P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v3i4.278

COMPARATION OF DECISION TREE MODEL AND SUPPORT VERCTOR MACHINE IN SENTIMENT ANALYSIS OF REVIEW DATASET SAMSUNG SSD 850 EVO AT NEW EGG SHOP

Muhammad Fahmi Julianto[1], Yesni Malau[2*)], Wahyutama Fitri Hidayat[3], Wawan Nugroho[4], Fintri Indriyani[5]

Computer Sience[1], Electrical Engineering[2], Software Engineering[3], Information System[4], Information Technology[5]
Universitas Bina Sarana Informatika
www.bsi.ac.id
fahmi.fjl@bsi.ac.id, yesni.ymu@bsi.ac.id, revelationtama.wfh@bsi.ac.id, wawan.wgh@bsi.ac.id, fintri.fni@bsi.ac.id
(*) Corresponding Author

Abstrak

Perkembangan teknologi nformasi saat ini berkembang sangat pesat, tidak terkecuali berdampak pada perangkat keras yang digunakan. Hal tersebut dapat dicontohkan dalam penggunaan hardisk yang mulai beralih ke SSD. Proses pemilihan produk SSD yang akan digunakan tidak terlepas dari sumber informasi yang terdapat dalam internet. Melalui internet setiap pengguna dapat memberikan ulasan baik itu ulasan positif maupun negative. Dengan banyaknya ulasan mengenai ulasah terhadap SSD Samsung 850 Evo pada NewEgg Store penulis menggunakannya untuk diproses menjadi sebuah informasi, yang akan memiliki pengetahuan baru. Berdasarkan hal itu, penulis membuat penelitian, dalam bentuk klasifikasi opini dengan menganalisis sentimen melalui pendekatan text mining. Dalam penelitian ini digunakan dua model klasifikasi yaitu Decision Tree dan Support Vector Machine. Hasil penelitan ini berupa perbandingan 2 model yang digunakan berdasarkan nilai akurasi dan AUC. Berdasarkan penelitian model Support Vector Machine lebih baik dibandingkan dengan Model decision Tree. Kesimpulan tersebut dapat dibuktikan dengan nilai akurasi model Support Vector Machine menghasilkan nilai sebesar 0,87 atau 87% sedangkan nilai akurasi model Decision Tree menghasilkan nilai 0,82 atau 82%. Selain itu niali AUC model Support Vector Machine menghasilkan nilai 0,87 dan mode Decision Tree menghasilkan nilai 0,82 atau dapat dikatakan nilai AUC model Support Vector Machine lebih baik dibandingkan dengan model Decision Tree.

Kata kunci: Sentimen Analisis, Decision Tree, Support Vector Machine

Abstract

The development of information technology is currently growing very rapidly, including the impact on the hardware used. This can be exemplified in the use of hard drives that are starting to switch to SSDs. The process of selecting an SSD product to be used cannot be separated from the sources of information found on the internet. Through the internet, every user can provide reviews, both positive and negative reviews. With the many reviews regarding the review of the Samsung 850 Evo SSD on the NewEgg Store, the author uses it to be processed into information, which will have new knowledge. Based on that, the author makes research, in the form of opinion classification by analyzing sentiment through a text mining approach. In this study, two classification models were used, namely Decision Tree and Support Vector Machine. The results of this study are in the form of a comparison of the 2 models used based on the accuracy and AUC values. Based on research, the Support Vector Machine model is better than the Decision Tree model. This conclusion can be proven by the accuracy value of the Support Vector Machine model resulting in a value of 0.87 or 87% while the accuracy value of the Decision Tree model produces a value of 0.82 or 82%. In addition, the AUC value of the Support Vector Machine model produces a value of 0.87 and the Decision Tree mode produces a value of 0.82 or it can be said that the AUC value of the Support Vector Machine model is better than the Decision Tree model.

Keywords: Sentiment Analysis, Decision Tree, Support Vector Machine

INTRODUCTION

As one of the largest E-Commerce in America, NewEgg is an online retail company that focuses on selling computer equipment, both hardware and software. Because NewEgg is in the sales field, they rely heavily on reviews of every product they sell. At this time the author discusses reviews on the Samsung Evo SSD product. On the other hand, Solid State Drive (SSD) is a data storage device that uses a series of ICs as memory that is used to store data or information.(Fadjar Efendi Rasyid, 2016).

