Journal of Applied Artificial Intelligence
https://jaai.sabapub.com
ISSN: 2709-5908

2022 Volume 3, Issue 2: 24–38

DOI: https://doi.org/10.48185/jaai.v3i2.667

A Framework for Predictive - Diagnosis of Prevalent Illness among University Students

Dauda Olorunkemi Isiaka [1,*], Joshua Babatunde Agbogun [2], Taiwo Kolajo[3]

[1,3]Department of Computer Science, Faculty of Science, Federal University Lokoja, Kogi State, 058, Nigeria.
[2] bDepartment of Computer Science and Mathematics, Faculty of Natural Sciences and Environmental Studies, Godfrey Okoye University, Enugu, 051, Nigeria.

Received: 26.11.2022 • Accepted: 20.12.2022 • Published: 30.12.2022 • Final Version: 31.12.2022

Abstract: The issue of identifying the prevalence of sickness that is linked to the population of a nation, state, neighborhood, organization, or school has not been taken into consideration by the majority of prior studies on the prediction of illness among populations. They frequently merely choose any sickness based on assumption, while those that determined the prevalence of the condition before developing their framework utilized survey data or data from web repositories, which removes idiosyncrasies from those data. In order to increase performance, this research suggests an enhanced data analytics framework for the predictive diagnosis of common illnesses affecting university students. In order to do this, exploratory data analysis (EDA) using a multivariate analytic technique was conducted using a high-level model methodology using CRISP-DM stages. When the suggested strategy was evaluated on support vector machines, ensemble gradient boosting, random forest, decision tree, K-neighbors, and linear regression machine learning models, experimental findings revealed that it outperformed current methods.
In comparison to other reviewed frameworks that used survey datasets, standardized or online repositories' dataset, the framework with emphasis on the ensemble Gradient Boosting classifier and regression had accuracy of 100% and mean absolute error of 0.18, respectively. It is also steady due to its ability to manage both small and big data sets without impacting the model's performance. The enhanced results through localized dataset demonstrate the benefit of including local data sources in the process of developing models for the diagnosis and prognosis of prevalent illnesses of any area with people.

Keywords: data analytics, framework, prevalent illness, machine learning, prediction

1. Introduction

The expense of caring for those who have common illnesses has increased significantly as a rising section of the global population does. A common disease like malaria, diabetes, or hypoglycaemia affects almost half of the population of a country, state, neighbourhood, organization, or school [1].

Common diseases are responsible for 70% of the distractions that prevent people from completing their daily obligations, careers, and education [2]. Since their health is more prone to swiftly deteriorate owing to situations like living alone for the first time, having irregular eating habits, being uninformed of a prevalent ailment, etc., students are particularly vulnerable patient

* Corresponding Author: dauda.isiaka@fulokoja.edu.ng

groups in educational settings [3]. Today, an increasing percentage of people worldwide experience many of these diseases. It impacts patients' quality of life and put a strain on their loved ones and caregivers.

Instead of focusing explicitly on evidence-proof to inform potential patients within a population of a prevalent illness and then encourage them to take precautions and safety measures while pursuing daily needs of life like education, sustenance, etc., previous researchers have done works regarding the impact on the quality and safety of patients' care. Furthermore, determining the frequency of sickness relative to the population of a nation, state, neighborhood, organization, or school has not been taken into account in the majority of earlier studies on the prediction of illness among communities. These vast amounts of data gathered from interactions between patients and healthcare professionals would aid in lowering costs and raising the standard of service. Additionally, it aids in disease surveillance, community health management, and clinical decision support [4].

Radiology, serious illness management, public health, and targeted therapy are just a few of the care-related domains where the use of data analytics has shown promise. This might increase the effectiveness of medical delivery, lessen administrative costs, and hasten the discovery of sickness or illness. Despite the fact that data analytics have increased the quality of patient care management, there are still several difficulties [5]. And also, It has never been discovered to be very useful in this respect to use controlled clinical trials and personal work experiences to reduce the principal effects of usual sickness occurrences [6].

These challenges stemmed from the substantial proof healthcare, hybrid data (Univariate, Multivariate and High Dimensional Data) and misleading outliers. According to Morgenstern et al. [7], supervised learning prediction models have been used to forecast a broad range of outcomes for population health. Many illnesses that are more common in certain areas or situations, such as the market or education, with limited access to conventional health data, may fall short of coverage.

Most researchers relied heavily on digital health records and investigator-generated data, including the use of relatively small study cohorts, they did not analyze a large number of observations and features. If machine learning methods are utilized to leverage unique data sources for research in these regions or environments, it could enable greater study of neglected diseases. It can be challenging to obtain large sample numbers or large feature data when using various data sources, which may have an influence on the performance of algorithmic machine learning. Since both health professionals and patients are important stakeholders in the administration of patient care, the data must be analyzed in order to gain clarity and support the decisions made by the care providers for the patients.

As a result, manually examining the information from these papers takes time and is inefficient because it frequently requires specialized knowledge. Likewise, data analytics might be used effectively to establish a framework for managing the medical care of those with common illnesses.

This study expands on earlier works [7, 8] that offered an abbreviated report. Based on the use of more complex localized data aspects, a strong conclusion on the significance of creating a strategy for the predictive-diagnosis of common disease among students of the university would be drawn.

Regarding [7], the following enhancements were made specifically:

i. The students' health records, a thorough and updated version of the localized collection, were utilized. Compared to the one in [7], which is an asymmetrical dataset, the new dataset has an additional 1,049 observations and is balanced, with an equal distribution of bachelor students at all levels, and also [8] whose dataset was from survey and Google forms. A more realistic foundation for the suggested framework's evaluation will be made possible by the expansion of the localized dataset's size and richness.

ii. To boost the output quality of a suggested framework, two tasks—classification and an admission's runtime prediction module—have been combined.

The purpose of this study is to devise a framework for identifying illnesses that are often experienced by college students and to create a machine learning-based predictive diagnosis engine.

