Tourist spending prediction using machine-learning techniques: A real case study from the international visitors to London

Abstract. *In the tourism sector, which makes a significant contribution to the country's economy, it is considered substantial to analyze tourist behaviors accurately and effectively in order to achieve greater tourist satisfaction. In the same manner, tourists' spending/expenditure prediction is crucial in conducting revenue and resource management for governments and organizations.*
The main purpose of this study is to provide a mechanism that will predict spending levels, which can help the organizations (governments, travel agencies, etc.) regarding the offers of additional services and excursions depending on different profiles of tourists. For this purpose, this paper provides three techniques for the classification and prediction including k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree whereas KNN, is the most suitable model to do the prediction on spending. The advantages of these methods are much faster, less computationally demanding, easier to interpret, and high accuracy. The dataset used in this paper was taken from the database of the Office for National Statistics (ONS) as a set of the number of international visitors to London data from different countries in a specific time period, and the corresponding tourists' information. We use variables as a predictor to train and test the prediction models based on three machine learning algorithms. After preprocessing the raw data by standardizing, dealing with missing data, and scaling, we conduct the prediction and compare each model through three metrics as a confusion matrix, accuracy score, and F1-score. The result shows that KNN has best performance in predicting expenditures because it has the greatest number of correct prediction samples and the highest accuracy score. The findings of this research will aid the tourism sector in developing strategies for financial planning of organizations.

Keywords: Tourism industry, Machine learning algorithms, k-Nearest Neighbor, Support Vector Machine, Decision Tree, Data Analysis

1. INTRODUCTION

The tourism industry, also known as the travel industry, is linked to the idea of the movement of people to other countries or places outside their usual environment, either domestically or internationally, for social, leisure, or business purposes (UNWTO, 2008). This industry is one of the main sectors of the world economy, which makes significant contributions to both the national and global economy, and the economies of many nations are driven, to a large extent, by their tourist trade. In addition to the significant share of cash to the formation of countries' GDP, the sector contributes to the increase in the number of additional jobs and the formation of a positive image by creating new employment areas and new opportunities. Due to the development of tourism, it is known that many travel activities such as festivals, cultural trips, business trips (e.g., congresses, meetings, etc.), fairs, sports, and entertainment are very significant in terms of social and economic development, and region and country-oriented promotion.

Since the economic impact of tourism flows is generally essential for nations, understanding how effectively the industry is functioning and identifying trends in the industry is critical (Disegna et al., 2016). That's why using methods of data analysis and statistical information are needed to determine the state of the industry, dependencies, and the predicted values of indicators. By choosing appropriate data and analysis tools to observe and improve the economic effects of tourism visits, the determinants of tourism expenditures should be examined, and also the spending behavior of tourists should be analyzed in depth (Hung et al., 2012).

In this regard, the purpose of this study is to examine the international tourists' behaviors in a general sense and draw outcomes from their tourism activity and especially tourist spendings by using some of the statistical data analysis methods. With these analysis results, it is aimed to improve the best marketing strategies to be able to attract more tourists, and as a result, to increase revenue generated from tourism. Also, the decision-making process of government and private organizations regarding new investment opportunities for the tourism sector can be assisted.

The paper starts by providing an overview of the tourism and travel industry. The rest of the paper is organized as follows: In Section 2, the relevant literature is briefly reviewed, and this literature review is mainly in two parts except for the tourism industry itself: literature on data analytics techniques that are used in this paper considering other industries, and literature on the applications of statistical data analysis techniques in the tourism industry. Next, the details about the data set used in the application part of the study, and the methods used in data analysis are explained in Section 3. The analysis and interpretation of data are given in Section 4. In the final section, the paper is summarized, and some implications are discussed and suggested for further research and real-life applications.

2. LITERATURE REVIEW

Tourism has become a sector whose socio-cultural and economic importance is increasing and developing over the past decades. Tourism, one of the fastest-growing industries, is a key force for the economy of many developed and developing countries (Manzoor et al., 2019). In parallel with the increase in the level of civilization, the development of international trade, and the increase in living standards, people have started to travel more today. As a natural result of these increasing travel trends, tourism has increased by spreading over a wider area geographically and has become a phenomenon experienced by the whole world (Alaeddinoğlu and Can, 2007). There is no doubt that the tourism sector, which is one of the building blocks of the service sector, and the travel sector are two concepts that cannot be considered separately from each other, and they have a major proportion in stimulating the economy of a country. Therefore, increases in tourism flows bring positive economic outcomes for countries, especially considering Gross Domestic Product (GDP) and new employment opportunities. In this regard, it is extremely crucial to analyze the travel activities and behaviors of tourists. In parallel with the developments in the sector in recent years, the analysis and research in this field have risen and the tourism sector has been the subject of many studies.

