IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727 PP 79-86
www.iosrjournals.org

A Comparative Study To Identify The Best Machine Learning Algorithms

P. Arumugam[1] A. PoomPavai[2], Manimannan G[3]

[1.] *Assistant Professor, Department of Statistics, Annamalai University, Chidambaram.*
[2.] *Assistant Professor, Department of Statistics, SDNB Vaishnav College for Women, Chrompet, Chennai.*
[3] *Assistant Professor, Department of Statistics, Apollo Arts and Science College Chennai, Guduvanchery, Chennai*

ABSTRACT
This paper aims to identify and evaluate three classification methods based on accuracy and kappa statistics, and to visualize them with different levels of rainfall data collected from the India Meteorological Department. The purpose of this research is to determine which classifier is the most effective. The accuracy of various classifiers for the southern states of India is compared and the sensitivity, specificity, accuracy, true positive rate, and false positive rate of each classifier for all states are calculated. Furthermore, a comparison of kappa statistics is conducted by using a confusion matrix. To analyze the performance of the most popular classification techniques, the training dataset is used to train the classifier using classification and regression. The accuracy of the Naïve Bayes approach, K Nearest Neighbor algorithm and SVM are tested on the test dataset, and the results show that the SVM model has the best performance. The Naïve Bayes Classifier has also performed well, but the KNN algorithm did not. The True Positive Rate and False Positive Rate table reveals a true positive rate of more than 70% and a false positive rate of less than 30% for the datasets. Finally, a Kappa statistic analysis between different classes is conducted, and a higher value of Kappa statistic is considered a good result.

Keywords: *Naïve Bayes Classifier, k-Nearest Neighbor Algorithm, Support Vector Machine, Confusion Matrix, Precision and Kappa Statistics.*

I. INTRODUCTION

Predicting rainfall is of utmost importance, as heavy and erratic rainfall can cause destruction of crops and farms, and damage to property. To reduce the risks to life and property, while managing agricultural farms, an effective predictive model is essential. Conventional methods are not as efficient and so, using Machine Learning techniques can produce more accurate results. This can be done by analyzing historical data of rainfall and predicting rainfall for future seasons. Different methods, such as classification and regression, can be used depending on the requirements. The model can then be evaluated by calculating the error between the actual and the prediction, and the accuracy. It is important to select the right algorithm and model it according to the requirements, as different techniques offer different levels of accuracy.

Classification is a discipline of machine learning that attempts to discover a function that effectively describes and differentiates between classes of data, which can then be used to predict objects of unknown classification. This process is defined as the determination of a set of models or functions to define labels of unknown objects and to differentiate between classes of concepts or data. Several machine learning-based modeling techniques have been developed to aid in the classification process, including Artificial Neural Networks (McCulloch and Pitts, 1943), Classification and Regression Trees (Brieman, Friedman, Olshen and Stone, 1984), Multivariate Adaptive Regression Splines (Friedman, 1991), K-Nearest Neighbors (Fix and Hodges, 1951), Support Vector Machines (Cortes and Vapnik, 1995), and more.

II. REVIEW OF LITERATURE

The paper by H Sine and H Kuzwando shows that Support Vector Machine (SVM) method can be used for classification of time series data. The results show that the accuracy ratio of the prediction results is high when it is compared to the training and test data. Therefore, classification of rainfall data using SVM method has very good performance. The aim is to develop a model for long-term rainfall forecasting from a training data set. Decision tree classification is one of the best machine learning algorithms which is structured like a tree. B.Revathi and C.Usharani used CART and IDA decision tree algorithms to forecast rainfall dataset. These algorithms provide the highest predictive accuracy when performance metrics are used.

They proposed to design cost-sensitive machine learning algorithms to model the learning and diagnosis process. Clinical tests, similar to attributes in machine learning, can have their values empirically

evaluated in order to determine the cost of various testing strategies. Ezekiel T. Ogidan, KamilDimililer, and YoneyKirsal Ever applied the results to real-world diagnostic tasks. Deepti Gupta and UdayanGhose discussed the advantages and limitations of all algorithms, making it difficult to decide which is the best. Naive Bayes and Decision Trees are easy to understand and work with, however, tree pruning using cost-complex methods require intensive computations and can be time consuming. Neural networks, however, provide better results than the other discussed algorithms.

