PREDICTION OF COMPATIBILITY BETWEEN LECTURERS AND THE SUBJECTS USING THE MACHINE LEARNING WITH NAIVE BAYES ALGORITHM

Munirah, Desriyanti

Abstract— The focus of this research is to build a machine that can learn to predict one case, namely how much is the degree of compatibility between lecturers who take the subject and those taught by the lecturer. This research involves a questionnaire set which will be filled in by at least 70 students from the informatics engineering and electrical engineering study programs related to the case. The results of this study are a representation of machine learning models that can predict the level / degree of compatibility in the cases raised in this study. Machine learning model that is produced using the Naive Bayes algorithm which is run by using a special mining programming language, the R language.

Index Terms— Prediction, *Machine Learning*, Algorithm, *Naive Bayes*, Compatibility, Lecturers and The Subjects.

1 INTRODUCTION

Datamining is a scientific tool that can be accounted for to obtain the data patterns in question and can be used to use those data patterns to make predictions. Harvesting (mining) data patterns is as simple as translating a record with n attributes into a point in n-dimensional space. Then from a set of records obtained a set of points in the dimension n space. The problem of harvesting data patterns is the problem of grouping points in space into groups of points where each group has its own characteristics. This collection of point groups with their respective characteristics is called a data pattern. On the other hand, what is meant by using data patterns to predict is if there is a new record in the future, this new note means a new record, this new record means a new point in space of dimension n, guess at which group of points the new point included is a prediction. The results of these predictions are interpreted back into the real world or phenomena as predictions of a record.

Datamining analysis has a difference with statistical analysis. Basically, statistical analysis takes a sample from the population. Then from the sample a statistical analysis was carried out. Based on the statistical analysis above the sample, people then generalize the results or assume they apply above the overall population. From that generalization, people make statistical predictions. Datamining analysis performs analysis directly to the population, not the sample, because datamining relies on computational power. Computing allows people to count all records in a population. Even datamining can continuously improve its predictions on populations that move or increase in time.

Datamining is a way to mine a number of patterns from a phenomenon. In the context of data, phenomena are first modeled into a data model. Modeling data models for a phenomenon is done by representing the phenomenon into a collection of attributes. Like for example the phenomenon of the teaching process that takes place every day in class. This phenomenon can be represented for example by a lecturer attribute with several other attributes in one case. Furthermore, the process that takes place every day in the class is recorded in a record where the record is built based on the attributes that represent the phenomenon. The record can be represented from moments or events every day that will later build a collection of records day to day, week to week, month to month and so on so as to build a table of moments or events that are sufficient to become a database or dataset. It is said that the data model represents a phenomenon because it is enough just to look at the lecture record in the dataset, everyone can see and understand what is happening in the class every day.

In conducting mining, many studies and studies have developed many algorithms through various approaches [1]. Data mining is based on existing datasets. When rules are generated using the rule miner algorithm, the next step is how to verify and validate the rules that have been produced to be implemented in real conditions.

A number of studies related to mining studies and models have been produced, including some that use learning machines with knowledge neural networks to extract a number of datasets as well as make improvements to a number of datasets produced with the concept of learning [2], [3]. This method was developed and optimized into Fuzzy Logical Rules based concepts using Gaussian uncertainty measurement [4].

- *Munirah, Informatics Engineering Department, Faculty of Engineering, Universitas Muhammadiyah Ponorogo, Indonesia. E-mail: munirah.mt@gmail.com*
- *Desriyanti, Electrical Engineering Department, Faculty of Engineering, Universitas Muhammadiyah Ponorogo, Indonesia. E-mail: yunandes@gmail.com*

The use of the rough set concept has also been developed to mine a number of datasets [5], [6]. There is also applied by using the dominance approach with the resulting algorithm that is DOMLEM which can induce a set of general decision rules that are consistent with the principle of dominance [7].

There are several ways in which naive bayes-based datamining is widely used, namely implementing the naive bayes algorithm to predict things. Even though this type of phenomenon is different from this research, it basically has the same or similar mechanism in making predictions based on Naive Bayes. Naive Bayes is used by people to filter e-mail whether it contains spam or not. As research Marsono et al who implemented the naive bayes classifier on the hardware that filters incoming email [8]. Naive Bayes was also conducted by Das and Kolya who used the algorithm for text mining and sentiment analysis of twetter statuses [9]. Naive Bayes is also used by Hsu et al for image classification [10], and also by Nirmala et al to detect glaucoma in patients [11].