This study aims to analyze tourists' behaviors on spending by using k-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Decision Trees classification methods and make meaningful inferences considering the analysis results.

KNN provides the calculation of the similarity of the data to be classified to the normal behavior data in the learning set; With the average of the k data, which are thought to be the closest, they are assigned to the classes according to the determined threshold value. The K value allows us to determine how much data and the closeness of the data to be classified will be measured (Xing and Bei, 2019; Zhang, 2020). KNN method can be applied to different fields. Xing and Bei (2019) used improved KNN in order to classify the medical health big data and compared it with the traditional KNN algorithm. Harrou et al. (2020) used KNN to detect road traffic congestion for improving safety and traffic management.

SVM is one of the supervised machine learning methods used in classification. In the algorithm, a line is drawn to separate the points placed on a plane. This line is intended to be at the maximum distance for the points of both classes (Chauhan, 2019). In literature, it is used for different purposes. Huang et al. (2018) used SVM as a classification tool for cancer genomics to discover new biomarkers and drugs and understand the cancer driver genes in future research. On the other hand, Liu et al. (2020) developed an SVM model to forecast short-term wind speed by using the Jaya optimization algorithm as well.

Decision trees, which can be easily integrated with other systems that are easily interpreted, and which are reliable, are the widely used classification algorithm in the literature. When we look at the structure of decision trees, they consist of roots, branches, and leaves. Decision trees that start with the root node divide data sets into small groups and branches as they go down. Each node is divided into two or more branches, and if there is no new problem after the node, the branching ends (Song and Ying, 2015; Somvanshi et al., 2016). Decision trees also have a wide range of application areas like other methods. Rizvi et al. (2019) used decision trees with the aim of exploring the impact of six demographic characteristics on academic outcomes in the online learning environment in the United Kingdom. Kavzaoğlu and Çölkesen (2010) developed a decision tree algorithm in order to classify the satellite images.

KNN, SVM, and Decision Trees which have already been used by a number of scientific research are usable also in the tourism and travel industry. When the studies in the literature are examined, there are studies that are used data analysis methods for the classification of tourists considering their lifestyles, analyzing tourist behavior, clustering of tourist destinations, analyzing the decision-making processes of travelers, analyzing and segmentation travel and tourism activities for tourism growth, increasing the efficiency of hotels and travel agencies, and many other areas.

The studies conducted in the tourism, travel, and hospitality industry in the last 5 years have been examined and some studies and the analysis methods used are given in the table below.

Table 1. Summary of some studies using data analysis methods

Reference	**Research Subject**	*Regression Analysis*	*Correlation Analysis*	*Cluster Analysis*	*Factor Analysis*	*SVM*	*KNN*	*Decision Trees*
Iswandhani and Muhajir (2018)	Classify the popularity of tourist destinations based on the number of likes from the most popular instagram account in the region.			✓				
Kür and Topal (2018)	Determine Chinese tourists Turkey preference's potential.						✓	
Ntaliakouras et al. (2019)	Forecasts for tourism demand in Greece by considering the contribution of explanatory variables.							✓
Han (2020)	Predicts the market turnover to provide quantitative analysis of environmental resources in tourism development					✓		
Wahyono et al. (2020)	Develop a smart map to determine tourist attractions based on tourist desires						✓	
Lusia and Alamsyah (2021)	Investigate the significance of destination attributes to travel decisions in the new normal	✓						
Wang et al. (2021)	Identify the choice of residents to tourism destination			✓				
Thi (2022)	Understand guests' demands considering the service experience and the service quality in the luxury hotel segment				✓			
Paramita et al. (2022)	Analyze the impact of bookings through offline travel agencies to increase room occupancy in the hotel	✓	✓					

3. DATA SET AND METHODS

3.1. Description of Dataset

The data set analyzed in this article and the variables used will be explained in this section.

