iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VII **Month of publication:** July 2023

DOI: https://doi.org/10.22214/ijraset.2023.54944

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

Performance Evaluation of Computer Science Students in Mathematical Courses Over Prgramming Language Courses

Sokunbi Michael A.[1], Akinsola Adeniyi F[2], Onadokun I. O[3], Uzor Chidinma J[4]
[1]*Yaba College Of Technology, Computer Tech Dept., Yaba, Lagos Nigeria*

Abstract: ***The low level of performance observed among many computer science students in mathematical and computer programming courses served as the motivation for this work. The prediction of computer science students' overall performance in mathematical courses over programming courses was done using WEKA and machine learning algorithms. Dataset of students performance from Yaba College of Technology, Lagos, Nigeria and Lagos State Polytechnic, Lagos, Nigeria, for both National Diploma and Higher National Diploma were used. These data set were studied and analyzed using WEKA and Random Forest, J48, Naïve Bayes and Logistic Regression algorithms. The algorithms were applied for the HND and ND dataset and there was comparison based on their accurancy, learning time and percentage of correctly classified instances. This comparison showed that there is direct relation between the execution time in building models and volume of data records. This shows that the predictor did not only predict the number of students that are likely to be in distinction, upper credit, lower credit, pass and fail but also show the relationship between having the knowledge of mathematics and programming language for an overall performance in computer science. The knowledge pattern represented further satisfies the exertion that it is imperative for students to have a standard knowledge of mathematics as this will help in being the best in their chosen profession.***
Keywords: Mathematics, Computer Programming, Educational Data Mining

I. INTRODUCTION

There has been an ongoing debate on the relevance of Mathematical courses as a pre-requisite to understanding computer programming. This debate also borders on the effect of the performance of Computer science students in mathematics over computer programming courses. Some computer science students consider mathematical courses as borrowed courses, believing it has no direct impact on their learning or on their profession outside the classroom. This work intended to do a scientific examination of this perception and shed more light on the relevance of the knowledge of mathematics as an aid to learning and understanding computer programming and becoming a programmer. Evaluating student performance is an essential part in higher institutions to students and the institute. This helps to rank the institute in the level of quality education based on the students' excellent academic performance.. This performance evaluation was achieved by obtaining students' learning assessment; Grade Point Average (GPA) scores over a period of time. There are many techniques that can be used to measure performance academic but data mining techniques happens to be the most used technique in evaluating students performance. The aspect of data mining concerned with this is Educational Data Mining (EDM). Educational data mining is used to extract useful informations and patterns from educational database. "EDM aims to predict students' potential learning behavior, explore the impact of educational support, and advance scientific knowledge about learning" (Intellipaat,2023). Some of the data mining techniques used are classification techniques which include decision tree algorithm, Bayesian classification, Logistic regression, Random forest, and so on.

II. LITERATURE REVIEW

There are several correlations between computer programming language and mathematics. These coreelations include: logical thinking (Grover & Pea (2013), problem solving (Bocconi et al, 2018) and functions and variables.

1) Logical Thinking and Reasoning: Grover & Pea (2013) logical reasoning and thinking is the ability of analyzing situations and finding a reasonable soliutions to it. So mathematics and programming language helps in the logical reasoning of a student i.e. the ability to think outside the box, analyze and being creative about situations.
2) *Problem Solving:* Bocconi et al (2018) programming language mainly iunvolved identifying a gap and finding solutions to it or improve on an existing problem. Mathematics mainly is all about problem solving that helps in our day to day activities. So both are essential in problem solving in the real world.

3) *Functions and Variables:* In programming language it is inevitable for a programmer not to write functions which make use of variables. Mathematical equation also includes functions and variables. These functions include add, subtract, divide etc and variables can be numeric or alphabetic.

Most students not knowing how correlated this courses are, end up focusing on only programming languages.

The field of data mining combines many other disciplines such as Databases Management, Datawarehousing, Statistics, Artificial Intelligence (AI) and Machine Learning (ML). Using programming in mathematics education is not a new concept. Papert (1980) developed a Logo environment that required students to program a computer to steer a turtle on a computer screen with the intention of providing a different environment for learning mathematics and motivating studenst to engage in in mathematics. Yelland (1995) examined "the potential of Logo to act as as a mathematical environment" based on Papert's Logo environment. Ke (2014); Lambic (2011) examined that programming has the potential to influence their attribute toward mathematics. It was discovered by La Paglia et al (2017) that using Logo Mindstorm robots improved learners' attitude towards mathematics.

