Prediction of SDG regions from sustainable development growth - 6 using Naïve Bayes Classification

1st Bipin Nair B J
(department of computer science)
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Mysuru, India
bipin.bj.nair@gmail.com

2nd Achaiah KK
(department of computer science)
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Mysuru, India
dishanachaiah@gmail.com

3rd Arjun S Pramod
(department of computer science)
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Mysuru, India
arjunspramod@gmail.com

4th Chethan L Reddy
(department of computer science)
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Mysuru, India
chethanlreddy@outlook.com

5th Darshan T C
(department of computer science)
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Mysuru, India
darshantejur@gmail.com

***Abstract*—The present study aims in Prediction of Sustainable Development Goal (SDG) regions based on various factors such as Countries, year, school levels, basic amenities, and other entities. SDG-6th goal focuses on ensuring the availability and sustainable management of water and sanitation for all, throughout every nations. In the proposed work, Naive Bayes classification algorithm is applied to predict the same. The secondary dataset (Drinking water, sanitation and hygiene in schools by country, 2000-2021) is collected from United Nations International Children's Emergency Fund (UNICEF) and the World Health Organization (WHO) repositories. The proposed work deals with Gaussian Naive Bayes classifier and has achieved highest accuracy at 92.15%.**

***Index Terms*—SDG 6, Naïve Bayesian, Machine Learning**

I. Introduction

Access to safe and sustainable water and sanitation is a key global concern and is the most basic human need for health and well-being, and is addressed in the United Nations' Sustainable Development Goals (SDGs). Billions of people will lack access to these basic services in 2030 unless progress quadruples. Demand for water is rising owing to rapid population growth, urbanization and increasing water needs from agriculture, industry, and energy sectors. Naive Bayes classification is a popular machine learning technique which is used for prediction which is a probabilistic algorithm that makes predictions based on the likelihood of each input feature belonging to class 'SDG regions'. Basically, there exists three different naive bayes classifiers, namely, Gaussian NB, Multinomial NB, and Bernouli's NB. The accuracies achieved are 92.15%, 58.61%, and 54.24% respectively. Hence Gaussian Naive Bayes classifier is used for the implementation.

II. Literature Review

P. Varalakshmi et al[1] worked on the Prediction of Water Quality Using Naive Bayesian Algorithm. The research gap is that they don't focus on its classification. The dataset used Around 100 samples were collected and analyzed from 6 municipalities. Once trained, the Naive Bayes model correctly 64 out of 68 cases, including cases with missing data. Fitriana Harahap et al[2]worked on the Implementation of the Naive Bayes Classification Method for Predicting Purchases. Limited research on its application to predicting consumer behavior in the automotive industry. The dataset used in the paper is a car purchase dataset. For the 20 car purchase data used in the test by the Naive Bayes method, a percentage of 75% was obtained for the accuracy of prediction.Fitriana Harahap et al[3]worked on the GPU-NB: A Fast CUDA-based Implementation of Naive Bayes.The paper argues that there is still a need for more efficient and effective algorithms[GPU-NB] that can handle the large volume of information to be processed in ADC. These collections are Medline, ACM, 20ng, Reuters ny, Webkb, and acl bin.They tested their method on 20 car purchase data and got a percentage of 75% K.Vembandasamy et al [4]worked on the Heart Diseases Detection Using the Naive Bayes Algorithm The lack of analyzing tools to provide effective test results with hidden information in the healthcare

