DATA ANALYTICS FRAMEWORK FOR THE DIAGNOSIS OF PREVALENT ILLNESS AMONG UNIVERSITY STUDENTS

[1]Dauda Olorunkemi Isiaka, [2]Joshua Babatunde Agbogun, [3]TaiwoKolajo
[1]Department of Computer Science, Federal University Lokoja, Nigeria.
[2]Department of Computer Science and Mathematics, Godfrey Okoye University, Enugu, Enugu State - Nigeria.
[3]Department of Computer Science, Federal University Lokoja, Nigeria.
Corresponding author: dauda.isiaka@fulokoja.edu.ng,

Received: January 18, 2023 **Accepted:** April 5, 2023

Abstract: This study proposes the development of a data analytics framework for diagnosis of prevalent illness among university students. A high-level model methodology with a Cross Industry Standard Process for Data Mining (CRISP-DM) steps was adopted. The findings showed 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. For Malaria to be prevalent there is need to deal with it in ways that would drastically reduce it occurrence or rate of hospitalisation, and also the distraction from lectures. Engagement of the model in diagnosis of prevalent illness will allow evidence-based awareness on the prevalent illness, personalised treatment; reduce human-prone errors using sample data from the Federal University Lokoja Health Centre. Furthermore, another finding showed that the Gradient Boosting Classifier had 100% accuracy, 100% precision and 100% recall as compared to other six algorithms.

Keywords: Data, Analytics, Framework, Prevalent illness, and Classification.

Introduction

A growing proportion of the world's population suffers from common sickness, and the cost of caring for such individuals has risen dramatically. Nearly half of people in a population suffer from a common sickness like malaria, diabetes, or hypoglycaemia (Liu and Kauffman, 2020). Seventy per cent (70%) of distractions when pursuing everyday living demands, profession, and education are caused by common disease (Thorpe, 2009). Students are some of the most vulnerable patients in the UK due to factors such as living alone for the first time and irregular eating patterns, being uninformed of a prevalent illness, and so on (Williams *et al.*, 2021). Many such illnesses are conditions which reduce patients' quality of life and stress their family members and caregivers.

Documentation can be used to make potential patients within a population become aware of a prevalent illness and then take precautionary measures and safety, while in the pursuit of daily needs of life. In order to collect medical information relevant to a patient's healthcare needs, documentation can be done using paper, electronics or both (Adeyemo and Olaogun, 2013). But most data are stored in a hard copy form (Ravikumaran *et al.*, 2020).

These massive quantities of data would help to improve the quality of health care delivery and reduce the cost. It also helps in clinical decision support, disease surveillance, and population health management (Ravikumaran et al., 2020).

The use of data analytics in medical profession has shown promise in improving a variety of areas of care, ranging from medical imaging to chronic disease or prevalent illness management to population health and precision medicine. This could improve care delivery efficiency, reduce administrative burdens, and speed disease/illness detection.

Although data analytics have improved the quality of patient care management, this still faces several challenges (Smiti, 2020), and using randomized controlled trials, and personal job experience aimed at lowering the major outcomes of common sickness occurrences have never been found to be highly helpful in this regard (Warner, 2021). Data analytics could be used to develop a framework for patient care management that could be done satisfactorily with data analytics. Manually analysing the data from these documents is time-consuming and inefficient, as often times, they rely on personal expertise. This paper aims to create a data analytics framework for diagnosing prevalent illness among university students.

Materials and Methods

Aims to design a framework to facilitate and prevent/reduce the prevalent illness among students of typical university. The dataset collected consists of 1048 students from the Federal University Lokoja Health centre. We would pre-process it to remove outliers, perform an Exploratory Data Analysis (EDA), and then build model for the diagnosis of prevalent illness.

Figure 1 describes the activities that take place in the framework, and shows the design of the models within the framework.

