Journal of Statistical Modelling and Analytics Vol 5(1),65-80. 2023
https://doi.org/10.22452/josma.vol5no1.5

Diagnosing Sexually Transmitted Disease from Some Symptoms Using Machine Learning Models

Nureni Olawale Adeboye[1*], Kehinde Adekunle Bashiru[2], Habeeb Afolabi[3] & Taiwo Ojurongbe[4]

Department of Statistics, Faculty of Basic and Applied Sciences, Osun State University, P.M.B 4494 Osogbo, Nigeria[1,2,3,4]

Corresponding author: nureni.adeboye@uniosun.edu.ng

Abstract

Sexually Transmitted Diseases (STDs) is of serious concern, especially among the youth. Efforts to eradicate such diseases are often frustrated due to different sociodemographic and clinical factors that usually lead to misdiagnosis. Thus, this paper applied six (6) distinct machine learning models to accurately analyze STD infections reported by 400 patients who attended Federal Polytechnic Ilaro Medical Center, Ogun State Nigeria. A range of relationships between weak to non-significant correlations was obtained between the 7 symptoms considered for the diagnoses and the diagnosis outcome, but no significant pattern was observed. However, the application of data mining tools revealed a hidden pattern that correctly predicted the outcome using the subjects' symptoms, age, and sex. Four out of the six machine learning models were adjudged to perform well using different performance metrics, of which logistic regression model was found to be the best. The model feature importance chart shows that vagina discharge and vagina itching have the highest and almost the same level of impact on the possibility of a diagnosed patient having STDs. Furthermore, a 100% performance of logistic regression implies that the model correctly predicted all the 309 true negatives and 101 true positives with a misclassification (misdiagnosis) of zero.

Keywords: Age, gender, machine learning, misclassification, sexually transmitted diseases.

1. Introduction

Sexually transmitted infections (STIs) remain a global burden due to their prevalent spread among youth and adults. According to World Health Organization, about 376 million infections have been reported globally which include the four most common curable STIs (chlamydia, gonorrhea, trichomonas and syphilis) (WHO, 2018). Many researchers proposed and performed various methods of extirpating the spread of STDs globally without success, hence appropriate STI diagnosis and treatment are crucial to

prevent the transmission of untreated infection (Maynud et al., 2004). Many adults are a carrier of STDs unknowingly and do not bother to subject themselves to medical diagnoses due to inadequate information about the factors that may be responsible for such infection, having suffered ignorantly, cases of misdiagnoses in the past. STD control using etiological diagnosis remains difficult due to limited access to laboratory diagnostics. Even when there is provision for facilities, tests result for people with suspected STDs take days or even weeks, making immediate treatment based on laboratory results unfeasible (Unemo et al., 2017). Most often, classification and clustering are needed to achieve accurate prediction and better treatment outcomes (Tosado et al., 2020). Improved technology, especially in the area of machine learning has increased the chances of diagnosis, classification, and prediction of diseases with some level of accuracy (Wong et al., 2019). This aids in early identification, quicker conclusion, and treatment of sicknesses.

Machine learning has been applied in the classification and prediction of the following diseases with some level of accuracies: parkinson's sickness (Gupta et al., 2020), alzheimer's illness (Shankar et al., 2020), skin malignant growth (Amin and Sharif, 2020), chest infection (Guan et al., 2020), diabetes mellitus (Ronoud & Asadi, 2019), fatigue from dried blood droplet (Hamadeh et al., 2020), maxillofacial injuries (Govindarajan et al., 2020), cancer from gene expression (Kumar et al., 2020), coronary heart disease (Ghiasi et al., 2020), survival outcome of patients following liver transplantation (Raji, 2016), dementia in old people (Lagunin et al., 2020), clinical score in alzheimer's disease (Lei, 2020), individuals at high risk of developing type 2 diabetes comorbidities (Dworzynski, 2020), prognosis of acute exacerbations chronic obstructive lung disease (Peng 2020), anticancer drug resistance (Chio, 2020), involuntary pathological hand tremor (Shahtalebi, 2020), heart attack (Sarkar et al., 2020), cardiovascular disease infection (Adeboye & Abimbola, 2020), gene expression analysis in breast cancer (Hassan et al., 2020), just to mention a few. As specified above, a mix of characterization, determination, and forecast of illnesses utilizing machine learning and information mining has been accounted for in the literature. The examinations are frequently refined or improved by utilizing streamlining strategies or transformative computational techniques.

Of recent, many researchers have developed machine learning models from electronic health records and different sources applying the multifactor classification approach to assess and predict risk groups for medical conditions and to identify major risk factors. Such targets include delirium occurrences (Corradi et al., 2018), alcohol use disorder (Kinreich et al., 2021), mortality in patients with liver disease (Lin Y et al., 2020), cardio-cerebrovascular events (Park et al., 2019), metabolic syndrome (Yu et al., 2020), and postpartum depression (Zhang et al., 2020). However, no specific model has been adjudged as the best machine learning model for diagnosing STDs in the literature.

This study attempts to research how STDs can be correctly diagnosed from some symptoms due to an increase in the rate of its transmission, contagious nature regardless of their age, gender, and date of

infection, as well as the challenges encountered in the application of the laboratory diagnostic test. The study proposed six unique information mining models to diagnose STDs based on the seven symptoms documented for the patients who reported for medical treatment at the medical center. The symptoms considered are dysuria, penile itching, vaginitis, candidiasis, virginal discharge, virginal itching and foul-smelling which literature has confirmed not to have been utilized in previous studies to evaluate STDs using machine learning, most especially in ameliorating the challenges of misclassification in the disease diagnoses.