A. Mining Academic Result Data

Mining academic result data is an Educational Data Mining (EDM) technique which is a sub field of data mining. Educational data mining deals with data that comes from different educational environment. Educational data mining focuses on the development of methods for discovering hidden patterns within the kinds of data that comes from educational settings (Ahmed and Elaraby 2014).

EDM aims at achieving some educational objectives amongst which are better understanding of students and the environment of learning which helps to improve student performance (Ahmed and Elaraby, 2014). EDM is important to enhance reading and learning process. The primary goal for using EDM methods in Student Academic Performance (SAP) prediction is to develop a prediction model for the overall performance of student in a selected course using their performance in prior courses as prediction parameter.

B. The Study Of Computer Science

Computer science is the study of computers and computing Newell, Perlis and Simon (1967). Theoretical and algorithmic foundations together with hardware, software and their uses for processing information all form the root of the study of computer science (ACM, 2006). In learning computer programming, students are expected to learn about algorithm and data structures, computer network designs, data modeling, information processessing, and now, artificial intelligence (AI). With computing being the core object of study, the following disciplines are inter-related with the study of computer science; computer engineering, computer science, information systems, information technology and software technology and software engineering. The major subfields of computer science include the traditional study of computer architecture, programming languages and software development. Interestingly, computer science draws some of its foundational knowledge from mathematics and engineering and therefore incorporate techniques from areas such as queuing theory, probability, statistics and electronic circuit design. Computer science emerged as an independent discipline in the early 1960s although electronic digit in the related fields of mathematics was invented some two decades earlier. It is also important to note that the root of computer sciece lies primarily in the related fields of mathematics, electrical engineering, physics and management information system (Allen Tucker, 2021).

C. Programming Languages In Computer Science

A Programming language is the language with which a programmer coomunicates with the computer and instruct the computer to get work done or to perform a task. The earliest form of programming language was assembly language which is a machine language that has binary encoded instruction directly executed by the computer. In the 1950s, programmers began to write codes using English-like high level languages such as FORTRAN (Formula Translator) and ALGOL (Algorithmic Language) which were the two first high level languages. These languages allows programmers to write algebraic expression and solve scientific computing problems. In the 1960s, a new relatively simpler language called BASIC (Beginner's All Purpose Symbolic Instruction Code) was developed. This allowed students in elementary school learn programming. Also COBOL (Common Business-Oriented Language) was developed to support business programming application. This was a commercial language that allows managing information stored in records and files. The goal of all these developments was to help develop programming languages that allows the programmer to communicate with the machine at a level higher than machine code. COBOL, FORTRAN, PASCAL and C were regareded as Procedural Languages because they allow programmers to develop and reuse procedures, subroutines and functions to avoid reinventing basic tasks for every new application. Other high level languages are called Functional languages. Functional languages view a program as a collection of mathematical functions and its semantics are very precisely defined.

Examples are LISP (List Processing) which in the 1960s was the mainstay programming language for Artificial Intelligence. Other successors to this in Artificial Intelligence include Scheme, Prolog, C and C++. Scheme is similar to LISP but has more mathematical definition. Prolog is used mainly for logic programming and its application in natural language and expert systems. C and C++ has been widely used in robotics, an application of Artificial Intelligence research (Allen Tucker, 2021).

In 1980s an additional support for data encapsulation was developed which gave rise to object oriented programming called Small talk, C++, VISUAL BASIC and Java. Java is unusual because its application are translated not into a particular machine language but into an intermediate language called Java Byte code which runs on Java Virtual Machine (JVM). Java is platform independent i.e. it can be executed on contemporary computer platform. At a higher level of abstraction, lies declarative and scripting programming language. They are strictly internet languages and often drive applications running in web browsers and mobile devices. Some declarative language allows programmers to conveniently occur and retrieve information from a database using queries. Examples are MySQL and SQLite. Another form of declarative languages is those that describe the layout of the webpage on the users screen e.g. HTML (Hyper Text Markup Language). The scripting language such as PHP glues the web page together with the database (Allen Tucker, 2021).

The requirement for learning programming language in computer science is the basic knowledge of programming language concepts which centres on developing programming logic. Learning programming also requires an understanding of technical concepts of algorithms, source code, compilers/compilation, data types, identifiers, transfer of controls, functions, classes and objects, and others.