2 IDENTIFICATION OF PROBLEMS

The fact that is also interesting to study especially around the world of lectures is often encountered by many students who complain of several lecturers who they think are not compatible in teaching for certain subjects. And this will certainly be very fatal if it happens repeatedly without any repairs.

By utilizing the concept of prediction using the datamining method, the researcher will build a prediction model that is used to analyze the level of compatibility of each lecturer profile in the campus environment (in this case data will be taken at least 70 record profiles of lecturers and subjects in the environment of technical study programs random informatics and electrical engineering).

While the collection of the dataset will be carried out by distributing questionnaires to a number of students for all lecturers supporting both courses in electrical engineering and informatics engineering during the even semester 2018/2019.

3 RESEARCH METHODS

The study begins with the construction of a data model for the case to be examined, which will produce metadata for the dataset to be processed later. After the metadata is generated, then a dataset is constructed that is taken from a questionnaire that is distributed to at least 70 students randomly to random subjects as well. The construction is poured into the lecturer and subject dataset tables.

After the 2 previous constructions were carried out then we step on how to construct the machine learning. Basically, machine learning will first learn based on the initial dataset and therefore it must be ensured that the dataset is really large in volume so that the portion of machine learning can be more accurate.

The dataset table is then imported into Rstudio. Data that has been imported is adjusted to the parameters of the Naive Bayes algorithm. The Naive Bayes algorithm was chosen to be used in the set of algorithms contained in the bnlearn library. Bnlearn library contains many algorithms based on Bayes theorem, especially Bayes Network and Naive Bayes and their variations. After the parameters are adjusted. The algorithm is run so that the results obtained are ready to be interpreted for predictions. The flow of research can be seen in the following figure :