The data set was gathered in 2019 from a survey with the purpose of collecting data and turning it into valuable information regarding tourism in London. This survey is provided by the Office for National Statistics (ONS). As mentioned earlier in the literature review, tourism has become the biggest source of income for many cities and even countries and a good example of this is the city of London.

In the dataset, the only consideration for the year is 2019 which is shown in the first column. The following column indicates in which quarter of the year the person surveyed visited London. Other columns show where the surveyed people come from, the way of arrival to London, the purpose of the visit, the total number of nights spent, and the total number of visits, respectively. As a target variable, total expenditure is used to categorize tourists by their spending levels.

Table 2. Variables and Descriptions

Variable	**Description**	**Labels**
Year	Year of interview	2019
Quarter	Period of Year	January – March April – June July – September October - December
Market	Country of provenance of the respondent	Argentina Australia Bahrain Chile China Egypt Hong Kong India Indonesia Israel Japan Kenya Kuwait Malaysia New Zealand Belgium Irish Republic Iceland Luxembourg Spain Romania Germany Russia Portugal Hungary Sweden Mexico Czech Republic Italy Turkey

		Nigeria	Canada
		Oman	Poland
		Other Africa	Brazil
		Other Asia	Austria
		Central&South America	France
		Other Middle East	Bulgaria
		Other Southern Africa	USA
		Pakistan	Western Europe
		Qatar	Serbia
		Saudi Arabia	Norway
		Singapore	Denmark
		South Africa	Eastern Europe
		South Korea	Finland
		Taiwan	Greece
		Thailand	Netherlands
		United Arab Emirates	Switzerland
Mode	Main method of travel	Air Sea Tunnel	
Purpose	Main purpose of visit	Holiday Business Study VFR (visit friends or relatives) Miscellaneous	
Visits	Number of total visits	Numerical data	
Spend	Number of total expenditures	Numerical data	
Nights	Number of total nights	Numerical data	

To make accurate the manipulation and analysis process of the dataset, the total number of visits and the total number of nights variables were normalized and recorded as visits_2 and nights_2.

3.2. Methodology

The main purpose of classification is to determine which class the objects belong to by checking the properties of the objects. There are many different classification types and algorithms. In this research, KNN, SVM, and Decision Tree methods are used. This paper estimates the spending levels of tourists considering the features given in the previous section.

KNN algorithm, also known as the k-Nearest neighbor algorithm, is one of the most known and used algorithms among machine learning algorithms (Kılınç et al., 2016). It's used in many different areas, such as handwriting detection, image recognition, or even video recognition, etc. KNN is most useful when labeled data is too expensive or impossible to obtain, and it can achieve high accuracy in a wide variety of prediction problems.

The algorithm is based on the local minimum of the target function which is used to learn an unknown function of desired precision and accuracy. It also finds the neighborhood of an unknown input, its range or distance from it, and other parameters. Classification is made by using the closeness between a selected feature and the closest feature. The value of K found here is expressed with a number.

The steps of the KNN algorithm are given below, respectively:

Step 1. *Select the number k of the neighbors*

First, the parameter k which represents the number of nearest neighbor points is determined. This parameter is the number of nearest neighbors to a given point. For instance, let k is 5, in this case, classification will be made according to the 5 nearest neighbors. To choose the value of k, take the $\sqrt{n}$, where n is the total number of data points, and if n is a decimal number, round it to the next odd number. In this article, the algorithm optimizes the accuracy and uses the k which gives a higher accuracy ratio.

Step 2. *Calculate the distance of k number neighbors*

The closeness/distance between the data points is calculated with the following distance measures namely: Euclidean, Manhattan, or Hamming distance, etc.

Step 3. *Take the k nearest neighbors as per the calculated Euclidean distance*

Based on the distance value, it is assigned to the class of the closest k neighbors according to the attribute values.

Step 4. *Among these k neighbors, count the number of the data points in each category*

Step 5. *Assign the new data points to that category for which the number of the neighbor is maximum*

SVM is a machine learning method that is used in regression but mostly in classification analysis. This method is based on a supervised learning model. It can be used in core functions depending on the type of data during the operation of the algorithm. In this way, both linear and nonlinear classification operations can be performed. It is used in fields such as face detection, text and hypertext classification, image recognition, bioinformatics, and environmental sciences (Mathur and Foody, 2008).