industry is mentioned in the paper. The data used in the paper was collected from a leading institute in Chennai. The paper states that the proposed model was able to classify 86%. Iftakhar Mohammad Talha et al[5]worked on the Human Behaviour Impact on to Use of Smartphones with the Python Implementation Using a Naive Bayesian, a problem that has not been fully explored.The data was collected from three major sections: physical methods, virtual methods, and medical reports. A vast data set was trained by data to compare methods. Naive Bayes' theorem was 71% accurate in indicating the negative impact of human behavior.Tedy Setiadi et al[6]worked on the Implementation Of Naive Bayes Method In Food Crops Planting Recommendation.The research gap use of cropping patterns.The results showed an accuracy of 85.71%. Soheli Farhana et al[7]worked on the research Classification of Academic Performance for University Research Evaluation by Implementing a Modified Naive Bayes Algorithm. The research gap needs an efficient and accurate method to classify and predict the research and academic performances of university staff before evaluating the Research Assessment. The dataset used in this research work consists of hundreds of thousands of research data points from a large university. Naive Bayes with an accuracy of 96.15% and 94.23%, respectively. Simon Prananta Barus et al[8] conducted a study titled "Implementation of Naive Bayes Classifier-based Machine Learning to Predict and Classify New Students at Matana University" to address the research gap of utilizing data mining to optimize the achievement of marketing targets, particularly in predicting and classifying prospective student data at Matana University. The study used the dataset of prospective students who have registered at Matana University, and the accuracy rate achieved was 73%, which was helpful for the head of marketing in making marketing strategies. R. Reza El Akbar et al [9] conducted a study about The Implementation of Naive Bayes Algorithm for Classifying Tweets Containing Hate Speech with Political Motives. Limited research on the identification and classification of political hate speech in social media. A collection of tweets that were automatically drawn using a Twitter scraper. The tweets were then filtered and labeled before being classified using the Naïve Bayes Algorithm. The average accuracy of the Naïve Bayes Algorithm for classifying tweets containing hate speech with political motives is 93.4%. Fadli Shadiqin Thirafi et al[10]conducted a study Implementation of the Naïve Bayes Classifier Algorithm to Categorize Indonesian Song Lyrics Based on Age. The research gap in this study is that while several types of research classify songs. The dataset used in this study consists of 400 titles of Indonesian song lyrics obtained through crawling on the lirik.kapanlagi.com website. The accuracy of the Naive Bayes Classifier (NBC) algorithm used to tested using training data of 60%, 70%, 80%, and 90% from the overall dataset with equitable distribution. The resulting accuracies were 62.5%, 65%, 67.5%, and 67.5% respectively. Beta Priyoko et al[11] conducted a study titled "Implementation of Naive Bayes Algorithm for Spam Comments Classification on Instagram" to address the research gap of the need for a model to detect spam comments on Instagram. A dataset of comments collected randomly using a scraping technique from posts by public figures in Indonesia was manually classified into two classes: spam and not spam. The model test result showed an F1-measure of 0.83, recall of 0.98, and precision of 0.72, indicating a successful classification of spam comments on Instagram in this research. Muhammad Ilham Insani et al [12] conducted a study about Implementation of an Expert System for Diabetes Diseases using Naive Bayes and Certainty Factor Methods. The research gap diagnoses diabetes using computer systems that have been entered with a knowledge base and a set of rules to solve problems like an expert. The data used results from medical records of 100 patients from 2016 and 2017 who suffer from particular types of diabetes in RSUD Brendan Pekalongan. The accuracy rate data 70 training data and 30 testing data is equal to 100% according to the doctor's diagnosis. Hajer Kamel et al [13]conducted a study Cancer Classification Using Gaussian Naive Bayes Algorithm. There is a need for more effective diagnostic tools that can aid in the early detection and classification of different types of cancer. The algorithm is tested by applying it to two datasets in which the first is Wisconsin Breast Cancer dataset (WBCD) and the second is lung cancer dataset. The evaluation results of the proposed algorithm have achieved 98% accuracy in predicting breast cancer and 90% in predicting lung cancer. Abbi Nizar Muhammad et al[14]conducted a study Sentiment Analysis of Positive and Negative YouTube Comments Using the Naive Bayes – Support Vector Machine (NBSVM) Classifier. There is no mention of how this method compares to other sentiment analysis methods, such as lexicon-based approaches or deep learning models. The type of dataset used YouTube video comments. The combination of Naive Bayes and Support Vector Machine produces better accuracy levels and stronger performance with the use of a 7:3 scale of data that is 70% training data and 30% testing data. A. Mansour et al[15]conducted a study Performance Evaluation of Machine Learning Techniques for Mustard Crop Yield Prediction from Soil Analysis. There is a lack of research focusing specifically on predicting the mustard crop yield in the Jammu region using machine learning algorithms. The data for the experimental set-up was collected from the Department of Agriculture Department. The dataset consists of 5000 instances with 11 input parameters representing the soil nutrient status of the Jammu region and one output attribute. It achieves the maximum (99%) detection. Abbi Nizar Muhammad et al [16]conducted a study Accurate Detection of Covid-19 patients based on the Feature Correlated Naive Bayes (FCNB) classification strategy. There is a research gap in developing machine-learning models that can overcome these limitations and improve the accuracy and speed of COVID-19 detection. The FCNB classification strategy selects only the most effective features among the extracted features from laboratory tests for both COVID-19 patients and non-COVID-19 people by using the Genetic Algorithm as a wrapper method. Experiments show that maximum detection accuracy of 99%. Bipin Nair Balakrishnan Jayakumari et al[17] conducted a study on the