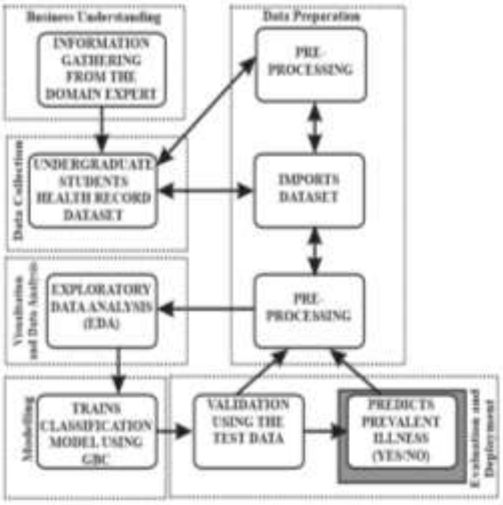

Business Understanding
INFORMATION GATHERING FROM THE DOMAIN EXPERT
Data Preparation
PRE-PROCESSING
Data Collection
UNDERGRADUATE STUDENTS HEALTH RECORD DATASET
IMPORTS DATASET
Visualisation and Data Analysis
EXPLORATORY DATA ANALYSIS (EDA)
PRE-PROCESSING
Modelling
TRAINS CLASSIFICATION MODEL USING GBC
VALIDATION USING THE TEST DATA
PREDICTS PREVALENT ILLNESS (YES/NO)
Evaluation and Deployment

Figure 1: High-Level Model of Proposed DAF

Methodology

For this proposed model, we are adopting the Cross Industry Standard Process of Data Mining (CRISP-DM) methodology, because of the data-driven nature of the research.

The following activities were deployed in the project work, as shown in Table 1 below:

Table 1: Description of CRISP-DM Methodology

Steps	**Process**	**Activities**
Business Understanding	The domain experts' view on the prevalent illness among students was gotten. Data from the medical record were collected.	Application to visit and request for dataset was done, and which was approved. Visits to the Federal university Lokoja, Health Centre for understanding of their mode of operations and collection of data. Electronic Google forms were used to get the views from the domain experts on the prevalent illness.
Data Understanding	Clarifications on the data collected. Pre-processing of dataset is done. The correct values for all missing values were treated. Using visualisation tool to represent analysis result	The right entries to the Wrong/missing entries from the health record were requested and filled in appropriately. No missing value was left out. We pre-processed the dataset by making every complaints and diagnosis to be on individual column. Exploratory data analysis was used to visualize and understand the relationship among all the features.
Data Preparation	Pre-processing of dataset is done. Only relevant data to this project is chosen. Transformation of the categorical data into numeric forms was done.	We pre-processed the dataset by checking if there was any missing value. We drop all irrelevant columns from the features and the target variable. Here we use One-hot encoding for the transformation. The dataset was divided into train and test data respectively using a ratio of 80:20.
Modelling	Smart techniques/algorithms are applied to create the models	We use GBC to create the Classification Model. We split our dataset into 80% for training and 20% for testing.
Evaluation	Finding interesting score of the models.	We use confusion matrix to evaluate our classification model and then root mean square error for the prediction model.
Deployment	The framework is easy for the user to use and also access the output in the same file format.	An input file in CSV format is imported and processed to produce an output of both models in a CSV file format, which is stored in the internal storage of the computer for the user to access and view the result.

Result

The original dataset collected from the University is shown in Table 2 below:

Table 2: Sample of original dataset collected

S/N	File No	Sex	Age	Type of Attendance	Complaints	Diagnosis	Outcome
1	250	F	23	NEW	CATARRH, COUGH, BODY PAIN, FEVER	MALARIA	TREATED
2	258	F	30	FOLLOW UP	FEVER, JOINT PAIN, HEADACHE	MALARIA	TREATED
3	262	M	23	NEW	FEVER,JOINT PAIN, HEADACHE	MALARIA	TREATED
4	257	F	20	NEW	JOINT PAIN, DYSENTRY STOOL, ABDOMINAL PAIN	DYSCENTRY	TREATED
5	265	F	18	NEW	PV ITCHING	VULVO-VAGINA	TREATED
6	261	F	18	NEW	FEVER	MALARIA	TREATED

Flow Process of the Data Analytic Framework

The development and flow process consists of the following steps:

