Journal of Data Applications
Issue: 2, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

RESEARCH ARTICLE

Multi-Class Classification with the Gaussian Naive Bayes Algorithm

Ayşe ÇINAR[1]

ABSTRACT

Classification is a data mining technique involving supervised machine learning and is the process of predicting the class of data or dataset whose class is unknown using existing data with defined class. Supervised learning occurs during this classification process as a result of how this technique parses the data according to predetermined outputs. The Naive Bayes classifier is a type of machine learning algorithm and an approach that adopts Bayes' theorem by combining theoretically obtained preliminary information with new information. The most obvious advantages of this classifier are its simple algorithm and high accuracy rate. The aim of this study is to create a classification model using the Gaussian Naive Bayes algorithm and to evaluate the obtained prediction results. For this purpose, the study first theoretically considers within its scope the Naive Bayes classifier and then carries out an application on a dataset using the Gaussian Naive Bayes algorithm as one of the types of this classifier in order to create a classification model, which is the subject of the study. Operations were carried out for the classification model using Python, an open-source programming language. The dataset used within the scope of the study was obtained from the University of California Irvine (UCI) Machine Learning Repository website. The purpose for creating the dataset is to determine the different types of Erythemato-squamous disease (ESD). In line with developing technologies, the number of studies demonstrating the ability to make fast and reliable disease prediction using machine learning techniques are increasing daily.

Keywords: Gaussian Naive Bayes, Supervised Machine Learning, Classification, Naive Bayes Classifier

DOI: 10.26650/JODA.1389471

[1]Marmara University, Retired Faculty Member, İstanbul, Türkiye

ORCID: A.Ç. 0000-0001-7321-5959

Corresponding author:
Ayşe ÇINAR,
Marmara University, Retired Faculty Member, İstanbul, Türkiye
E-mail: ayse.cinar@gmail.com

Submitted: 11.11.2023
Accepted: 25.12.2023
Published Online: 07.03.2024

Citation: Çınar, A. (2023). Multi-class classification with the gaussian naive bayes algorithm. *Journal of Data Applications, 2*, 1-13. https://doi.org/10.26650/JODA.1389471

Introduction

The classification method known as supervised machine learning involves a process for distinguishing data classes in order to predict the class of an observation whose class label is unknown.

A dataset considered for classification purposes is generally divided into two sets: the training and test datasets. A dataset can also be divided into three sets to create a validation dataset in addition to these two. The training dataset, which is created with respect to a certain ratio of the whole data, is used to train classification algorithms. The training process occurs by using known values in the training dataset to predict a new observation or class of observations.

One of the popular methods used to classify data is the Naive Bayes classifier. The Naive Bayes algorithm is based on Bayes' Theorem and creates a probability set by calculating the frequency and combination of the values of the features in a given dataset (Wibawa et al., 2019). While determining the class of a new observation, the algorithm calculates conditional probabilities for each class label by taking into account the values of all the features of this observation. The highest calculated posterior probability and the class label with this value are selected and evaluated as the class of the new observation.

Naive Bayes algorithm has been proven to be effective and potentially good in many applications, including text classification, medical diagnosis, and system performance management (Aggarwal & Kaur, 2013; Wibawa et al., 2019; Salmi & Rustam, 2019). This study uses the Gaussian Naive Bayes classifier to process numerical data.

Literature Review

Many studies are found to have been conducted on the dermatology dataset using different techniques in order to accurately diagnose Erythemato-Squamous Diseases (ESD).