This research makes a contribution by suggesting an enhanced strategy for pre-processing student health records via;

i. gather a lot of elements,

ii. balance lacking values, notations, jargons, and synonymous terms in students' health records, and

iii. use localized sources of data to develop business intelligence methodology for determining prevalent illness among any population space.

2. Related Work

In addition to making predictions about depression in university freshmen, [8]'s research sought to understand why university students in Bangladesh, and undergraduates in particular, experience depression disease. A survey was used to gather the information for their study, and it was conducted on paper and via a Google survey form. Gradient Boost Algorithm and Deep Learning both have an F-Measure of 63%, making them the two algorithms with the fewest false negatives. The other four technologies utilized for comparison were Generalized Linear Model, Random Forest, K-Nearest Neighbor, and Support Vector Machine. Utilizing data from a localized dataset might improve the performance of the authors' model because using a survey approach to obtain data could be inaccurate and biased., as most responders will not give the true response to certain questions.

The authors [9] used a framework to develop and evaluate machine learning (ML) classification models for the prediction of diabetes patients, including Logistic Regression, KNN, SVM, and RF. The Pima Indian Diabetes Database (PIDD), which comprises 768 rows and 9 columns, was subjected to machine learning techniques.

Using two different datasets, the authors of [10] created a diagnosis system that focuses on Random Forest, Support Vector Machine, Naïve Bayes and Decision Tree predictions algorithm models to predict diabetes (Frankfurt Hospital in Germany and PIDD provided by the UCI machine learning repository). This work might use some improvements, such applying an ensemble method, and utilizing a localized dataset to forecast diabetes would help us get better outcomes. Only the COVID-19 patient's geographic, social, and economic circumstances, clinical risk factors, medical reports, and demographic information are included in the proposed model to predict rescue and death. The healthcare records of 1,028 people with type 2 diabetes and 1,028 people without the disease were taken from de-identified data in order to estimate the probability

of developing type-2 diabetes. The experimental findings reveal the models' effectiveness with an Area under Curve (AUC) varied from 0.79 to 0.91.

[11] created a machine learning model utilizing characteristics from the current year to forecast the prevalence of Type 2 Diabetes in the future year (Y + 1). (Y). The authors used ensemble machine learning methods to apply logistic regression, RF, SVM, XGBoost, and other algorithms to predict the result of prediabetes, diabetes, and non-diabetes. According to the experimental findings, the logistic regression model had a low accuracy of 71 percent and the RF had a high accuracy of 73 percent.

[12] used feature selection to improve the model's accuracy when classifying the diabetes dataset using SVM and NB algorithms. For analysis, PIDD is downloaded from the UCI Repository. The K-fold cross-validation model was used by the authors for training and testing; the SVM classifier performed better than the NB technique, providing about 91 percent accurate predictions. The authors do recognize that in order to improve the model, they must use the most recent dataset, which will include more characteristics and rows.

To identify heart illness at an early stage, the authors of [13] developed the unsupervised machine learning method K-means clustering for the UCI heart disease dataset. Dimensionality reduction using PCA was employed, and the method's results show 94.06 percent accuracy in predicting early heart illness. To improve the generalization or acceptability of the outcome, the authors should implement the suggested approach utilizing more than one algorithm.

[14] The logistic regression classification approach was used by the authors to build a prediction model for the categorization of diabetes data. 459 patients make up the training data in the dataset, whereas 128 instances make up the testing data. The prediction accuracy using logistic regression was attained at 92 percent. The key disadvantage of this research is that the authors have not tested the model with other diabetes prediction algorithms and therefore it cannot be validated. This work might use some improvements, such applying the ensemble method and employing a localized and generic collection of patient health records to forecast for better outcomes.

In order to address health-related issues by enabling medical professionals to identify diseases at an early stage, the authors of [15] created a prediction model that analyzes the user's symptoms and forecasts the disease using machine learning algorithms (DT classifier, RF classifier, and NB classifier). A sample of 4920 patient records with diagnoses for 41 illnesses makes up a dataset.

3. Materials and Methods

This contains the following sub-sections: Description of existing system, High Level Model of the Proposed DAF, and implementation method. This section describes the implementation method used to realize the Data Analytics Framework for the Predictive-Diagnostics of Prevalent Illness (DAFPDPI). Below is a full description of part all of the sub-sections;

3.1. Description of Existing System

The process of prevalent illness diagnosis is to identify if an individual from the population sample (undergraduate students) is suffered with the identified prevalent illness. The class of diagnosis could be either you are suffered by the prevalent illness disease or other illness. The existing system focuses on disease endpoint using symptomatic approach only, rather than considering also risk factors like age, level, semester, and even vital signs.

3.2. High Level Model of the Proposed DAFPDPI

Knowing that the objective is to create a predictive-diagnostic model, with the capacity to categorize patients as being afflicted by the prevalent sickness or not, as well as anticipate the length of time that patients with such illness would need to be admitted, we advocated the use of electronic inputs and storage utilizing CSV files and data analytics strategy.

The Federal University Lokoja Health Center's 1048 students make up the dataset. We would pre-process the formalized dataset, which has roughly 90 characteristics, to get rid of misfits, perform an Exploratory Data Analysis (EDA), and then build model for the diagnosis of prevalent illness, and also further predict the admission duration of patients.

Figure 1 illustrates the model design within the framework and lists the activities that take place there.