Product reviews are a form of conveying consumer opinions and sentiments towards a product online. Product reviews today have a very important role in influencing consumer interest in a product(Siringoringo & Jamaludin, 2019). Where reviews from buyers who have purchased the product will appear in the review column. At this stage, prospective buyers will be able to see reviews from previous buyers. This condition is generally called sentiment analysis and can determine the interest of the prospective buyer because the previous buyer's review column has given a perspective on the SSD. Sentiment Analysis is a study consisting of Natural Language processing, linguistic computing, and text analysis, so that it can assist in identifying the opinions of a product submitted by users, where the reviews are usually divided into two classes, namely positive and negative.(Fanissa, Fauzi, & Adinugroho, 2018).Detection of classification patterns in this review is an interesting object of research using text mining. Text mining has become an interesting field of research because of the large amount of text that exists on the web. Text mining is an important field in the context of data mining to find interesting patterns in textual data(Said A. Salloum, Ahmad Qasim Al Hamad & Shaalan, 2017). Text mining also involves all activities in finding information and other important data from various textual sources(Hashimi, Hafez, & Mathkour, 2015).

Based on previous research by regarding sentiment analysis conducted by(Irene, 2017)from the scenario test, it can be seen that the Support Vector Machine algorithm can be used for film review cases with an F1-Score value of 84.9%. In addition, other research conducted by(Mardiana, Syahreva, & Tuslaela, 2019) The test results with the confusion matrix obtained an accuracy value of 83% for Neural Network, 52% for K-Nearest Neighbor, 83% for Support Vector Machine, and 81% for Decision Tree. This study shows that the Support Vector Machine and Neural Network methods are best for classifying positive comments and negative related to franchising.

Sentiment analysis using text mining can help the process of understanding textual data extraction and processing to obtain information contained in an opinion sentence, this process aims to obtain interesting models and relationships and can be presented in large volumes of data. (Ronen Feldman, Bar-Ilan University, Israel, James Sanger, ABS Ventures, Boston, 2006). Based on the description of the background that has been described, the problems found are whether reviews can affect the selection of the Samsung Evo SSD product, the author will conduct an analysis that aims toto perform sentiment analysis using the Decision Tree algorithm and Support Vector Machine to analyze sentiment problems related toreview of the Samsung Evo SSD for sale at the NewEgg Store.

RESEARCH METHODS

This research is a sentiment analysis process to classify positive and negative user reviews on the sale of the Samsung Evo SSD at the NewEgg Store and determine the accuracy results using 2 methods, namely Decision Tree (DT) and SVM (Support Vector Machine).

Research subject

The subject of this research is to use positive and negative review data taken from the evaluation of Samsung Evo SSD sales at the NewEgg Store.

Types of research

This research is a type of experimental research. This method tests the success of the hypothesis and relates it to the research problem. This experimental model aims to classify sentiment analysis about reviews of Samsung Evo SSD sales at the NewEgg Store. data collection to obtain the data source used is a public data collection method by taking data on the sales review of the Samsung Evo SSD at the NewEgg Store. Data collection on Samsung Evo SSD sales at the NewEgg Store is carried out with the help of a Python tool to collect review data.

Time and Place of Research

a. Problem identification and needs analysis

At this stage, a search for problems related to positive and negative reviews of Samsung Evo SSD sales is carried out at the NewEgg Store.

b. Data collection

Collecting the necessary data from the Samsung Evo SSD sales review data on the NewEgg Store.

c. Experiment

This stage determines the model used to enter training data into the model and tests using Python tools for the method used.

d. Implementation

Applying the proposed KNN, SVM and Naive Bayes methods to determine the accuracy of the predictions used by the user.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v3i4.278

e. Evaluation

To measure whether the model that has been developed is successful or not, an evaluation is carried out. Evaluation is used to measure the accuracy achieved by the model.

f. Writing

Writing in the form of research reports is carried out simultaneously or in parallel with other steps to be effective and efficient.

