International Journal of Computer Applications Technology and Research
Volume 9–Issue 09, 257-264, 2020, ISSN:-2319-8656

Applicability of Naïve Bayes Model for Automatic Resume Classification

Patrick Nyanumba Mwaro
Student SCIT, JKUAT
Nairobi, Kenya
E-mail:
patrickmwaro@gmail.com

Dr. Kennedy Ogada
Senior Lecturer SCIT, JKUAT
Nairobi, Kenya
E-mail:
kenogada@gmail.com

Prof. Wilson Cheruiyot
Professor SCIT, JKUAT
Nairobi, Kenya
E-mail:
wilchery68@gmail.com

Abstract: Resume selection and classification is a very important function of Human Resource Department of every institution. Due to increased use of technology and online job application, this department receives large volumes of resumes which has made resume selection and classification a complex process in terms of information processing, time taken and transparency in the selection process. In this research, a machine learning model is proposed to assist resume selection and classification. Naïve Bayes model was developed to select and classify resumes. The predictive accuracy attained will be recorded and compared to predictive accuracy of homogeneous Ensemble classifier model developed by using different data sets. Naïve Bayes classifier models obtained from different data sets was used as base classifiers to develop Ensemble Naïve Bayes Classifier. It was observed that the new model produced a better predictive accuracy compared to the original Naïve Bayes Classifier model. The original Naïve Bayes classifier model gave an average predictive accuracy of 89.8148% while Ensemble Naïve Bayes Classifier model attained an overall accuracy of 94.4444%.

Keywords: classification; machine learning; naïve Bayes; selection process; predictive accuracy

1. INTRODUCTION

Human resource department is very important in both public and private institutions. Bringing the right employees on board is a challenging task in all institutions. Employing people with good talent in their areas of specialization is a challenge and therefore the first important stage of getting the right employee is resume selection and classification. Recruitment process is a product of resume selection. Recruitment process is the process of identifying job vacancies, analyzing job requirements, reviewing applications, screening, shortlisting and selecting the right candidates. The purpose of this research is to use machine learning to automate the process so that we can achieve efficiency in resume selection and classification process in the human resource department. Research has been done to investigate the effectiveness of online recruitment and selection process in terms of use of the internet. In this research it was found that online recruitment is effective in reducing the cost and time of recruitment and selection [1]. Machine learning has been used to show that recruiting asymptomatic individuals in clinical trials can obtimize the cost of clinical trials. In the research which was done proposed the method of recruiting asymptomatic Amyloid positive individuals in clinical trials where by two step process was employed to select subset individuals more likely to be amyloid positive based on automatic analysis of data acquired in a routine clinical practice [2]. Technology has greatly impacted how information is being accessed in everyday life. It has changed the way people communicate in daily bassis. In this modern days, we cannot imagine how life could be without the World Wide Web. Every one uses the internet for different purposes such as looking for information or posting information in the internet. The information can be easily posted by people in form of blogs, forums, social networks, feedbacks can be given in particular web pages. Today, there are many sites that provide business and product review such as Amazon is an e-shop where customers can publish feedback about various products. This gives rise to opinions from various customers and therefore necessitating creation of automated system for seaching and classifying opinions [3]. Due to improvements in technology and use of internet, government and private institutions use websites to share information of job adverntisement and recruitment of new workers. This information overflows in various sites with various attributes and criteria making the selection process to be complex due to limited time. To simplify the process, reseach has been done to construct and collaborate web scrapping technique and classification using Naïve Bayes on search engine and it resulted in effective and efficient application for users to seek pontetial jobs that fits their interests [4]. Soft skills are impotant factor to consider for a certain set of jobs. To get soft skills from job applicants is not an easy task, to do this, Bayesian network has been employed in some research to assist inidentification of soft skills because Bayesian is suitablle for reasoning and making decisions under uncertainity. In the research, the Bayesian was trained using a dataset collected through extracting information from advertisements and also through interviews with few selected experts [5]. The most popular challege in data management is to support the construction and maitenance of machine learning models over large data which is multi-dimensional and evolving. Bayesian network can be employed to support the task of managing large volumes of data in terms of processing and classification [6]. Bayesian network classifier have demostrated good classification accuracy in variours applications however it is error prone in high confidence labels. To address the problem, the label-driven learning framework was proposed which incorporated instance based learning and ensemble learning [7]. From these various applications of machine learning techniques, it has been shown that machine learning methods can be applied to solve realy world problems such as online job application, job recommender systems, sentment analysis and spam filter sytems for identifying and filtering spam emails.