Bilgin and Çifçi (2021) used support vector machines (SVMs), ensemble learning algorithms (ELAs), decision trees (DTs) and k-nearest neighbors (k-NN) algorithm on the same dataset. According to their obtained results, they achieved the highest accuracy value of 99.73% with the SVM algorithm; their study did not include patient age in the model. Luukka (2011) proposed a classification method in which the data is first pre-processed using a new non-linear fuzzy robust principal component analysis (NFRPCA) algorithm to transform the data into a more suitable form; after the preprocessing step, Luuka obtained a prediction using a similarity classifier with a 99.02% accuracy rate. Verma et al. (2020) applied the bagging, adaptive boosting (AdaBoost), and gradient boosting (GBoost) methods and reached their highest accuracy rate (99.68%) with the gradient boosting method. Applying six different classification methods, Rashid et al. (2020) achieved their highest accuracy

rate (97.54%) with the Naive Bayes and random forest methods. Putatunda's (2020) study on dermatology applied many methods to the same dataset, in which Derm2Vec as a hybrid deep learning model consisting of autoencoders and deep neural networks (DNNs) had the highest accuracy (96.92%), followed by the DNN (96.65%) and the extreme gradient boosting classifier (95.80%) methods. Shastri et al. (2021) reached a high accuracy rate of 99.45% with the GBoost method on the dermatology dataset.

Method

This study constructs a classification model using the Gaussian Naive Bayes algorithm, for which an application was implemented in Python. A dataset named *dermatology.data* was used to construct the model and contains 366 observations. The purpose of this dataset is to identify the types of ESD (İlter et al.,1998).

The application phase evaluates the first 10 observations in the dataset as validation data. Of the remaining 356 observations, 70% are considered the training dataset and 30% the testing dataset. Using this method for dividing the dataset enables the classification model to predict the class of 10 new observations it had never seen during the training phase for the 110 observations in the testing dataset.

Dataset

The dataset was downloaded from the University of California Irvine (UCI) Machine Learning Repository website and consists of 366 observations, 34 feature variables, and one target variable as shown in Table 1 (İlter & Güvenir, 1997). In the dataset, the feature named *family history* has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The feature named *age* represents the age of the patient. All other features are given a rating ranging from 0 to 3. For these features, 0 indicates that the feature is not present, 3 indicates the largest amount possible, and 1, 2 indicate relative intermediate values (İlter & Güvenir, 1997). Within the scope of this study, the feature named *class* was considered as the target (output), as shown in Table 2. The result from the classification model is determined by the class label of the target.

In the dataset, the classes of the target are represented by numerical values. The distributions of these classes in the dataset are shown in Table 3 along with their codes.

Table 1. *Information about the variables in the dataset.*

Name	Explanation	Type
Erythema	Erythema, reddening of the skin as a result of blood accumulation in the capillaries.	Integer
scaling	Skin flaking	Integer
definite_borders	strict borders	Integer
itching	Itching	Integer
koebner_phenomenon	Small, red blisters that form on the skin.	Integer
polygonal_papules	polygonal papules	Integer
follicular_papules	follicular_papules	Integer
oral_mucosal_involvement	Oral mucosal involvement	Integer
knee_and_elbow_involvement	Knee and elbow involvement	Integer
scalp_involvement	Skull skin involvement	Integer
family_history	Family history	Integer
melanin_incontinence	Inability to retain melanin	Integer
eosinophils_infiltrate	Eosinophil infiltration	Integer
PNL_infiltrate	PNL infiltration	Integer
fibrosis_papillary_dermis	Fibrous degeneration of blistered skin	Integer
exocytosis	Exocytosis	Integer
acanthosis	Acanthosis	Integer
Hyperkeratosis	Hyperkeratosis	Integer
parakeratosis	Parakeratosis	Integer
clubbing_rete_ridges	Clubbing of the rete protrusions	Integer
elongation_rete_ridges	Lengthening of rete processes	Integer
thinning_suprapapillary_epidermis	Thinning of the epidermis with high bubbles	Integer
spongiform_pustule	cancellous boil	Integer
munro_microabcess	Munro microabscess	Integer
focal_hypergranulosis	Focal hypergranulosis	Integer
disappearance_granular_layer	Disappearance of rough layer	Integer
vacuolisation_damage_basal_layer	Vacuuming and damage of the base layer	Integer
spongiosis	spongiosis	Integer
saw_tooth_appearance_retes	The rete has a sawtooth appearance	Integer
follicular_horn_plug	Follicular horn plug	Integer
perifollicular_parakeratosis	Perifollicular parakeratosis	Integer
inflammatory_mononuclear_infiltrate	Inflammatory mononuclear infiltrate	Integer
band_like_infiltrate	Band-shaped infiltration	Integer
Age	Age	Float
class	Type of ESD	Integer