D. Mathematics In Computer Science

Mathematics is an area of study concerned with logical study of numbers, shapes, arrangement, quantity, measure and many related concepts. Computer science continues to have strong mathematical roots. It is the source of the key components in the development of computer science, the understanding that all information can be represented as sequences of zero and ones and the abstract notion of it being a 'stored program'. In binary number system, numbers are represented by a sequence of binary digit 0 and 1 and in mathematical formula the decimal system are represented using digits 0 to 9. For example, computer science undergraduates must study discrete mathematics (logic, combination and elementary graph theory) as a selective course. Some may require students to have knowledge of numerical analysis, calculus, statistics and algebra to complete their course field. In computer science, mathematical measure of complexity allows student to predict timing before writing the code. This will show how fast an algorithm will run and how much memory it will require. Such predictions are important guidelines for programmers implementing and selecting algorithms for real world application (Allen Tucker, 2021).

The requirement for learning Mathematics in computer science are those basic knowledge of mathematics. These knowledge are needed in excelling in the more difficult computer science profession. These include Calculus, Discrete mathematics, Linear algebra, Number theory, Statistics and Probability, Graph theory. However, not all computer scientists use mathematics every day.

E. Benefits Of Mathematics In Learning Programming Language

Computer science has a great affinity with the related fields of mathematics. Programming is all about logical reasoning and problem solving and this can be said for mathematics as well. Mathematics is one of the tool a programmer need to develop sophisticated application and without the knowledge of this a programmer is said to be handicapped. Some of the benefits include:

1) It enhances student's ability to think logically.
2) It helps in developing clarity and precision of thought.
3) It increases preciseness in problem analysis and modeling.
4) It helps to develop student's ability to apply formal techniques in design and specification etc.
5) It provides confidence in using symbols, mathematical notation and abstraction.

III. REVIEW OF RELATED WORK

Ahmed et al. (2015) designed a framework to predict the performance of first year bachelor students of computer science course. The dataset was taken from 8 years data starting from July (2006/2007 – 2013/2014). The classifiers used include Decision tree, Naives Bayes and Rule based classifiers. The data collected contained various information about the students' previous academic records, demographic and family background. The classifiers that showed the highest accurancy was the Rule based classifier and it was 71.3% accurate. The limitation of this research was that the teacher had no prior knowledge about the students' previous background. The issue of small size of data due to incomplete and missing value in the collected data.

Sadiq et al (2018) used WEKA tools to evaluate academic performance of students from three different colleges in Assam, Indian. The data collected were academic and personal data of the student. There were 300 instances of data and 24 features were collected after data cleaning. After using feature selection, 12 highly influential attribute were discovered. Some of the features include students age, gender, parent qualification, end of semester result, class assessment etc. The classification technique used include J48 classifier, BayesNet Classifier, Random forest classifier and PART classifier. Also, Apriori algorithm was appied to the dataset to find the best rules. It was discovered that random forest was the best having an accurancy of 99% for the 12 attributes and 84.33% for the 24 attributes. The study is limited in the fact that the author called for an improvement in the study. This improvement includes extracurricular activities and technical skills of the students with the academic performance in class. It also includes working on different courses studied by the students and checking the success rate of each course.

Aderibigbe et al (2019) used ORANGE data mining tools and regression analysis to evaluate the relevance of ethnicity in predicting graduating student set (2008 – 2013) academic performance. The case study of Covenant university in Nigeria. The research was carried out to identify the hidden knowledge and vital statistical trends for students of the six geo-political zones in Nigeria to understand the impact of ethnicity on their performance. Datas of 2413 students were collected and these include the ststistical figures of Jamb score, the graduation CGPA and the geopolitical zones of the student. The geopolitical zone are North Central, North west, North east, South south, South east and South west. The algorithm used include Classification tree, Neural network, Naïve bayes, Random forest algorithm and multilinear regression. For data mining algorithm, class grade was used and CGPA was ignored while for regression analysis CGPA was used while class grade was skipped. Considering the geopolitical zones increased the accurancy of the random forest and it shows that pre-admission academic performance is a complete predictor for student performance. The non academic factor such as social lifestyle, internet addictions, class attendance and games shape the performance of student once admitted. However, this research is limitedto the fact that the university in concern was in the South west geo political zone and the author would like to find out what it would look like using universities from other geopolitical zones. The use of other analytical techniques and alternative tool would influence the outcome.