The choice of machine learning techniques in this study has a robust place in literature where it has been established that the technique is an improved technology which has increased the chances of diagnosis, classification, and prediction of diseases with some level of accuracy (Wong et al., 2019).

2. Materials and Methods

2.1 Data Sources and Method of Collection

The target population for this study comprised all the records of students, staff, and non-staff who registered their health information at Federal Polytechnic Ilaro Medical Center, and were diagnosed with symptoms of STDs between the years 2018-2021. This study was approved by the ethics committee of the Federal Polytechnic Ilaro Medical Centre. The data contains information from a convenient sample of 400 patients for sufficient coverage, of which 361 are females and 39 are males; factors of age, sex and the seven symptoms were taken into consideration, being the maximum number of symptoms available in the hospital records as at the time of collecting the data. The age ranges between 9 and 60 years while gender was encoded in nominal form as “1” for females and “0” for males. Other factors are encoded in integers (“0” for non-presence and “1” for the presence of the symptoms). The data was split into two, with 80% of the data used as training, and the remaining 20% used as testing. The focus is on the fitting of six machine learning models based on classification algorithms using a grid search of 10-fold cross-validation.

2.2 Machine Learning Models

Five different stages were involved in the analytical process which are; data coding, data processing, data splitting into training and testing sets, model creation and model evaluation. Six classification algorithms were used in the diagnosis and prediction of STDs infection with five evaluation metrics to validate the best model which captures the goodness of fit.

The proposed inferences were carried out as easy to use tool stash, utilizing Jupyter I Python Note pad. The tool stash utilizes an extensive variety of superior execution figuring bundles to handle input information, create elements, train and test models. Besides, the created Jupyter I Python Note pad gives

an enabling platform to scientists to show a custom prescient model with extraordinary adaptability in highlighting features. The classification models and performance metrics are discussed below.

2.2.1 Logistic Regression

The technique is a particular type of strategic relapse which is most utilized when the information being referred to has a twofold result such as 0 or 1 (Adeboye & Adesanya, 2022). Suppose there are n independent observations $y_1, y_2, \cdots, y_n$ and that the i^{th} observation can be treated as a realization of a dichotomous random variable Y_i, then the logit (ℓ) of the underlying probability P_i for the predictors X_i is given as

$$\ell(P_i) = X_i'\beta \tag{1}$$

where β is the regression coefficient, Y_i is the patient status and $X_i = (X_{i1}, X_{i2}, \cdots, X_{i7})$ represents the symptoms. By taking an exponential of equation (1), the log-odds for the i^{th} observation becomes

$$\frac{P_i}{1-P_i} = e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots\beta_k X_{ik}} \tag{2}$$

By simple algebraic manipulation, the probability P_i in equation (2) becomes

$$P_i = \frac{e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots\beta_k X_{ik}}}{\mathbf{1}+ e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}\cdots\beta_k X_{ik}}} \tag{3}$$

2.2.2 Naïve Bayes Classifier

This is a regulated learning procedure that is utilized for grouping activities. It is mostly utilized in logical and prescient issues when the dimensionality of the data sources is high. In spite of the effortlessness, this is many times utilized in more refined order techniques (Adeboye & Abimbola, 2020). Suppose A represents the STDs status and B represents any of the predictor variables (symptoms), then

$$\mathrm{P}(A/B) = \frac{P(B/A)P(A)}{P(B)} \tag{4}$$

In equation (4), B is the proof and A is the speculation. $P(A|B)$ is the posterior of class (A, target) given indicator (B, variable). $P(A)$ is the class prior likelihood while $P(B|A)$ is the possible odd of the likelihood of an indicator in a given class and $P(B)$ is the prior likelihood of predictor variables.

2.2.3 Decision Tree

The algorithm works by dividing the observation into at least two homogeneous sets in view of the highly important variables making the study distinct. In fact, the model is similar to stratified technique in sampling. According to Pouriyeh et al. (2017), the mathematical expression is given as;

$$\mathrm{P} = \frac{n_A}{n_A+n_B} \tag{5}$$

2.2.4 Random Forest

The algorithm is the ensemble of a large number of individual decision trees, and the final classification or prediction is obtained from the class with the most votes. The idea is that a large collection of uncorrelated models (trees) operating as an attribute will outperform any of randomly selected individual constituent models. According to Adeboye and Abimbola (2020), random forest has nearly the same hyperparameters as a decision tree or bagging classifier. The importance for each feature on the constructed decision tree is then calculated as:

$$fi_i = \frac{\Sigma_{J:node\ J\ splits\ on\ feature\ i} ni_J}{\Sigma_{k \in all\ nodes} ni_k} \tag{6}$$

where fi_i is the importance of feature i and ni_J is the importance of node j.

This is then normalized to a value $0\ and\ 1$ by dividing equation (6) with the sum of all features' importance values given as:

$$normfi_i = \frac{fi_i}{\Sigma_{J \in all\ features} fi_J} \tag{7}$$

Thus, the final feature importance at the Random Forest level, is its average over all the trees given as:

$$\mathrm{RF}fi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T} \tag{8}$$

where, RFfi_i is the importance of feature i calculated from all trees in the random forest model; $normfi_{ij}$is the normalized feature importance for i in tree j and T is the total number of trees.

2.2.5 K-Nearest Neighbors

KNN is a supervised algorithm that can also be used for both classification and regression tasks. In KNN, there is no learning required as the model stores the entire dataset and classifies data points that are similar to it. It makes predictions based on the training data only. The distance functions in KNN are given as

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{9}$$

$$\text{Manhattan} = \sum_{i=1}^{k} |x_{i-} y_i| \tag{10}$$

$$\text{Minkowski} = [\sum_{i=1}^{k} (|x_{i-} y_i|)^q]^{1/q} \tag{11}$$

where x_i and y_i are as defined in equation (1).