Classification of heterogeneous Malayalam documents based on structural features using deep learning models. There is a Research gap method to classify the documents based on content, as well as spectral features, that can be developed. This could be considered a research gap that can be explored further. Data collected from documents classified include the agreement. documents, notebook images, and palm leaves. The models attained high accuracies of 99.7%, 96%, and 95%, respectively.

III. Proposed Methodology

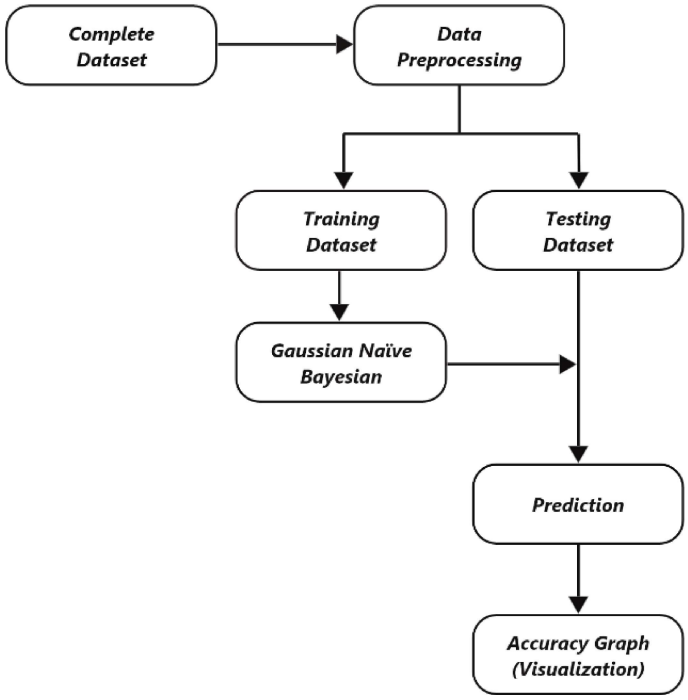

Fig. 1. Block diagram

The workflow of the project is represented in the above block diagram, which starts from Data acquisition phase that has **complete dataset** with redundant, null, and inconsistent values. It then undergoes **Data Pre-processing** phase at which dataset is filtered which is then split into **Training dataset** and **Testing dataset**. The model is trained with training dataset using **Gaussian Naive Bayes** classifier for **Prediction**. And the results are represented visually in graph and table.

A. Data acquisition

The data set used here for the implementation is a **Secondary data set** that is **'Drinking water, sanitation and hygiene in schools by country, 2000-2021'** which is collected from United Nations International Children's Emergency Fund (UNICEF) and the World Health Organization (WHO) repositories. The same repository has multiple data sets about Water, Sanitation and Hygiene with respect to healthcare facilities, schools, and households based on Countries, and regions.

```
Data columns (total 23 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   COUNTRY                         3780 non-null   object
 1   Year                            3780 non-null   float64
 2   School age population (thousands)  3780 non-null   object
 3   % urban                         3780 non-null   float64
 4   % pre-primary                   3780 non-null   float64
 5   % primary                       3780 non-null   float64
 6   % secondary                     3780 non-null   float64
 7   national_basic                  3780 non-null   float64
 8   national_limited                3780 non-null   float64
 9   national_no                     3780 non-null   float64
 10  PRIMARY_basic                   3780 non-null   float64
 11  PRIMARY_limited                 3780 non-null   float64
 12  PRIMARY_no                      3780 non-null   float64
 13  SECONDARY_basic                 3780 non-null   float64
 14  SECONDARY_limited               3780 non-null   float64
 15  SECONDARY_no                    3780 non-null   float64
 16  SDG region                      3780 non-null   object
 17  WHO Region                      3591 non-null   object
 18  UNICEF Programme Region         3066 non-null   object
 19  UNICEF Reporting Regions        3675 non-null   object
 20  UNESCO UIS Regions              3780 non-null   object
 21  UNESCO GEMR Regions             3738 non-null   object
 22  iso3 code                       3780 non-null   object
dtypes: float64(14), object(9)
```