In the SVM classification process, if fully separable data is used, usually all data can be classified with a hyperplane. However, if data that cannot be fully separated is used, it is often possible to classify with a single plane of the same size. Therefore, different kernel functions are used.

The steps of the SVM algorithm are given below, respectively:

Step 1. *Gathering the data and splitting dataset into training and testing*

Step 2. *Constructing the objective function*

It is also known as the cost function. Objective function that it is tried to minimize or maximize to achieve our objective. In SVM, it is aimed to find a hyperplane with the largest margin while keeping the misclassification as low as possible.

Step 3. *Obtaining the gradient cost function to optimize weights by calculating the gradients*

Step 4. *Minimizing the cost function*

It can be achieved by two ways: minimizing $||w||^2$ which maximizes margin ($2/||w||$) or minimizing the sum of hinge loss that minimizes misclassifications.

Step 5. *Stop the training when the current cost hasn't decreased much as compared to the previous cost (Yu and Kim, 2012)*

Decision trees are one of the most used supervised learning algorithms. They are generally applicable to solving classification and regression problems. A decision tree is a structure used to divide a dataset containing a large number of records into smaller sets by applying a set of decision rules. In other words, it is a structure used by dividing large amounts of records into very small record groups by applying simple decision-making steps.

Algorithm selection is based on the type of target variable. The most frequently used algorithms in decision trees; Entropy, Gini, Classification Error for categorical variables; for continuous variables, it is the Least Squares method (Daniya et al., 2020).

The steps of the general concept of the decision tree algorithm are given below, respectively:

Step 1. *Selection of the features and target attributes for each branch in split*

Step 2. *Calculation percent branch represents for each class in branch*

Step 3. *Calculation probability of class in the given branch*

Step 4. *Taking the squares of the class probabilities*

Step 5. *Summing the squared class probabilities*

Step 6. *Subtracting the sum from 1*

Step 7. *Weighting each branch based on the baseline probability*

Step 8. *Summing the weighted Gini index for each split (Tangirala, 2020)*

The application of these three models, whose steps are theoretically explained is given in Section 4. Python language and Jupyter Notebook integrated development environment are used for coding of models.

4. APPLICATION

As mentioned earlier, tourists coming to London from abroad are considered in this paper. Tourism is one of the most important factors in economic development. Therefore, estimating the financial returns of tourism plays an important role in planning the future actions that the country will take. In this study, we developed KNN, SVM, and Decision Tree models for spending level estimation and compare the confusion matrix, accuracy, and F1 score for each method. Moreover, when we examined certain parts of the dataset, we observed how much accuracy each model gave.

First, the dataset used in the paper is gathered from the database of the Office for National Statistics (ONS). After gathering the data, raw data is preprocessed by standardizing, dealing with missing data, and scaling. This dataset contains 2966 records for 8 columns (variables) as explained in the section 3. During the application, country, the quarter of the year, the way of arrival, the purpose of the travel, the total number of visits, and the total number of nights stayed were used as inputs, and spending level used as an output (estimated variable).

Expenditures were classified into 3 classes: below £0.9 million as low, between £0.9 million and £3.8 million as medium, and above £3.8 million as high. The data distribution of each class is given in Figure 1.

