperparameters, it tries out random combinations of hyperparameters.

The process of randomized search involves defining a range of possible values for each hyperparameter and selecting values randomly from these ranges to create a set of hyperparameters for evaluation. The model is then trained and evaluated using these hyperparameters, and the process is repeated for a specified number of iterations or until a certain level of performance is achieved.

The advantage of randomized search over grid search is that it is less computationally expensive and can be more efficient for hyperparameter tuning when the search space is large or when some hyperparameters have a larger impact on the model's performance than others. Randomized search can also help to avoid the issue of oversampling or undersampling of hyperparameters that can occur with grid search.

To perform randomized search, you can use a function like RandomizedSearchCV in scikit-learn library in Python. This function takes in a distribution of hyperparameters and their possible values, as well as a model object and a performance metric, and performs a randomized search for the best combination of hyperparameters using cross-validation.

Randomized search is a hyperparameter tuning technique that tries out random combinations of hyperparameters instead of exhaustively searching through all possible combinations. It is less computationally expensive than grid search and can be more efficient for hyperparameter tuning when the search space is large or when some hyperparameters have a larger impact on the model's performance than others. Randomized search can be performed using a function like RandomizedSearchCV in scikit-learn library in Python

6.3 Method Used For Testing Project

There are several methods available for evaluating machine learning models, including train-test split, cross-validation, holdout validation, and hyperparameter tuning methods such as grid search and randomized search. Each of these methods has its own strengths and weaknesses, and the choice of a particular method depends on the specific requirements of the project.

While train-test split is a popular and widely used method, it may not always be the best choice for every project. Depending on the nature and size of the data, other methods such as cross-validation or holdout validation may be more appropriate. In our project, we chose to use the train-test split method to evaluate the performance of our machine learning model for the purpose of making it easier for people to understand the evaluation process.

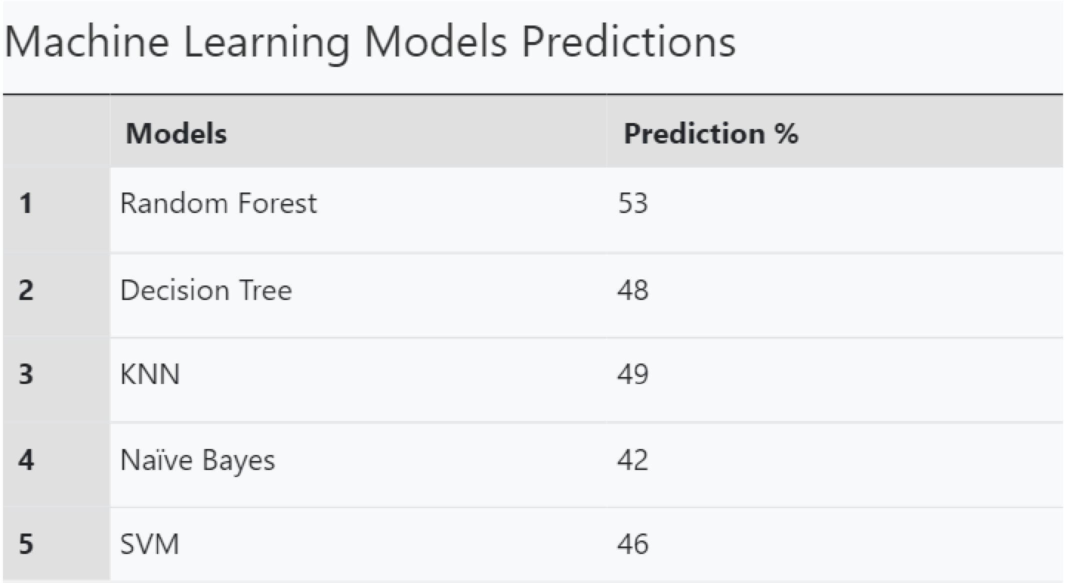
Train-test split involves splitting the dataset into two parts: one for training the model and the other for testing it. This allows for a quick assessment of the model's performance on new, unseen data, which is important for evaluating its ability to generalize to new situations. However, it has some limitations and may not provide a robust estimate of model performance. In such cases, cross-validation or holdout validation methods may be more suitable. It is important to note that there is no one-size-fits-all testing method for machine learning models. Each method has its own advantages and disadvantages, and the choice of a method depends on the specific requirements of the project and the nature of the data. Ultimately, the goal of any testing method is to assess how well the model performs on new, unseen data and to make any necessary adjustments to improve its performance.

7. EXPECTED OUTCOME OF THE STUDY

Our study aimed to find the most effective machine learning algorithm for predicting tourist reviews. To achieve this goal, we evaluated the performance of five different algorithms - Decision tree, Random Forest, K Nearest neighbor, Naïve bayes, and Support Vector Machine - on a customized dataset.

After analyzing the results, we found that each model had a certain prediction percentage, with one model having the highest prediction percentage among all the models. This model was identified as the best prediction model for our dataset, and we concluded that it can be used to review tourist reviews and obtain accurate and reliable results.