III. DATABASE

Data for the study was collected from the Department of Economics and Statistics in Chennai during the period 1901 to 2020, with the parameter of twelve months. The Southern States of India Rainfall Database from 1901 to 2020 is an extensive collection of data on rainfall in the Southern States of India, containing information on the quantity of rainfall in each state, the average annual rainfall, the number of monsoon months, and the seasonal distribution of rainfall. These secondary sources of data were gathered from various sources such as the Indian Meteorological Department, the Central Water Commission, and the Ministry of Earth Sciences. This data is being used to evaluate the impact of climate change on the region, and to inform agricultural and water management decisions. The database is updated annually and is an invaluable resource for researchers and policymakers.

IV. METHODOLOGY

4.1 Classification Training and Testing Model

The process of classification starts by using a learning algorithm on the training dataset, forming classification rules. These rules are then used to test data in order to assess the accuracy of the algorithm. The accuracy of a classifier is the percentage of samples or instances that have been correctly classified(Figure 1).

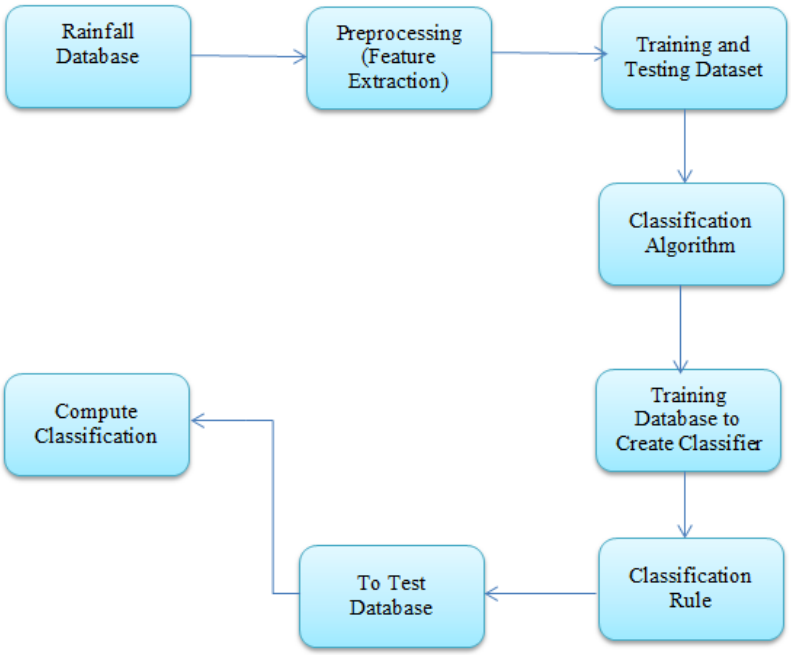

Figure 1. Classification Training and Testing Model

4.2NAÏVE BAYES ALGORITHM

The Naïve Bayes Algorithm is composed of two terms, Naïve and Bayes. Naïve implies that the occurrence of a particular feature is assumed to be independent of the occurrence of other features. Bayes refers to the principle of Bayes theorem, also known as Bayes' rule or Bayes' rule, which is used to calculate the probability of a hypothesis given prior knowledge. It is based on the concept of conditional probability, which is expressed in the following formula:

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right) * P(A)}{P(B)}$$

Where,$P\left(\frac{A}{B}\right)$is the posterior probability, $P\left(\frac{B}{A}\right)$is the probability of a hypothesis being true given the evidence$P(A)$is the prior probability, and $P(B)$) is the marginal probability.

4.3SUPPORT VECTOR MACHINE ALGORITHM

Support vector machines (SVMs) are a group of supervised learning algorithms used for classification, regression, and outlier detection. Unlike other classification algorithms, SVMs aim to produce a boundary which maximizes the distance between the closest data points of all categories. This boundary is referred to as the maximum margin classifier or maximum margin hyperplane. The SVM classifier works by drawing a line between the two classes. This means that each data point on one side of the line represents one class and data points on the other side of the line represent another class. SVMs are used for tasks such as handwritten digit recognition, intrusion detection, face recognition, email classification, gene classification, and medical diagnosis (Figure 2).