Table 2. *The target.*

Name	Explanation	Type	Value
class	Type of ESD	Integer	1-6

Table 3. *Distribution rate of class labels of the target in the original dataset.*

Code	Class Labels	Frequency	Percentage
1	psoriasis	112	31%
2	seborrheic dermatitis	61	17%
3	lichen planus	72	20%
4	pityriasis rosea	49	13%
5	chronic dermatitis	52	14%
6	pityriasis rubra pilaris	20	5%

Data Preprocessing

The following data preprocessing steps were carried out before establishing the classification model.

As seen in Table 1, all features in the dataset have numerical values. The target (*class*) is also included in the dataset, with numerical codes representing class values. Within the scope of the study, these features are transformed into a categorical value by assigning the class names shown in Table 3.

The dataset has eight missing values that were found to belong to the feature named *Age*. This problem was resolved by assigning the mean value of the relevant feature instead of leaving the values blank. Moreover, when considering the class distribution ratio of the target (Table 3), the number of observations with the class named *psoriasis* was observed to be quite high compared to the other classes, while the number of observations with the class named *pityriasis rubra pilaris* was quite low. This situation reveals that the dataset under consideration is unbalanced. In order to balance the dataset, an oversampling process was applied using synthetic minority over-sampling technique (SMOTE), which enables the production of synthetic data.

Building a Classification Model

The Gaussian Navie Bayes algorithm, a type of Navie Bayes classifier, was applied to create a classification model.

Navie Bayes Classifier

The Naive Bayes classifier is a statistical classification method based on Bayes's theorem as developed by Thomas Bayes in the 18th century and works in accordance with the conditional

probability principle discussed in Bayes' theorem (Meiriza et al., 2019). Based on Bayes' theorem, the resulting events (observations) are expected to be discrete events that are independent of one another (Khuda, 2021).

The Naive Bayes classifier reveals which class has the higher value in the form of a probability before making any classification. When estimating the class of new data with this method, a class with the highest probability value is considered as the class of new data. (Sarkar, 2023). Advantages and disadvantages of the Naive Bayes classifier can be listed as follows (Vadapalli, 2022):

Advantages:

It is a very fast working algorithm.

If the independence of the features assumption is valid, it performs better than other models with less training data.

It works effectively on multi-class predictions.

Disadvantages:

If a feature in the test dataset has a value that cannot be observed in the training dataset, it returns a probability value of 0, meaning it cannot make a prediction. This phenomenon is called zero frequency, and correction techniques such as Laplace estimation are used to resolve this issue (Wu et al., 2013).

All features are assumed to be independent. Although this seems possible in theory, finding a set of independent features is impossible in real life.

Bayes' Rule:

The Naive Bayes classifier is based on Bayesian Decision Theory . The probability equation of an observation whose class will be determined is shown below (Orhan & Adem, 2012).

- Posterior Probability:

$$P(C_i \backslash \mathrm{X}) = \frac{P(C_i) * P(X \backslash \mathrm{C}_i)}{P(X)} \qquad \textit{Eq. 1}$$

$$P(X) = \sum_{i=1}^{n} P(C_i) * P(X \backslash C_i) \qquad \textit{Eq. 2}$$

Since the denominator *P(X)* will be equal for each class, only the numerator values are considered.

$$P(C_i \backslash \mathrm{X}) = P(C_i) * P(X \backslash \mathrm{C}_i) \qquad \textit{Eq. 3}$$