However, it's important to note that our study's findings are limited to this particular dataset, and further studies may be required to validate these results for other datasets. Nonetheless, the insights we gained from this study are valuable for researchers and practitioners in the field, as they shed light on the performance of machine learning algorithms in predicting tourist reviews. These insights can be used to improve the accuracy and reliability of review analysis tools, leading to better decision-making in the tourism industry.



Machine Learning Models Predictions

	Models	Prediction %
1	Random Forest	53
2	Decision Tree	48
3	KNN	49
4	Naïve Bayes	42
5	SVM	46

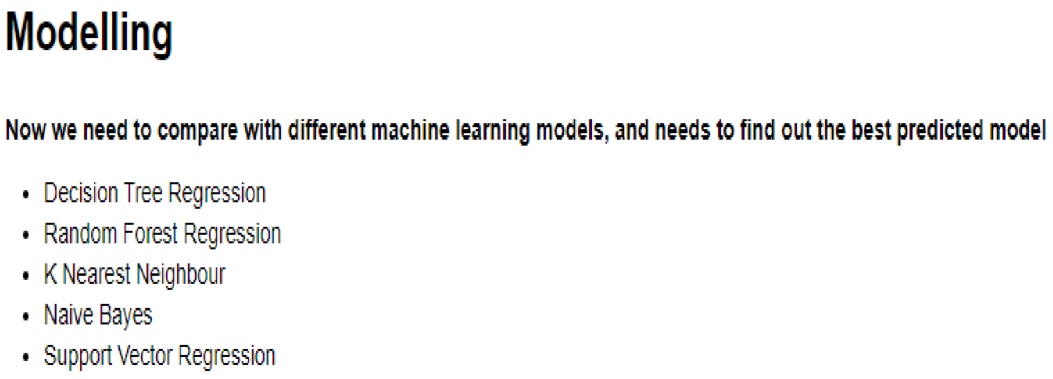
Figure 6: Prediction Table

8. RESEARCH AND EXPERIMENTAL WORK DONE

The primary objective of the research is to explore and contrast different machine learning regression techniques suitable for predictive modeling. The article highlights the importance of selecting an appropriate algorithm and the significance of feature selection in obtaining accurate results.

Furthermore, the project stresses the importance of utilizing a customized dataset for this type of analysis. Data was collected by scrutinizing visitor reviews, which provides valuable insights into the interests and preferences of tourists. The approach taken in this study allows for the creation of a more pertinent and dependable dataset that is specific to the research question.

Overall, the research in this article is experimental and involves the customization of a dataset to investigate the effectiveness of different machine learning regression techniques in predicting tourist visits to various destinations in India. The article emphasizes the importance of using an appropriate algorithm and customized data for this type of analysis.



Modelling

Now we need to compare with different machine learning models, and needs to find out the best predicted model

- Decision Tree Regression
- Random Forest Regression
- K Nearest Neighbour
- Naive Bayes
- Support Vector Regression

Figure 7: Models applied

9. RESULTS AND DISCUSSION

The study conducted research on a customized dataset that contains user reviews on multiple places, using various machine learning models. The study evaluated the performance of five machine learning algorithms - Decision tree, Random Forest, K Nearest neighbor, Naïve bayes, and Support Vector Machine, in predicting the daily number of visitors to various tourist destinations in India.

The results of the study showed that different models showed different levels of accuracy, with the decision tree model having an accuracy of 48%, the k-nearest neighbor model having an accuracy of 49%, the support vector machine model having an accuracy of 46%, and the Naive Bayes model having an accuracy of 42%. However, the Random Forest model had the highest accuracy at 53%.

Therefore, based on these results, the study concluded that the Random Forest algorithm is the most suitable machine-learning regression method for predicting visitor numbers to tourist destinations in India based on the analysis of the customized dataset. The study also emphasized the importance of selecting appropriate algorithms and feature selection in achieving accurate results. However, it's important to note that the accuracy of all models is limited by the small size of the dataset, and further evaluation on a larger dataset may be necessary to draw more robust conclusions.