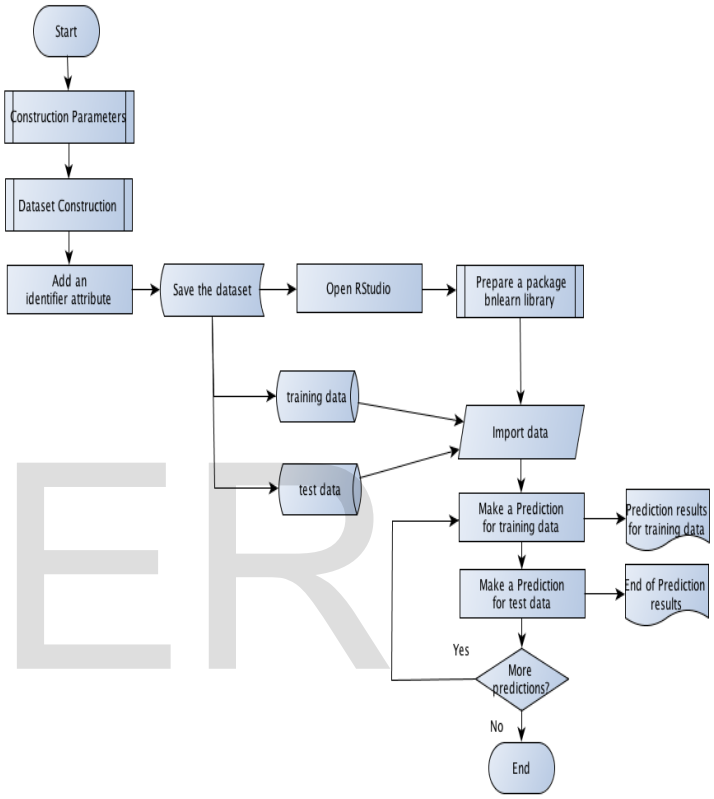


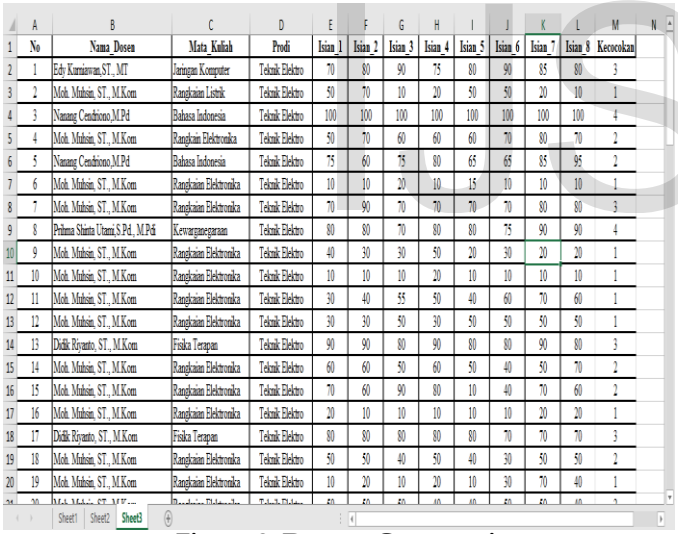
Figure 1. Research Flow

4 Discussion

The initial results achieved were the results of the construction of 8 (eight) parameters formed into the survey form along with the results of the data recapitulation of the survey form filling that had previously been shared with several related lecturers, namely lecturers in the study program in informatics and electrical engineering. The eight parameters are :

1. Learning styles are easily understood by students,
2. The learning settings are clear from the content to the assessment,
3. Mastering the subjects details to the end,
4. Bringing the subjects with a simple and practical,
5. Can harmonize cases with real conditions,
6. There is a transfer of knowledge given to students,
7. Provide space for students to ask questions,
8. Satisfaction in answering student questions thoroughly.

From the eight parameters formed, one column attribute is added as an identifier attribute, which is the match column. In the match column filled with numbers 1 to 4, with the category if filled in number 1, it means that the match level is less, if number 2 then the match level is moderate, if number 3 then the match level is high and if filled with number 4 then the match level is very high. The results of the whole dataset construction can be seen in Figure 2 below :



No	Nama_Dosen	Mata_Kuliah	Prodi	Isian_1	Isian_2	Isian_3	Isian_4	Isian_5	Isian_6	Isian_7	Isian_8	Kecocokan
1	Edy Kurniawan,ST., MT	Jaringan Komputer	Teknik Elektro	70	80	90	75	80	90	85	80	3
2	Moh. Muhsin, ST., M.Kom	Rangkaian Listrik	Teknik Elektro	50	70	10	20	50	50	20	10	1
3	Nanang Cendriono,M.Pd	Bahasa Indonesia	Teknik Elektro	100	100	100	100	100	100	100	100	4
4	Moh. Muhsin, ST., M.Kom	Rangkain Elektronika	Teknik Elektro	50	70	60	60	60	70	80	70	2
5	Nanang Cendriono,M.Pd	Bahasa Indonesia	Teknik Elektro	75	60	75	80	65	65	85	95	2
6	Moh. Muhsin, ST., M.Kom	Rangkaian Elektronika	Teknik Elektro	10	10	20	10	15	10	10	10	1
7	Moh. Muhsin, ST., M.Kom	Rangkaian Elektronika	Teknik Elektro	70	90	70	70	70	70	80	80	3
8	Prihma Shinta Utami,S.Pd., M.Pdi	Kewarganegaraan	Teknik Elektro	80	80	70	80	80	75	90	90	4
9	Moh. Muhsin, ST., M.Kom	Rangkaian Elektronika	Teknik Elektro	40	30	30	50	20	30	20	20	1
10	Moh. Muhsin, ST., M.Kom	Rangkaian Elektronika	Teknik Elektro	10	10	10	20	10	10	10	10	1
11	Moh. Muhsin, ST., M.Kom	Rangkaian Elektronika	Teknik Elektro	30	40	55	50	40	60	70	60	1
12	Moh. Muhsin, ST., M.Kom	Rangkaian Elektronika	Teknik Elektro	30	30	50	30	50	50	50	50	1
13	Didik Riyanto, ST., M.Kom	Fisika Terapan	Teknik Elektro	90	90	80	90	80	80	90	80	3
14	Moh. Muhsin, ST., M.Kom	Rangkaian Elektronika	Teknik Elektro	60	60	50	60	50	40	50	70	2
15	Moh. Muhsin, ST., M.Kom	Rangkaian Elektronika	Teknik Elektro	70	60	90	80	10	40	70	60	2
16	Moh. Muhsin, ST., M.Kom	Rangkaian Elektronika	Teknik Elektro	20	10	10	10	10	10	20	20	1
17	Didik Riyanto, ST., M.Kom	Fisika Terapan	Teknik Elektro	80	80	80	80	40	70	70	70	3
18	Moh. Muhsin, ST., M.Kom	Rangkaian Elektronika	Teknik Elektro	50	50	40	50	40	30	50	50	2
19	Moh. Muhsin, ST., M.Kom	Rangkaian Elektronika	Teknik Elektro	10	20	10	20	10	30	70	40	1