2. RELATED WORK

2.1 Artificial intelligence

Artificial intelligence is relevant for the recruitment process due to many human challenges faced by human resource professionals in the recruitment process. Some of the

challenges include large volumes of reumes received, time constraits and lack of transparency in the resume selection process. Despite this, limitted research has been done on this topic. One of the research which has been done used and repricated data from the Boston consulting Group, CV Library, Linkedln, MIT Sloan Management Review, Software Advice, Statista and Tractica [8]. In the research, analysis was performed and estmates were made regarding utilization of artificial intelligence and automation in interviewing and assessment of candidates. Approaches to building artificial intelligence related skills and effect of it on human resource. The artificial intelligence (AI) has started managing the recruitment process in the human resource department. Providing the right human resource duringrecruitment process is one of the important function of human resource management [9].

2.2 Email Spam Filtering

Due to 5G era, the email application has become more extensive and spam messages have caused serious problems. There are various spam filtering methods. Probability based bayesian classification algorithm has been employed in the past as a simple and efficient method to filter spam messages to avoid serious problems caused by spam messages [10]. The accuracy so far attained by bayesian email spam filtering was 90%. The research bayesian model based on naïve bayesian classification model included advantages and disadvantages of the effectiveness of the model.

2.3 Predicting Diseases

Naïve bayes has been used as a classification method for predicting diseases in haemoglobin protein sequence. In the research which was done it was shown that the technique can be successively applied to unveil the structures, functions and evolutionally relationship in protein sequences and predict diseases based on their sequence information [11]. The use of data mining method in protein analysis provided efficient way of examing the proteins to identify their characteristics and provided away for designing better drugs. In the research Naïve Bayes attained an Accuracy of 85%.

It has also been demostrated that machine learning algorithms can be used in medical sector. Heart disease prediction using classification technique such as Naïve Bayes have been employed in the past. A research was done presentig six classification models where indipendent features were used to build the model [12]. The classification models which were used included K-nearest neighbors(KNN), support vector machine(SVN), and Naïve Bayes. The feature selection was applied to improve prediction accuracy. In the same study, Naïve Bayes produced an accuracy of 88.16%. Health care is a task to be done in human beings. Effective decissions about human disease can be handled by applying data mining techniques. Different tests can be done to detect cardiovascular diiseases in patients. Naïve bayes is one of data mining technique which can be employed to serve as diagnosis of heart disease in patients. Paramenters were analyzed and predictions made on heart disease and therefore heart disease prediction system was proposed [13]. Using the same approach, machine learning algorithms can be applied in resume selection and classification process to make it easy, simple and straight forward. It has been shown that machine learning applications in realy world task eliminates time wastage and makes processes efficient in terms of time taken to complete some human task. Naïve Bayes classifier has also been used in predicting the risk of heart disease in diabetic patients. The clinical analysis on datasets was done to detect diseases and diagnosis based on data and attributes and a predictive accuracy of 89.41% was attained in the prediction of risk of heart disease in diabetic patients [14]. This is a human task just like resume selection and classification. And therefore Naïve Bayes algorithm is intended to be used to perform the task of resume selection and classification from job applicants.

2.4 Customer Churn Analysis

In customer Churn Analysis and Prediction, customer past behaviour are analyzed to find the cause of churn and predicting whether the churn will happen in the future. The aim of the study was to use machine learning Prediction to predict, retain and avoid customer churns. Company growth entiry depends on customers. Customers play an important role and therefore companies should understand their customers in terms of their behaviour and requirements. Various machine learning algorithms such as Naïve Bayes, decision trees, random forest have been applied in feature selection for use in Customer Churn Analysis and [15]. Therefore machine learning techniques has become very important tool of making good predictions with highest accuracy in getting correct information in most of human task in the modern technology. Equaly machine learning can be applied in resume selection and classification process and this is the main purpose of this research work.