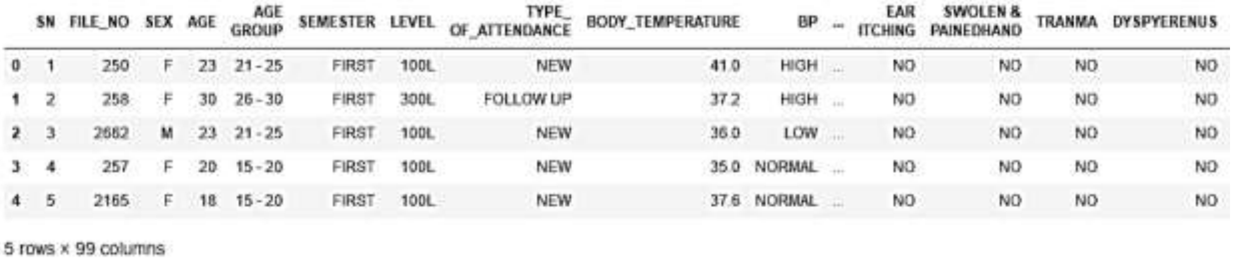
Step 1: Data Collection

We collected the patients' record for the undergraduate students of Federal University Lokoja from inception to date as compiled by the Health Records Unit from the University Health Centre, and we use this information for the visualisations.

Step 2: Data Cleansing and Transformation

The data we collected from the University Health Centre was checked for missing values, impurities, and a cleaning process was also carried out to ensure there are no missing values, and also ensured the correctness of the data, so as to be usable for the visualisation. The cleaned and transformed dataset is shown in Table 3 below:

Table 3: Transformed dataset



	SN	FILE_NO	SEX	AGE	AGE GROUP	SEMESTER	LEVEL	TYPE_ OF_ATTENDANCE	BODY_TEMPERATURE	BP	...	EAR ITCHING	SWOLEN & PAINEDHAND	TRANMA	DYSPYERENUS
0	1	250	F	23	21 - 25	FIRST	100L	NEW	41.0	HIGH	...	NO	NO	NO	NO
1	2	258	F	30	26 - 30	FIRST	300L	FOLLOW UP	37.2	HIGH	...	NO	NO	NO	NO
2	3	2662	M	23	21 - 25	FIRST	100L	NEW	36.0	LOW	...	NO	NO	NO	NO
3	4	257	F	20	15 - 20	FIRST	100L	NEW	35.0	NORMAL	...	NO	NO	NO	NO
4	5	2165	F	18	15 - 20	FIRST	100L	NEW	37.6	NORMAL	...	NO	NO	NO	NO

5 rows × 99 columns

Step 3: Data Exploration and Preparation

In this step, a data understanding was carried out through the exploratory data analysis to report what the dataset entails by tabulating all the necessary parameters and also visualize the behaviours within the dataset.