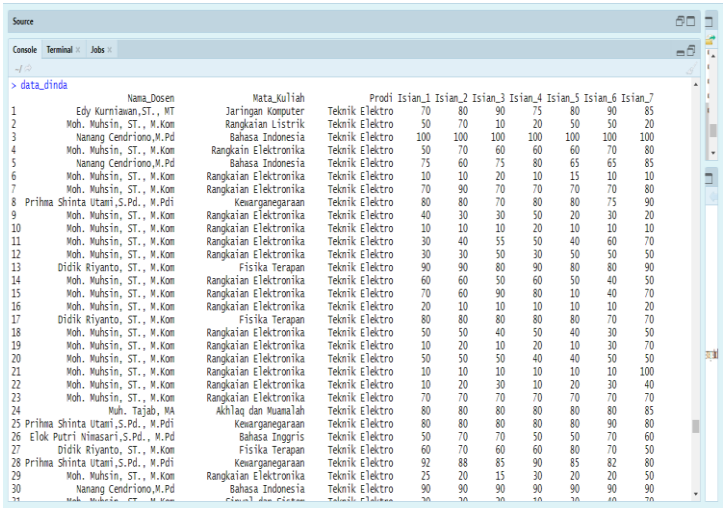
Figure 2. Dataset Construction

Then enter the processing stage of the dataset into RStudio, using the bnlearn library package in R Language with the following command :

```
# installing the package bnlearn
> install.packages("bnlearn")

# load package bnlearn
  >              library(bnlearn,quietly=TRUE,ver-
bose=FALSE,warn.conflicts = FALSE)
```

Then import the dataset above, then the result of importing the dataset can be seen in the following figure 3 :



```
> data_dinda
                    Nama_Dosen                 Mata_Kuliah          Prodi Isian_1 Isian_2 Isian_3 Isian_4 Isian_5 Isian_6 Isian_7
1            Edy Kurniawan,ST., MT             Jaringan Komputer Teknik Elektro      70      80      90      75      80      90      85
2            Moh. Muhsin, ST., M.Kom            Rangkaian Listrik Teknik Elektro      50      70      10      20      50      50      20
3              Nanang Cendriono,M.Pd             Bahasa Indonesia Teknik Elektro     100     100     100     100     100     100     100
4            Moh. Muhsin, ST., M.Kom          Rangkain Elektronika Teknik Elektro      50      70      60      60      60      70      80
5              Nanang Cendriono,M.Pd             Bahasa Indonesia Teknik Elektro      75      60      75      80      65      65      85
6            Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      10      10      20      10      15      10      10
7            Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      70      90      70      70      70      70      80
8  Prihma Shinta Utami,S.Pd., M.Pdi                Kewarganegaraan Teknik Elektro      80      80      70      80      80      75      90
9            Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      40      30      30      50      20      30      20
10           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      10      10      10      20      10      10      10
11           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      30      40      55      50      40      60      70
12           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      30      30      50      30      50      50      50
13         Didik Riyanto, ST., M.Kom                Fisika Terapan Teknik Elektro      90      90      80      90      80      80      90
14           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      60      60      50      60      50      40      50
15           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      70      60      90      80      10      40      70
16           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      20      10      10      10      10      10      20
17         Didik Riyanto, ST., M.Kom                Fisika Terapan Teknik Elektro      80      80      80      80      40      70      70
18           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      50      50      40      50      40      30      50
19           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      10      20      10      20      10      30      70
20           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      50      50      50      40      40      50      50
21           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      10      10      10      10      10      10     100
22           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      10      20      30      10      20      30      40
23           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      70      70      70      70      70      70      70
24                   Muh. Tajab, MA                Akhlaq dan Muamalah Teknik Elektro      80      80      80      80      80      80      85
25 Prihma Shinta Utami,S.Pd., M.Pdi                Kewarganegaraan Teknik Elektro      80      80      80      80      80      90      80
26  Elok Putri Nimasari,S.Pd., M.Pd                Bahasa Inggris Teknik Elektro      50      70      70      50      50      70      60
27         Didik Riyanto, ST., M.Kom                Fisika Terapan Teknik Elektro      60      70      60      60      80      70      50
28 Prihma Shinta Utami,S.Pd., M.Pdi                Kewarganegaraan Teknik Elektro      92      88      85      90      85      82      80
29           Moh. Muhsin, ST., M.Kom       Rangkaian Elektronika Teknik Elektro      25      20      15      30      20      20      50
30             Nanang Cendriono,M.Pd             Bahasa Indonesia Teknik Elektro      90      90      90      90      90      90      90
```