2.2.6 Adaptive Boosting

This technique can also be utilized in classification and regression works by joining numerous powerless classifiers into a solitary stronger classifier. Adaboost frequently enhances the outcome acquired by individual classifiers. This can be expressed as given in the following equation.

$$F(x) = \sum_{t=1}^{T} \alpha_t \, h_t(x) \tag{12}$$

where $h_t(x)$ is the output of weak classifier t for input x (symptoms) and α_t is the weight assigned to the classifier and it is evaluated as $\alpha_t = 0.5 \times \ln\left(\frac{1-\varepsilon}{\varepsilon}\right)$. The weight of the classifier is straightforward, it is based on the error rate ε.

2.3 Assessment Methods

The evaluation metrics utilized to assess the classification results obtained by the above-listed techniques are as previously discussed in Adeboye and Abimbola (2020).

2.3.1 Area Under Curve (AUC)

The region under receiver operator characteristic bend estimated the nature of expectations regardless of the picked characterization limit. AUC closes to one is alluring.

2.3.2 Classification Accuracy (CA)

CA is the capacity to accurately forecast the right class. It is the proportion of a number of right forecasts to the all-out number of information tests. At default, CA has the accompanying definition:

$$\text{Classification accuracy} = \frac{Total\ number\ of\ True\ predictions}{Total\ number\ of\ predictions} \tag{13}$$

For binary classification, accuracy can also be estimated in terms of positive and negatives as follows:

$$\text{Classification accuracy} = \frac{TP-TN}{TP-TN-FP-FN} \tag{14}$$

2.3.3 F1 Score

$F1$ score is the reciprocal of the mean of the reciprocal of sensitivity and precision. F1 is needed to strike a balance between precision and sensitivity.

$$\text{F-Measure} = \frac{2(Precision \times Recall)}{Precision + Recall} \tag{15}$$

2.3.4 Precision

Precision quantifies the number of positive class predictions that actually belong to the positive class. Therefore, it is a better metric when the number of FALSE POSITIVE is high.

$$\text{Precision} = \frac{TP}{TP-FP} \tag{16}$$

where TP (true positives); FP (false positives); TN (true negatives); FN (false negatives) are the available class predictions.

2.3.5 Recall

Recall estimates the level of expectations that were accurately grouped. The review helps when the expense of bogus negatives is high and it is determined as follows:

$$\text{Recall} = \frac{TP}{TP-FN} \tag{17}$$

3. Results and Discussion

3.1 Results from Exploratory Analysis

A total of 400 patients' medical records of STDs were obtained from 2018 to 2021. Exploratory analysis of the data was carried out to highlight the pattern of the data for initial results. The mean age of the patients was 23.19 years old (SD = 6.24) while the modal age is 22 years old with the youngest age recorded to have STD is as early as 9 years old. Majority of the patients were females and most of the patients that are diagnosed for STI are between the ages of 13 and 36 years old. Details investigation was carried out on the dataset and it was found out that there are no missing values or outliers in the dataset.

Table 1: Descriptive statistics of the respondents

Estimate	Age	Sex	Dysuria	Peniel itching	Vagina discharge	Vaginal itching	Candid-iasis	Vagin-itis	Foul smelling	Status
Count	400	400	400	400	400	400	400	400	400	400
Mode	22.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2: Frequency and percentage analysis of the diagnosis outcome based on 7 symptoms.

Symptoms	Absent	Present
Vaginal itching	229 (57.25%)	171 (42.75%)
Dysuria	319 (79.75%)	81 (20.25%)
Vaginal discharge	256 (64%)	144 (36%)
Penile itching	385 (96.25%)	15 (3.75%)
Candidiasis	366 (91.5%)	34 (8.5%)
Foul-smelling	356 (89%)	44 (11%)
Vaginitis	397 (99.25%)	3 (0.75%)
Diagnosis	Negative	Positive
Severe STD	299 (74.75%)	101 (25.25%)

The baseline symptoms of STDs are shown in Table 2. Vaginal itching has the highest frequency and percentage of presence among the diagnosed patients. On the average, it can be inferred that 42.75% of the treated patients were positive of STDs. Overall, an aggregate of 101 (25.5%) patients were diagnosed to

https://doi.org/10.22452/josma.vol5no1.5

have suffered from severe STDs while the remaining percentage of 74.75% were confirmed not to be severely infected.

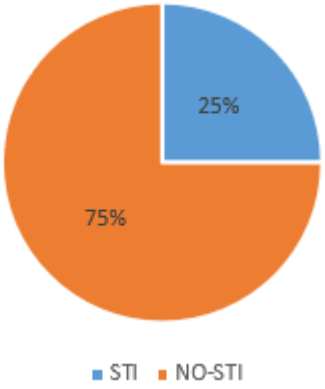

Figure 1: Distribution of STI classified status

The outcome of the target variable classified as "1" for the presence of STDs and "0" for its absence is represented in Figure 1. The pie chart shows that 75% of patients were not infected while 25% of the patients were infected according to the status classification. Further investigation shows that most of the patients that are diagnosed with STDs are between the ages of 13 and 36 years old while patients that are of age 22 years old were found to be the most infected as depicted in Figure 2.