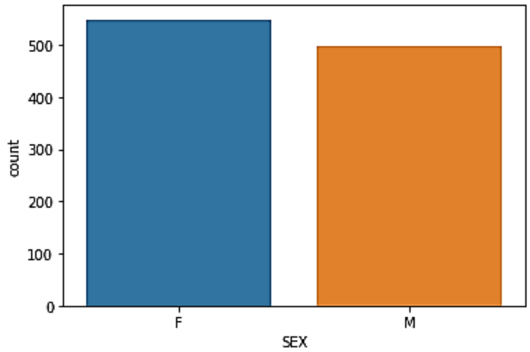


Figure 2: Gender

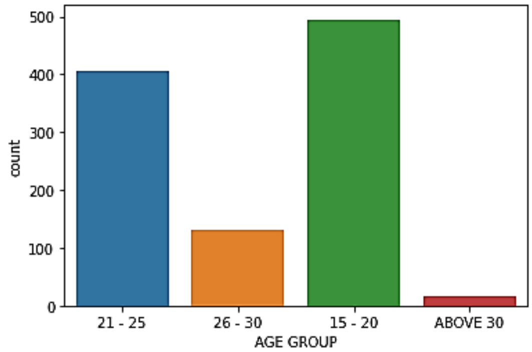


Figure 3: Age Group

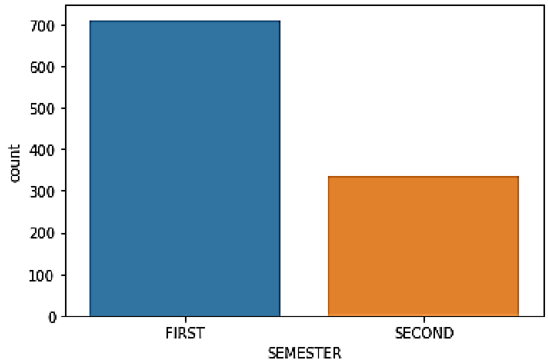


Figure 4: Semester

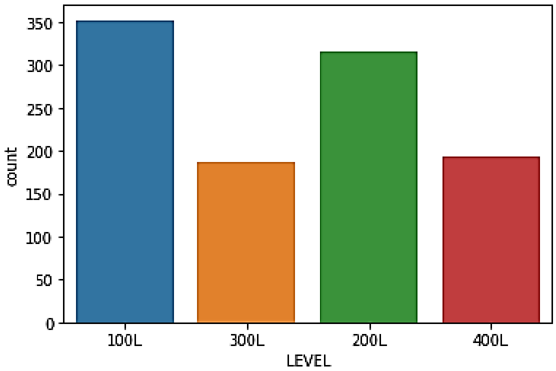
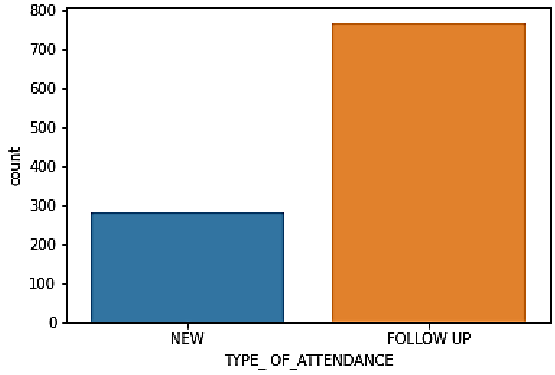


Figure 5: Level



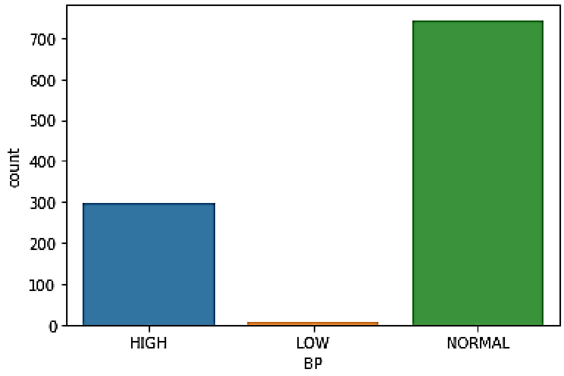
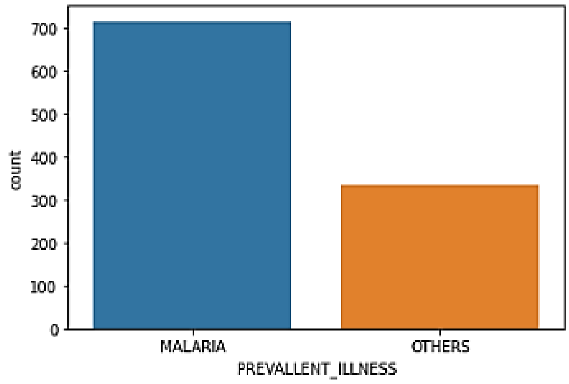