Fig. 2. Original data set information

B. Pre-processing

Initially, the data set was consisting of 3781 rows X 32 columns.Cleaned the dataset using the following methods:

- Removal of null values and irrelevant columns.
- Applied mean for essential columns containing null values
- Converted string values into numerical values using dictionary by indexing, for the data in 5 columns such as Country, WHO Region, UNICEF Programme Region, UNICEF Reporting Regions, and SDG region in order to avoid conflicts that were going to be present otherwise.
- Limited the values in the dataset from the year 2016 to 2021 since there existed occurrences many null values in most of the columns between the years 2001 and 2015 which would have caused irregularities in the dataset, and inefficiency of the model.
- It resulted in a data size of 510 rows X 19 columns with zero null values.

	COUNTRY	Year	School age population (thousands)	% urban	% primary	% secondary	national_basic	national_limited	national_no
16	0	2016.0	12465	25.0	49.0	42.0	65.0	8.000000	25.0
17	0	2017.0	12703	25.0	48.0	42.0	65.0	9.000000	24.0
18	0	2018.0	12907	25.0	48.0	42.0	65.0	11.000000	23.0
19	0	2019.0	13092	25.0	48.0	43.0	65.0	12.000000	21.0
20	0	2021.0	13417	26.0	47.0	43.0	65.0	12.000000	21.0
58	1	2016.0	8501	71.0	42.0	47.0	92.0	15.381579	7.0
59	1	2017.0	8791	72.0	43.0	46.0	92.0	15.381579	7.0
60	1	2018.0	9115	72.0	44.0	45.0	92.0	15.381579	7.0
61	1	2019.0	9465	73.0	44.0	45.0	91.0	15.381579	7.0
62	1	2021.0	9921	74.0	45.0	48.0	90.0	1.000000	7.0

Fig. 3. Filtered data set part 1

SECONDARY_basic	SECONDARY_limited	SECONDARY_no	WHO Region	UNICEF Programme Region	UNICEF Reporting Regions	SDG region
75.0	16.292857	20.820513	0	0	0	0
75.0	16.292857	20.820513	0	0	0	0
75.0	16.292857	20.820513	0	0	0	0
75.0	16.292857	20.820513	0	0	0	0
75.0	16.292857	20.820513	0	0	0	0
98.0	16.292857	1.000000	1	1	1	1
96.0	1.000000	1.000000	1	1	1	1
95.0	3.000000	1.000000	1	1	1	1
94.0	4.000000	1.000000	1	1	1	1
92.0	6.000000	1.000000	1	1	1	1

Fig. 4. Filtered data set part 2

C. Proposed algorithm

The implementation is carried out on pre-processed data in the following steps:

- Step 1: Imported essential Python libraries such as pandas (to clean the dataset), numpy (to deal with multi-dimensional arrays), sklearn (to perform Naïve Bayes classification), and matplotlib (for providing visualization), accuracy_score for finding out the accuracy of the predictions made.
- Step 2: Converted 4 columns' data (objects) into numerical labels as per requirements using the LabelEncoder library.
- Step 3: Split the data into training and testing data in different sizes using sklearn.model_selection.train_test_split() for random selection of data from the dataset.
- Step 4: Trained the model using the Gaussian Naïve Bayes classifier, since it is optimal for this implementation, and it yielded better accuracy than Bernoulli and Multinomial naïve Bayes classifiers.
- Step 5: Tested the classifier on the testing set, and obtain accuracy.
- Step 6: For creating the accuracy graph, we had to test the data with varying training and testing data size proportions from 10% to 90%. The highest accuracy of 92.15% is achieved at a 30% size split.
- Step 7: Using matplotlib. pyplot library, the visualization is achieved.