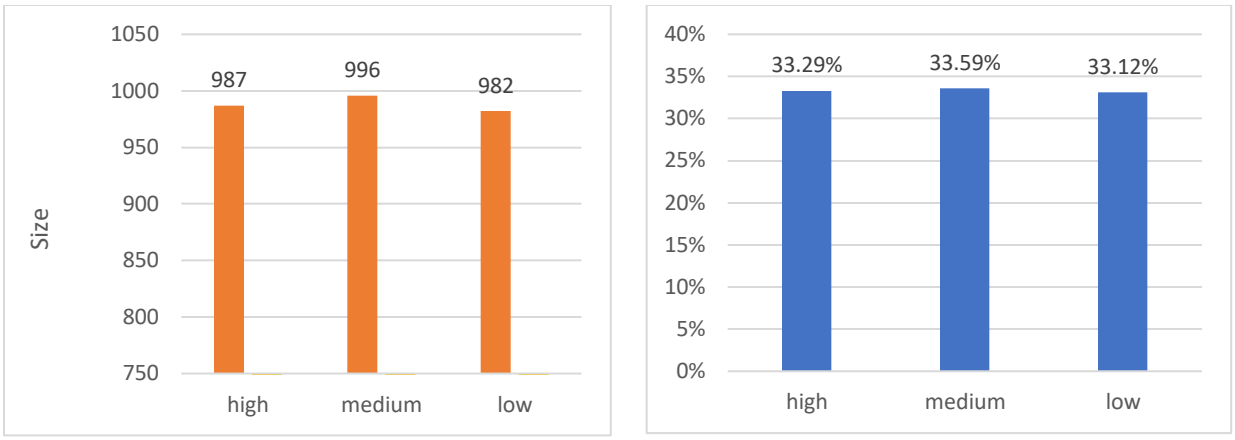


Figure 1. Data distribution for spending levels (size and percentages)

During the application, 3 models were used to estimate spending levels according to given inputs: KNN, SVM, and Decision Trees. According to the accuracy and cross-validation accuracy results given in Figure 2, the KNN model has the highest estimation accuracy with 69% cross-validation accuracy and 70% accuracy (with 74 neighbors), and the decision tree has the lowest cross-validation and accuracy value as 60% and 58%, respectively. This indicates that there is sufficient training data for KNN the data set is not easily separable using the decision planes that let SVM use and there is not enough categorical value for decision trees to perform better.

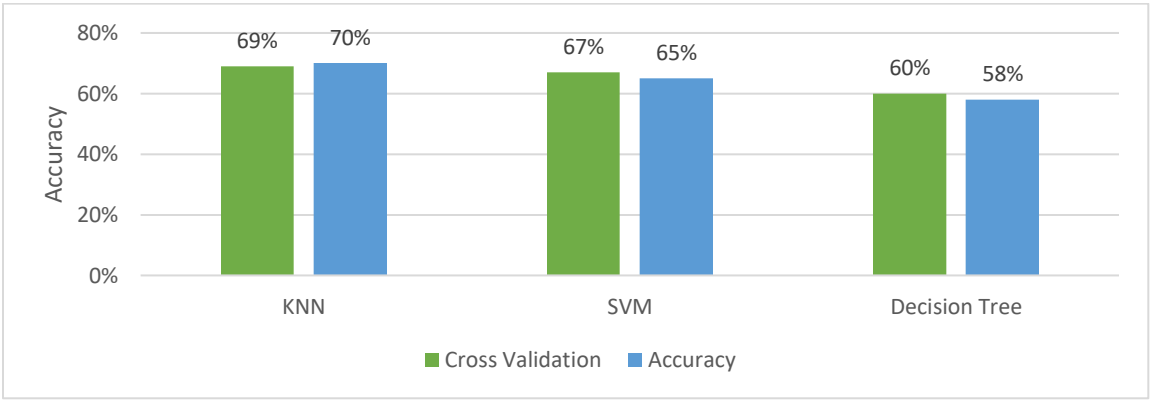


Figure 2. Accuracy and cross-validation accuracy results

Confusion matrices and F1-scores for each method were calculated in order to investigate which class was estimated more accurately. Table 2 and Figure 3 show the precision, recall, and F1-score values in each category for the developed models. It is observed that high and low classes have F1 scores that are close to each other for each model. On the other hand, the medium class has the lowest value in each model.

Table 3. Precision, Recall and F1-score values for spending levels

		Precision	Recall	F1-score
KNN	**high**	81%	72%	76%
	medium	57%	63%	60%
	low	73%	75%	74%
SVM	**high**	82%	68%	74%
	medium	53%	45%	49%
	low	62%	82%	82%
Decision Tree	**high**	67%	67%	67%
	medium	44%	45%	44%
	low	64%	63%	64%