Evaristus et al. (2021) implement the use of big data to determine student academic performance and learning effectiveness. The research was carried out to check how big data can be applied in helping teachers analyze what students know and the techniques most effective for learning. The data mining algorithms used include Naïve Bayes, Decision tree and K-means clustering. The data set was friom Kaggle entitled ‘student performance data set’. The result show that big data can improve student performance by imitating the ways of learning methods, environment, health, school, parenting and others in accordance with existing data. The study was limited to the concern for data security, privacy protection and access rights in accessing private digital data.

Aderibigbe et al (2019) used KNIME tool to predict if the performance of student within the first three years in university would determine the overall performance of the student. The research was carried out in Covenant University in Nigeria and is limited to the seven programs offered in the engineering department of the school and admission year (2002/2003 – 2009/2010). The data mining algorithm used include Probabilistic Neural Network (PNN), Random Forest, Decision tree, Naïve Bayes, Tree Ensemble and Logistic regression. The data set used was obtained by Popoola et al. (2018). The most influential feature was the third year CGPA followed by the second and first year. The third year result was influential because it was observed that the fourth and final year’s work became more robust and intensive. This is due to the fact that it involves the student core courses and the first three years were like an introductory approach to the main program. The logistic regression has the highes regression followed by the Tree Ensemble and the least accurant was the PNN. The limitation in the study was the fact that other factors like non academic factors, technical skills and extracurricular activities were not taken into consideration. Also the notion that there will be difficulty in improving on the academic performance in the last two years if the student fails to peform well in the first three years.

Hafez Mousa, Ashraf Maghari (2017) conducted a study to predict the model that is suitable to determine student performance using data mining classification techniques. The research was implemented to determine which classifier performs better with the collected educational data. The classifiers include Naives bayes, Decision tree, and K-NN classifiers. The Decision tree classifiers gave the best accurancy when used with student’s data (academic and social features). The social features had little impact on the student performance. The limitation of this research includes the fact that it determine the student that will fail but not the reason for the failure. The reason for such failure may be social features since it has little impact on the academic performance. Student behavior to learning may also affect their academic performance.

Khasanah et al. (2017) conducted a study to determine how high influence attribute may be selected carefully to predict student performance. The feature selection was used before classification techniques. The data was collected from the Department of Industrial Enginneering University Islam Indonesia. The feature selection method showed that students’ attendance and the GPA of the first semester topped the list of features.

The classifeirs used were the Bayesian Network and Decision tree. It was observed that the Bayesian network has the highest accurancy than the decision tree. The limitation in this study is the fact that social factors, age and gender were not evaluated to give a comprehensive report on the study.

Hilal Almarabeh (2017) used WEKA tool to evaluate the performance of university students. Different data mining classifiers were used to evaluate this performance. There were 225 instances and ten features were selected. The classifiers used under WEKA include Naives Bayes, Neural Network, Bayesian Network, ID3 and J48. The study showed that the Bayesian Network has the highest accurancy in evaluating the performance of university students. The study is limited in the sense that more datasets instance will be collected, compared and analyzed with other data mining technioques such as associative and clustering should be used.

Aderibigbe et al. (2018) used KNIME and ORANGE tool to predict the performance of students and the extent of the relationship between the academic results at the point of admission based on the university admission entry requirements. The datasets contains results of students of Covenant University.

The parameters used were students students' entry age, aggregate WAEC score, Jamb score, Putme score and CGPA for the first year. The data mining tool used were KNIME and Orange. The data mining algorithm used include Neural network, Decision tree, Naïve Bayes, Logistic regression, Resilient back propagation, Random forest, Tree Ensemble and Multilayer perception algorithm. Using the ORANGE tool it was discovered that Neural network has the highest accurancy and regression was used to further validate the accurancies. Using the KNIME tool, it was discovered that Neural network has the highest accurancy. Also checking the percentage of accurancies of both tools, ORANGE tool has the highest percentage (51.9%) while KNIME has (50.23%). The accurancy level was low due to the expectation of the common assumption that the performance of students based on the entry requirement is a strong indicator of the performance of a student once in admitted into a university. The limitation of the study is that other factors were not included like non academic factors and psychological factors may affect the performance of students. This calls for area of further research of this study.