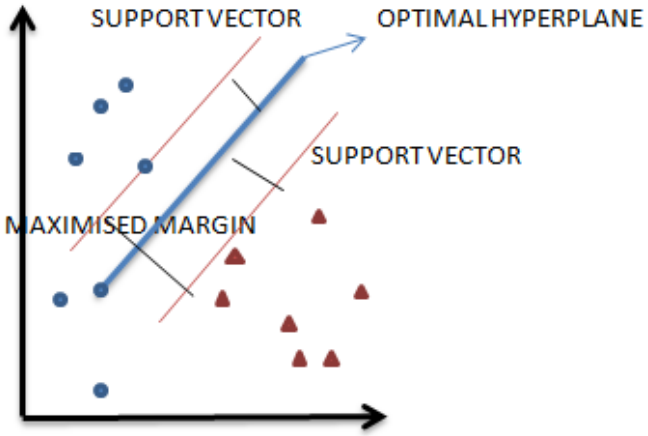

Figure 2. Support Vector Machine Algorithm

4.4 k-NEAREST MEIGHBOR ALGORITHM

K-Nearest Neighbor is one of the simplest machine learning algorithms based on supervised learning. The k-NN algorithm looks at the similarity between the new data point and the existing data points and assigns the new data point to the most similar group. The k-NN algorithm can be used for regression and classification, but primarily it is used for classification problems. k-NN is a non-parametric algorithm, meaning it does not make any assumptions about the underlying data. It is also known as a lazy learning algorithm because it does not learn from the training set immediately, but instead, during the training phase, it stores the data set, and when new data is received, it classifies that data into the group most similar to itself.

V. RESULT AND DISCUSSION

A unique confusion matrix was derived to calculate sensitivity, specificity, and accuracy. This matrix is a representation of the classification results, with the upper left cell indicating the number of samples that were correctly classified as true (i.e., TP). The upper right cell indicates the number of samples that were incorrectly classified as true (i.e., FN), and the lower left cell indicates the number of samples that were incorrectly classified as true (i.e., FP). (Table 1)

Table 1.Confusion Matrix

Actual/predicted	**0**	1
0	TP	FN
1	FP	TN

5.1SENSITIVITY, SPECIFICITY AND ACCURACY

The present work used the formulae of sensitivity, specificity and accuracy to calculate the performance analysis of five different data mining classifiers for rainfall datasets of southern states of India.

Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

Accuracy = (TP + TN) / (TP + FP + TN + FN)

The results were obtained in the R language environment in the form of tables and graphs and were defined by parameters such as sensitivity, precision and kappa statistic.

5.2COMPARISON OF SENSITIVITY AND SPECIFICITY

The sensitivity and specificity of three different machine learning algorithms for South Indian states are visualized and tabulated below in Table 2 and Figure 3..

ALGORITHMS	STATES	SENSITIVITY	SPECIFICITY
KNN ALGORITHM	ANDAMAN & NICOBAR	0.862	0.984
	COASTAL ANDHRA PRADESH	0.406	0.926
	COASTAL KARNATAKA	1.000	1.000
	KERALA	0.882	0.996
	LAKSHADWEEP	0.824	0.938
	NORTH INTERIOR KARNATAKA	0.436	0.941
	SOUTH INTERIOR KARNATAKA	0.471	0.942
	TAMIL NADU	0.800	0.971
	TELANGANA	0.500	0.935
NAÏVE BAYES	ANDAMAN & NICOBAR	0.889	0.949
	COASTAL ANDHRA PRADESH	0.591	0.929
	COASTAL KARNATAKA	0.941	1.000
	KERALA	0.683	0.987
	LAKSHADWEEP	0.857	0.953
	NORTH INTERIOR KARNATAKA	0.697	0.967
	SOUTH INTERIOR KARNATAKA	0.595	0.966
	TAMIL NADU	0.938	0.996
	TELANGANA	0.649	0.975
SVM	ANDAMAN & NICOBAR	0.955	0.969
	COASTAL ANDHRA PRADESH	0.583	0.932
	COASTAL KARNATAKA	0.941	1.000
	KERALA	0.757	0.987
	LAKSHADWEEP	0.905	0.957
	NORTH INTERIOR KARNATAKA	0.750	0.971
	SOUTH INTERIOR KARNATAKA	0.710	0.967
	TAMIL NADU	0.912	1.000
	TELANGANA	0.675	0.987