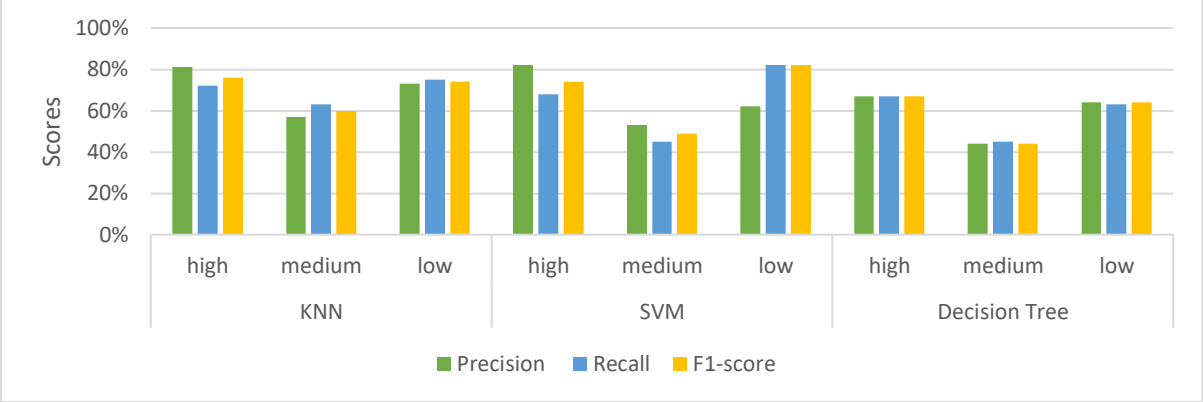
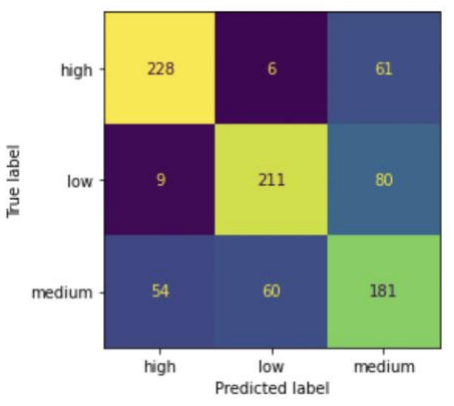
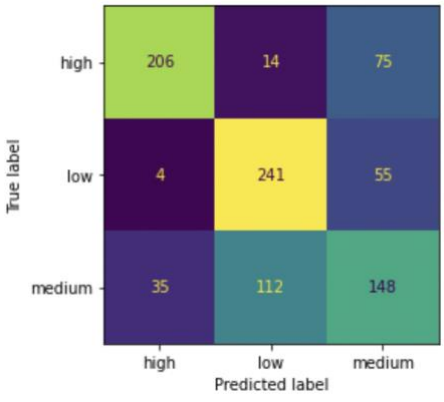


Figure 3. Precision, Recall and F1-Score values for spending levels

When we check the confusion matrices for KNN, SVM, and Decision Tree in Figure 4-6 respectively, it is also observed that medium classes are not classified as well as the other classes. It indicates that models do not perform well for medium classes. The score can be increased by adding more medium-class data.





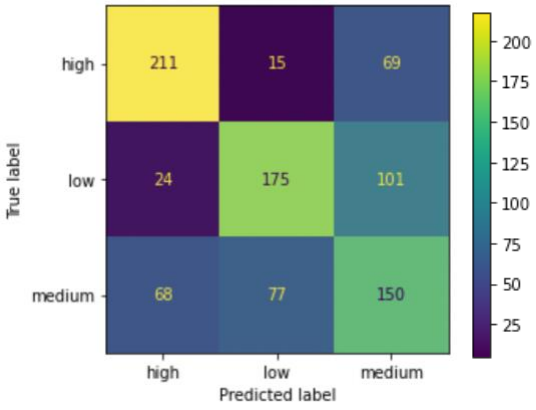


Figure 4. KNN Confusion Matrix **Figure 5.** SVM Confusion Matrix **Figure 6.** Decision Tree Confusion Matrix

As a final step, the percentages of 20%, 40%, 60%, and 80% of the random dataset were taken and the cross-validation accuracy rates were compared. According to the results, 20% of the dataset gives the highest accuracy value for the KNN, 40% of the dataset gives the highest accuracy value for the SVM, and 60% and 40% of the dataset gives the highest accuracy value for the decision trees.

