Data Mining Using Support Vector Machine Model for Baturraden Tourism Visitor Satisfaction Prediction

Suci Damayanti, Luzi Dwi Oktaviana, Trias Bratakusuma*

Faculty of Computer Science, Information System Department, Amikom Purwokerto University, Purwokerto, Indonesia
Email: 1.sucid7888@gmail.com, 2.oktaviana@amikompurwokerto.ac.id, 3.brata@amikompurwokerto.ac.id

Abstract—The tourism industry is a significant economic sector and a driver of local and national economic growth. Tourism not only contributes economically, but also plays an important role in introducing and preserving the cultural and natural wealth of an area. One of them is Baturraden tourism, a tourist destination located in Indonesia is experiencing a rapid increase in the number of tourist visits. Baturraden is a tourist destination located in the highlands at the foot of Mount Slamet, Indonesia, precisely in Banyumas Regency, Baturraden District. Tourists who visit every year are increasing but tour managers have not realized whether the tourists are satisfied or not so research is needed to measure the level of satisfaction of tourists so that the Baturraden tourism is better. To measure the level of satisfaction, an algorithm is needed, in this study the algorithm used is a support vector machine (SVM) to collect data that will be used as a dataset by taking reviews on google maps manually then the data is grouped into groups of satisfied and dissatisfied tourists, as many as 100 data are taken and processed. So that the final result obtained an accuracy value of 86.00%, and for reviews tend to be positive or satisfied tourists visiting the baturraden tourist area.

Keywords: Baturraden; Tour; Support Vector Machine; Level Of Accuracy; User Reviews

1. INTRODUCTION

The tourism industry is a significant economic sector and a driver of local and national economic growth. Tourism not only contributes economically, but also plays an important role in introducing and preserving the cultural and natural wealth of an area [1]. In this study, one of them is Baturraden tourism, a tourist destination located in Indonesia, has experienced a rapid increase in the number of tourist visits. Baturraden is a tourist destination located in the highlands at the foot of Mount Slamet, Indonesia, precisely in Banyumas Regency, Baturraden District. The region fascinates visitors with its natural and fresh natural beauty, making it an attractive destination for nature lovers and mountain tourism enthusiasts alike [2].

Baturraden's main attraction involves its cool atmosphere and clean air, which are highly desired by those looking for an escape from the hustle and bustle of urban life. In addition, the presence of natural hot springs in this area adds to its attractiveness as a place to relax and heal the body. Baturraden is also known for the diversity of parks and flower gardens that adorn its territory [3].

Beautiful gardens, such as Taman Bunga Nusantara and Taman Kyai Langgeng, offer an alluring visual experience and become a favorite place for tourists who love the beauty of nature. The harmonious meeting of nature, local culture, and the coolness of the climate in Baturraden makes it a suitable destination for families, couples, or individuals looking for tranquility and pristine natural beauty. In addition, various activities such as hiking, cycling, and enjoying local cuisine can also be enjoyed here, adding to the variety of experiences for visitors [4].

Therefore, the presence of a prediction system that can anticipate the level of tourist satisfaction is very important. So this research was conducted as a useful reference for tour managers in predicting the level of visitor satisfaction, both for free and commercial. With an effective prediction system in place, tour managers can take proactive steps to improve the visitor experience, thereby increasing tourist attraction and visits [5].

Because there are so many tours in Baturraden District, for this study only took data on tourists who visited the tour called lokawisata baturraden. Baturraden tourism is the peak of existing tourism and the most crowded visitors, in it there are various types of photo spots, children's playgrounds, various culinary specialties of baturraden or banyumas [6].

Basically, tourist satisfaction not only creates a positive experience for them, but is also key to maintaining the attractiveness of tourist destinations and ensuring the sustainability of the tourism industry in the region. Today, data is a valuable asset in decision making. Therefore, the utilization of technology and data analysis methods has become very relevant. One approach that can be used to analyze tourist visit data is Data Mining. Data mining is the process of extracting useful and valuable information from large and complex data.

Within the framework of Data Mining, the Support Vector Machine (SVM) has proven to be a reliable model for analysis and prediction [7]. SVM is a machine learning model capable of handling classification and regression, even on data that is of high complexity and is not linear. The application of SVM to tourist visit data in Baturraden can provide deep insight into the factors that affect tourist satisfaction [8].

This study aims to develop a special SVM model to predict the level of tourist satisfaction of baturraden tourism. By utilizing historical data on visits and traveller feedback, the study sought to identify patterns and relationships that conventional analysis methods might miss. The results of the study are expected to make a positive contribution to tourism managers, stakeholders, and local governments in making strategic decisions to...
improve the quality of tourism services and experiences in Baturraden, while maintaining the sustainability of the tourism industry.