Table 1 below shows that the research schedule is carried out for four weeks by carrying out activities that have been arranged according to the planned schedule. Table 1. Research Time

N O	ACTIVITY	WEEK			
		1	2	3	4
1	Search and selection of research objects	■			
2	Study object of research	■			
3	Research problem formulation	■			
4	Topic determination	■			
5	Reference material collection	■			
6	Preparation of the framework / rationale	■			
7	Preparation of research methodology/design	■			
8	Preparation of research proposal manuscript	■			
9	Submission of research proposal		■		
10	Implementation of sample data collection			■	
11	Data analysis			■	
12	Preparation of the final research manuscript				■

Research Target / Subject

The purpose of this study is to analyze problems related to several reviews about the sale of SSD Samsung Evo at the NewEgg Store globally to determine whether it tends to be positive or negative and to show accuracy results based on the two methods used. In addition, it is also used to find the best method used by comparing the accuracy value and the AUC value.

Data, Instruments, and Data Collection Techniques

In this study, the dataset used is a review of the sale of the Samsung Evo SSD at the NewEgg Store which was obtained from the website https://www.kaggle.com/abdulrahmanalqannas/ssd-reviews which consists of 3108 data consisting of pros and cons labels.

Data analysis technique

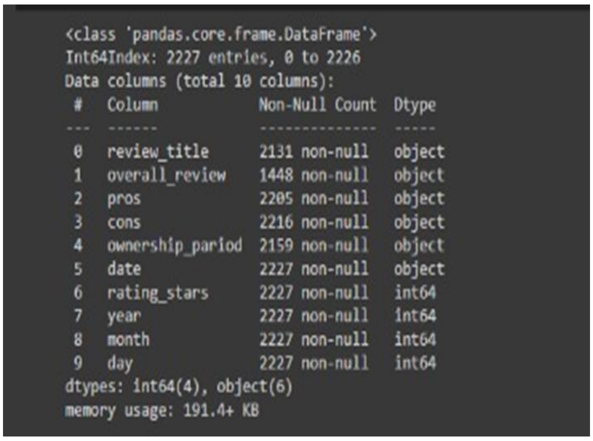
The data analysis technique used is tekt mining using python programming where sentiment analysis is used to identify positive and negative opinions. In addition, a comparison of the best methods is proposed, namely Decision Tree and Support Vector Machine. Steps used:

1. Definition of dataset
   The dataset is defined by using pros and cons labels.
2. Pre-processing
   At this stage, two pros and cons labels are combined which are then labeled into positive and negative statements and then stopwords are removed.
3. Transformation
   The weighting of textual data, the process used is TF-IDF.
4. Classification
   Text classifiers usually use the Decision Tree and SVM methods.
5. Interpretation/Evaluation
   At this stage an evaluation is carried out to calculate the accuracy value and AUC value.

RESULTS AND DISCUSSION

1. Definition of Dataset

The dataset used in this study is public data regarding sales of the Samsung Evo SSD at the NewEgg Store which consists of 10 data attributes. However, only 2 attributes are used in this study, namely pros which consists of 2205 data and cons which consists of 2216 data.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2227 entries, 0 to 2226
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   review_title      2131 non-null   object
 1   overall_review    1448 non-null   object
 2   pros              2205 non-null   object
 3   cons              2216 non-null   object
 4   ownership_pariod  2159 non-null   object
 5   date              2227 non-null   object
 6   rating_stars      2227 non-null   int64
 7   year              2227 non-null   int64
 8   month             2227 non-null   int64
 9   day               2227 non-null   int64
dtypes: int64(4), object(6)
memory usage: 191.4+ KB
```

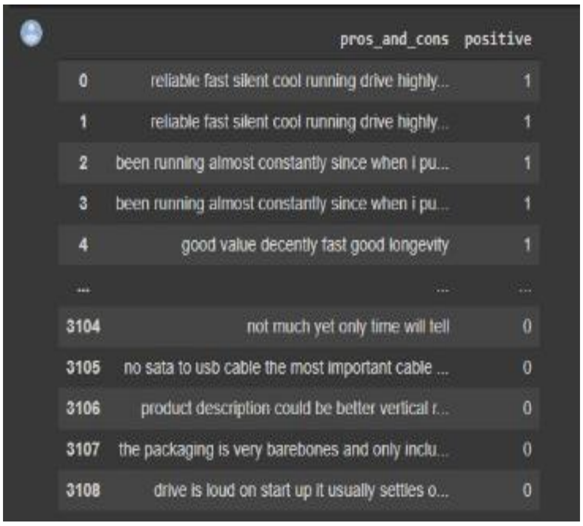
Figure 1. Dataset Description

Based on Figure 1 above, it explains the number of words used in each label, but only pros and cons labels are used. To add a dataset and view the info, use the following code:

```
df- pd.read_csv('ssd_reviews.csv', index_col=0
df.info()
```

2. Pre-processing

The pre-processing stage in data mining is to transform data into a format so that the process is easier and more effective(Meilina, 2015). Preprocessing can be done in two conditions, the first is to form training data and the second is to detect intrusions(Jacobus & Winarko, 2014). In this study, pre-processing, the data that has been collected is first processed by labeling, and carrying out the stopwords removal process.



	pros_and_cons	positive
0	reliable fast silent cool running drive highly...	1
1	reliable fast silent cool running drive highly...	1
2	been running almost constantly since when i pu...	1
3	been running almost constantly since when i pu...	1
4	good value decently fast good longevity	1
...	...	...
3104	not much yet only time will tell	0
3105	no sata to usb cable the most important cable ...	0
3106	product description could be better vertical r...	0
3107	the packaging is very barebones and only inclu...	0
3108	drive is loud on start up it usually settles o...	0

Figure 2. Data Pre-processing Process

Figure 2 describes the process of determining each review used in this study where 0 indicates that the review is negative while 1 indicates that the review is positive. The code used is as follows:

```
df_cons = df[['cons']].dropna()
df_cons['positive'] =0
df_cons.drop(df_cons[df_cons['cons'].isin(['none',
'none so far', 'non'])].index, inplace=True)
df_cons.rename(coloums=('cons':'pros_and_cons'),
inplace=True
df_pros =df[['pros']][:1562].dropna()
df_pros.rename(coloums=('pros':'pros_and_cons'),
implace=True)
df_pros['positive'] =1
merged_df = pd.merge(left=df_pros, right=df_cons',
'positive'], right_on=['pros_and_cons', 'positive'],
how='outer']
```

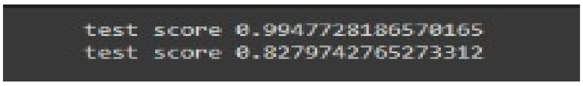
3. Transformation

The process of transforming data into a certain format so that the data is in accordance with the data mining process (June Arta, Indrawan, & Dantes, 2017). At this stage, TF-IDF is used in weighting the data that has been obtained. The weighting process with TF-IDF uses the following code:

```
tfidf_vectoriser =Tfidfvectorizer(stop_word='english')
tfidf_f = tfidf_vectoriser.fit(X['pros_and_cons'])
tfidf_transform = tfidf_f.transform(x['pros_and_cons'])
```

4. Classification Decision Tree

The first classification process used the Decision Tree model. The reason for choosing the decision tree model is that this algorithm produces a model that can predict data categories by determining categories based on data features(Ceballos, 2019)(Ochiai, Masuma, & Tomii, 2019). The following is the result of data processing using Decision Tree.

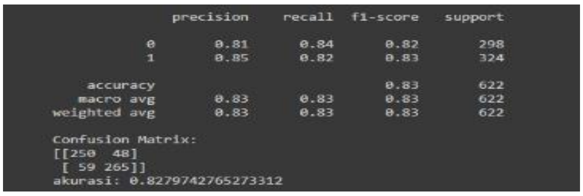


```
test score 0.9947728186570165
test score 0.8279742765273312
```

Figure 3. Value of x test and y train Decision Tree Model

For the x train and y train processes, the following code is used:

```
Tree = DecisionTreeClassifier()
Tree.fit(tf_x_train, y_train)
print('test score', tree.score(tf_x_train, y_train))
print('test score', tree.score(tf_x_test, y_test))
y_pred = tree.predict(tf_x_test)
```