2.5 Naïve Bayes

Naïve Bayes Classifier is a supervised machine learning algorithm making use of the Bayes' Theorem where features are statistically indipendent. A research was done where simple machine learning model was developed from a set of attributes (training examples) in relation to response variables [15]. The Bayesian principle states that the probability of something happening in future can be estimated by calculating how often it has happened [10]. In spam filtering study, Bayesian algorithm was utilized to filter spam from asample size of 5574 and cross validation was used.

The essence of Bayes theorem is the conditonal probability where conditional probability is given as the probability of a given event happening given that some other events have already occurred. By using conditional probability, probability that a given event will occur given the knowledge of previous event can given by equation (1) below

$$P(A/B) = \frac{P(B/\ B)P(A)}{P(B)} \quad (1)$$

Where

P(A/B)= Posterior probability, Probability of event A happening given the value of B.

P(B/A)= Likelyhood of B given A is true

P(A)= Prior Probability, Probability of event A

P(B)= Marginal Probability, Probability of event B

The Naïve Bayes Classifier formula using Bayes theorem isgiven as equation (2) below.

$$P(y/x_1, \dots x_j) = \frac{P(x_1, \dots x_j/P(y)}{P(x_1, \dots x_j)} \quad (2)$$

$P(y/x_1, \dots x_j)$ = Posterior probability, probability of the data included in the class of y given their features $x_1\ up\ to\ x_j$

$P(x_1, \dots x_j/P(y)$ = Likelyhood of features value given that their class is y.

$P(y)$ = Prior probability.

$P(x_1, \ldots x_j)$ = Marginal probability.

Naïve Bayes Model is easy to develop which makes it useful for large datasets. Despite its simplisity, it performs well than other sophisticated machine learning methods [16].

3. METHODS AND EXPERIMENTS

This section describes data pre-processing, experimental setup, application of Naïve Bayes Classifier in resume classification and accuracy improvement through homogeneous Bayesian ensemble classifier.

3.1 Data preprocessing

Data preprocessing is the process of transforming raw data into understandable format to increase the validity of the data. Therefore, before clasification, the datasets must be prepared to optimize the data [17]. A dataset contains and providents concise and unumbiguos definition of items related to the phenomenon under study [13]. The data set used in this research was obtained from resumes of job applicants and most of it was obtained from linkedln. There were a total of 250 resumes which were used and experiments were used as the main research methodology for this study. During classification, three major classes were considered as follows: employable resumes(ER), waiting resumes(WR) and not employable resumes(NER). Employable resumes are those resumes with all the required attributes which meets the recruitment criteria. Waiting resumes are those resumes which meet the minimum requirement for recruitment but lacked some important attributes for the recruitment, while not employable are those resumes which did not meet the minimum recruitment criteria.

3.2 Data and Input Attributes

There are two major terminologies associated with data sets. The first term is an Instance, an instance is an example in the training set. In this research an instance is a particular resume in a group of resumes. The second terminology is an attribute which refers to a feature of a group of characteristics which makes up one instance in a dataset. Therefore attributes are called features in machine learning. A class is a label given to an instance of a class. In this research, there are three classes namely: Employable, waiting and not employable. The following attributes were taken into consideration during resume classification process:

1. Level of education
2. Course (area of specialization)
3. Work experience
4. Skill (special skills)
5. Year of graduation
6. Length of stay after gradution
7. Quaality of the certificate(e.g. first class)

Some attributes were combined due to close relationship to each other. For example year of graduation and length of stay after graduation were combined to give one attribute.

The dataset was divided into five subsets. This was done because the dataset was small and the major objective of this research was to improve predictive accuracy of Naïve Bayes classifier by combning four homogeneours Naïve Bayes models deveoped from the four data subsets. The fifty data subset was used as testing dataset while the other four datasets were used to develop four base classifiers for use in developing ensemble classifier.