Table 2. Comparison of Sensitivity and Specificity

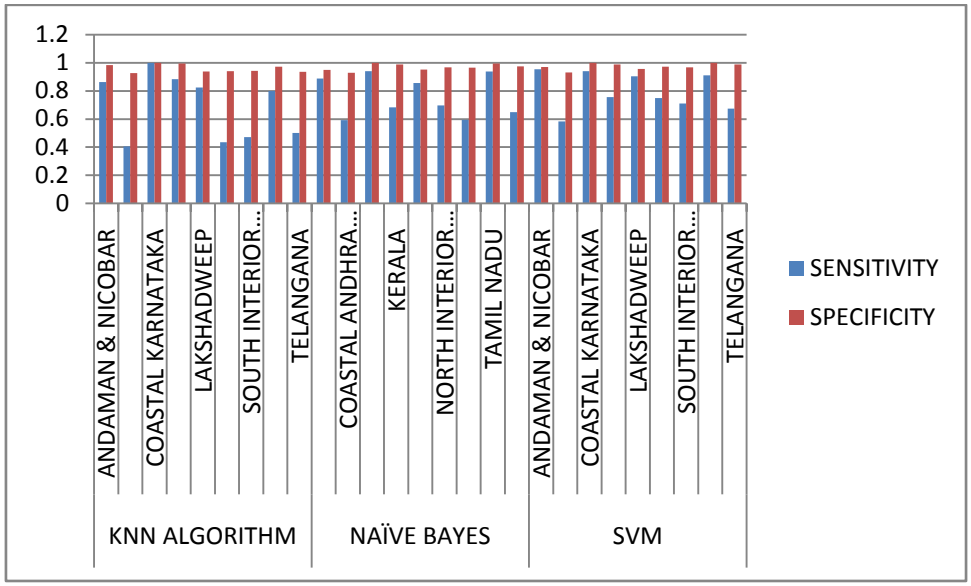

Figure3. Comparison of Sensitivity and Specificity

The researcher compared three data mining classifiers on a rainfall database dataset based on their sensitivity, specificity and accuracy, and found that the SVM classifier had the best classification accuracy (Table 3 and Figure 4).

ALGORITHMS	ACCURACY	SENSITIVITY	SPECIFICITY
KNN ALGORITHM	0.6727	0.686749	0.959199
NAÏVE BAYES	0.7491	0.759862	0.96914
SVM	0.7927	0.798571	0.974596