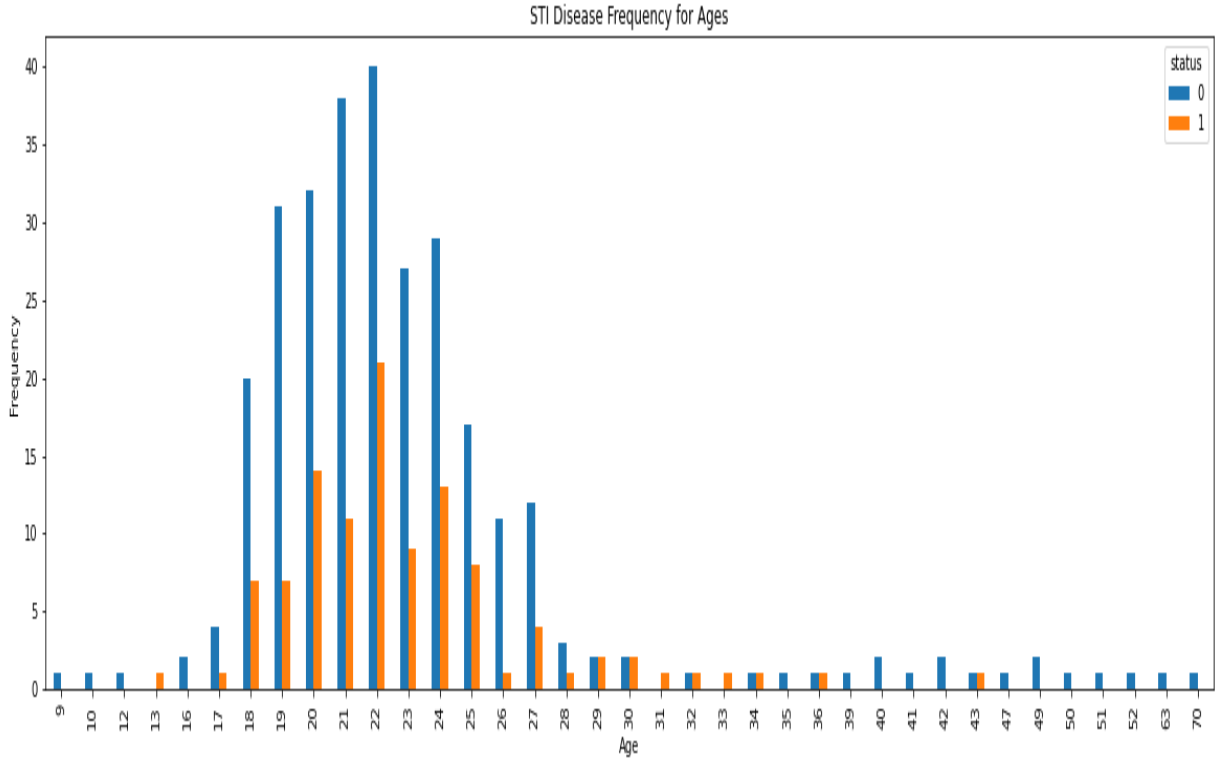

Figure 2: Distribution of STI status and age

https://doi.org/10.22452/josma.vol5no1.5

Figure 3 presents the correlation heatmap of the 2D correlation matrix of the features in the dataset. It shows both the pairwise relationships between the seven features contained in the dataset. The correlation map shows that there exist both positive and negative relationships among the features. The findings indicate that there are ranges of correlation observed between features. However, we do not encounter any strong relationship among features. The heatmap shows positive relationship between the age of patients with the symptoms of dysuria, peniel itching and foul-smelling, yet, it does not serve as an indicator for being affected with STDs. The gender of patients was found to be positively correlated with the majority of the symptoms except dysuria and peniel itching, and also with the possibility of a patient being infected with STDs. This gender has been traced to be female as it constitutes 90.25% of the diagnosed population.