3.3 Experimental setup

The experimental data was organized into a feature vector representing the recruitment data with various attributes. The recruitment data was obtained from job applicant resumes. Matlb 2016a was used to conduct experiments to design four Bayesian Classification models from the four data subsets. Supervised machine learning algorithm(Naïve Bayes) was used for the classification. The total number of resumes for each class was organized into a vector of instances as follows:

$$\text{Employable Resumes (ER)} = \begin{bmatrix} ER_1 \\ ER_2 \\ ER_3 \\ \cdot \\ \cdot \\ \cdot \\ ER_n \end{bmatrix}$$

$$\text{Waiting Resumes (WR)} = \begin{bmatrix} WR_1 \\ WR_2 \\ WR_3 \\ \cdot \\ \cdot \\ \cdot \\ WR_n \end{bmatrix}$$

$$\text{Not Emplyable Resumes (NER)} = \begin{bmatrix} NER_1 \\ NER_2 \\ NER_3 \\ \cdot \\ \cdot \\ \cdot \\ NER_n \end{bmatrix}$$

Employable resumes consitituted 32.4%, waiting resumes 40.8% and not employable resumes 26.8% of the overal recruitment data. Each resume also contained a vector of features called attributes. Each attribute was assigned different weights depeding on the importance contributed to the overal recruitment criteria. Some features in the recruitment process are more import and therefore they are assigned more weight compared to others. All 250 resumes contained equal number of vector features. Therefore organizing the resumes against there individual features, they form a rectangular matrix of the order 250 X 4. This means that each resumes contais four recruitment attributes which contributes to the overal score of each resume. The 75% of the total resumes with seven features were given to the original Bayesian classifier as input for training the model while 25% of the total resumes were used for testing the Naïve Bayes model. Table (2) below shows percentage datasets for the three classes.

Table 2. Percentage datasets

Class	Number	Percentage
Employable	81	32.4%
Waiting	102	40.8%
Not Employable	67	26.8%
Total	250	100%

3.4 Naïve Bayes Classification

Classification is a supervised machine learning technique used to predict the class of a given data points. Classification is best used when the outputs or targets are known. In this research a nonlinear model naïve bayes algorithm is used to classify job applicant resumes into three classes namely: employable, waiting and not employable.

The naïve classifier was designed as follows:

1. Let T be a training set of rsumes and there associated classes employable(CE), waiting(CW) and not employable(CN). Each resume(record) contains various attributes forming a vector $Y=(y_1,y_2,..y_{n-1},y_n)$
2. Let k be a number of classes for prediction, $C_1,C_2,\ldots C_K$. Given the record Y, the naïve bayes classifier will predict Y belong to the class which have the highest posterior probability. i.e.

$$P(Ci/Y) > P(Cj/Y) \ for \ 1 \leq k \ and \ j \neq i \qquad (3)$$

Therefore maximizing $P(Ci/Y)$, the bayes theorem becomes

$$P(Ci/Y) = \frac{P(Y/Ci)P(Ci)}{P(Y)} \qquad (4)$$

3. To predict the class label for Y, $P(Y/Ci)P(Ci)$ was evaluated for each class Ci where the naïve bayes classifier predicted that the class label for Y was the class Ci if and only if

$$P(Y/Ci)P(Ci) > P(Y/Cj)P(Ci) \ for \ 1 \leq k, j \neq i \qquad (5)$$

This was to say that, the predicted clsss label was the class Ci in which the $P(Y/Ci)P(Ci)$ was maximum. In this study the datasets were preprocessed and then subdivided into two datasets namely training dataset and testing datasets. The overal dataset was normalized. The training dataset was preprocessed and the correct class labels were given for the purpose of conducting supervised machine learning classification. Nonlinear naïve bayes classification was done by developing a classification model based on Bayes rule.