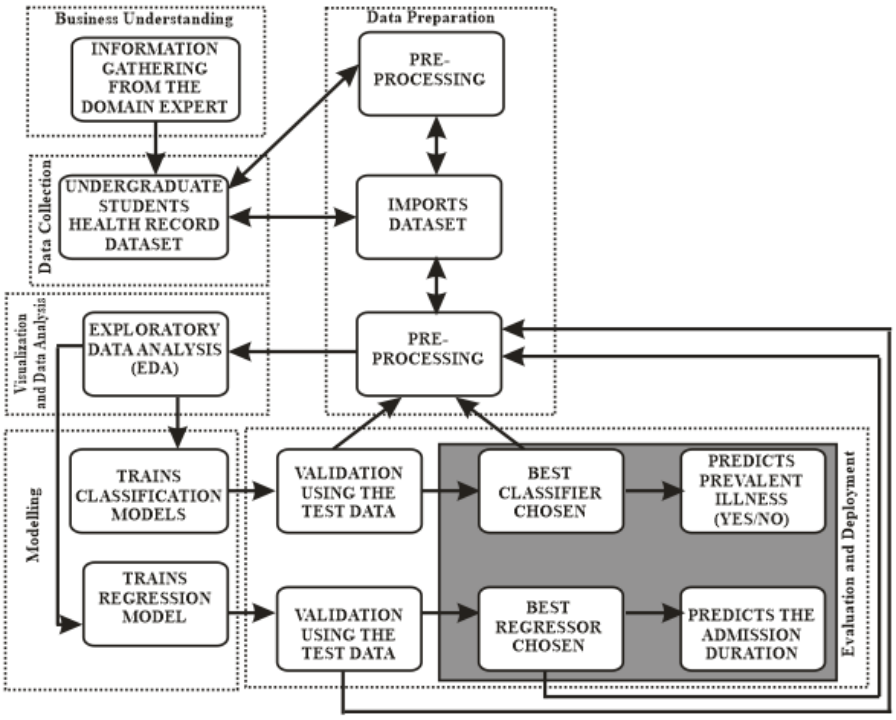


Figure 1. DAFPDPI High-Level Model being proposed.

The general model and its constituent parts are shown in Figure 1 above.

i. Business Understanding: In this stage of the model, we would meet with subject-matter experts to get their opinions on the most common ailment affecting students and conduct in-person interviews. Additionally, additional data would be obtained in order to help the model.

ii. Data Gathering: Information from health professionals as well as a dataset of student health records with all of their features will be acquired in order to have a deeper understanding of the various aspects and their significance.

iii. Data Preparation: The acquired dataset is pre-processed, imported, converted, and organized in order to make it suitable for additional processing in the next step.

iv. Data Analytics and Visualization: To perform feature engineering and get deeper visual understanding of the dataset through the presentation of graphs and tables, an exploratory data study utilizing both unilateral and multivariate analytic approaches will be used.

v. Modeling: To forecast and classify the most common sickness (task 1) and the length of hospitalization, the models would be constructed using the suitable supervised machine learning methods (task 2).

3.3. Methodology

Because the research is data-driven, we are using the Cross Industry Standard Process of Data Mining (CRISP-DM) paradigm for this suggested model.

3.4. An explanation of Dataset

The data for the network training and testing, which are composed of the characteristics to train the model and the target variable to categorize the output of the classification network together with their values and measurements, are contained in the table, which has 1048 sets of observations. To develop models for classification and regression, respectively, which result in the values of the dependent variable, the characteristics may be categorical, nominal, or ordinal in type.

Table 1 below displays the initial data set gathered from the university:

Table 1. A representative original dataset from the Federal University of Lokoja's Health Center

S/N	File No	Sex	Age	Type of Attendance	Complaints	Diagnosis	Outcome
1	72	F	19	FOLLOW UP	HEADACHE, BODY WEAKNESS, BODY PAIN, NECK PAIN	PVD	TREATED
2	325	F	16	FOLLOW UP	HEADACHE, BODY WEAKNESS, BODY PAIN, ABDOMINAL PAIN	BODY ACHE	TREATED
3	446	F	18	NEW	CATARRH, CHEST PAIN, COUGH	PVD	TREATED
4	640	F	20	FOLLOW UP	HEADACHE, CATARRH, COUGH	URTI	TREATED
5	14	F	21	FOLLOW UP	FEVER, ABDOMINAL PAIN, DIARHEA, VOMITTING	HTN	TREATED
6	131	M	17	FOLLOW UP	FEVER, COUGH, ABDOMINAL PAIN	PLASMODIASIS	TREATED
7	365	M	18	FOLLOW UP	FEVER, HEADACHE	MALARIA	TREATED
8	122	M	31	FOLLOW UP	FEVER	MALARIA	TREATED
9	72	F	19	FOLLOW UP	HEADACHE, BODY WEAKNESS, BODY PAIN, NECK PAIN	MALARIA	TREATED

4. Result

4.1. Data Transformation and Cleaning

In order to make sure there are no null data, the data gathered from the Health Center was cleaned by checking for missing values, contaminants, and data imputation. An interval scale “age group” feature was generated from the “age” column, and also, the “complaints” column was broken into all the possible complaints as features towards data exploration and model building.

4.2. Exploration and Preparation of Data

In this stage, an interpretation of the data was used to comprehend the data in order to describe what the dataset contains by tabulating all relevant parameters and also to depict the dataset's behaviors using both univariate and multivariate analysis techniques. Furthermore, Pearson correlation coefficient was used to find out whether there are correlations amongst the features of the dataset.