- Class-Conditional Probability:

The Naive Bayes classifier assumes the conditional independence of the feature values given the class as follows (Berrar, 2018):

$$P(X \backslash C_i) = \prod_{k=1}^{n} P(x_k \backslash C_i) \qquad \textit{Eq. 4}$$

- Classification Method for a New Data According to the Calculated Maximum of a Posteriori Probability (MAP):

Finally, by selecting the class with the highest probability value, the algorithm estimates the probability that a new observation with known feature values belongs to a class with this highest value. This expression is represented mathematically as:

$$Predicted\_Class = Argmax\, P(C_i) * \prod_{k=1}^{n} P(x_k \backslash C_i) \qquad \textit{Eq. 5}$$

$\mathrm{X} = \{x_1, x_2, \ldots, x_n\}$ Dataset with unknown class membership.
$x_1, x_2, \ldots, x_n$ The value of the features
$\mathrm{C} = \{C_1, C_2, \ldots, C\}$ Class labels

There are three types of Naive Bayes classifiers (Ismail et al.,2020). Gaussian Naive Bayes should be used for features with continuous values. Gaussian Naive Bayes assumes that the features follow a normal distribution. The Gaussian Naive Bayes classification used within the scope of this study is discussed here. Multinomial Naive Bayes is mostly used to classify documents. Each document in the dataset should contain features with discrete positive integer values that express the frequency of words. Bernoulli Naive Bayes should be used for features with binary or Boolean values such as true/false or 0/1.

Classification with the Gaussian Naive Bayes Algorithm

Unlike with the Naive Bayes classification algorithm, normal (i.e., Gaussian) distribution curves are obtained for features with continuous values associated with each class using the Gaussian Naive Bayes algorithm. For this purpose, the distribution is summarized by estimating the mean and standard deviation values for each class from the training data. The obtained results obtained are considered as the conditional probability value for each class of features of an observation.

- Conditional Probabilities of the Features:

$$P(x_i \backslash C) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2}\right) \qquad \textit{Eq. 6}$$

- Class-Conditional Probability:

$$P(X \backslash C_i) = \prod_{k=1}^{n} P(x_k \backslash C_i) \qquad \textit{Eq. 7}$$

- Classification Method According to the Maximum a Posteriori Probability (MAP):

$$Predicted\_Class = Argmax\, P(C_i) * P(X \backslash C_i) \qquad i \in \{1, .., n\} \qquad \textit{Eq. 8}$$

Experimental Results

Confusion Matrix

Table 4 shows the confusion matrix obtained for the Gaussian Naive Bayes classification model created in this study. By considering the confusion matrix, various performance metrics are obtained for the classification models. One of these is the accuracy value, which is the correct prediction rate of the class label of the observations in the model's testing dataset. The number of correct predictions for each class is shown in bold in Table 4. The accuracy value of the model constructed in this study is found to be 98.18%, as seen below.

$$\text{Accuracy} = \frac{\text{total number of correct predictions}}{total\ number\ of\ observations} = \frac{13+20+22+5+31+14}{107} = 0.9813$$

This classification model has a high accuracy rate and made incorrect predictions for the class of only two observations. Accordingly, although the class of both observations was *seb-dermatitis*, the model predicted it as *pityriasis rosea*.

Table 4. *Six-class confusion matrix.*

Actual / Predicted	chronic dermatitis	lichen planus	pityriasis rosea	pit_ rubra pilaris	psoriasis	seb_ dermatitis
chronic dermatitis	**13**	0	0	0	0	0
lichen planus	0	**20**	0	0	0	0
pityriasis rosea	0	0	**22**	0	0	2
pit_ rubra pilaris	0	0	0	**5**	0	0
Psoriasis	0	0	0	0	**31**	0
seb_ dermatitis	0	0	0	0	0	**14**

Performance metrics apart from model accuracy, sensitivity (recall), specificity, precision, and harmonic precision-recall mean (F1 score) are calculated based on a positive class of

the target. The calculation of these values was carried out with the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) frequency values included in the Confusion Matrix.