```
              precision    recall  f1-score   support

           0       0.81      0.84      0.82       298
           1       0.85      0.82      0.83       324

    accuracy                           0.83       622
   macro avg       0.83      0.83      0.83       622
weighted avg       0.83      0.83      0.83       622

Confusion Matrix:
[[250  48]
 [ 59 265]]
akurasi: 0.8279742765273312
```

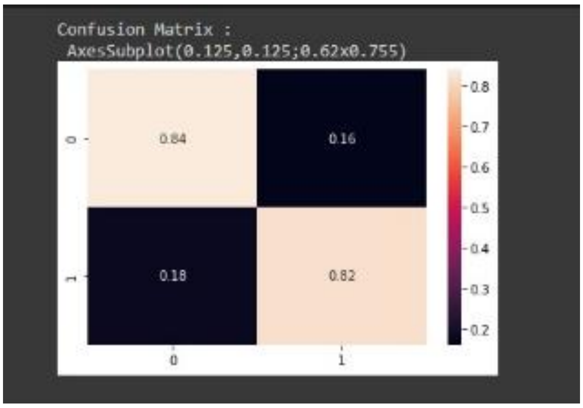
Figure 4. Classification Accuracy Value With Decision Tree Model

As for the classification process with the decision tree model, the following code is used:

```
print(classification_report(y_test, y_pred))
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))
accuracy_entropy=metrics.accuracy_score(y_test,
y_pred)
print("accuracy:" accuracy_entropy)
```

Based on Figure 3 and Figure 4 above, it can be stated that the classification results using the Decision Tree test score model with x tran and y train yields a value of 0.99 while the test score using x test and y test produces a value of 0.82. The results of the accuracy value using the Decision Tree model produce a value of 0.82. The resulting confusion matrix value is shown in the following figure:

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v3i4.278

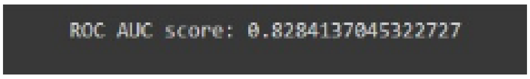


```
Confusion Matrix :
 AxesSubplot(0.125,0.125;0.62x0.755)
```

Figure 5. Confusion Matrix Model Decision Tree

To display the confusion matrix of the decision tree model classification process, the following code is used:

```
Print("Confusion Matrix : \n",sns,
hetmap(confusion_matrix(y_test, y_pred,
normalize='true'). Annot = True))
```



```
ROC AUC score: 0.8284137045322727
```

Figure 6. The AUC Value of the Decision Tree Model

Meanwhile, to display the ROC AUC score of the decision tree model, the following code is used:

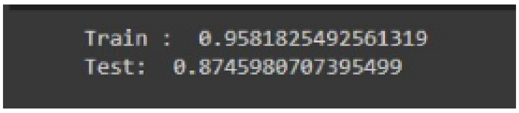
```
def_multiclass_roc_auc_score(y_test, y_pred,
average="macro"):
    lb = preprocessing.LabelBinarizer()
    lb.fit(y_test)
    y_test = lb.transforms(y_test)
    y_pred = lb_transform(y_pred)

return roc_auc_score(y_test, y_pred.
average=average)
print('ROC AUC score:',
multiclass_roc_auc_score(y_test,y_pred)
```

Based on Figure 5 and Figure 6 the confusion matrix Decision Tree model produces a true positive value of 0.84, a false negative of 0.16, a false positive of 0.18, and a true negative of 0.82. Meanwhile, the Area Under Curve (AUC) score is 0.82.

5. Classification Support Vector Machine

The second classification process used the Support Vector Machine model. The reason for using the SVM model is that this model uses a kernel function to determine the feature space and where the classifier function will be searched(Parapat, Furqon, & Sutrisno, 2018). In addition, SVM uses a hypothetical space in the form of linear functions in a high-dimensional feature and is trained using a learning algorithm based on optimization theory.(Puspitasari, Ratnawati, & Widodo, 2018). The following is the result of data processing using the Support Vector Machine.

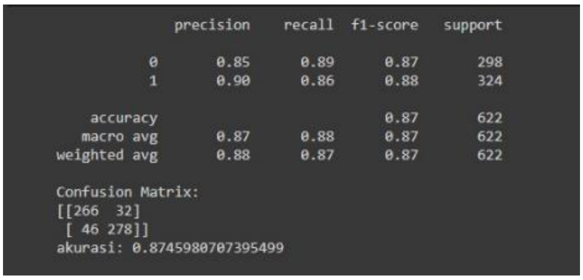