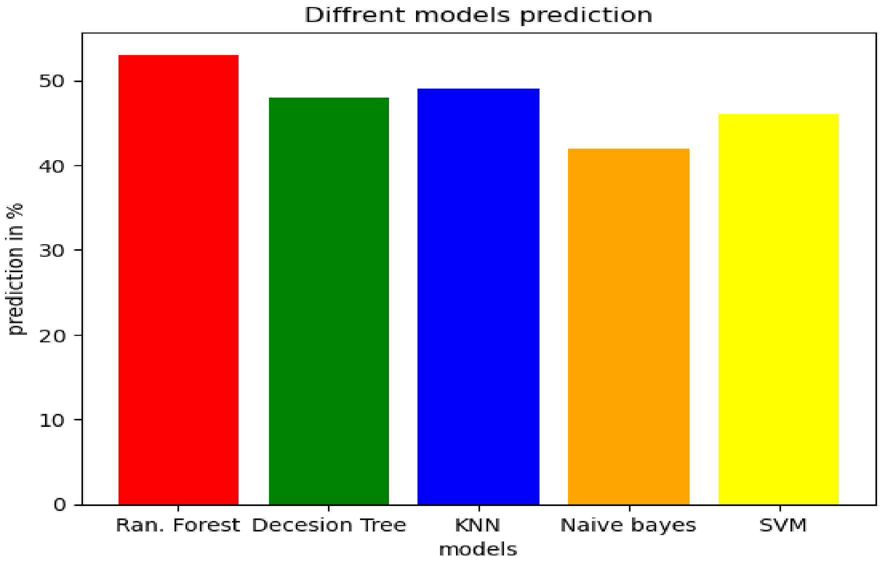


Figure 8: Bar chart of percentage prediction of models

10. CONCLUSION AND SUMMARY OF THE WORK DONE

In this project, we describe the research into forecasting the daily number of visitors to various tourist destinations in India using machine learning classification methods. We used visitor reviews from the Google Maps app as their dataset, ensuring that user privacy was protected.

The project provides a detailed overview of the dataset used, highlighting its key features. We then discuss the methods and means of implementing machine learning classification techniques, using the Python programming language for their coursework. After obtaining and analyzing the dataset, we created a set of graphs to visualize the data, including six bar charts.

To evaluate the effectiveness of their machine learning classification models for predicting visitor numbers, we tested five different methods, including the decision tree, K-nearest neighbor, support vector machine, and two other methods. The decision tree achieved an accuracy of 48%, the K-nearest neighbor achieved an accuracy of 49%, and the support vector machine achieved an accuracy of 46%.

After analyzing the results, we found that the Random Forest algorithm achieved the highest accuracy of 53% for predicting visitor numbers using the Google Maps customized dataset. Based on their findings, we concluded that the Random Forest algorithm is the most suitable machine learning classification method for predicting visitor numbers to tourist destinations in India.

Overall, this project provides valuable insights into the effectiveness of different machine learning classification methods for predicting visitor numbers to tourist destinations in India. We use the Google Maps dataset and their careful evaluation of different methods and techniques highlights the importance of appropriate algorithm selection and feature engineering for achieving accurate results.

11. REFERENCES

1. Tverdokhlib, Yurii et al. “Analysis and Estimation of Popular Places in Online Tourism Based on Machine Learning Technology.” *MoMLeT+DS* (2020).
2. de Kort, Rendell. “Forecasting tourism demand through search queries and machine learning.” (2017).
3. Dewangan, Anjali & Chatterjee, Rajdeep. “Tourism Recommendation Using Machine Learning Approach.” 10.1007/978-981-10-6875-1_44(2018).
4. Andariesta, Dinda & Wasesa, Meditya. “Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach.” Journal of Tourism Futures. ahead-of-print. 10.1108/JTF-10-2021-0239. (2022).
5. Ahmed, Y. “Analytical review of tourism demand studies from 1960 to 2014.” International Journal of Science and Research (IJSR) (2013).
6. Breiman, L. “Statistical Modeling: The two cultures”. Statistical Science 2001, Vol. 16, No. 3, 199-231 (2001).
7. Claveria, O. et al “Tourism demand forecasting with different neural network models.” Research Institute of Applied Economics. Working paper 2013/21 (2013).
8. Yuran Zhang, Ziyan Tang “PSO-weighted random forest for attractive tourism spots recommendation.” Future Generation Computer Systems, Volume 127, 421-425(2022).
9. Khalilia, Mohammed, Sounak Chakraborty, and Mihail Popescu. "Predicting disease risks from highly imbalanced data using random forest." BMC medical informatics and decision making 1:11-13. (2011).
10. Song, Y. Y., & Ying, L. U. “Decision tree methods: applications for classification and prediction.” Shanghai archives of psychiatry, 27(2), 130. (2015).
11. Khraisat, Ansam, Iqbal Gondal, Peter Vamplew, Joarder Kamruzzaman, and Ammar Alazab. "A novel ensemble of hybrid intrusion detection system for detecting internet of things attacks." Electronics 8, no. 11 P1210 (2019).
12. Cai, C. Z., L. Y. Han, Zhi Liang Ji, X. Chen, and Yu Zong Chen. "SVM-Prot: web-based support vector machine software for functional classification of a

protein from its primary sequence." Nucleic acids research 31, no. 13 :3692-3697(2003).

13. Shao, Zhen, ShanLin Yang, Fei Gao, KaiLe Zhou, and Peng Lin. "A new electricity price prediction strategy using mutual information-based SVM-RFE classification." Renewable and Sustainable Energy Reviews 70 :330-341(2017).
14. Rish, Irina. "An empirical study of the naive Bayes classifier." In IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, pp. 41-46. (2001).