TP (True Positive): The number of observations that are positive and have been correctly classified as positive.

TN (True Negative): The number of observations that are negative and have been correctly classified as negative.

FP (False Positive): The number of observations that are negative but have been incorrectly classified as positive.

FN (False Negative): The number of observations that are positive but have been incorrectly classified as negative.

Within the scope of the study, performance metric values have been calculated by considering each class of the target as a positive class. The number of observations (T) of the test dataset considered in the calculation phase is 107.

The calculation method and obtained performance metric values of the positive classes called *chronic dermatitis* (Figure 1) and *pityriasis rosea* (Figure 2) are presented below as examples.

Positive class: « ***chronic dermatitis*** »

Actual / Predicted	chronic dermatitis	lichen planus	pityriasis rosea	pit_ rubra pilaris	psoriasis	seb_ dermatitis
chronic dermatitis	TP	FP	FP	FP	FP	FP
lichen planus	FN	TN	TN	TN	TN	TN
pityriasis rosea	FN	TN	TN	TN	TN	TN
pit_ rubra pilaris	FN	TN	TN	TN	TN	TN
psoriasis	FN	TN	TN	TN	TN	TN
seb_ dermatitis	FN	TN	TN	TN	TN	TN

Actual / Predicted	chronic dermatitis	lichen planus	pityriasis rosea	pit_ rubra pilaris	psoriasis	seb_ dermatitis
chronic dermatitis	13	0	0	0	0	0
lichen planus	0	20	0	0	0	0
pityriasis rosea	0	0	22	0	0	2
pit_ rubra pilaris	0	0	0	5	0	0
psoriasis	0	0	0	0	31	0
seb_ dermatitis	0	0	0	0	0	14

Figure 1. *Six-class confusion matrix, where the positive class is chronic dermatitis.*

The total TP, FN, FP and FN values obtained with the test dataset for *chronic dermatitis*, which is considered as a positive class, are as follows.

Total(TP)=13

Total(TN)=94

Total(FP)=0

Total(FN)=0

Performance Metrics:

$$\text{No Information Rate} : \frac{\text{Total(actual(chronic dermatitis))}}{T} = 0{,}12$$

$$\text{Sensitivity(Recall)} : \frac{TP}{TP+FN} = 1$$

$$\text{Specificity} : \frac{TN}{TN+FP} = 1$$

$$\text{Precision} : \frac{TP}{TP+FP} = 1$$

Positive class: « ***pityriasis rosea*** »

Actual / Predicted	chronic dermatitis	lichen planus	pityriasis rosea	pit_ rubra pilaris	psoriasis	seb_ dermatitis
chronic dermatitis	TN	TN	FN	TN	TN	TN
lichen planus	TN	TN	FN	TN	TN	TN
pityriasis rosea	FP	FP	TP	FP	FP	FP
pit_ rubra pilaris	TN	TN	FN	TN	TN	TN
psoriasis	TN	TN	FN	TN	TN	TN
seb_ dermatitis	TN	TN	FN	TN	TN	TN

Actual / Predicted	chronic dermatitis	lichen planus	pityriasis rosea	pit_ rubra pilaris	psoriasis	seb_ dermatitis
chronic dermatitis	13	0	0	0	0	0
lichen planus	0	20	0	0	0	0
pityriasis rosea	0	0	22	0	0	2
pit_ rubra pilaris	0	0	0	5	0	0
psoriasis	0	0	0	0	31	0
seb_ dermatitis	0	0	0	0	0	14

Figure 2. *Six-class confusion matrix, where the positive class is pityriasis rosea.*

The total TP, FN, FP and FN values obtained with the test dataset for *pityriasis rosea*, which is considered as a positive class, are as follows.