```
Train :  0.9581825492561319
Test:  0.8745980707395499
```

Figure 7. Value of x test and y train Model Support Vector Machine

For the x train and y train processes, the following code is used:

```
svm_linear = svm.SVC(kernel='linear')
svm_linear.fit(tf_x_train, y_train)
print('Train ', svm_linear.score(tf_x_train, y_train)
print('Test ', svm_linear.score(tf_x_test, y_test))
y_pred = svm_linear.predict(tf_x_test)
```



```
              precision    recall  f1-score   support

           0       0.85      0.89      0.87       298
           1       0.90      0.86      0.88       324

    accuracy                           0.87       622
   macro avg       0.87      0.88      0.87       622
weighted avg       0.88      0.87      0.87       622

Confusion Matrix:
[[266  32]
 [ 46 278]]
akurasi: 0.8745980707395499
```

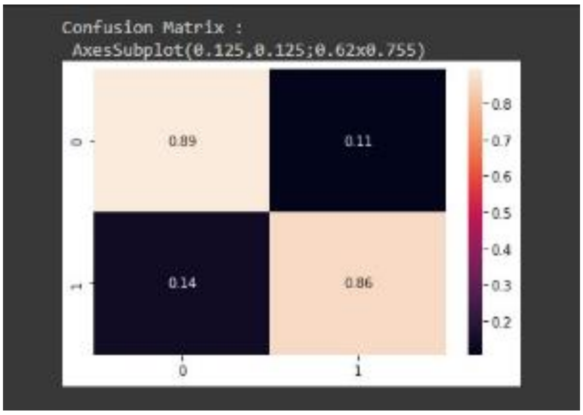
Figure 8. Accuracy Value of Support Vector Machine

As for the classification process with the SVM model, the following code is used:

```
print(classification_report(y_test, y_pred))
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))

accuracy_entropy=metrics.accuracy_score(y_test,
y_pred)
print("accuracy:" accuracy_entropy)
```

Based on Figure 7 and Figure 8 above, it can be stated that the classification results using the Support Vector Machine test score model with x tran and y train yields a value of 0.95 while the test score using x test and y test produces a value of 0.87. The results of the accuracy value using the Decision Tree model produce a value of 0.87. The resulting confusion matrix value is shown in the following figure:

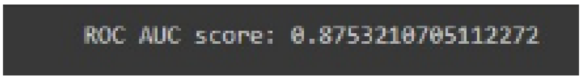


```
Confusion Matrix :
 AxesSubplot(0.125,0.125;0.62x0.755)
```

Figure 9. Confusion Matrix Model Support Vector Machine

To display the confusion matrix of the SVM model classification process, the following code is used:

```
print("Confusion Matrix : \n",sns,
hetmap(confusion_matrix(y_test, y_pred,
normalize='true'). Annot = True))
```



```
ROC AUC score: 0.8753210705112272
```

Figure 10. ROC AUC Value of Support Vector Machine Model

Meanwhile, to display the ROC AUC score of the decision tree model, the following code is used:

```
def_multiclass_roc_auc_score(y_test, y_pred,
average="macro"):
    lb = preprocessing.LabelBinarizer()
    lb.fit(y_test)
    y_test = lb.transforms(y_test)
    y_pred = lb_transform(y_pred)