The application of the SVM model in this study is expected to provide in-depth analysis of tourist visit data. SVM has the advantage of handling non-linear data and has a high degree of accuracy, making it an ideal choice for analyzing complex patterns in tourism data. By utilizing the important features found by SVM, we can identify the variables that are most influential on traveller satisfaction levels.

In addition, the study will also consider factors such as weather, season, and special events that might affect tourist satisfaction. This analysis is expected to provide a more holistic and contextual insight into the factors that can increase or decrease the level of tourist satisfaction in Baturraden. In the face of technological developments and digital transformation, this research also integrates the concept of Big Data to improve the accuracy of analysis. The utilization of Big Data in the context of tourism allows handling large and complex volumes of data, thereby obtaining deeper and more accurate insights. In addition, cooperation with relevant parties, including private parties, local communities, and government agencies, will be an integral part of this research. This collaborative approach is expected to ensure the sustainability of research results and the implementation of the resulting recommendations.

Thus, this research is not only an academic contribution, but also a foundation for the development of evidence-based tourism policies and sustainable destination management strategies in Baturraden. Through the integration of technology, sophisticated data analysis, and cross-sector cooperation, it is hoped that this research can provide holistic and innovative solutions to increase tourist satisfaction and support sustainable tourism development in Baturraden.

Before conducting this study, researchers conducted a literature review as reference material for several similar studies that became literature reviews. First, the research conducted by Husada et al obtained the results of opinion classification carried out with a machine learning approach utilizing a multi-class Support Vector Machine (SVM) algorithm. The data used in the study were opinions in English from Twitter users of airlines. Based on the tests that have been done, the best classification results were obtained using SVM kernel RBF at parameter values C (complexity) = 10 and γ (gamma) = 1, with accuracy values of 84.37% and 80.41% when using 10-fold cross validation[9]

The two studies conducted by Darwis et al obtained the results of 1890 data and 3846 terms / words from the results of preprocessing and then calculated the value of the occurrence of words for labeling that produced positive, negative and neutral sentiments. Based on the test results produced, the application of the SVM method resulted in an Accuracy value of 82% and resulted in sentiment with a greater negative label with a total of 77%, a positive label of 8% and a neutral label of 25% [10]

The three studies conducted by Rahma et al obtained evaluation results using confusion matrix on aspect and sentiment classification, respectively resulting in precision of 0.94 and 0.86, recall of 0.6 and 0.98, accuracy of 0.88 and 0.86, and f-measure of 0.73 and 0.92 [11] Therefore, it can be concluded that the use of the Support Vector Machine (SVM) method in this study is the right method, judging from the various review literature that has been described earlier.

2. RESEARCH METHODOLOGY

2.1 Types of research

According to David, laboratory experiments can be interpreted as a scientific investigation that examines variations of all or almost all independent variables that may have an influence, while variables that are not related to the research problem are sought to have as little impact as possible. Business process referred to as data mining is an activity that is carried out to investigate large amounts of data with the aim of finding patterns that have meaning and regularity. This research is an experiment that aims to measure the level of accuracy of the SVM algorithm. The source of data used in the study came from the results of google maps reviews.

2.2 Data Collection Techniques

Data collection is a crucial stage in research methods. In this study, this step was implemented by taking visitor review data derived from google maps. This study took data from google maps because of the short time of research so that if you use a questionnaire it will not be enough time. Therefore, the data used in the research process uses existing reviews only.

2.3 Data Analysis Techniques

The data applied in this study came from the results of a survey conducted directly in the field, where questionnaires were distributed to visitors to Tabebuya Park with the help of security guards. The method applied in this study is using the SVM algorithm model. The following are the steps of the research procedure summarized in figure 1.
In this research process, the steps began with data collection by researchers manually through google maps on baturraden lokawisata, with a set number of 100 review data. Furthermore, the data will be used as an initial dataset to analyze the level of visitor satisfaction of the baturraden tourist workshop. The results of the review are calculated to determine the number of satisfied and dissatisfied respondents, please note that data collection is carried out randomly or randomly so as not to sort or select positive or negative review data. The next steps involve selecting the data to be used, cleaning the data of irrelevant or incomplete values, and calculating the number of satisfied and dissatisfied using the SVM algorithm model. The data processing process uses Rapid Miner, with testing carried out three times using the sharing of training data and different testing data.

The results of this study are applied to make decisions regarding visitor satisfaction of Baturraden Tourism. Evaluation and validation are carried out to measure the level of accuracy of customer satisfaction data using several techniques contained in the Rapid Miner framework.