Total(TP)= 22

Total(TN)= 83

Total(FP)= 2

Total(FN)=0

Performance Metrics:

No Information Rate : $\frac{\text{Total(actual(pityriasis rosea))}}{T}$ = 0,21

Sensitivity(Recall) : $\frac{TP}{TP+FN}$ = 1

Specificity : $\frac{TN}{TN+FP}$ = 0,98

Precision : $\frac{TP}{TP+FP}$ = 0,92

F1-test : $\frac{2*Precision*Sensitivity}{Precision+Sensitivity}$ = 0,96

The performance metrics obtained by the classification model created with the Gaussian Naive Bayes algorithm separately for each class of the target are shown in Figure 3.

```
In [11]: print(metrics.classification_report(y_test,y_pred_tuned))
                          precision    recall  f1-score   support

      chronic dermatitis       1.00      1.00      1.00        13
           lichen planus       1.00      1.00      1.00        20
         pityriasis rosea      0.92      1.00      0.96        22
pityriasis rubra pilaris       1.00      1.00      1.00         5
               psoriasis       1.00      1.00      1.00        31
     seboreic dermatitis       1.00      0.88      0.93        16

                accuracy                           0.98       107
               macro avg       0.99      0.98      0.98       107
            weighted avg       0.98      0.98      0.98       107
```

Figure 3. *Performance metrics obtained with Gaussian Naive Bayes algorithm.*

Conclusion

This study has constructed a classification model using the Gaussian Naive Bayes algorithm. A dermatology dataset was used for this model to make predictions about patients' types of ESD. Numerous studies have attempted to apply machine learning techniques to dermatology datasets. Some of those that have applied Naive Bayes and other classification algorithms are found within this study's Literature Review section.

The study carried out within its scope some data preprocessing on the dataset prior to building the model and getting the data ready for analysis. Additionally, the dataset was

divided into three groups: the training, testing, and validation datasets. While the validation group includes the first 10 observations from the dataset, of the remaining observations, 30% made up the testing and 70% the training data. Training was carried out by constructing a classification model using the training data, and the accuracy and other performance metrics of the model were obtained using the testing data. The study then applied model tuning to improve the model's performance metrics. The most appropriate performance metrics for the dataset were achieved by building the model using the best parameter values obtained by applying this process. As a result, the accuracy rate of the model was 98.13%. According to this result, the classification model built with the Gaussian Naive Bayes algorithm was observed to have achieved results close to other studies using the same dataset.

After tuning the model, the tuned classification model predicted a class of 10 new observations in the validation dataset, and the model was seen to correctly predict the class of all new observations.

Ethics Committee Approval: Since the dermatology dataset used in my study was not obtained by me through a survey, I think that there is no need for ethics committee approval for my study. It is a public file published on the internet.

Peer-review: Externally peer-reviewed.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

References

Aggarwal, S., & Kaur, D. (2013). Naive Bayes Classifier with Various Smoothing Techniques for Text Documents. *International Journal of Computer Trends and Technology (IJCTT)*, 4(4), 873-876. https://ijcttjournal.org/Volume4/issue-4/IJCTT-V4I4P187.pdf

Ahmed, M., Kashem, M. A., Rahman, M., & Khatun, S. (2020). *Review and Analysis of Risk Factor of Maternal Health in Remote Area Using the Internet of Things (IoT),* Part of the Lecture Notes in Electrical Engineering book series (LNEE), volume 632, Springer.

Berrar, D.(January 2018), *Bayes' Theorem and Naive Bayes Classifier,* Reference Module in Life Sciences, https://www.researchgate.net/publication/324933572_Bayes'_Theorem_and_Naive_Bayes_Classifier#fullTextFileContent

Bilgin, G., & Çifçi A. (2021). Eritematöz Skuamöz Hastalıkların Teşhisinde Makine Öğrenme Algoritmaları Performanslarının Değerlendirilmesi. *Zeki Sistemler Teori ve Uygulamaları Dergisi*, 4(2), 195-202.