return roc_auc_score(y_test, y_pred.
average=average)
print('ROC AUC score:',
multiclass_roc_auc_score(y_test,y_pred)
```

Based on Figure 9 and Figure 10 the confusion matrix Support Vector Machine model produces a true positive value of 0.89, a false negative of 0.11, a false positive of 0.14, and a true negative of 0.86. Meanwhile, the Area Under Curve (AUC) score is 0.87.

6. Interpretation/Evaluation

The interpretation and evaluation stages are used to compare the accuracy results and the AUC 2 models used are Decision Tree and Support Vector Machine.

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the discussion and research objectives, it can be concluded that the sales dataset of the Samsung 850 Evo SSD at the NewEgg Store with the classification of the Support Vector Machine model is better than the decision tree model. This conclusion can be proven by the accuracy value of the Support Vector Machine model resulting in a value of 0.87 or 87% while the accuracy value of the Decision Tree model produces a value of 0.82 or 82%. In addition, the AUC value of the Support Vector Machine model produces a value of 0.87 and the Decision Tree mode produces a value of 0.82 or it can be said that the AUC value of the Support Vector Machine model is better than the Decision Tree model.

Suggestion

Suggestions for further research, this research can be used as a reference or previous research and use a classification model other than the one already used. So that later it can be used as a comparison of research that has been done.

REFERENCES

Ceballos, F. (2019). Scikit-Learn Decision Trees Explained. Retrieved from https://towardsdatascience.com/scikit-learn-decision-trees-explained-803f3812290d

Fadjar Efendi Rasyid, SK (2016). Solid State Drive (SSD) Data Storage Media.

Fanissa, S., Fauzi, MA, & Adinugroho, S. (2018). Tourism Sentiment Analysis in Malang City Using Naive Bayes Method and Selection of Query Expansion Ranking Features. Journal of Information Technology and Computer Science Development, 2(8), 2766–2770.

Hashimi, H., Hafez, A., & Mathkour, H. (2015). Selection criteria for text mining approaches. Computers in Human Behavior, 51, 729–733. https://doi.org/10.1016/j.chb.2014.10.062

Irene, AF (2017). Sentiment Classification of Movie Reviews Using the Support Vector Machine Sentiment Classification of Movie Reviews Using Algorithm Support Vector Machine, 4(3), 4740–4750.

Jacobus, A., & Winarko, E. (2014). Application of Support Vector Machine Method in Real-time Intrusion Detection System. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 8(1), 13. https://doi.org/10.22146/ijccs.3491

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v3i4.278

Juni Arta, IK, Indrawan, G., & Dantes, GR (2017). Data Mining Recommendations for Outstanding Student Candidates at STMIC Denpasar Using the Technique for Others Reference By Similarity To Ideal Solution Method. JST (Journal of Science and Technology), 5(2), 792. https://doi.org/10.23887/jst-undiksha.v5i2.8549

Mardiana, T., Syahreva, H., & Tuslaela, T. (2019). Comparison of Classification Methods in Franchise Business Sentiment Analysis Based on Twitter Data. Journal of Pilar Nusa Mandiri, 15(2), 267–274. https://doi.org/10.33480/pilar.v15i2.752

Meilina, P. (2015). Application of Data Mining with Classification Method Using Decision Tree and Regression. Journal of Technology, University of Muhammadiyah Jakarta, 7(1), 11–20. Retrieved from journal.ftumj.ac.id/index.php/jurtek

Ochiai, Y., Masuma, Y., & Tomii, N. (2019). Improvement of timetable robustness by analysis of drivers' operation based on decision trees. Journal of Rail Transport Planning and Management, 9(March), 57–65. https://doi.org/10.1016/j.jrtpm.2019.03.001

Parapat, IM, Furqon, MT, & Sutrisno. (2018). Application of the Support Vector Machine (SVM) Method on the Classification of Deviant Growth in Children. Journal of Information Technology and Computer Science Development, 2(10), 3163–3169. Retrieved from https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/2577

Puspitasari, AM, Ratnawati, DE, & Widodo, AW (2018). Classification of Dental and Oral Diseases Using the Support Vector Machine Method. J-Ptiik, 2(2), 802–810. Retrieved from http://j-ptiik.ub.ac.id

Ronen Feldman, Bar-Ilan University, Israel, James Sanger, ABS Ventures, Boston, M. (2006). The Text Mining Handbook.

Said A. Salloum, Ahmad Qasim Al Hamad, MA-E., & Shaalan, and K. (2017). A Survey of Arabic Text Mining. Studies in Computational Intelligence.

Siringoringo, R., & Jamaludin, J. (2019). Text Mining and Sentiment Clustering in Online Store Product Reviews. Journal of Technology and Computer Science Prima (JUTIKOMP), 2(1), 41–48. https://doi.org/10.34012/jutikomp.v2i1.456