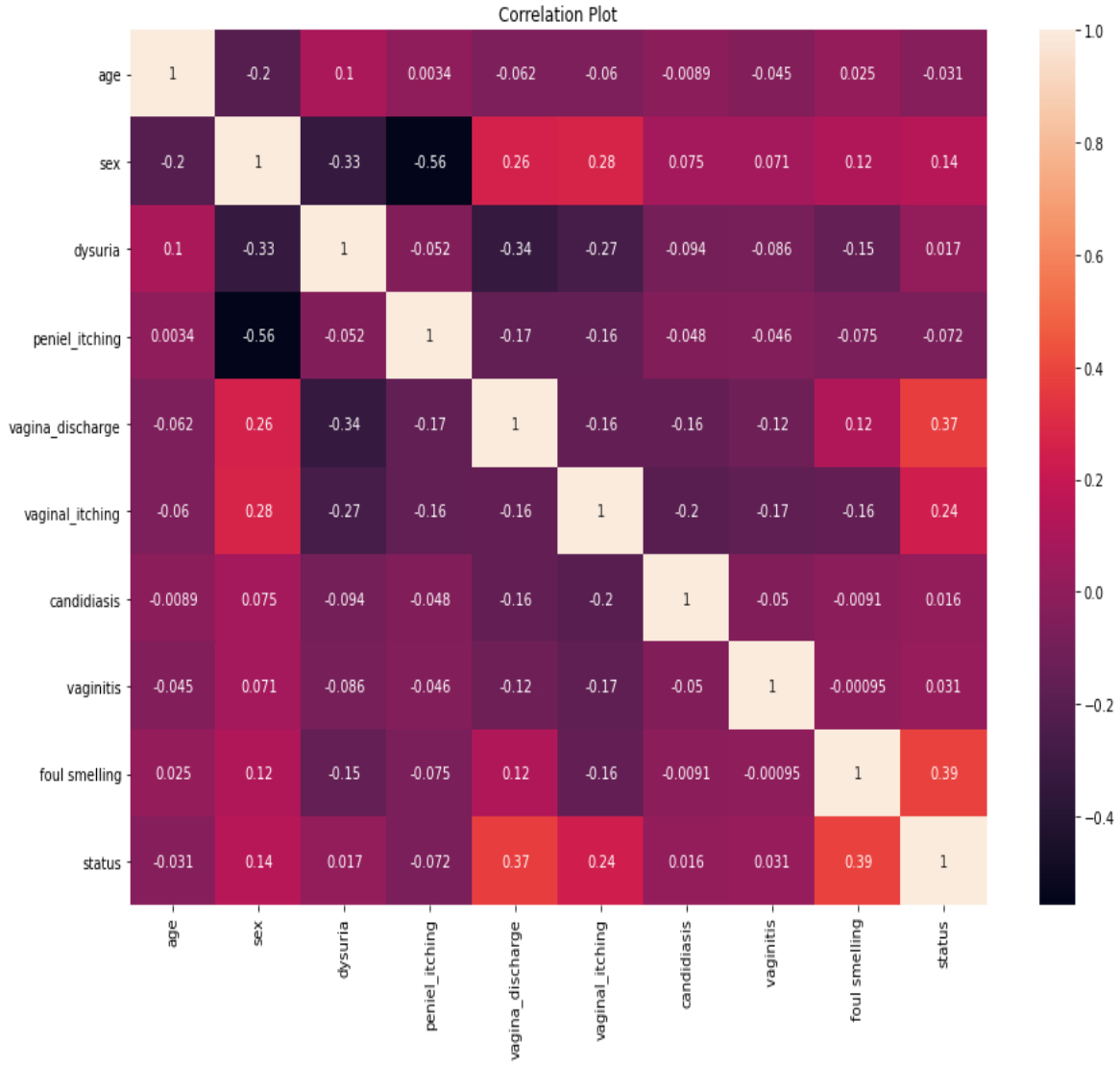

Figure 3: Correlation map of features

3.2 Model Assessment Results

Here, the data was split into 80% of training set and 20% of test set. The training set was fed to the model in other to learn the patterns in the data, thereafter, the test set was used for predicting the model accuracy. The model was trained using 10-fold revalidation with the base hyper parameters and the values of the predictions were then used to compute the metrics shown in Table 3 below. As earlier stated, the inferences are carried out as an easy to use tool stash, utilizing Jupyter I Python Note pad. The tool stash utilizes an extensive variety of superior execution figuring bundles to handle input information, create elements, train and test models.

Table 3: Model Assessment Metrics

Classifier	AUC	Accuracy	Recall	Precision	F-Score
Logistic Regression	**0.946708**	**0.95**	**0.939394**	0.885714	**0.911765**
K-Nearest Neighbors	0.530303	0.741667	0.060606	1	0.114286
Decision Tree	0.895507	0.916667	0.848485	0.848485	0.848485
Naive Bayes	0.766458	0.825	0.636364	0.7	0.666667
Random forest	0.907001	0.933333	0.848485	0.848485	0.875
AdaBoost	0.931557	0.941667	0.909091	0.909091	0.895522

The performance of the model as shown in Table 3 shows that the logistic regression model outperforms the other models with a classification accuracy of about 95%, an AUC of 94.6%, recall of 93.9, and an F-score of about 91.1%. This shows that the logistic regression model is a better classifier for the proper diagnosis of patients with STDs symptoms.

To understand the predictive capacity of the fitted models, the receiver operating characteristic curve (ROC) was constructed for each of the models. The ROC curve is a graphical plot which illustrates the diagnostic ability of a binary classifier system as its discrimination threshold varies.

As observed in the plot, it was seen that logistic regression has a curve that largely tends towards one which makes it outperform other machine learning models used for the analysis as shown in Figure 5 below.