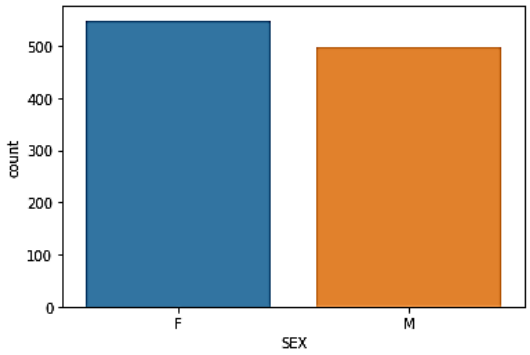


Figure 2. Gender

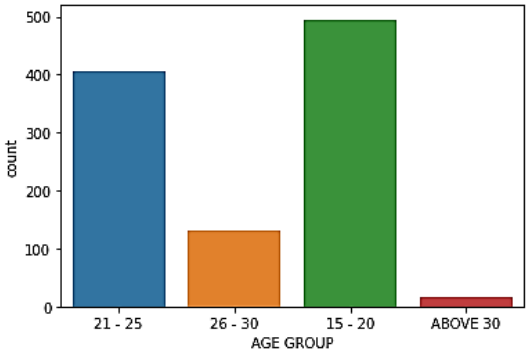


Figure 3. Age Group

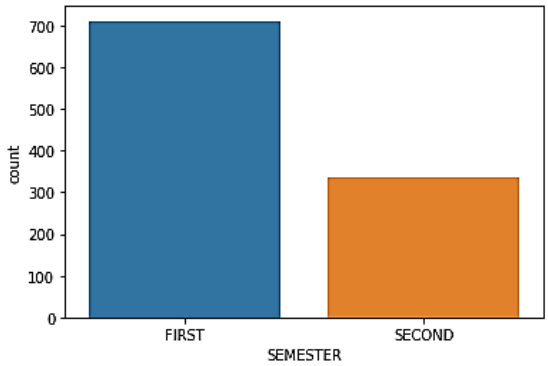


Figure 4. Semester

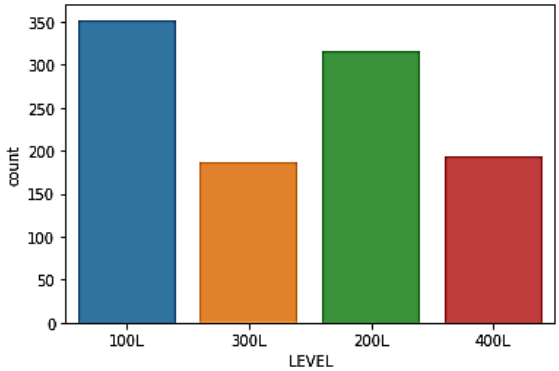


Figure 5. Level

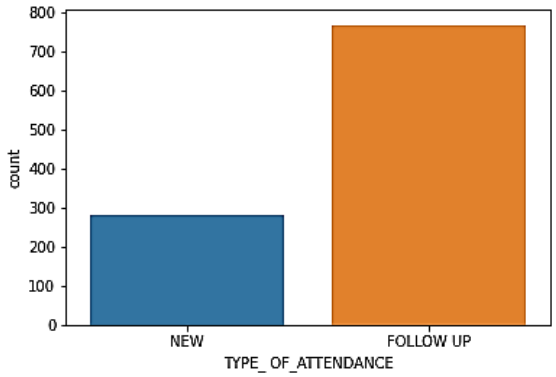


Figure 6. Attendance Type

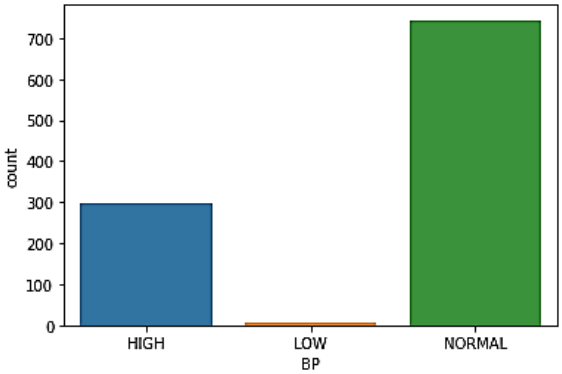


Figure 7. Blood Pressure

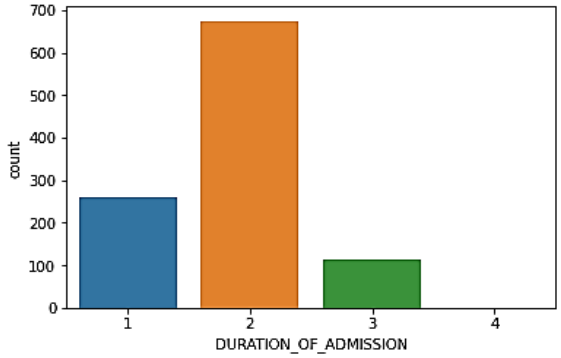


Figure 8. Admission Duration

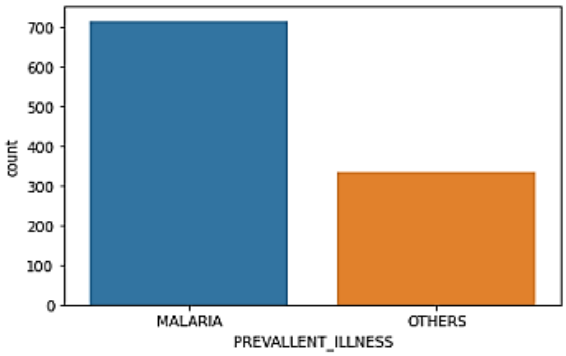


Figure 9. Prevalent illness

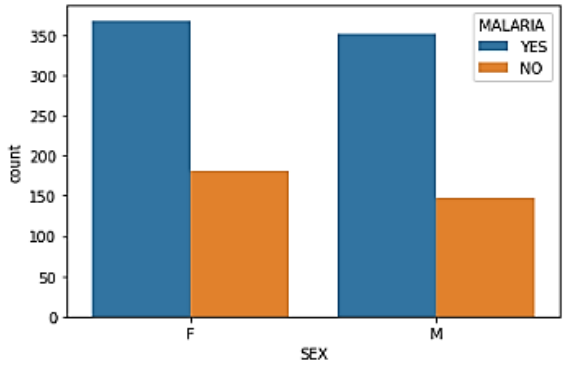


Figure 10. Sex versus Malaria

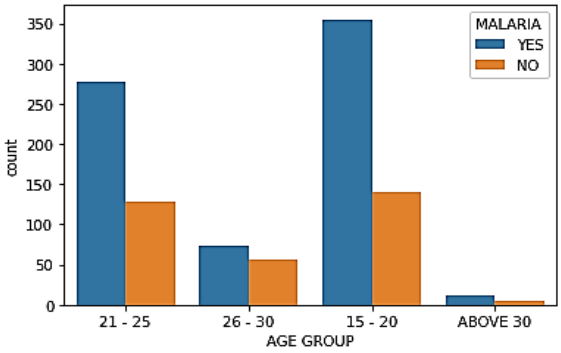


Figure 11. Age Group versus Malaria

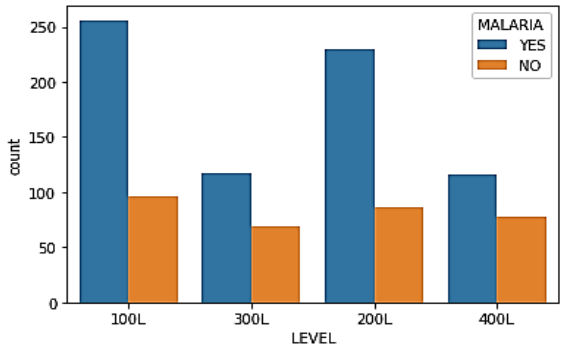


Figure 12. Level versus Malaria

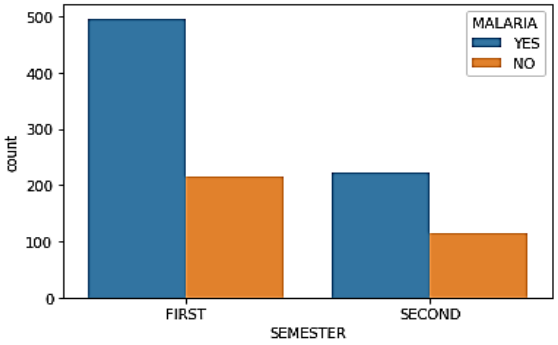


Figure 13. Semester versus Malaria

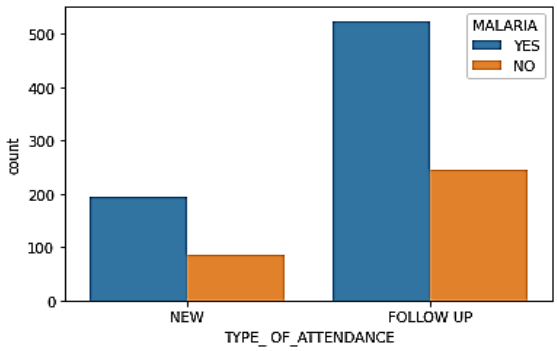


Figure 14. Attendance Type versus Malaria

4.3. Model Training and Implementation