Table 3.Comparison of Machine Learning Algorithms

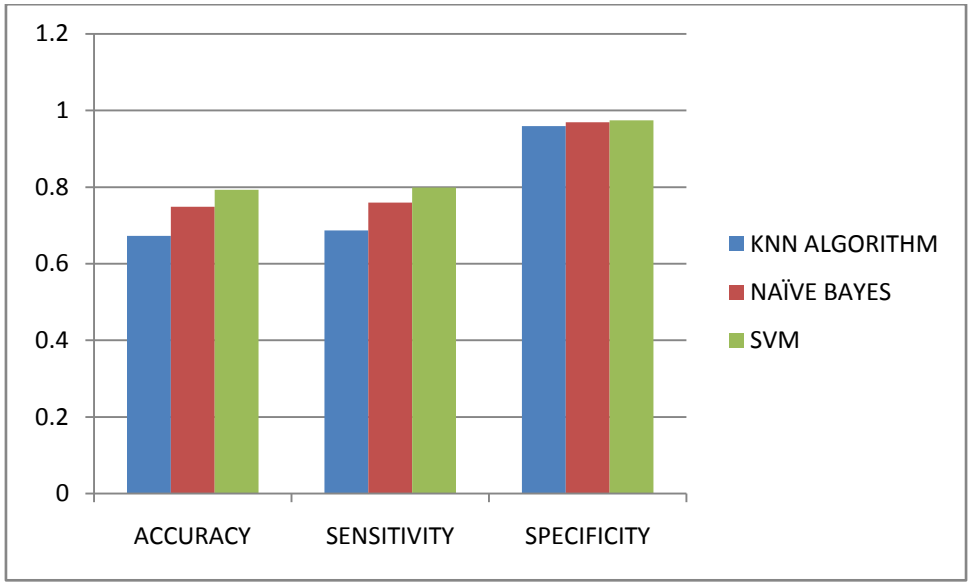

Figure 4.Graphical representations of sensitivity, specificity and accuracy

5.3 ASSOCIATION OF TRUE POSITIVE RATE AND FALSE POSITIVE RATE

Table 4 and Figure 5 provide a graphical comparison of the true positive rate and false positive rate for the southern states of India with different algorithms.

ALGORITHMS	STATES	TRUE POSITIVE RATE	FALSE POSITIVE RATE
KNN ALGORITHM	ANDAMAN & NICOBAR	0.862	0.138
	COASTAL ANDHRA PRADESH	0.406	0.594
	COASTAL KARNATAKA	1.000	0.000
	KERALA	0.882	0.118
	LAKSHADWEEP	0.824	0.176
	NORTH INTERIOR KARNATAKA	0.436	0.564
	SOUTH INTERIOR KARNATAKA	0.471	0.529
	TAMIL NADU	0.800	0.200
	TELANGANA	0.500	0.500
NAÏVE BAYES	ANDAMAN & NICOBAR	0.889	0.111
	COASTAL ANDHRA PRADESH	0.591	0.409
	COASTAL KARNATAKA	0.941	0.059
	KERALA	0.683	0.317
	LAKSHADWEEP	0.857	0.143
	NORTH INTERIOR KARNATAKA	0.697	0.303
	SOUTH INTERIOR KARNATAKA	0.595	0.405
	TAMIL NADU	0.938	0.063
	TELANGANA	0.649	0.351
SVM	ANDAMAN & NICOBAR	0.955	0.045
	COASTAL ANDHRA PRADESH	0.583	0.417
	COASTAL KARNATAKA	0.941	0.059
	KERALA	0.757	0.243
	LAKSHADWEEP	0.905	0.095
	NORTH INTERIOR KARNATAKA	0.750	0.250
	SOUTH INTERIOR KARNATAKA	0.710	0.290
	TAMIL NADU	0.912	0.088
	TELANGANA	0.675	0.325

Table 4. Comparison of True Positive Rate (TPR) and False Positive Rate (FPR)

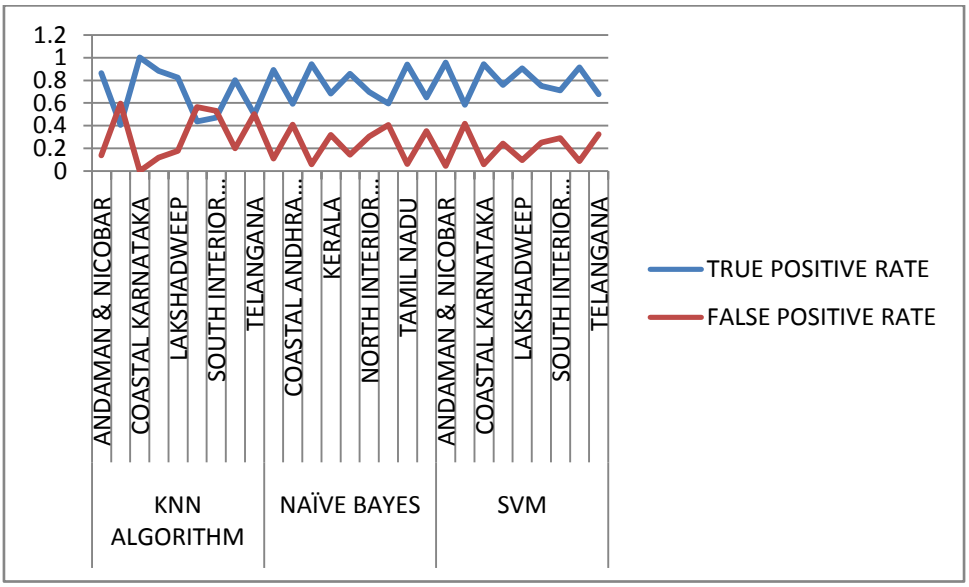

Figure 5. Graphical Comparison of True Positive Rate and False PositiveRate

Table 5 displays the True Positive Rate and False Positive Rate for the Support Vector Machine (SVM), Naive Bayes, and KNN Algorithms.