Figure 6: Attendance Type

Figure 7: Blood Pressure



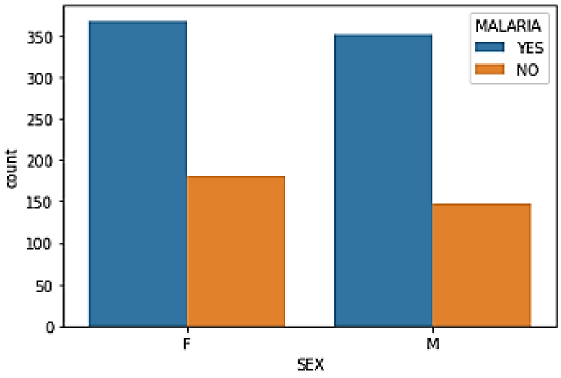
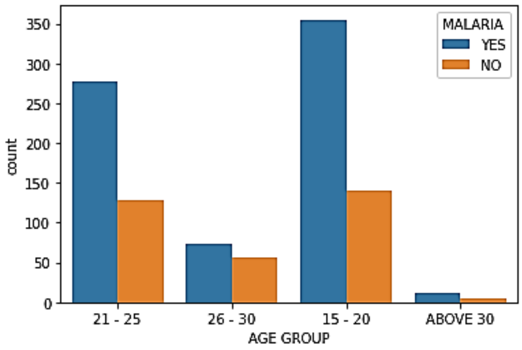


Figure 8: Prevalent illness

Figure 9: Sex versus Malaria



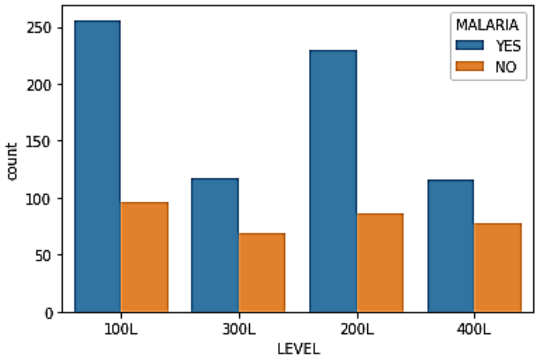
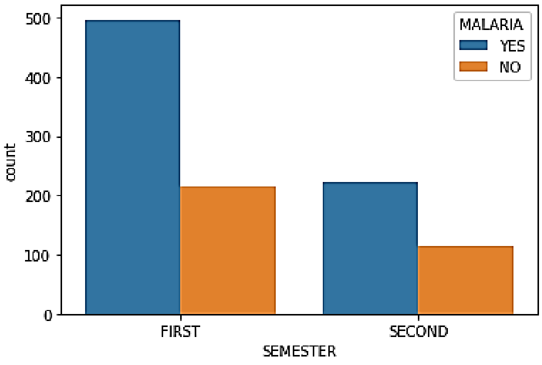


Figure 10: Age Group versus Malaria

Figure 11: Level versus Malaria



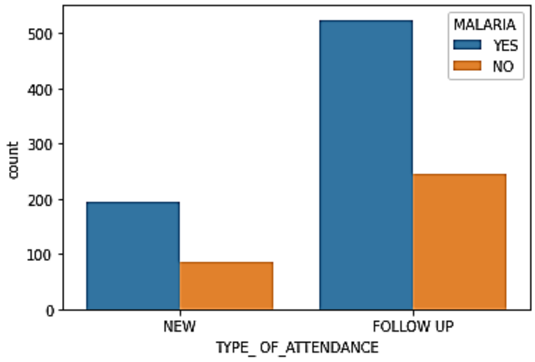
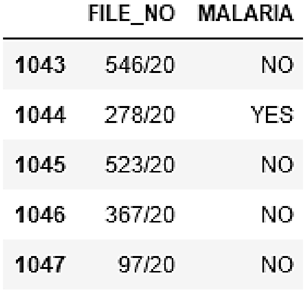


Figure 12: Semester versus Malaria *Figure 13: Attendance Type / Malaria*

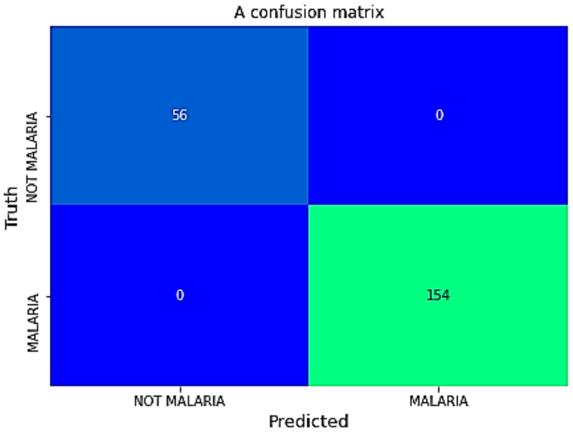
Step 4: Model Training or Implementation
To train the models of the data using the Random Forest Classifier, Decision Tree Classifier, K-Neighbor Classifier, Gradient Boosting Classifier and Support Vector Classifiers from the "modelselection" library of the "SKlearn" package installed into the Python. The result for the classified target variable is shown in Table 4 below:

Table 4: The output of the predicted prevalent illness from the test data



	FILE_NO	MALARIA
1043	546/20	NO
1044	278/20	YES
1045	523/20	NO
1046	367/20	NO
1047	97/20	NO

Step 5: Model Evaluation
We used a Confusion Matrix to determine the accuracy, error rate, precision, recall and F1 score of the seven (7) classifiers algorithms used to build the model, as shown in Figures 14 - 20 below:



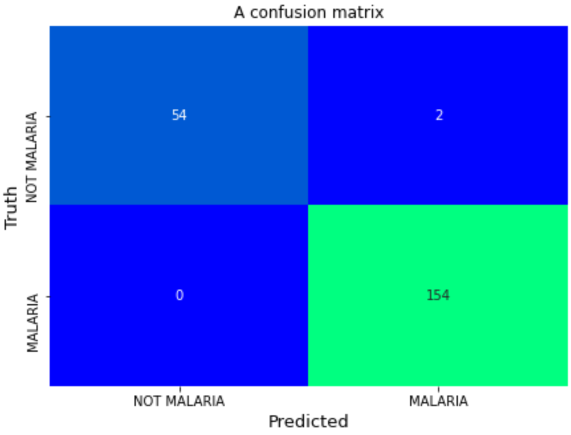
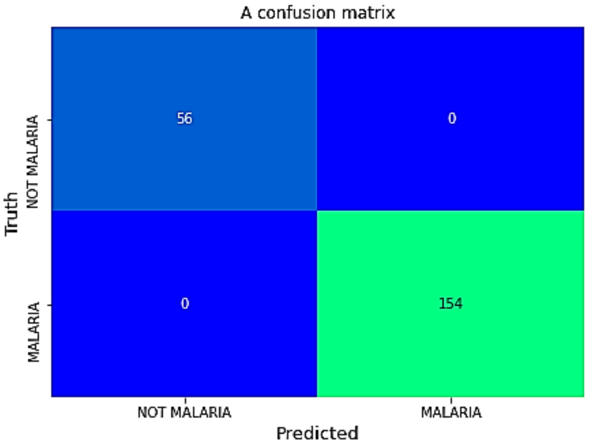


Figure 14: Gradient Boosting Classifier *Figure 15: Random Forest Classifier*



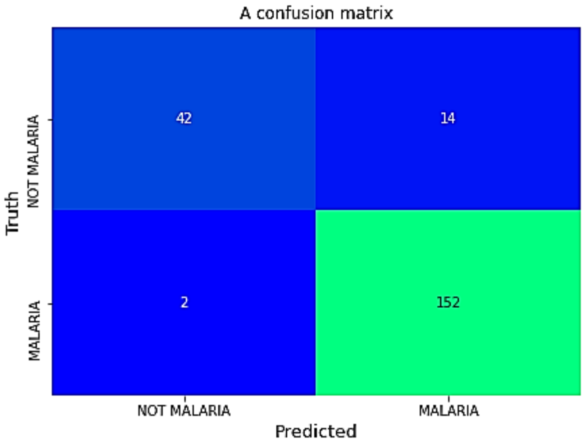
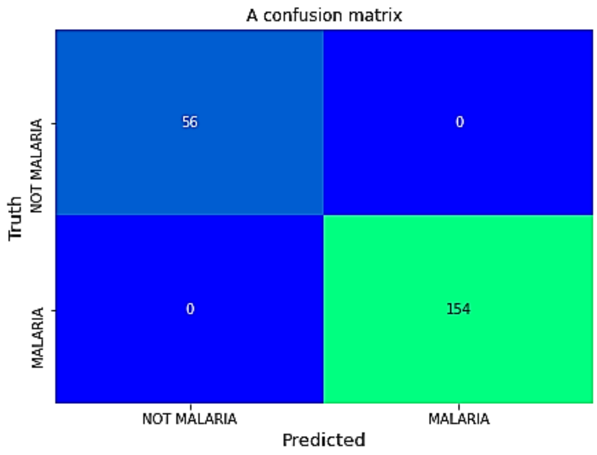