We trained the model for the first task of the framework using the K-Neighbor Classifier, Random Forest Classifier, Gradient Boosting Classifier, Decision Tree Classifier and Support Vector Classifiers from the "modelselection" library of the "SKlearn" package installed into the Python.

We trained the model for the second task of the framework using the Linear Regression, Gradient Boosting Regression, Random Forest Regression, Decision Tree Regression and Support Vector Regression from the "modelselection" library of the "SKlearn" package installed into the Python.

Table 2. The output of the predicted admission duration for the prevalent illness.

	FILE_NO	MALARIA	DURATION_OF_ADMISSION
0	250	YES	2.127195
1	258	YES	2.167702
2	2662	YES	2.158686
3	257	NO	0.969604
4	2165	NO	2.079713

4.4. Step 7: Model Evaluation

The accuracy, margin of error, precision, recall, and F1 score of the seven (7) classifiers used to create the model were evaluated using a confusion matrix, and their respective results were plotted on Figure 15 – Figure 21 as shown below:

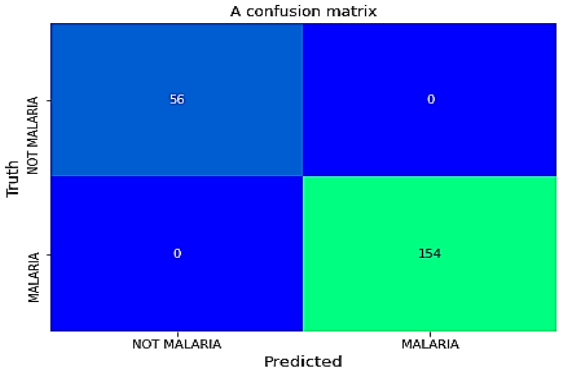


Figure 15. Gradient Boosting Classifier

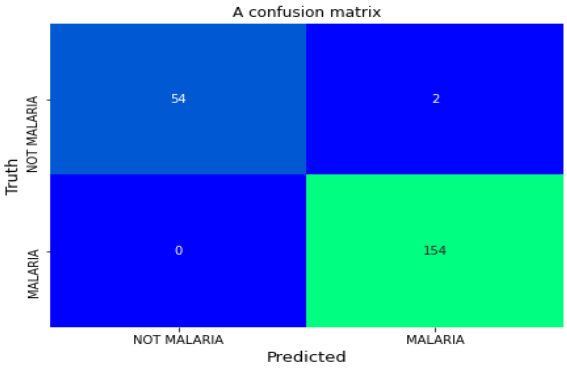
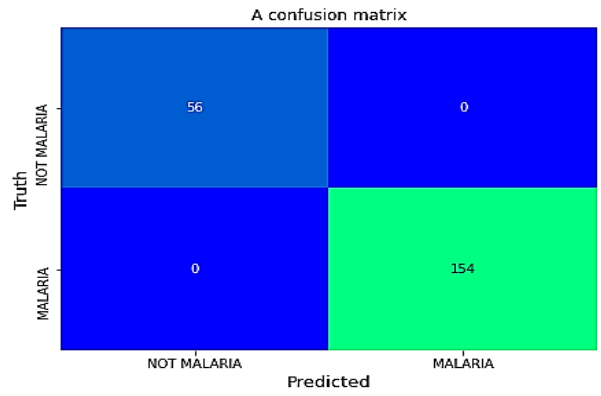