ALGORITHMS	TRUE POSITIVE RATE	FALSE POSITIVE RATE
KNN ALGORITHM	0.686749	0.313251
NAÏVE BAYES	0.759862	0.240138
SVM	0.798571	0.201429

Table 5.True Positive Rate and False Positive Rate

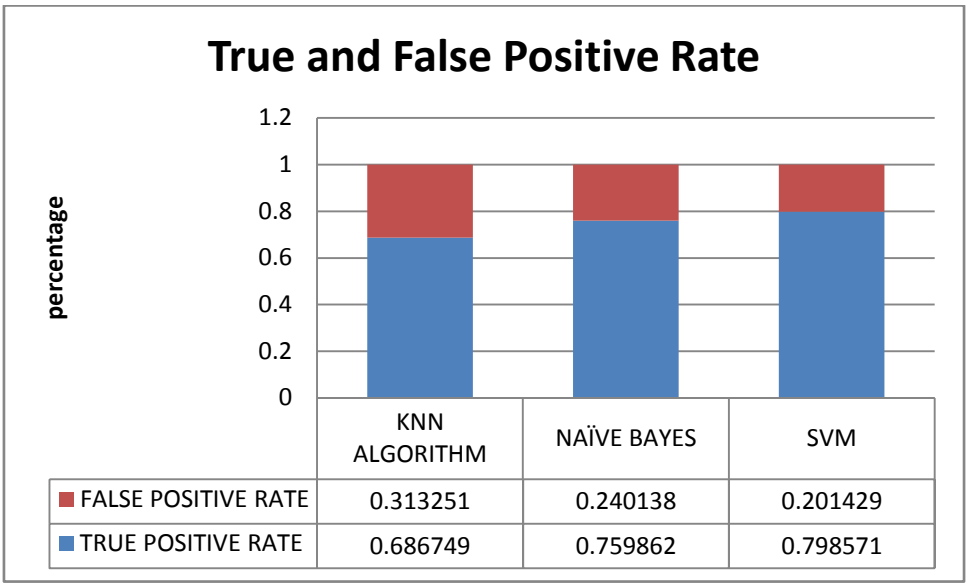

Figure6. Graphical Representation of TPR and FPR

The results indicate that SVM outperforms naive Bayes models in terms of parameters such as sensitivity, specificity, accuracy, and error rates. Additionally, Figure 6 provides a graphical representation of the true positive rate and false positive rate.

5.4EVALUATION OF ACCURACY AND BALANCED ACCURACY

Table 6 shows the comparison of precision and balance accuracy, along with their summary statistics.

ALGORITHMS	ACCURACY	BALANCED ACCURACY
KNN ALGORITHM	0.6727	0.822978
NAÏVE BAYES	0.7491	0.864507
SVM	0.7927	0.886523

Table 6.Comparison of Accuracy and Balanced Accuracy

5.5CALCULATION OF KAPPA STATISTICS

The following table shows the compression of Kappa statistics for k-NN, SVM and Naïve Bayes classifier results. Kappa statistics values closer to 1 indicate that the model is good. In this study, the three different ML algorithms achieved 0.631, 0.717 and 0.766 respectively. This result shows that the Kappa Statistics for SVM and Naive Bayes is high and k-NN is moderate. Overall, the ML Model is effective and good.