D. Mathematical model

1. Bayes' Theorem:

$$P(y|x) = P(x|y) * P(y)/P(x) \quad (1)$$

- $P(y \mid x)$ is the posterior probability of class y given the features x
- $P(x \mid y)$ is the likelihood of observing features x given class y
- P(y) is the prior probability of class y
- P(x) is the probability of observing features x

E. Experimental result

Prediction using Gaussain NB has achieved its highest accuracy of 92.15% at splitting proportion of 30% size of the dataset for testing. And in following proportions, it shows quite promising accuracies until 60%-70% proportions because of even distribution of class in the dataset, and classes might be easier to predict for the model based on the given features. Basically there are 6 SDG regions namely, 'Central and Southern Asia, Northern Africa and Western Asia, Sub-Saharan Africa, Latin America and the Caribbean, Europe and Northern America, Eastern and South-Eastern Asia, and Oceania'. As the result, the countries of testing dataset and their predicted class respectively are provided, based on which the accuracy is calculated as an evaluation metric and is represented in a tabular form and graphically as well in figures 6 & 7 respectively. A sample of results are represented the figure 5. **Sample of predicted output** in which Country of a predicted class along with the respective predicted SDG regions are mentioned. Hence, based on these, a evaluation metric 'Accuracy' is used for model evaluation.

```
Predicted SDG region for ['Afghanistan']: ['Eastern and South-Eastern Asia']
Predicted SDG region for ['Burkina Faso']: ['Sub-Saharan Africa']
Predicted SDG region for ['Angola']: ['Eastern and South-Eastern Asia']
Predicted SDG region for ['India']: ['Eastern and South-Eastern Asia']
Predicted SDG region for ['Antigua and Barbuda']: ['Northern Africa and Western Asia']
Predicted SDG region for ['Mozambique']: ['Eastern and South-Eastern Asia']
Predicted SDG region for ['Botswana']: ['Northern Africa and Western Asia']
Predicted SDG region for ['Jordan']: ['Northern Africa and Western Asia']
```

Fig. 5. Sample of Predicted output

Training data size (in %)	Testing data size (in %)	Accuracy
90	10	90.1961
80	20	91.1765
70	30	92.1569
60	40	89.7059
50	50	89.0196
40	60	91.1765
30	70	91.5966
20	80	81.8627
10	90	53.159

Fig. 6. Table representing accuracy with respect to different split proportions.

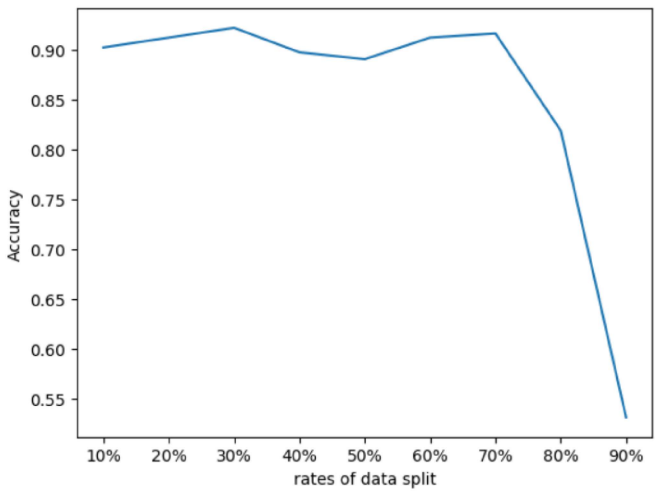

Fig. 7. Graphical representation.

IV. Conclusion

The proposed implementation of Naïve Bayes classification on the SDG 6 dataset has achieved an accuracy of 92.15%. The drawback of this model as it works better for numeric data, in case of categorical data it must be converted to numerical data. Further research could focus on improving the accuracy of our model and exploring other machine-learning approaches to predict the success of interventions aimed at achieving SDG 6.

Acknowledgment

We would like to express our sincere thanks to Mr. Bipin Nair B J, Asst. Professor, Department of Computer Science, for giving us the opportunity to work on the project and for his excellent guidance and valuable suggestions rendered throughout the project.