Figure 16. Random Forest Classifier



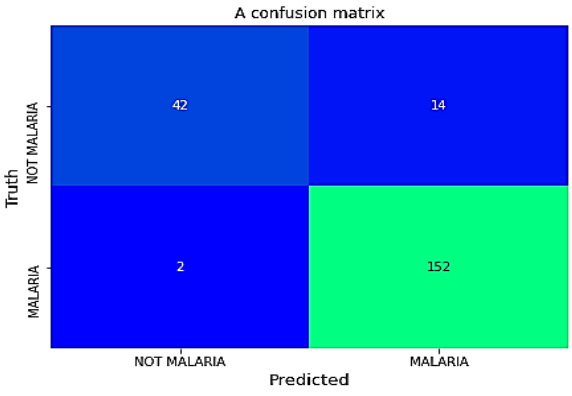


Figure 17. Decision Tree Classifier

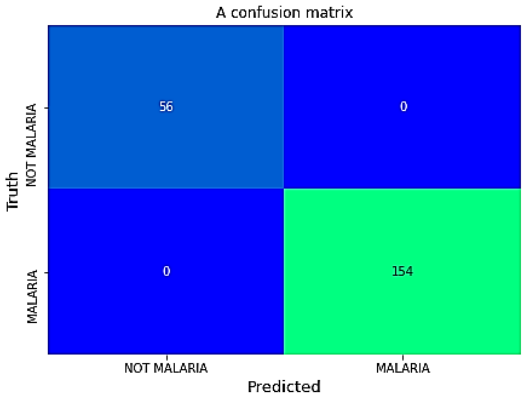


Figure 18. K-Neighbors Classifier.

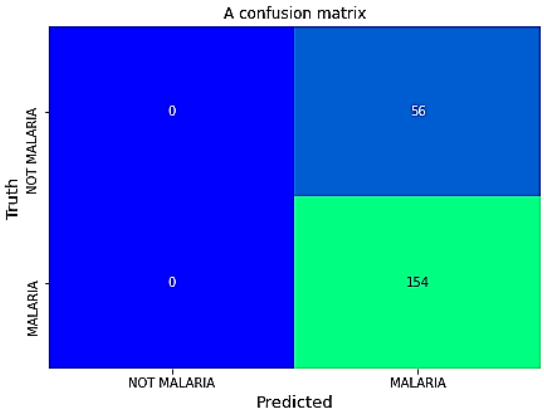


Figure 19. Linear Kernel Support Vector Classifier

Figure 20. RBF Support Vector Classifier

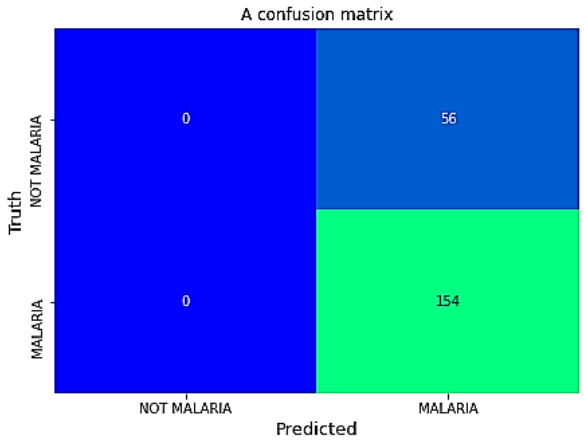


Figure 21. Support Vector Classifier

Below are the mathematical assessment metrics formulae used for each of the classifiers mentioned above:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \text{ X } 100 \quad (1)$$

$$\text{Error Rate} = \frac{FP+FN}{TP+FP+TN+FN} \text{ X } 100 \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \text{ X } 100 \quad (3)$$

Where;

TP = True Positive

FP = False Positive

FN = False Negative

TN = True Negative

The following table lists the relevant parameters that were measured along with their values:

Table 3. Summary of the seven (7) classifiers' assessment outcomes

ML Algorithm	Accuracy	Error Rate	Precision
Gradient Boosting	100%	0%	100%
Random Forest	100%	0%	100%
Decision Tree	100%	0%	100%
K-Neighbors	92.4%	76%	75%
Support Vector Machine	73.3%	26.7%	0%
RBF Support Vector	73.3%	26.7%	0%
Linear Support Vector	100%	0%	100%

The individual algorithms for the classifiers are listed in the table above. These algorithms' accuracy, margin of error, precision, recall, and F1-score were measured. These numbers were calculated using the produced values from each confusion matrix to complete the formulas for accuracy, margin of error, precision, recall, and F1-score. They have a range of 0 to 1 for their actual or acceptable value, which has been converted to a percent (multiples of 100) like this: 0 to 100.

The evaluation metrics formulae used for each of the corresponding Regression methods are provided below mathematically:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \tag{4}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \tag{5}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \tag{6}$$

Where,

$\hat{y}$ — predicted value of y

Table 4. Summary of the five (5) regressions' assessment outcomes

ML Algorithm	Root Mean Square Error (RMSE)	Mean Absolute Error (MAE)
Gradient Boosting	0.305	0.180
Decision Tree	0.422	0.180
Random Forest	0.321	0.180
Support Vector Machine	0.385	0.239
Linear Regression	10847232407.645	987481166.076

The corresponding methods for the regression model are listed in the table above, and their Mean Absolute Error and Root Mean Square Error were calculated. These values were calculated using the derived values for the relevant metrics used to assess regression models. The rating with the lowest value is frequently preferable to those with greater values.

5. Discussion

From figure 2, the graph shows that female group has the highest frequency with figure of 549 as against the male counterpart that have 499. The variance between them is 50, which is not much as such, it can be said that both genders have the same rate of hospitalization, and also are all vulnerable to fall sick while in school.