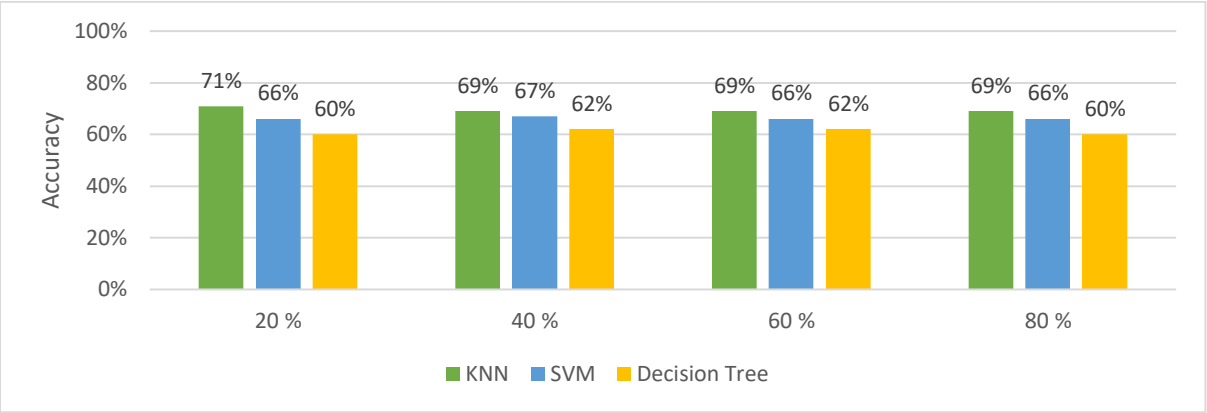


Figure 7. Cross-validation accuracy results of randomly selected data

5. CONCLUSION

The tourism sector is one of the most important key factors in economic development. Therefore, estimating the financial returns of tourism plays an important role in planning the future actions that the country will take. In this study, we developed three machine learning models to estimate spending levels by analyzing tourist behaviors accurately and effectively going to London in order to achieve greater tourist satisfaction. Factors affecting the spending level were determined as countries, the quarter of the year, way of arrival, the purpose of the travel, the total number of visits, and the total number of nights stayed.

During this study, we developed KNN, SVM, and Decision Tree models for spending level estimation and compared the confusion matrix, accuracy, and F1-score for each method. According to the results, the KNN algorithm achieved the best results with 70% accuracy compared to the SVM and Decision Tree models, and the best results on KNN are obtained when the k value is 74. This indicates that the dataset is not easily separable using the decision planes that SVM use and there is not enough categorical value for decision trees to perform better. For further research, more categorical data should be gathered because the current data set includes enough quantitative data but not enough categorical data for decision trees. Also, increasing the size of the dataset and adding more features to the model might enhance the performance of SVM.

Another issue to discuss is the medium spending class was the worst classified in each method. While the high and low spending classes have the F1-scores around 70%, the medium class has 60% in KNN. Therefore, it is recommended to add more medium-classed data in order for the model to perform well.

To summarize, we present KNN, SVM, and decision tree models for the estimation of spending levels to analyze tourist behaviors. According to the results obtained it is observed that the KNN

algorithm outperformed the other methods with 70% accuracy. In future studies, the accuracy of the models can be increased by using more data and features. Thus, the expenditure estimation of such tourists can be used in revenue and resource management for governments and organizations.

6. REFERENCES

Alaeddinoğlu, F., & Can, A. S. (2007). Türk turizm sektöründe tur operatörleri ve seyahat acentaları. *Gazi Üniversitesi Ticaret ve Turizm Eğitim Fakültesi Dergisi*, (2), 50-66.

Chauhan, V. K., Dahiya, K., & Sharma, A. (2019). Problem formulations and solvers in linear SVM: a review. *Artificial Intelligence Review*, *52*(2), 803-855.

Daniya, T., Geetha, M., & Kumar, K. S. (2020). Classification and regression trees with Gini index. *Advances in Mathematics: Scientific Journal*, *9*(10), 8237-8247.

Disegna, M., Durante, F., & Foscolo, E. (2016, September). A multivariate analysis of tourists' spending behaviour. In *International Conference on Soft Methods in Probability and Statistics* (pp. 187-195). Springer, Cham.

Han, J. (2020, September). Application of SVM model to environmental resource analysis in tourism development. In *Journal of Physics: Conference Series* (Vol. 1629, No. 1, p. 012007). IOP Publishing.

Harrou, F., Zeroual, A., & Sun, Y. (2020). Traffic congestion monitoring using an improved kNN strategy. *Measurement*, *156*, 107534.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer genomics & proteomics*, *15*(1), 41-51.