2.4 Data Mining

Data mining is a systematic and comprehensive process of analysis performed on large datasets to identify valuable patterns and useful information [12]. The main goal of data mining is to uncover new knowledge or trends that may not be clearly visible through direct examination of raw data [13]. The process of data mining involves a series of steps, from the selection and collection of relevant data, data processing to clean and prepare it, to the application of various analytical techniques such as classification, regression, clustering, and association. This analysis aims to extract patterns that can provide insight into relationships and characteristics in the data [14]. Applications of data mining can be found in a variety of industries and fields, including business, science, health, security, and more. The main advantage of data mining is its ability to identify complex patterns or relationships that may not be visible manually, allowing for more informational and strategic decision making [15]. The challenges in data mining include the need for a deep understanding of the data being processed, the selection of appropriate analysis techniques, and handling the potential complexity and high dimensions of large datasets [16]. In essence, data mining serves as an exploratory tool that helps uncover valuable knowledge from the complexity of abundant data. For more details about data mining, researchers summarize in figure 2.
2.5 Support Vector Machine (SVM)

Support Vector Machine (SVM) is an algorithm in machine learning used for classification and regression tasks [17]. SVM works by finding the best line or hyperplane that can separate two or more classes in feature space. The goal is to achieve optimal separation between the classes. The essence of SVM lies in the concept of "support vectors". It refers to the data points closest to the hyperplane separating the classes. SVM strives to find the hyperplane that has the maximum margin, that is, the largest distance between the hyperplane and the support vectors. The importance of maximal margins is to increase the generalizability of the model to new data that has never been seen. SVM models that have larger margins tend to be more resistant to overfitting [18]. SVM can be applied in a variety of dimensional spaces, including high dimensions, and has the flexibility to handle nonlinear data using kernel functions. Kernel functions allow SVM to map data into higher dimensions, where linear separation may become easier to achieve. In the context of classification, each data point will be labeled a class based on which side they are located on against the selected hyperplane. SVM can also be applied in regression tasks, where the goal is to predict continuous values rather than discrete categories [19]. The advantages of SVM include its ability to handle high-dimensional data sets, resistant to overfitting, and can perform well even in the case of non-linear data. However, setting the right parameters and handling large datasets can be challenging in SVM implementation. In the figure 3 is a Step Step of the SVM algorithm process in general [20].

![Figure 3. Step of the SVM algorithm process in general](image)

2.6 KDD

Knowledge Discovery in Databases (KDD) is a systematic approach to extracting valuable knowledge from large and complex data. The KDD process includes a series of in-depth steps to transform raw data into meaningful and useful information. The first step in KDD is the selection of data relevant to the purpose of the analysis. The selected data then undergoes a preprocessing stage to clean it up and prepare it to be ready for further analysis. A process of data transformation follows, in which the data is transformed into a form more suitable for analysis, including the selection of the most relevant attributes or features. The essence of KDD lies in the data mining step, where special algorithms are applied to explore patterns or relationships in the data [21]. The type of algorithm used varies depending on the desired purpose of analysis, such as classification, regression, clustering, or association. The results of the data mining process are further evaluated to ensure their accuracy and relevance. This evaluation involves the use of appropriate metrics, which depend on the type of analysis performed. The information or knowledge found is presented clearly through reports, graphs, or other visual forms for the end user to understand. Furthermore, the findings are evaluated in the context of business objectives or research more broadly. A deep understanding of the implications and relevance of findings is crucial for effective decision
making. The knowledge discovered from KDD is then applied in a broader context, including business decision making, operational improvement, or as a basis for further research [22]. The KDD process requires a combination of expertise from a variety of disciplines, including statistics, computer science, and domain-specific knowledge of the data being processed. Iterative characteristics are often inherent in the KDD process, where processing and analysis are returned to an early stage based on the results of the evaluation and understanding that develops as the process progresses. To be clearer about KDD, the KDD process can be seen in the figure 4.

![KDD Process Diagram](image)

Figure 4. KDD process

3. RESULT AND DISCUSSION

### 3.1 Data Retrieval

In the process of data retrieval is the first step before calculations are carried out using the Support Vector Machine (SVM) algorithm. In this study, the dataset used amounted to 100 data taken manually through reviews in the Google Playstore. Because of the many comments and the brevity of the study this time so that researchers took data randomly. For the results of the data retrieval process can be seen in Table 1. In addition, in Table 1 the data has been grouped into data satisfied with the Baturraden tourism or dissatisfied with the Baturraden tourism. For Attributes taken only username and comments only.