İlter, N., & Güvenir, H. (1997). Dermatology, https://archive.ics.uci.edu/dataset/33/dermatology

İlter, N., Güvenir, H., & Demiroz, G. (1998). Learning Differential Diagnosis of Erythemato-Squamous Diseases Using Voting Feature Intervals. *Artificial Intelligence in Medicine*, 13(3), 147-165.

Ismail, M., Hassan,N., & Bafjaish,S.S. (2020). Comparative Analysis of Naive Bayesian Techniques in Health-Related for Classification Task. *JSCDM-Journal of Soft Computing and Data Mining*, 1(2), 1-10.

Khuda, I. E. (2021). Innovative Teaching Pedagogy for Teaching and Learning of Bayes'Theorem. *Journal of Science and Engineering, CUSJE*, 18(1), 61-71, Çankaya University, 63-71.

Luukka P. (2011). A New Nonlinear Fuzzy Robust PCA Algorithm and Similarity Classifier in Classification of Medical Datasets. *International Journal of Fuzzy Systems*, 13(3), 153-162.

Meiriza, A., Lestari, E., Putra, P., Monaputri, A. & Lestari, D.A. (2019). *Prediction Graduate Student Use Naive Bayes Classifier*, Advances in Intelligent Systems Research Series. Volume 172, Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019), Atlantis Press.

Orhan, U & Adem, K. (2012, October 29-December 01). *Naive Bayes Yönteminde Olasılık Çarpanlarının Etkileri* [Conference presentation]. Elektrik - Elektronik ve Bilgisayar Mühendisliği Sempozyumu, 29 Kasım - 01 Aralık 2012, Bursa.

Putatunda, S. (2020, July 2-4). *A Hybrid Deep Learning Approach For Diagnosis Of The Erythemato-Squamous Disease*[Conference presentation]. IEEE International Conference On Electronics, Computing And Communication Technologies, Bangalore, India, 1-6.

Rashid, A.N.M.B., Ahmed, M., Sikos, L.F., Haskell-Dowland, P. (2020). A Novel Penalty-Based Wrapper Objective Function for Feature Selection In Big Data Using Cooperative Co-Evolution. IEEE Access, 8, 150113-150129.

Salmi, N. & Rustam, Z. (2019). *Naive Bayes Classifier Models for Predicting the Colon Cancer*, 9th Annual Basic Science International Conference 2019 (BaSIC 2019), IOP Conf. Series: Materials Science and Engineering 546 (2019) 052068, IOP Publishing.

Sarkar P. (2023). *Naive Bayes in Machine Learning [Examples, Models, Types]*, https://www.knowledgehut.com/blog/data-science/naive-bayes-in-machine-learning

Shastri, S., Kour, P., Kumar, S., Singh, K., Mansotra, V. (2021). GBoost: A Novel Grading-AdaBoost Ensemble Approach for Automatic Identification of Erythemato-Squamous Disease. *International Journal of Information Technology*, 13(3), 959-971.

Vadapalli,P. (2022). *Naive Bayes Classifier: Pros & Cons, Applications & Types Explained*, https://www.upgrad.com/blog/naive-bayes-classifier/

Verma, A.K., Pal, S., Kumar, S. (2020). Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method—A Comparative Study. *Applied Biochemistry and Biotechnology*, 190(2), 341-359.

Wibawa, A.P., Kurniawan, A.C., Murti, D.M.P., Adiperkasa, R.P., Putra, S.M., Kurniawan, S.A. & Nugraha, Y.R. (2019). Naïve Bayes Classifier for Journal Quartile Classification. *iJES*, 7(2), 91-99, https://doi.org/10.3991/ijes.v7i2.10659

Wu, J., Cai, Z. & Zhu, X. (2013, August 4-9). *Self-adaptive probability estimation for Naive Bayes Classification*[Conference presentation]. International Joint Conference on Neural Networks (IJCNN), U.S.A.