Figure 3. Import Dataset Results to R

The next step is to build a prediction model using naïve Bayes in R Language, and begin by translating all data types into factors, then create training data for prediction based on the identifier attribute. The plot results of the naïve bayes model can be seen in the following figure 4 :

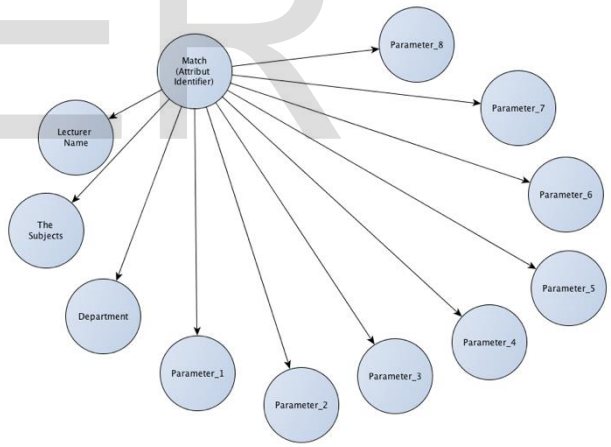


Figure 4. Naïve Bayes Model Plot

Furthermore, the Naive Bayes bn model can be used to predict the level of compatibility with the courses taught. The prediction is done with the following command :

```
pred = predict(bn, data_dinda_baru)      [ENTER]
```

To see how accurate the prediction results can be seen in two ways, namely the table () command and the cbind () command.

```
> table(pred, data_dinda_baru[, "Kecocokan"])
```

Then the prediction results can be seen in Figure 5 below:

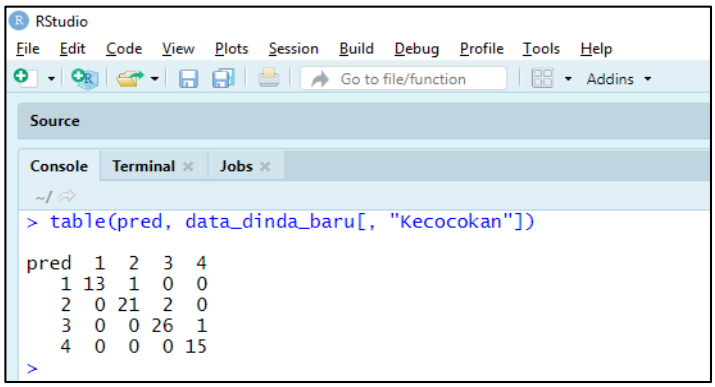


Figure 5. Overall Dataset Prediction Results

The picture above explains that:

a. From the original value 1 obtained 13 correct answers with the results of the prediction also 1. But there is 1 prediction result that is not 1 that is 2.
b. From the original value 2, 21 correct answers were obtained with the predicted results also 1. But there are 2 predictive results that are not 2 namely 3.
c. From the original value 3 obtained 26 correct answers with the results of the prediction also 3. But there is 1 prediction result that is not 3 that is 4.
d. From the original value 4 obtained 15 correct answers with the results of the prediction as well 4. There are no wrong prediction results.