https://doi.org/10.22452/josma.vol5no1.5

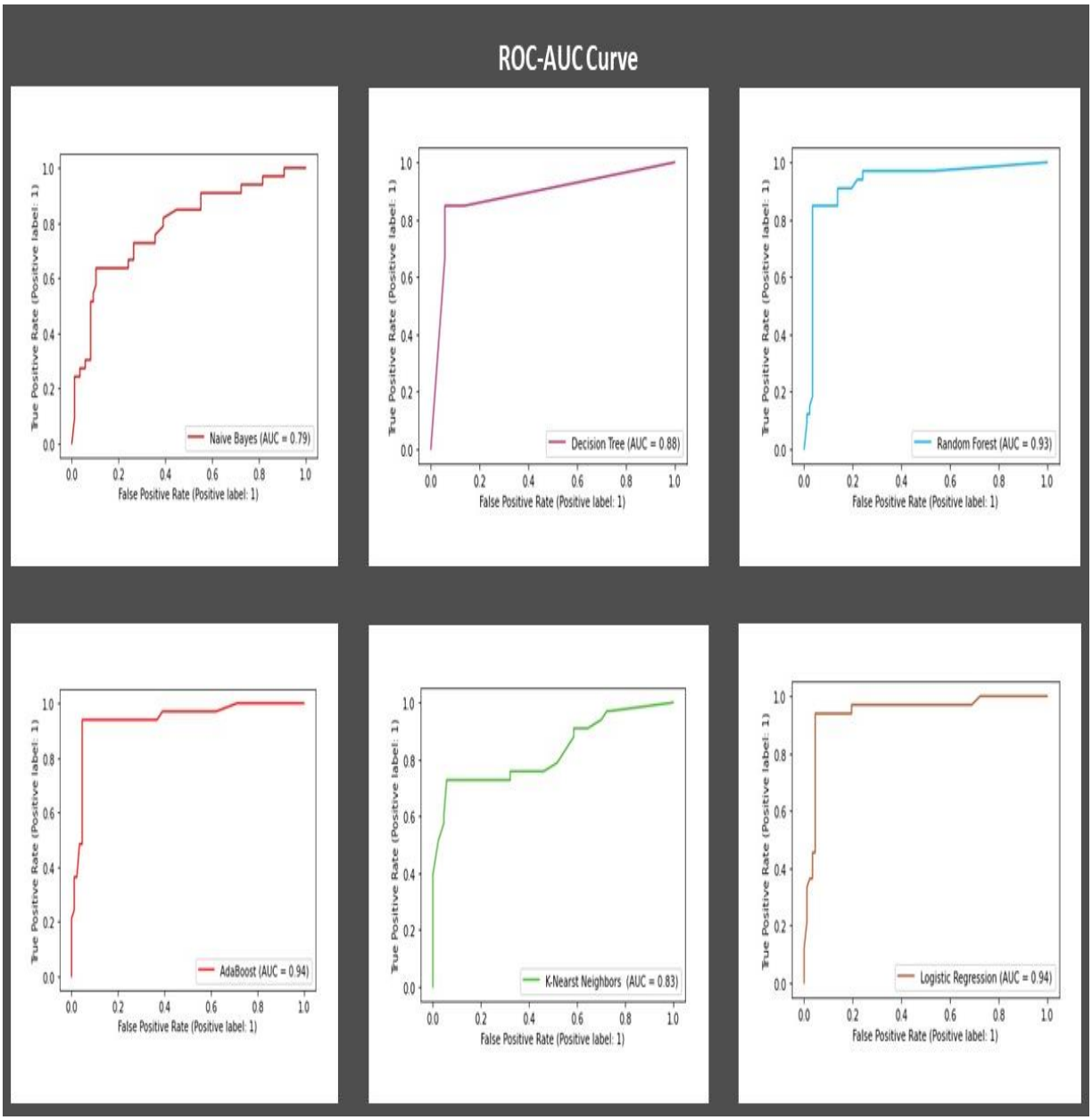

Figure 4: ROC curve

Based on the identified best model (logistic regression), a feature importance chart was constructed for the model as shown in Figure 5. The chart on a scale of 0 to 2 helps to visualize the best features contributing to the performance of the model and also to the major features contributing to the likelihood of a patient having STD.

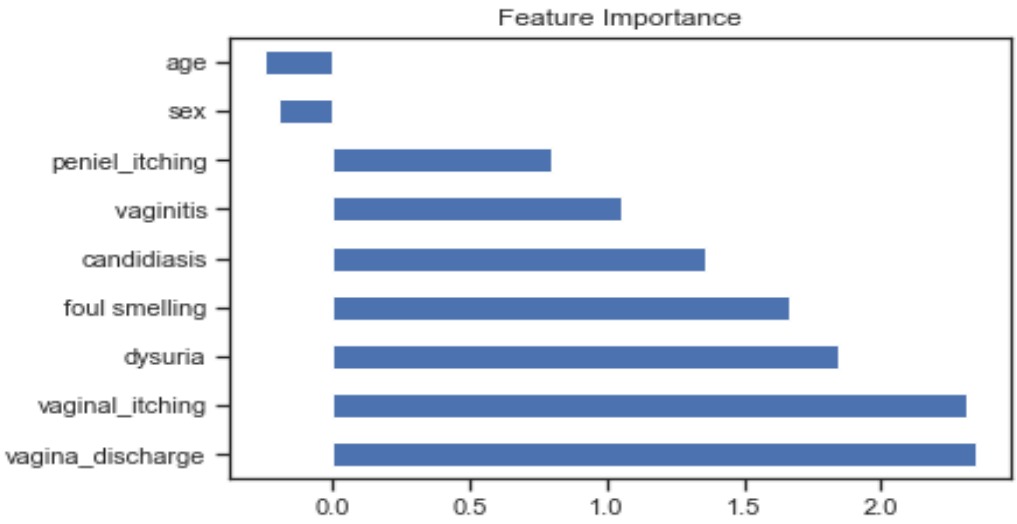

Figure 5: Feature importance of the best model

It was observed from the feature importance chart that vagina discharge and vagina itching have the highest scale and similar level of impacts in the possibility of a diagnosed patient having STDs. Dysuria, foul smelling, and candidiasis are among the major impacts while peniel itching and vaginitis are not of much significant impact in diagnosing a patient with STDs. The contributions of age and sex cannot be overemphasized, though very important medical records but cannot be said to be major diagnostic features. From the analysis, it is identified that young adults (13 and 36 years old) especially women are at high risk of STDs. This is in tandem with the work by van Teijlingen et al. (2020) where it was opined that STDs are significantly more common in women less than 25 years and in this age group. The finding that vagina discharge and vagina itching have the highest impacts in the possibility of a diagnosed patient having a STDs, followed by the impacts of dysuria, foul smelling, and candidiasis are in tune with most laboratory tests results as presented in literature (Korenromp et al., 2018).

Logistic regression model was found to be the best diagnostic technique based on the evaluation metrics considered in the research. This outcome buttressed the rationale why most research in health care investigations usually considered logit models (Adeboye and Adesanya, 2022; Adeniyi et al., 2022).

The challenges in diagnosing STDs in women is emphasized in the results, as the female gender constitute the largest population observed to have sought medical attention about the disease.