References

[1] Varalakshmi, P., Vandhana, S., & Vishali, S. (2017, January). Prediction of water quality using Naive Bayesian algorithm. In 2016 Eighth International Conference on Advanced Computing (ICoAC) (pp. 224-229). IEEE.

[2] Harahap, F., Harahap, A. Y. N., Ekadiansyah, E., Sari, R. N., Adawiyah, R., & Harahap, C. B. (2018, August). Implementation of Naïve Bayes classification method for predicting purchase. In 2018 6th International Conference on Cyber and IT Service Management (CITSM) (pp. 1-5). IEEE.

[3] Andrade, G., Viegas, F., Ramos, G. S., Almeida, J., Rocha, L., Gonçalves, M., & Ferreira, R. (2013, October). GPU-NB: a fast CUDA-based implementation of naive bayes. In 2013 25th International Symposium on Computer Architecture and High-Performance Computing (pp. 168-175). IEEE.

[4] Subbalakshmi, G., Ramesh, K., & Rao, M. C. (2011). Decision support in heart disease prediction system using naive bayes. Indian Journal of Computer Science and Engineering (IJCSE), 2(2), 170-176

[5] Talha, I. M., Salehin, I., Debnath, S. C., Saifuzzaman, M., Moon, N. N., & Nur, F. N. (2020, July). Human behavior impact to use of smartphones with the Python implementation using naive Bayesian. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[6] Setiadi, T., Noviyanto, F., Hardianto, H., Tarmuji, A., Fadlil, A., & Wibowo, M. (2020). Implementation of naïve bayes method in food crops planting recommendation. Int. J. Sci. Technol. Res, 9(02), 4750-4755.

[7] Farhana, S. (2021). Classification of Academic Performance for University Research Evaluation by Implementing Modified Naive Bayes Algorithm. Procedia Computer Science, 194, 224-228.

[8] Barus, S. P. (2021, March). Implementation of Naïve Bayes Classifier-based Machine Learning to Predict and Classify New Students at Matana University. In Journal of Physics: Conference Series (Vol. 1842, No. 1, p. 012008). IOP Publishing.

[9] El Akbar, R. R., Shofa, R. N., & Paripurna, M. I. (2019, August). The implementation of Naïve Bayes algorithm for classifying tweets containing hate speech with political motive. In 2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC) (pp. 144-148). IEEE.

[10] Thirafi, M. F. S., & Rahutomo, F. (2018, November). Implementation of naïve bayes classifier algorithm to categorize indonesian song lyrics based on age. In 2018 International Conference on Sustainable Information Engineering and Technology (SIET) (pp. 106-109). IEEE

[11] Priyoko, B., & Yaqin, A. (2019, July). Implementation of Naive Bayes algorithm for spam comments classification on Instagram. In 2019 International Conference on Information and Communications Technology (ICOIACT) (pp. 508-513). IEEE.

[12] Insani, M. I., Alamsyah, A., & Putra, A. T. (2018). Implementation of Expert System for Diabetes Diseases using Naïve Bayes and Certainty Factor Methods. Sci. J. Informatics, 5(2), 185-193.

[13] Kamel, H., Abdulah, D., & Al-Tuwaijari, J. M. (2019, June). Cancer classification using gaussian naive bayes algorithm. In 2019 International Engineering Conference (IEC) (pp. 165-170). IEEE.

[14] Muhammad, A. N., Bukhori, S., & Pandunata, P. (2019, October). Sentiment analysis of positive and negative of youtube comments using naïve bayes–support vector machine (nbsvm) classifier. In 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE) (pp. 199-205). IEEE.

[15] Pandith, V., Kour, H., Singh, S., Manhas, J., & Sharma, V. (2020). Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. Journal of scientific research, 64(2), 394-398.

[16] Mansour, N. A., Saleh, A. I., Badawy, M., & Ali, H. A. (2022). Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy. Journal of ambient intelligence and humanized computing, 1-33.

[17] Balakrishnan Jayakumari, B. N., & Kavana, A. T. (2023). Classification of heterogeneous Malayalam documents based on structural features using deep learning models. International Journal of Electrical & Computer Engineering (2088-8708), 13(1).