<table>
<thead>
<tr>
<th>No</th>
<th>Username</th>
<th>Reviews</th>
<th>Result Satisfied</th>
<th>Not Satisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Agus Priyo Purwanto</td>
<td>a sense of accomplishment healed by the beauty of this place. Hire tickets 25K, Dec 2023, cheap compared to its natural beauty. Hungry thirsty ... No need to worry</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Nur Hadi Triana</td>
<td>There are hot springs and swimming pools. River view, glass bridge but quiet / empty no one went up maybe the impact of the broken glass accident in Banyumas. There is a prayer room inside quite clean, many toilets are paid. INPUT: Please for music players to move the location not near the prayer room, because it is very disturbing for worshippers</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3</td>
<td>Fajar Isnain Hasan</td>
<td>It's really exciting to go to baturaden, enjoy the extraordinary view in the water tour, there is a tall fountain</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Samuel Sibagariang</td>
<td>Tourist attractions are wide with a variety of photo spots, there are waterfalls and swimming pools in the area inside. Pay an adult ticket at a price of 25 thousand, but can only use water bikes and swimming pools for free</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Eti haryati</td>
<td>Amazing beautiful scenery, there are many rides</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Aldwin</td>
<td>Humm 80% of our time was spent in vain when we came here, because even the rides when we came, there were no guards, the counter was closed, or there was another charge.</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>7</td>
<td>Asih Cute</td>
<td>The place is comfortable, many places for selfies, many places to soak in sulfur hot springs</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>…</td>
<td>Novita Ambarwati</td>
<td>Come on weekends, quite crowded, food prices are not too expensive, many places to rest. The place is cool, there are many additional rides, just walk2 around is also okay, always</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
No | Username | Reviews | Result | Satisfied | Not Satisfied
---|-----------|---------|--------|-----------|-------------
100 | Nindhyta | bring flip-flops and umbrellas, because usually it always rains. Fun for a family holiday |  |  | √
 | | Not as expected. I thought it would be cool but it turned out to be hot. And obviously when an emergency has to go to the clinic, oh zona bgt.ga there is a medical team. The p3k box is not complete.complete has me personally.severe anyway.pay expensive but don't understand what you want to enjoy. |  |  |

When viewed in table 1 after collecting data and grouping review data for the results, the majority of tourists who visit Baturraden tourism are satisfied due to several factors such as, the number of spots for photos, the extent of tours that can be used for picnics and tourist destinations with family, besides that the entrance ticket is cheap and there are lots of food and drinks at affordable prices. In addition, there are visitors who are dissatisfied due to several things such as many rides that are closed, many pungli and no equipment that supports health if what happens at the Baturraden tourist area.

### 3.2 Support Vector Machine (SVM)

Training data can be seen in table 2 used to assess the level of visitor satisfaction when visiting Baturraden Tourism. This satisfaction prediction was carried out using the SVM classification method, and the kernel model was obtained through Rapid Miner with corresponding attribute weights, namely \( w[X1] = -1.249, \) \( w[X2] = -0.506, \) \( w[X3] = -0.791, \) and \( w[X4] = -0.753. \) The main purpose of this analysis is to determine whether visitors are satisfied or dissatisfied with their experience at Lokawisata Baturraden.

<table>
<thead>
<tr>
<th>Atribut</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W[X1] )</td>
<td>-1.249</td>
</tr>
<tr>
<td>( W[X2] )</td>
<td>-0.506</td>
</tr>
<tr>
<td>( W[X3] )</td>
<td>-0.791</td>
</tr>
<tr>
<td>( W[X4] )</td>
<td>-0.753</td>
</tr>
</tbody>
</table>

### 3.3 Model Testing

The accuracy of the model that has been developed is evaluated by testing test data derived from training data. Given that the amount of data available after going through the preprocessing process is only 100 data, the cross-validation method is used to measure the level of accuracy. The evaluation results showed that the accuracy of the model for the SVM method reached 86.00%.

### 3.4 Confusion Matrix Algorithm SVM

Table 3 contains the calculation of the accuracy of training data using the SVM algorithm. From a total of 100 training data, through the SVM algorithm method, the following classification results were obtained: 49 satisfied prediction data that matched the actual situation, 7 satisfied prediction data that turned out to be dissatisfied, 7 dissatisfied prediction data that was even satisfied, and 36 dissatisfied prediction data that matched the actual situation.