4. Conclusion

This article has investigated several important problems regarding sexually transmitted infection in patients. The principal motivation was to provide a framework for analyzing administrative healthcare data to generate significant features that assist machine learning techniques in correctly classifying patients with

sexually transmitted infections. This was achieved by building six classification models around the dataset, and logistic regression happened to be the best based on the evaluation metrics discussed in the methodology. The feature importance chart of the best model shows that vagina discharge and vagina itching have the highest scale and similar level of impacts in the possibility of a diagnosed patient having a STDs. Other features including Dysuria, foul smelling, and candidiasis provide major impact while penile itching and vaginitis are giving less impact in diagnosing a patient with STDs. The contributions of age and sex cannot be overemphasized, though they are very important medical records, the evidence obtained indicate that both features cannot be said to be major diagnostic features. Further investigations are required to gain some insights for their contributions.

It is important to get tested for people that are exposed to unprotected sex. More so, medical research has confirmed availability of sufficient medical treatment for patients diagnosed with STDs. Diagnosis and prediction of diseases appear to be vital in treatment and prevention. Most often, classification and clustering are needed to achieve accurate prediction and hence arrive at a better treatment outcome as already established in previous studies (Tosado et al., 2020). It is recommended that there should be healthy conversations about sex education among young people. STDs do not always showcase symptoms; thus it is possible to have an infection without knowing.

Logistic regression model is recommended for use in designing an internet based mobile devices capable of being use as a decision diagnostic tool in the diagnosis of sexually transmitted diseases. The application of this research results has potentials in the reduction of misdiagnosis incidences, reducing the mortality due to Sexually transmitted diseases and improving the overall spreads of the disease from one person to another.

5. Acknowledgements

The authors are grateful to Federal Polytechnic Ilaro Medical Centre for making the data available and the institutional bioethics committee for given approval for the use of the data.

6. References

Adeboye, N. O. & Adesanya, K. K. (2022). On the survival assessment of diabetic patients using machine learning techniques. *International Journal of Research and Innovation in Applied Science.* 7(1): 69-75.

Adeboye, N. O. & Abimbola, O. V. (2020). An overview of cardiovascular disease infection: a comparative analysis of boosting algorithms and some single based classifiers. *Statistical Journal of the IAOS. 36(4): 1189-1198*.

Adeniyi, O. I., Afolabi, N. B., Akinrefon, A. A., Omekam, I. V. & Olonijolu, I. R, (2022). Factors influencing the choice of place of delivery of a first child among Nigerian women. *Tanzania Journal of Science,* 48(2): *324-334.*

Amin, J., Sharif, A., Gul, N., Anjum, M. A., Nisar, M. W., Azam, F. & Bukhari, S.A, C. (2020). Integrated design of deep features fusion for localization and classification of skin cancer. *Journal of Pattern Recognition Letters,* 131: 63-70.

Banik, P. P., Saha, R. & Kim, K. D. (2020). An automatic nucleus segmentation and convolutional neural network model-based classification method of white blood cell. *Journal of Expert Systems with* Applications, 149: 113-211.

Choi, J., Park, S., & Ahn, J. (2020). A reference drug based neural network for more accurate prediction of anticancer drug resistance. *All Science Journal,* 10(1): 1861.

Corradi, J. P., Thompson, S., Mather, J. F., Waszynski, C. M., & Dicks, R. S. (2018). Prediction of incident delirium using a random forest classifier. *J Med Syst,* 42(12): 261.

Dworzynski, P., Aasbrenn, M., Rostgaard, K., Melbye, M., Gerds, T. A., Hjalgrim, H., & Pers, T. H. (2020). Nationwide prediction of type 2 diabetes comorbidities. *Journal of* Science, 10(1):1776.

Feldman, J., Thomas-Bachli, A., Forsyth, J., Patel, Z. H., & Khan, K. (2019). Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise. *Journal of Medical Information*, 26(11): 1355.

Ghiasi, M. M., Zendehboudi, S. A., Mohsenipour, A. A. (2020). Decision tree-based diagnosis of coronary artery disease: CART model. *Journal of Classification and Regression Tree*, 192: 105400.

Govindarajan, P., Soundarapandian, R. K., Gandomi, A.H., Patan, R., Jayaraman, P., & Manikandan, R. (2020). Classification of stroke disease using machine learning algorithms. *Journal of Sciences*, 32(3): 817–828.

Guan, Q, Huang, Y, Zhong, Z, Zheng, Z, Zheng, L, & Yang, Y. (2020). Thorax disease classification with attention guided convolutional neural network. *Journal of Pattern Recognition Letter, 131: 38*–45.

Gupta, U, Bansal H, & Joshi, D. (2020). An improved sex-specific and age-dependent classification model for Parkinson's diagnosis using handwriting measurement. *Computer Methods Programs Biomed.* 189:105305.

Hamadeh, L., Imran, S., Bencsik, M., Sharpe, G. R., Johnson, M.A., Fairhurst, D. J. (2020). Machine learning analysis for quantitative discrimination of dried blood droplets. *Science Textbook.* 10(1):3313.

Hassan, Z. A., Alsabi, Q., Ramirez-Vick, J. E., & Nosoudi, N. (2020). Characterizing basal-like triple negative breast cancer using gene expression analysis: a data mining approach. *Textbook Expert Systematic Application.* (148):113253.

Kamel. S. R., Yaghoub, Z. R., Kheirabadi, M. (2019). Improving the performance of support-vector machine by selecting the best features by Gray Wolf algorithm to increase the accuracy of diagnosis of breast cancer. *A Journal of Entomology Big Data*, 6(1): 90.

Kinreich, S., Meyers, J. L., Maron-Katz, A., Kamarajan, C., Pandey, A. K, & Chorlian, D B, (2021). Predicting risk for alcohol use disorder using longitudinal data with multimodal biomarkers and family history: a machine learning study. *A journal of* Psychiatry, 26(4): 1133-1141.