Figure 16: Decision Tree Classifier *Figure 17: K-Neighbors Classifier*



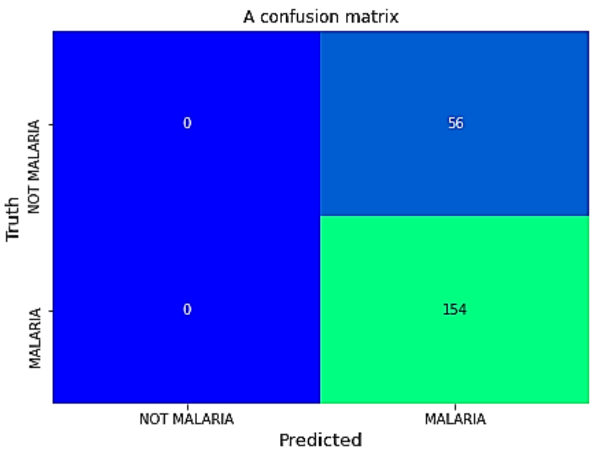


Figure 18: Linear Kernel Support Vector *Figure 19: RBF Support Vector Classifier*

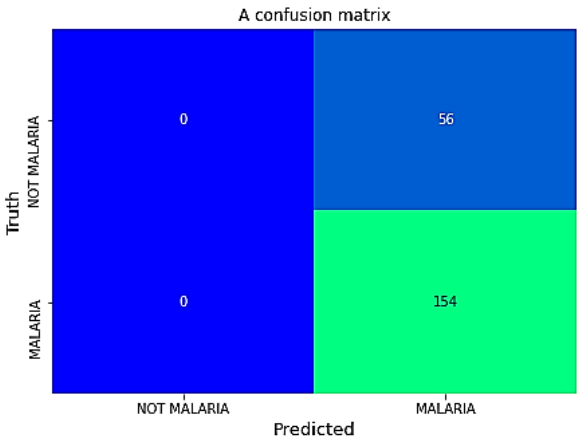


Figure 20: Support Vector Classifier

Mathematically, the evaluation metrics formulas applied on all the above respective classifiers/algorithms are given in the equations (i) – (v) below:

i. Accuracy = $\frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}}$

ii. Error Rate = $\frac{\mathrm{FP+FN}}{\mathrm{TP+FP+TN+FN}}$

iii. Precision = $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$

iv. Recall = $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$

v. F1 Score = $\frac{(2 \text{ X Precision X Recall})}{(\text{Precision+Recall})}$

Where;
TP = True Positive
FP = False Positive
FN = False Negative
TN = True Negative
The respective parameters measured and their values are shown in Table 5 below:

Table 5: Summary of the evaluation results on the seven (7) classifier algorithms

Classifier	Accuracy	Error Rate	Precision	Recall	F1 Score
Gradient Boosting	1.0	0.0	1.0	1.0	1.0
Random Forest	1.0	0.0	1.0	1.0	1.0
Decision Tree	1.0	0.0	1.0	1.0	1.0
K-Neighbors	0.924	0.076	0.75	0.955	0.84
Support Vector Machine	0.733	0.267	0	0	0
RBF Support Vector	0.733	0.267	0	0	0
Linear Support Vector	1.0	1.0	1.0	1.0	1.0

The table above shows how well each of the algorithms used in the classification model for Formula 1 cars perform in terms of their accuracy, error rate, precision, recall and F1-score. They have their true or accepted value ranging from 0 – 1, and can also be in percentage (multiples of 100) like this 0 - 100.