From figure 3, the graph shows that age group “15 – 20” has the leading frequency with 495, followed by the age group “21 – 25” with 406. Meanwhile, age group “26 - 30” and “above 30” have the least frequencies with 131 and 16 respectively. The age group “15 – 20” been within adolescent age and as human during this period reflects that you are young, even though you presumed to be an adult, and couple with the fact that most of them are leaving their homes for the first time to stay on their own, they need more orientation on self-care, adaptation and stress management.

From figure 4, it has shown us that, first semester has over double of second semester frequencies. This means that, first semester is when students fall sick most, which can be attributed to change of environment after long holiday. They might even resume to school with some kind of illness, which might thereby manifest during the early period of this semester or some other periods. Furthermore, the stress of registration, and not securing decent accommodation early enough due to inadequate hostel accommodation within the campus can also increase the hospitalization rate within first semester.

From figure 5, among the various levels of the undergraduate programme, 100 Level and 200 Level have the highest frequencies of been hospitalised, with figures of 352 and 316 respectively. The 300 Level and 400 Level have almost equal frequencies, that is about half of what was obtained from the lower levels. This can be attributed to the challenges of coping with the learning conditions for courses with large number of students like general courses which are usually the dominance of the courses to be offered for those lower levels, and so, they need to be aware of this truth and take precautionary measures against it and also, the university management should build more large halls that can cater for that population or break those course into groups that can be reasonably managed.

In figure 6, the undergraduate students that visit the sick-bay for treatments for the first time are fewer than those that re-visit for follow-up or re-hospitalization. The rate of the re-hospitalization is almost thrice the first hospitalization. This implies that, if a student fell sick and was admitted due to any illness for the first time, he or she might still be re-hospitalised. This can be due to not been treated properly on the side of the health workers, or was treated properly by the workers, but refuse to adhere to instructions given him or her. This situation might lead you to be diagnosed and treated for the previous illness or new illness as the case may be.

From figure 7, the blood group graph shows that those with low blood pressure are insignificant in number, while students with high blood pressure are very high with a total number of 299, which still call for concern. If intervened, it can add up to the figure obtained for those with normal blood pressure.

From figures 8, most students are hospitalised for two (2) days or above. Just few patients have stayed for four (4) days on admission before been discharged. These numbers of days signifies that, when you are sick, you would be away from academic pursuit, which can be within a critical period of studies like lectures, continuous assessment or even during examinations. The implication is that, you would not be able to recover those things done during your absence and it would affect your overall academic performance. Furthermore, it is about 1:3 for those admitted for 2 days or above to those that were hospitalised just for a day. This means that, if four (4) students are sick, such that they were taken to the health centre, the possibility that three (3) out of the four (4) would be hospitalised beyond a day is high. This implies that, health facilities would be overstretched, because they only have provision for two (2) patients each in the male and female wards respectively at a time.

Figure 9 shows that all the various illnesses that were diagnosed were treated. It was discovered that “Plasmodiasis” also known as “Malaria” has the highest occurrence, followed by “body pain” and then “Flu”, while the forty-five (45) other illnesses have less or insignificant amount of occurrence. As it has been established that Malaria has the highest frequency, even when we add up the other illnesses with 333 as total, they are nowhere close to the frequency of Malaria. So, Malaria was now determined to be prevalent or common illness among the undergraduate students. For Malaria to be prevalent there is need to deal with it in ways that would drastically reduce it occurrence or rate of hospitalization, and also the distraction from lectures, continuous assessment or even examinations, so as to boost their academic performance.

After determining the prevalent illness among undergraduate students, figure 10 shows the relationship between “Sex” and “Malaria”. It was crystal clear that, the rate of hospitalization of male and female due to Malaria illness is almost the same. So, both genders are vulnerable and prone to Malaria attack, and as such, they should be given same level of orientation on precautionary measures against it.

Figure 11 shows the relationship between “Age Group” and “Malaria”, which has it that, age group “15 – 20” are the most attacked by Malaria, followed by those of age group “21 - 25”. Meanwhile, age groups “26 - 30” and “above 30” are least affected by Malaria. So, the vulnerable need to be given more orientation on precautionary measures against malaria attack, as shown in this evidence-based graph. These measures could range from eating well, not skipping meals, regular pattern of eating, healthy lifestyles, often been hydrated and even sleep under treated mosquito nets.

Figure 12 shows that out of the four levels of the undergraduate programme, 100 Level and 200 Level are most attacked by malaria, while malaria attacks on the upper level are appreciably minimal. It thus shows that, the vulnerable levels are 100 and 200 Levels, and for that singular reason, they need to be aware of this and build their immune against it. Also, their lectures can be grouped in a considerable number due to lack of spacious lecture hall that can contain them, or build more large halls that can contain them, so that students don’t receive lecture while standing throughout the lecture hours, to improve their learning conditions and in turn help their health too.

In figure 13, the graph shows that first semesters have the highest rate of hospitalization resulting from Malaria attacks. The high rate of this hospitalization can be attributed to the ups and downs during registration ahead of the new session for 100 and 200 levels, and partial adaptation to the new environment. Furthermore, a lot of irregular eating patterns happens within the early periods of the first semester, anxiety, mosquito bites and infections from the toilet facilities which are not in total good conditions, and also not securing a decent accommodation early enough to settle down. The fumigation of all facilities, and putting the toilets system in good condition before every first semester would help to drastically reduce the malaria attack on undergraduate students.

Figure 14 is a graph, which shows that the possibility of a student being re-hospitalized is very high after a malaria attack on any undergraduate student. So, it is necessary to be diagnosed and treated properly on the part of the health workers, while on the part of the students, adhering to instructions given by the health workers are to be taken and executed properly. All these, is to avoid re-hospitalization, and which could lead to overstretching of the present health facilities.

These results show the capacity of Exploratory Data Analysis (EDA) to mine hidden information and gain insights from a dataset, which was also a quality of using the relevant attributes and properly pre-processing your data.