<table>
<thead>
<tr>
<th>Accuracy: 86.00% +/- 12.000% (mikro 85.86%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True satisfied</td>
</tr>
<tr>
<td>Satisfied predictions</td>
</tr>
<tr>
<td>Dissatisfied predictions</td>
</tr>
<tr>
<td>Class recall</td>
</tr>
</tbody>
</table>

Then for the calculation of the confusion matrix value, which is as follows:

\[
\text{akurasi} = \frac{TP+FN}{TP+TN+FN+FP} = \frac{49+36}{49+36+7+7} = 0.860 = 86.0 \%
\]

\[
\text{precision} = \frac{FN}{FP} = \frac{36}{7+36} = 0.8372 = 83.72 \%
\]

\[
\text{recall} = \frac{FN}{TP+FN} = \frac{36}{49+36} = 0.8372 = 83.72 \%
\]
3.5 Kurva ROC

The calculation results are visualized with the ROC curve, can be seen in figure 5 which is the ROC curve for the SVM algorithm.

Figure 5. Roc Curve

3.6 Final Results

Models generated through the application of SVM methods are evaluated using a cross-validation approach. A comparison of the values of accuracy, precision, sensitivity, and recall can be observed in the following table 4.

<table>
<thead>
<tr>
<th>SVM</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>86.00%</td>
<td>0.947</td>
</tr>
</tbody>
</table>

Table 4 records the accuracy and AUC comparison of each method. It can be seen that the accuracy value reflects the corresponding result. In the context of data mining classification, AUC values can be classified into the following ranges:

1. 0.90-1.00 = Excellent classification
2. 0.80-0.90 = Good classification
3. 0.70-0.80 = Sufficient classification
4. 0.60-0.70 = Bad classification
5. 0.50-0.60 = Incorrect classification

By referring to the categorization and information mentioned in table 4, it can be concluded that the SVM method can be classified as a very good classification. This is evident by the AUC (Area Under Curve) value which ranges from 0.90-1.00, according to excellent classification standards. Based on the grouping above and the data contained in Table 4, it can be concluded that the SVM method is included in an excellent classification because based on the accuracy data recorded in Table 4, the SVM algorithm proves to be the best choice in classifying the level of visitor satisfaction of the Banyumas Baturraden Tourist Workshop. The SVM algorithm stands out with the highest degree of accuracy, reaching a percentage of 86.00%.

4. CONCLUSION

This study aims to explore and apply the Support Vector Machine (SVM) method in the context of Data Mining to predict the level of tourist visitor satisfaction in Baturraden. The SVM model that has been developed undergoes a series of quality tests using the cross-validation method, where the average accuracy reaches 86.00%. The evaluation process using confusion matrix and AUC with ROC revealed the impressive performance of this model. From the evaluation results, it can be concluded that SVM is an effective tool in predicting tourist visitor satisfaction. The success of this model lies primarily in its ability to provide consistent and accurate predictions, as reflected in its high degree of accuracy. The Confusion Matrix provides deeper insight into how the model responds to different situations, while a high AUC value indicates the model's ability to distinguish between satisfaction and dissatisfaction categories. Looking at the results of reviews from tourist visitors, there is a positive trend that characterizes their experience in Baturraden. The majority of visitors tend to be satisfied, and SVM models are consistently able to correctly predict visitor satisfaction. This indicates that this tourism destination has succeeded in providing positive services and experiences for its visitors. The importance of these results is not only limited to the model's ability to predict satisfaction, but also to its potential applications in tourism destination management. Information obtained from the SVM model can help management to better understand visitor
preferences and expectations, so as to improve service quality, design a more satisfying experience, and optimize marketing strategies. In conclusion, the study underlines that SVM methods in Data Mining are not only powerful predictive tools, but also a valuable source of insights for sustainable development in the tourism sector. Implications include the potential to increase the attractiveness and competitiveness of tourist destinations, paying special attention to factors that increase visitor satisfaction. Furthermore, for research that will continue this research, research can be carried out by multiplying datasets and adding training data so that the results are more accurate.

ACKNOWLEDGMENT

With humility, I would like to express my sincere gratitude to all those who have contributed to the writing of this journal. Without your support and collaboration, the research would not have been able to achieve the level of success achieved. Thank you to fellow researchers and fellow authors who have provided their valuable insights, suggestions, and contributions. Team involvement and collaboration in brainstorming and analyzing data has been a strong foundation for the journal's success. I would also like to express my gratitude to the editors who have provided critical and constructive feedback, ensuring that this paper meets academic standards and has significant added value. Not to forget, thanks to the institutions or organizations that have provided financial support and facilities that support the course of research and writing this journal. All the contributions and hard work of each individual involved, either directly or indirectly, really means something for the smooth and successful writing of this journal. Hopefully the results of this research can make a positive contribution in the scientific world and bring benefits to readers and the wider community. Thank you for your great dedication and cooperation.

REFERENCES