3.5 Improving Accuracy of Naïve Bayes

There are many ways of improving Naïve Bayes prediction accuracy such as data preprocessing, feature selection, ensemble method and the fisher method. In this study data preprocessing, feature selection and homogeneous ensemble method was employed to improve predictive accuracy of Naïve Bayes Classifier. During data preprocessing, the data was converted into a form which can be accepted by the machine learning algorithm used. The input data was prepared with respect to the output expected. Feature selection was done where only those attributes which contributed to the overal classification accuracy were considered in the development of the classifier. Then, lastly ensemble method was employed where the overal dataset was divided into four data subsets. The four datasubsets were used to develop four Naïve Bayes Classifier models which formed base classifiers. The base classifiers were then combined to develop Ensemble Naïve Bayes Classifier(ENBC). Both the original Naïve Bayes and Ensemble Naïve Bayes Classifier were used to classify resume data and their accurancies were recorded and compared. It was noted that the predictive accuracy of Ensemble Naïve Bayes Classifier was better than the original Naïve Bayes Classifier.

3.6 Confusion Matrix

A confusion matrix is a table used to describe the performance of a classification model on a set of data in which the true values are known [18]. It gives the summary results of the classification model and it is used for model evaluation in terms of accuracy, precision and recall.

Table 3. Confusion Matrix

	Predicted NO	**Predicted YES**	**Total**
Actual NO	TN	FP	TN+FP
Actual YES	FN	TP	FN+TP
Total	TN+FN	FP+TP	TN+FN+FP+TP

$$Accuracy = \frac{TN + TP}{TN + FN + FP + TP} \qquad (6)$$

$$Precision = \frac{TP}{FP + TP} \qquad (7)$$

$$Recall = \frac{TP}{FN + TP} \qquad (8)$$

Where
TN = True Negative
FP = False Positive
FN = False Negative
TP = True Positive

4. RESULTS AND DISCUSSION

This section presents the experimental results and discussions. Both results are represented in the form of tables. The first part represents results in the form of confusion matrix for model evaluation and the second gives the experimental results summary.

4.1 Confusion Matrix Results

This were the results based on the confusion matrix which showed correctly classified resumes and those which were miss classified by the models as shown in the tables below.

Table 4. Confusion Matrix Dataset_1

	Predicted			Total
Actual	15	2	0	**17**
	1	21	0	**22**
	0	3	12	**15**
Total	**16**	**26**	**12**	**54**

Table 5. Confusion Matrix Dataset_2

	Predicted			Total
Actual	17	0	0	**17**
	5	17	0	**22**
	0	2	13	**15**
Total	**22**	**19**	**13**	**54**

Table 6. Confusion Matrix Dataset_3

	Predicted			Total
Actual	16	1	0	**17**
	2	20	0	**22**
	0	1	14	**15**
Total	**18**	**22**	**14**	**54**

Table 7. Confusion Matrix Dataset_4

	Predicted			Total
Actual	14	3	0	**17**
	1	21	0	**22**
	0	1	14	**15**
Total	**15**	**25**	**14**	**54**

Table 8. Ensemble Confusion Matrix

	Predicted			Total
Actual	16	1	0	**17**
	1	21	0	**22**
	0	1	14	**15**
Total	**17**	**23**	**14**	**54**

4.2 Experimental Results summary

The experimental results are subdivided into three. The first result was obtained when the original Naïve Bayes Classifier (NBC) was used to classify the job applicant resume data and it attained an accuracy of 89.8148%. The second results was obtained from the four base classifiers i.e. Naïve Bayes Classifier-1(NBC-1), Naïve Bayes Classifier-2, Naïve Bayes Classifier-3(NBC-3) and Naïve Bayes Classifier-4(NBC-4). The results was recorded in tables (9) given below. The third result was obtained when the ensemble Naïve Bayes Classifier(ENBC) was used to classify the same dataset and the results were recorded in table (10) as shown below.