Subsequently, we discovered that the component of our approach and data that predicts malaria sickness responds best to Gradient Boost Classifier (GBC). In comparison to the other six (6) algorithms, it possesses 100% accuracy, 100% precision, and 100% recall. Additionally, GBC is steady from a non-accuracy based assessment since it can manage both little and massive data without degrading the model's performance instead of improving it.

Gradient Boost Regression (GBR) turns out to be the most effective method for it and the data when it comes to our framework's other prediction of admission length. According to measurements, it has the lowest RMSE of 0.57 and MAE of 0.423 when compared to the other four (4) algorithms. Additionally, GBR is considered robust from a non-accuracy based evaluation since it can manage data containing exceptions without impairing the performance of the models.

These findings demonstrate GBC and GBR's capability in the data analytical framework for coordinating the patient care of common illnesses, and they also demonstrate that changing the accurateness requires the application of the proper features and data pre-processing.

Finally, GBC and GBR allow us to create an ensemble machine learning model using basic models and achieve excellent results that are comparable to those of models that consume a lot of resources, such neural networks.

5.1 Final Thoughts and Future Work

This essay offers a method that promotes comprehension, determination and development of a framework for the predictive-diagnosis of prevalent illness among university students' population according to the usage of more complex features of a localized data. It is a development of the idea in (Morgenstern et al., 2020), but was made stronger by the application of a comprehensive and newer iteration of the localized dataset (students' health record) to provide a stronger conclusion.

Also, the framework has been enhanced with the combination of two tasks; classification and prediction module, resulting in a boost in the precision of algorithms built upon them.

The following are the paper's main contributions:

i. Capturing much features, and the true translation of reconciling missing values, abbreviations, acronyms and synonymous terms that are used in students' health records.

ii. Introduction of localized data sources to develop data analytics framework for determining prevalent illness among any population space, and build predictive-diagnostic model using machine learning a manner that guarantees better and more accurate symptomatic health care interpretation.

Further research must concentrate on a more effective method of using cutting-edge machine learning algorithms to categorize the types of malaria that are common among undergraduate students of a college or university. If the prediction is accurate, this will eliminate the need to further identify the type of malaria. Since users may interact with the system through a user-friendly user interface, further research should be focused on integrating established models into a web application. And also that, only a few techniques have taken this into account thus far, the idea of aligning different data formats and sources for predictive diagnosis remains an unexplored field of study that is deserving of future investigation.

Acknowledgment

We appreciate management of Federal University Lokoja for providing the dataset used in this work. We also appreciate the unrelenting efforts of the Software Laboratory Unit of the Department of Computer Science, Federal University Lokoja for providing the enabling platforms and tools for the development, analysis and testing of the framework.

References

[1] Liu N. & Kauffman R. J., (2020). Enhancing Healthcare Professional and Caregiving Staff Informedness with Data Analytics for Chronic Disease Management, *Information and amp; Management.* doi: https://doi.org/10.1016/j.im.2020.103315.

[2] Thorpe, K. E., (2009). Chronic Disease Management and Prevention in the U.S. *Eurohealth*, 15 (1), 5-7.

[3] Williams, E., Gartner, D. & Harper, P., (2021). A survey of OR/MS models on care planning for frail and elderly patients. *Operations Research for Health Care*, 31, 100325, ISSN 2211-6923, https://doi.org/10.1016/j.orhc.2021.100325.

[4] Ravikumaran, P., Vimala Devi, K., Kartheeban, K., & Narayanan Prasanth, N. (2020). Health Data Analytics: Framework & Review on Tool & Technology. *Materials Today: Proceedings*. doi:10.1016/j.matpr.2020.10.131

[5] Smiti, A. (2020). When machine learning meets medical world: Current status and future challenges. *Computer Science Review*, 37, 100280. doi:10.1016/j.cosrev.2020.100280

[6] Warner K. (2021). Statistical Methods to Support Difficult Diagnoses. *Journal of Medical Diagnostic Methods*. 10:349.2021.10.349.

[7] Morgenstern J.D., et al., (2020). Predicting population health with machine learning: a scoping review. *BMJ Open* 2020;10:e037860. doi:10.1136/bmjopen-2020-037860

[8] Ahnaf, A. C., Rezwan, H. K., Nabuat, Z. N. & Sadid, R. T. (2018). Predicting Depression in Bangladeshi Undergraduates Using Machine Learning. Bangladesh: BRAC University, Dhaka

[9] Krishnamoorthi R, Joshi S, Almarzouki HZ, Shukla PK, Rizwan A, Kalpana C, & Tiwari B.(2022). A novel diabetes healthcare disease prediction framework using machine learning techniques. *Journal of Healthcare Engineering*.

[10] Edeh M.O., et al., (2022). A classification algorithm-based hybrid diabetes prediction model. *Frontiers in Public Health.* 10.

[11] Deberneh H.M., & Kim I. (2021). Prediction of Type 2 diabetes based on machine learning algorithm. *International journal of environmental research and public health.* 18(6):3317.

[12] Gupta S, Verma H.K., & Bhardwaj D. (2021). Classification of diabetes using Naive Bayes and support vector machine as a technique. *Operations Management and Systems Engineering*. 365–376. Singapore: Springer.

[13] Islam M.T., Rafa S.R., & Kibria M.G. (2020). Early prediction of heart disease using PCA and hybrid genetic algorithm with k means. *23rd International Conference on Computer and Information Technology (ICCIT)*. 1–6. IEEE.

[14] Qawqzeh Y.K, Bajahzar A.S., Jemmali M, Otoom M.M., & Thaljaoui A. (2020). Classification of diabetes using photoplethysmogram (PPG) waveform analysis: Logistic regression modeling. *BioMed Research International.*

[15] Grampurohit S, & Sagarnal C. (2020). Disease prediction using machine learning algorithms. *International Conference for Emerging Technology* (INCET). 1–7. IEEE.