If you see the prediction results of the test data, namely records 51 to 79 (for records 1 to 50 have been used as training data), we get a comparison between the original value and the prediction results in Figure 6 below:

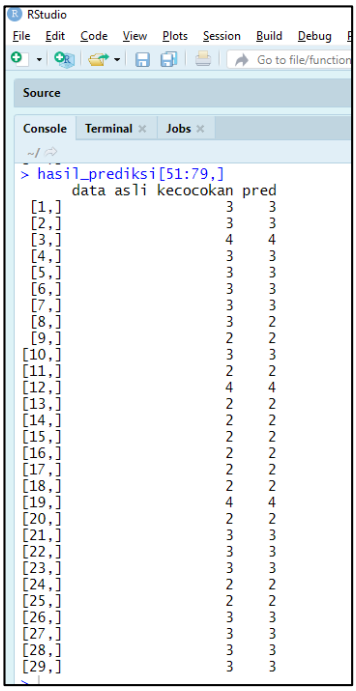


Figure 6. Test Data Prediction Results

5 Conclusion

This research was conducted by starting to construct existing data to be predicted by building a machine learning model with the Naïve Bayes algorithm to predict the suitability of lecturers with courses based on 8 (eight) indicators that have been made. In its processing, the data is predicted by using 4 (four) range of compatibility levels, then training the data for some initial data. After there is training data, all data will be predicted using machine learning. Prediction results there is only 1 error from 79 predicted data. This means, that the level of prediction accuracy of machine learning built there are many matches.

Acknowledgment

The authors thanks to LPPM Muhammadiyah University Ponorogo for supporting both the funds and morals for the ongoing research.

References

[1] Han, J., Kamber, M. & Pei, J., 2012. *Data Mining; Concepts and Techniques* 3rd ed. J. P. Jiawei Han, Micheline Kamber, ed., Waltham, USA: Morgan Kaufmann Publishers. Available at: www.elsevierdirect.com.
[2] Towell, G.G. & Shavlik, J.W., 1993. Extracting Refined Rules from Knowledge-Based Neural Networks. *Machine Learning*, 13(1), pp.71–101.
[3] Craven, M.W. & Shavlik, J.W., 1994. Using sampling and queries to extract rules from trained neural networks. *Machine Learning: Proceedings of the Eleventh International Conference*, pp.37–45.
[4] Duch, W., Adamczak, R. & Gra□bczewski, K., 2001. A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Transactions on Neural Networks*, 12(2), pp.277–306.
[5] Pawlak, Z., 1996. Rough sets and data analysis. *Soft Computing in Intelligent Systems and Information Processing. Proceedings of the 1996 Asian Fuzzy Systems Symposium*, 147, pp.1–12.
[6] Grzymala-busse, J.W., 2005. Rough Set Theory with Applications to Data Mining. *Real World Applications of Computational Intelligence*, Volume 179, p.pp 221-244.
[7] Greco, S. et al., 2001. An algorithm for induction of decision rules consistent with the dominance principle. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2005, pp.304–313.
[8] Marsono, M. N., El-Kharashi, M. W. and Gebali, F. (2006) 'Binary LNS-based na?ve Bayes hardware classifier for spam control', *Proceedings - IEEE International Symposium on Circuits and Systems*, pp. 3674–3677.
[9] Das, S. and Kolya, A. K. (2017) 'Sense GST: Text mining & sentiment analysis of GST tweets by Naive Bayes algorithm', *2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pp. 239–244. doi: 10.1109/ICRCICN.2017.8234513.
[10] Hsu, S.-C., Chen, I.-C. and Huang, C.-L. (2016) 'Image classification using pairwise local observations based Naive Bayes classifier', *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2015*, (December), pp. 444–452. doi: 10.1109/APSIPA.2015.7415311.
[11] Nirmala, K., Venkateswaran, N. and C, V. K. (2017) 'HoG Based Naive Bayes Classifier for Glaucoma Detection'. Proceedings of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017, pp. 2331–2336.