Table 9. Models Performance Measures

Model	Accuracy	Precision	Recall
NBC-1	88.8889%	0.9375	0.8824
NBC-2	87.0370%	0.7727	1.0000
NBC-3	92.5926%	0.8889	0.9412
NBC-4	90.7407%	0.9333	0.8235

Table 10. Ensemble Performance Measures

Model	Accuracy	Precision	Recall
ENBC	94.4444%	0.9412	0.9412

4.3 Result Discussion

The overall dataset was divided into five datasets. The first four were used to develop four naïve bayes models for resume classification. The fifty dataset was used for testing the models. Dataset 1 was used to train NBC-1 model which was later tested by dataset 5 and it classified 15 resumes correctly as employable while it misclassified 2 resumes as waiting but they were actually in the class employable. The model also classified 21 resumes correctly as waiting resumes while it misclassified 1 resume as employable. Lastly it classified 12 resumes correctly as not employable and 3 resumes were misclssified as waiting. These results are given in table (4) above. NBC-1 model attained a predictive accuracy of 88.8889% as shown in table (9) above.

When NBC-2 model was used, it classified a total of 47 resumes correctly. A total of 7 resumes were missclassified where 5 were missclassified to belong to the class employable while they were actually in the class of waiting. Another 2 resumes were also missclassified as waiting when they were actually in the class of not employable. The result of this model is given in table (5) above. The model attained a predictive accuracy of 87.0370% as indicated in table (9) above.

NBC-3 model classified 50 resumes correctly to belong to their actual classes and only 4 resumes were misclassified. 2 of misclassified belonged to the class of waiting resumes but they were classified as employable resumes. The other 2 resumes were misclassified as waiting and not employable respectively as shown in table (6) above. The model attained an accuracy of 92.5926% as indicated in table (9) above.

The fourth model was NBC-4 which classified an overall of 49 resumes correctly while misclassified 5 resumes. 3 resumes were misclassified as waiting but they were actually employable resumes. Another 1 resume was placed in employable resumes but it was actually in the class of waiting resumes. Lastly 1 resume was also misclassified as waiting but its actual class was not employable. The overall classification for this model is given in table (7) above. The predictive accuracy for this model was 90.7407% as given in table (9) above.

Ensemble Naïve Bayes Classifier (ENBC) classified 51 resumes correctly while it misclassified only three resumes as shown in table (8) above. The overall predictive accuracy of ENBC was 94.4444% as indicated in table (10) above.

5. CONCLUSSION

During resume selection process, there are many factors which influence resume selection for consideration for the job opening. The factors can be considered during selection process to ensure that job applicants are selected to take over the jobs that best fit their qualification and skills they possess. It has been noted in many occasions that selecting the best resume for a job opening is the main factor in the recruitment of new employees. Selecting a wrong resume for a job opening can misslead the process. Therefore it is vital to automate the process of resume selection inorder to save time, cost and make the resume selection process transparent hence eliminating personal interrest in the selection process. The anlysis of the five models showed that the individual naïve bayes classifier models gave different results and different predictive accuracies when used to classify the same recruitment data. It was also observed that when the four models were combined to form ensemble naïve bayes classifier, model stability improved together with predictive accuracy. Future work will involve using heterogeneous base classifiers to investigate the overal effects on the predictive accuracy of the model.

6. ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my supervisors for guiding me throughout my research time since I started to the end. It is through their tireless support that I have completed my research work successively. My first gratitude goes to my two supervisors Professor Wilson Cheruiyot of Jomo Kenyatta University of Agriculture and Technology (JKUAT), department of Computing and Information Technology. He guided me on how to design my research and gave me important corrections to my research. Secondly I would like to thank Doctor Kennedy Ogada who tirelessly guided me to come up with a technical research area and refining my research to focuss on scientific problem in machine learning field. He took his time to read and correct my research. I thank you and may God bless you all.

7. REFERENCES

[1] A. Gopalia, "Effectiveness of online recruitment and selection process: a case of Tesco," *World Applied Sciences Journal*, vol. 20, no. 8, pp. 1152-1158, 2012.

[2] M. a. E. Ansart, "Reduction of recruitment costs in preclinical AD trials: validation of automatic pre-screening algorithm for brain amyloidosis," *Statistical Methods in Medical Research*, vol. 19, no. 1, pp. 151-164, 2020.

[3] O. Shepelenko, "Opinion mining and sentiment analysis using Bayesian and neural networks approaches," 2017.