Discussion

From Figure 2, the graph shows that female group have 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 age group "15 – 20" have the leading frequency with 495, followed by the age group '21 – 25" and '26 – 30". Age group '26 - 30' and 'above 30' have the least frequencies with 131 and 16 respectively. This is because they need more orientation on self-care, adaptation and stress management as they are leaving their homes for the first time.

From Figure 4, 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. 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, 100 Level and 200 Level have the highest frequencies of being hospitalized, 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.

In Figure 6, the rate of 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-hospitalized. This can be due to not being treated properly on the side of health workers, or was treated properly by the workers, but refuse to adhere to instructions given him or her.

From Figure 7, the blood pressure 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.

Figures 8 shows 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. 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 undergraduate students.

The rate of hospitalization of male and female due to Malaria illness is almost the same. As such, both genders should be given same level of orientation on precautionary measures against it. After determining the prevalent illness among undergraduate students, Figure 9 shows the relationship between "Sex" and "Malaria".

Figure 10 shows the relationship between "Age Group" and "Malaria". Age group "15 – 20" are the most attacked by Malaria, followed by those of age group "21 - 25". The vulnerable ones need to be given more orientation on precautionary measures against malaria attack. These measures could range from eating well, not skipping meals, regular pattern of eating, often been hydrated and even sleep under treated mosquito nets.

Figure 11 shows that Students need to be aware of the fact that they are most susceptible to malaria. Out of the four levels of the undergraduate programme, 100 Level and 200 Level are most attacked by the disease. Lectures can be grouped in a considerable number due to lack of spacious lecture hall that can contain them, or large halls can be built to contain them. This will improve their learning conditions and in turn help their health too.

In Figure 12, 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. 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 13 show 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 adhering to instructions given by them. All these, is to avoid overstretching of the present health facilities.

The results show the capacity of Exploratory Data Analysis (EDA) to mine hidden information and gain insights from a dataset, which was also a characteristic of applying the right features and appropriately pre-process your data. GBC allow us to build an ensemble machine learning model using simple models and yet get great scores which are at par with the resource-hungry models like neural networks.

Conclusion

Data analytics is an emerging and contemporary topic in Computer Science that has been applied in the domain of healthcare. The data were obtained from undergraduate students/patients' health record spread across all levels, and semesters, as the learning environment is unique in its nature; the environmental factor, living conditions, pressure, anxiety towards academic pursuit. Results showed that there is possibility of diagnosing using only patient information with AI.

Acknowledgements

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

Conflict of Interest

The authors declare no conflict of interest.

References

Adeyemo FO & Olaogun, AA (2013). Factors affecting the use of Nursing Process in Health Institutions in Ogbomoso Town, Oyo State. International Journal of Medicine and Pharmaceutical Sciences, 89-96.

Kotepui, M & Kotepui, KU, (2019). Prevalence and laboratory analysis of malaria and dengue co-Journal Pre-proof infection: a systematic review and meta-analysis, BMC Public Health, 428(19) 1148. https://doi.org/10.1186/s12889-019-7488-4.

Liu N & Kauffman RJ, (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 .

Mahmoudi, S, Mamishi, S, Banar, M, Pourakbari, B & Keshavarz, H (2019). Epidemiology of echinococcosis in Iran: a systematic review and meta-analysis, BMC Infect Dis, 19, 929 https://doi.org/10.1186/s12879-019-4458-5.

Pierce, D *et al*, (2019). Safety and tolerability of experimental hookworm infection in humans with metabolic disease: study protocol for a phase 1b Randomised Controlled Clinical Trial, BMC EndocrDisord, 19, 136. https://doi.org/10.1186/s12902-019-0461-5.

Ravikumaran, P, Vimala DK, Kartheeban, K, & Narayanan PN (2020). Health Data Analytics: Framework & Review on Tool & Technology. Materials Today: Proceedings. doi:10.1016/j.matpr.2020.10.131

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

Warner, K (2021). Statistical Methods to Support Difficult Diagnoses.Journal of Medical Diagnostic Methods. 10:349.2021.10.349.

Williams, E., Gartner, D. and 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.