ALGORITHMS	KAPPA STATISTIC
KNN ALGORITHM	0.6317
NAÏVE BAYES	0.7176
SVM	0.7667

Table 7.Kappa statistic ML Algorithm

VI. CONCLUSION

This section focuses on various seasons of rainfall database from 1901 to 2020 by using Machine Learning (ML) classification Algorithm and studying each of them. Three classification methods are based on accuracy and kappa statistics and they are visualized with different levels of rainfall data collected from the India Meteorological Department. The purpose of this research is to determine which classifier is the most effective. The accuracy of various classifiers for the southern states of India is compared and the sensitivity, specificity, accuracy, true positive rate, and false positive rate of each classifier for all states are calculated. Furthermore, a comparison of kappa statistics is conducted by utilizing a confusion matrix. To analyze the performance of the most popular classification techniques, the training dataset is used to train the classifier using classification and regression. The accuracy of the Naïve Bayes approach, K Nearest Neighbor algorithm and Support Vector Machine (SVM) are tested on the test dataset, and the results show that the SVM model has the best performance. The Naïve Bayes Classifier has also performed well, but the KNN algorithm did not. The True Positive Rate and False Positive Rate table reveals a true positive rate of more than 70% and a false positive rate of less than 30% for the datasets.

REFERENCES

[1]. Chaovalitwongse W and Pardalos P M 2008 Cybernetics and Systems Analysis 44 125–138.

[2]. Orsegino C and Vercellis C 2010 Pattern Recognition 43 3787–3794.

[3]. Sain H and Purnami S Wulan 2015 In Procedia Computer Science 72 59-66.

[4]. B. EmilcyHern´andez, *et. al*. ," Rainfall Prediction: A Deep Learning Approach", this work is partially supported by the MINECO/FEDERTIN2012-36586-C03-01 of the Spanish Government,2012.

[5]. Ezekiel T. Ogidan, KamilDimililer, YoneyKirsal Ever," Machine Learning for Expert Systems in Data Analysis", DOI 10.1109/ISMSIT.2018.8567251,Oct 2019.

[6]. Deepti Gupta, UdayanGhose, "A Comparative Study of Classification Algorithms for forecasting rainfall", 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO),IEEE 2015.

[7]. H.Sain, H. Kuswanto, S.W. Purnami, and S.P. Rahaya, "Classification of rainfall data using support vector machine", Journal of physics: Conference Series, Volume:1763, The 2nd International Seminar on Science and Technology 2020, 16-17.

[8]. Nana Kofi AhoiAppiah-Badu, Yaw MarfoMissah,et al, "Rainfall Prediction using Machine Learning Algorithms for the various Ecological zones of Ghana" , IEEE Access 2022, pages: 5069-5082.

[9]. B.Revathi, C. Usharani, "Rainfall Prediction Using Machine Learning Classification Algorithms", International Journal of Creative Research Thoughts (IJCRT), Volume 9, Issue 1, January 2021.

[10]. OzlemTerzi, (2012), "Monthly Rainfall Estimation Using Data-Mining Process", Hindawi Publishing Corporation Applied Computational Intelligence and Soft Computing Volume 2012, 6 pages.

[11]. Edwin A. Roehl, Andrew M. O'Reilly, Paul A. Conrads, and Ruby C. Daamen, (2012), "Data Mining to Predict Climate and Groundwater Use Impacts on the Hydrology of Central Florida", South Carolina Water Resources Conference, 10–11 October 2012 Columbia, South Carolina.

[12]. SuhailaZainudin, Dalia Sami Jasim, and Azuraliza Abu Bakar, (2016), "Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction", International Journal of Advance Science Engineering Information Technology, Vol.6 (2016) No. 6 1148-1153.

[13]. RikoHerwanto,PPRosyanaFitriaPurnomo,PPSriyanto, (2017), "Rainfall Prediction Using Data Mining Techniques", 3rd International Conferences on Information Technology and Business (ICITB) , 7th Dec 2017 188-193.

[14]. Marwa Farouk M.Ali, Somia A. Asklany, M. Abd El-wahab, M.A.Hassan, (2019), "Data Mining Algorithms for Weather Forecast Phenomena : Comparative Study", IJCSNS International Journal of Computer Science and Network Security, VOL.19 No.9, September 2019 76-81.

[15]. Ting Zhang, SoungYueLiew, Xiao Yan Huang, How Chinh Lee, Dong Hong Qin, (), "Research Trend Analysis Of Artificial Intelligence Rainfall Prediction Algorithms Based On Knowledge Networks" 4th International Symposium on Green and Sustainable Technology (ISGST 2021) IOP Conf. Series: Earth and Environmental Science 945 (2021), 1-7.