Hung, W., Shang, J., & Wang, F. (2012). Another look at the determinants of tourism expenditure. *Annals of Tourism Research*, *39*(1), 495-498.

Iswandhani, N., & Muhajir, M. (2018, March). K-means cluster analysis of tourist destination in special region of Yogyakarta using spatial approach and social network analysis (a case study: post of@ explorejogja instagram account in 2016). In *Journal of Physics: Conference Series* (Vol. 974, No. 1, p. 012033). IOP Publishing.

Kavzoğlu, T., & Çölkesen, İ. (2010). Karar ağaçları ile uydu görüntülerinin sınıflandırılması. *Harita Teknolojileri Elektronik Dergisi*, *2*(1), 36-45.

KILINÇ, D., BORANDAĞ, E., YÜCALAR, F., TUNALI, V., ŞİMŞEK, M., & ÖZÇİFT, A. (2016). KNN algoritması ve r dili ile metin madenciliği kullanılarak bilimsel makale tasnifi. *Marmara Fen Bilimleri Dergisi*, *28*(3), 89-94.

Kür, M., & Topal, İ. (2018, October). Classification of Tourist Area Selections by Chinese Tourists using k Nearest Neighbor Algorithm. In *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (pp. 1-6). IEEE.

Liu, M., Cao, Z., Zhang, J., Wang, L., Huang, C., & Luo, X. (2020). Short-term wind speed forecasting based on the Jaya-SVM model. *International Journal of Electrical Power & Energy Systems*, *121*, 106056.

Lusia, E., & Alamsyah, D. P. (2021). The Impact of Destination Attributes to Traveling Decision in New Normal.

Manzoor, F., Wei, L., Asif, M., Haq, M. Z. U., & Rehman, H. U. (2019). The contribution of sustainable tourism to economic growth and employment in Pakistan. *International journal of environmental research and public health*, *16*(19), 3785.

Mathur, A., & Foody, G. M. (2008). Multiclass and binary SVM classification: Implications for training and classification users. *IEEE Geoscience and remote sensing letters*, *5*(2), 241-245.

Ntaliakouras, N., Vonitsanos, G., Kanavos, A., & Dritsas, E. (2019, July). An apache spark methodology for forecasting tourism demand in Greece. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1-5). IEEE.

Paramita, P. W. P., Astawa, I. P., Mataram, I. G. A. B., & Sudira, I. P. (2022). Implementation of Offline Travel Agent Promotion Model to Increase Room Occupancy.

Rizvi, S., Rienties, B., & Khoja, S. A. (2019). The role of demographics in online learning; A decision tree based approach. *Computers & Education*, *137*, 32-47.

Somvanshi, M., Chavan, P., Tambade, S., & Shinde, S. V. (2016, August). A review of machine learning techniques using decision tree and support vector machine. In *2016 international conference on computing communication control and automation (ICCUBEA)* (pp. 1-7). IEEE.

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, *11*(2), 612-619.

Thi, H. D. (2022). *An assessment of how mindfulness affects service quality and service experience in Vietnam's luxury hotels* (Doctoral dissertation, Northumbria University).

UNWTO. (2008). *Glossary of tourism terms | UNWTO*. United Nations World Tourism Organization. Retrieved June 1, 2022, from https://www.unwto.org/glossary-tourism-terms

Wahyono, I. D., Asfani, K., Mohamad, M. M., Aripriharta, A., Wibawa, A. P., & Wibisono, W. (2020, August). New smart map for tourism using artificial intelligence. In *2020 10th Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)* (pp. 213-216). IEEE.

Wang, L., Wang, S., Yuan, Z., & Peng, L. (2021). Analyzing potential tourist behavior using PCA and modified affinity propagation clustering based on Baidu index: Taking Beijing city as an example. *Data Science and Management*, *2*, 12-19.

Xing, W., & Bei, Y. (2019). Medical health big data classification based on KNN classification algorithm. *IEEE Access*, *8*, 28808-28819.

Yu, H., & Kim, S. (2012). SVM Tutorial-Classification, Regression and Ranking. *Handbook of Natural computing*, *1*, 479-506.

Zhang, S. (2020). Cost-sensitive KNN classification. *Neurocomputing*, *391*, 234-242.