Korenromp, E. L., Mahiané, S. G., Nagelkerke, N. Taylor, M. M., Williams, R., Chico, R. M., Pretorius, C., Abu-Raddad, L. J., & Rowley, J. (2018). Syphilis prevalence trends in adult women in 132 countries–estimations using the Spectrum Sexually Transmitted Infections model. *Sci Rep*, 8:11503.

Kumar, A., & Halder, A. (2020). Ensemble-based active learning using fuzzy-rough approach for cancer sample classification. *A Textbook of Engineering Applications of Artificial Intelligence*, 91:103591.

Lagunin, A. A., Ivanov, S. M., Gloriozova, T. A., Pogodin, P. V., Filimonov, D. A., Kumar, S., & Goel, R. K. (2020). Combined network pharmacology and virtual reverse pharmacology approaches for identification of potential targets to treat vascular dementia. *A journal of Science Reports*, 10(1): 257.

Lei, B., Yang, M., Yang, P., Zhou, F., Hou, W., Zou, W., Li, X., Wang, T., Xiao, X., & Wang, S. (2020). Deep and joint learning of longitudinal data for Alzheimer's disease prediction. *A Textbook of Pattern Recognition*, 102: 107247.

Lin, Y. J., Chen, R. J., Tang, J. H., Yu, C. S., Wu, J. L., Chen, L. C., & Chang, S. S. (2020). Machine learning monitoring system for predicting mortality among patients with non-cancer end-stage liver disease: Retrospective study. *JMIR Med Inform*, 8(10): e24305.

Mayaud P., & Mabey, D. (2004). Approaches to the control of sexually transmitted infections in developing countries: old problems and modern challenges. *Sexually Transmitted Infections*; 80**:**174-182.

Park, J., Kim, J., Ryu, B., Heo, E., Jung, S. Y., & Yoo, S. (2019). Patient-level prediction of cardio-cerebrovascular events in hypertension using nationwide claims data. *A Journal of Medical Internet*, 21(2): 11757.

Peng, J., Chen, C., Zhou, M., Xie, X., Zhou, Y., & Luo, C. H. (2020). A machine learning approach to forecast aggravation risk in patients with acute exacerbation of chronic obstructive pulmonary disease with clinical indicators. *A journal of Science*, 10(1): 3118.

Pouriyeh, S., Sara V., Giovanna S., Giuseppe D., Hamid A., & Juan G. (2017). A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. *IEEE Symposium on Computers and Communications (ISCC)*, doi:10.1109/iscc.2017.8024530.

Raji, C. G., & Chandra, S. S. V. (2016). Predicting the survival of graft following liver transplantation using a nonlinear model. *Journal of Public Health*, 24(5): 443–52.

Ronoud, S., & Asadi, S. (2019). An evolutionary deep belief network extreme learning-based for breast cancer diagnosis. *A journal of Soft Computing*, 23(24): 13139–59.

Sarkar, B. K. (2020). Hybrid model for prediction of heart disease. *A journal of Soft Computing*, 24(3): 190325.

Shahtalebi, S., Atashzar, S. F., Samotus, O., Patel, R. V., Jog, M. S., & Mohammadi, A. (2020). Characterization and deep mining of involuntary pathological hand tremor using recurrent neural network models. *A journal of Science*, 10(1): 2195.

Shankar, K., Lakshmanaprabu, S. K., Gupta, D., Maseleno, A., & de Albuquerque, V. H.C. (2020). Optimal feature-based multi-kernel SVM approach for thyroid disease classification. *Journal of Supercomputing*, 76(2): 1128–43.

Tosado, J., Zdilar, L., Elhalawani, H., Elgohari, B., Vock, D. M., Marai, G. E., Fuller, C., Mohamed, A.S.R., & Canahuate, G. (2020). Clustering of largely right-censored oropharyngeal head and neck cancer patients for discriminative groupings to improve outcome prediction. *A journal of Science*, 10(1): 3811.

Unemo, M., Bradshaw, C. S., Hocking, J. S., de Vries, H. J. C., Francis, S. C., & Mabey, D. (2017). Sexually transmitted infections: challenges ahead. *A journal of Lancet Infectious Diseases*, 17(8): 235–79.

Van Teijlingen, N. H, Helgers, L. C., Zijlstra-Willems, E. M., Van Hamme, J. L., Ribeiro, C. M., Strijbis, K., & Geijtenbeek, T. B. (2020). Vaginal dysbiosis associated-bacteria Megasphaera elsdenii and Prevotella timonensis induce immune activation via dendritic cells. *Journal of Reproductive Immunology*, 138: 103085.

WHO (2018). *World Health Organization.* Report on global sexually transmitted infection surveillance. Geneva: WHO; Available from: *https://www.who.int/reproductivehealth/publications/stis-surveillance-2018/en*

Wong, Z. S. Y., Zhou, J., & Zhang, Q. (2019). Artificial intelligence for infectious disease big data analytics. *A journal of Infectious Disease Health*, 24(1): 44–8.

Yu, C., Lin, Y., Lin, C., Wang, S., Lin, S., & Lin, S. H. (2020). Predicting Metabolic Syndrome with Machine Learning Models Using a Decision Tree Algorithm: Retrospective Cohort Study. *Journal of Medical Information*, 8(3): 17110.

Zhang, W., Liu, H., Silenzio, V. M. B., Qiu, P., & Gong, W. (2020). Machine learning models for the prediction of postpartum depression: Application and comparison based on a cohort study. *Journal of Medical Information*, 8(4): 15516.

Zhang, X., Zhang, Y., Zhang, Q., Ren, Y., Qiu, T., Ma, J., & Sun, Q. (2019). Extracting comprehensive clinical information for breast cancer using deep learning methods. *International Journal of Medical Information*, (132): 103985.