[4] S. Cepy, A. Rian, M. D. Sa'adillah, S. Suhendar, D. Wahyudin and R. M. Ali, "Web scraping and Naive Bayes classification for job search engine," 2018.

[5] B. A. Abu and T. Choo-Yee, "Soft skills recommendation systems for IT jobs: A Bayesian network approach," in *2011 3rd Conference on Data Mining and Optimization (DMO)*, IEEE, 2011, pp. 82-87.

[6] Z. Yu, T. Srikanta and C. Graham, "Learning graphical models from a distributed stream," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, IEEE, 2018, pp. 725-736.

[7] S. Yi, W. Limin and S. Minghui, "Label-Driven Learning Framework: Towards More Accurate Bayesian Network Classifiers through Discrimination of High-Confidence Labels," *Entropy*, vol. 19, no. 12, p. 661, 2017.

[8] H. Rodney, K. Valaskova and P. Durana, "The artificial intelligence recruitment process: how technological advancements have reshaped job application and selection practices," *Psychosociological Issues in Human Resource Management*, vol. 7, no. 1, pp. 2-47, 2019.

[9] N. Nawaz, "Artificial Intelligence interchange human intervention in the recruitment process in Indian Software Industry," volume, 2019.

[10] J. Li, "Application and Research of Naive Bayes Algorithm in Spam Filtering," *International Journal of Computer Applications Technology and Research*, vol. 9, no. 08, 2020.

[11] S. Vijayarani and S. Deepa, "Naive Bayes Classification for Predicting Diseases in Haemoglobin Protein

Sequence," *International Journal of Computational Inteeligence and Informatics,* vol. 3, no. 4, 2014.

[12] G. Akansh, K. Lokesh, J. Rachna and N. Preeti, "Heart Disease Prediction Using Classification (Naive Bayes)," in *Proceedings of First International Conference on Computing, Communications, and Cyber-Security*, 2020.

[13] K. Vembandasamy, R. Ssipriya and E. Deepa, "Heart Disease Detection Using Naive Bayes Algorithm," *International Journal of Innovative Science, Engineering and Technology,* vol. 2, no. 9, 2015.

[14] C. V.Verma, D. C. Raman and D. S. M. Ghosh, "Prediction of Heart Disease in Diabetic patients using Naive Bayes Classification Technique," *International Journal of Computer Applications Technology and Research,* vol. 7, no. 7, 2018.

[15] A. Kulkarni, A. Patil and M. Patil, "Customer Churn Analysis and Prediction," *International Journal of Computer Applications Technology and Research,* vol. 8, no. 09, 2019.

[16] C. V.Verma and S. M. Ghosh, "Prediction of Heart Disease in Diabetic patients using Naive Bayes Classification Technique," *International Journal of Computer Applications Technology and Research,* vol. 7, no. 7, 2018.

[17] A. P. Wibawa, A. C. Kurniawan and D. M. Prawidya, "Naïve Bayes Classifier for Journal Quartile Classification," *International Journal of Recent Contributions from Engineering, Science & IT (iJES),* vol. 7, no. 2, pp. 91-99, 2019.

[18] M. Kevin, "Simple guide to confusion matrix terminology," *Data school,* vol. 25, 2014.

Mr. Patrick Nyanumba Mwaro

Patrick Nyanumba Mwaro Received bachelor's degree in Mathematics and Computer Science from Jomo Kenyatta University of Agriculture and Technology, Kenya in 2011. Currently pursuing Masters Degree in Computer systems at Jomo Kenyatta University of Agriculture and Technology. Research interest include Machine Learning and Artificial Intelligence

Dr. Kennedy Ogada

Dr. Kennedy Ogada is a Seniour Lecturer Jomo Kenyatta University of Agriculture and Technology, Department of Computing and Information Technnology. His research interest is in machine learning and artificial intelligence.

Professor Wilson Cheruiyot

He is a professor Jomo Kenyatta University of Agriculture and Technology, department of computing and information Technology. His research areas include: machine learning, Multimedia systems and communications, Information retrieval, image processing, Semantic Web, Distributed databases and internet, Data warehousing and Theory of computation.