Strecht et al. (2015) predicted the outcome of students result and their grade. The study was carried out to predict students that will pass or fail due to the grades of the student. The use of classification and algorithm models were employed. The datasets contain 700 courses student data at the University of Porto. The classification algorithm used include K-NN, Random forest, AdaBoost, Classification and regression tree (CART), Support vector machine, Naïve Bayes and Ordinary Least Square. Positive results were obtained in predicting which student will pass or fail while predicting the grade of student was bad. Limitation of the study will be addition of new features like academic goals, personal interest, time management skills, sports activities, sleep habits will be an area of further research to encourage a worthwhile investigation.

IV. METHODOLOGY

The research design employed is Quantitative design. Quantitative design deals with numbers and statistics. The research method employed the use of WEKA tool, an open source tool. WEKA machine is a collection of visualization tools and algorithms for data analysis and predictive modeling.

A. Overview Of The Research

The research process was majorly divided into four phases which are:

1) *Data Collection And Integration:* This process comprised of collecting and combining multiple data. The data relevant to the studywere selected. The data collected includes the academic results of students in their first academic year. The secondary data include the students' gender and other non-academic data. The collectd data were stored in Microsoft Excel worksheet.
2) *Data Transformation:* This process was sub divided into three stages; data selection, data cleaning and normalization. It involved selection of relevant data based on the features and cleaning of the data selected to remove any incomplete or missing data. The cleaned data were then normalized and fed into the data mining algorithm to extract the meaningful data pattern which helps in prediction model.
3) *Pattern Extraction:* This subdivided into phases which are clean data training, pattern, testing and result evaluation. This process used specialized algorithm and analytical tool to discover the trend of pattern in the data provided. These patterns are further tested and the result evaluated to determine the knowledge being represented in the study.
4) *Knowledge Representation:* This phase was the final stage and drived the decision making. This stage determines if a pattern truly exists and the relationship within the dataset. This process is represented in isualized forms such as tables, forms etc.

The research stages are represented with the diagram below.

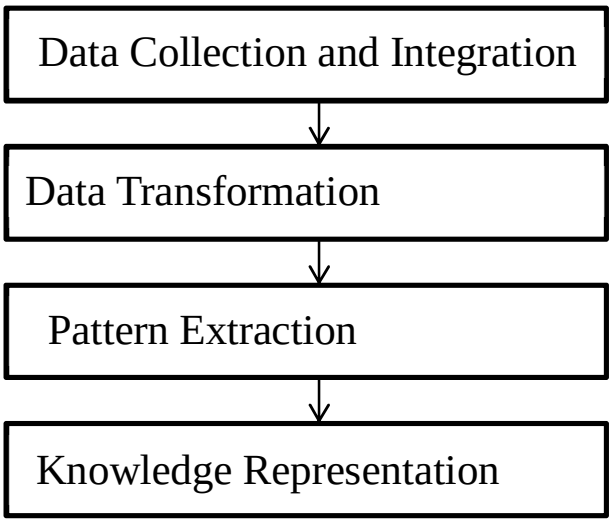


Fig 3.1Research Design

B. Population Of The Study

The population of the dataset were 411 data sets broken into 161 National Diploma (ND) Students and 250 students for Higher National Students (HND) National Diploma and Higher National Diploma students of Computer Science Departments Yaba College of Technology, Lagos and Lagos State Polytechnic, Lagos.

C. Data Selection And Parameters

Data selection in data mining is the process where the most relevant data is selected from a specific domain. The data selected will be used informative and facilitate learning within the domain. The dataset used was obtained from Department of Computer Science, Yaba College of Technology. Irrelevant fields of the data were cleaned and removed to enhance prediction accurancy. On the basis of the information obtained the attributes listed for the dataset include age, gender, CGPA, grade in C++, C, Java, Visual Basic, Calculus, Algebra and Statistics. This is shown in the table below:

Table 3.1. Student Performance Dataset Description.

S/N	ATTRIBUTE	DATA TYPE
1	Name	String
2	Gender	Character
3	Matric Number	Polynomial
4	Java	Numeric
5	C	Numeric
6	C++	Numeric
7	Calculus	Numeric
8	Algebra	Numeric

D. Applied Algorithms

WEKA tool is a data mining tool that supports several tasks like data preprocessing, clustering, classification, regression, feature selection and selection Sunita and Lobo (2011). For this research the proposed algorithm used are:

1) Logistic regression
2) Naives Bayes
3) Random forest algorithm
4) Decision tree (J48)

This different predictive algorithm used will enhance the prediction accurancy of the research.

V. RESULTS AND DISCUSSION

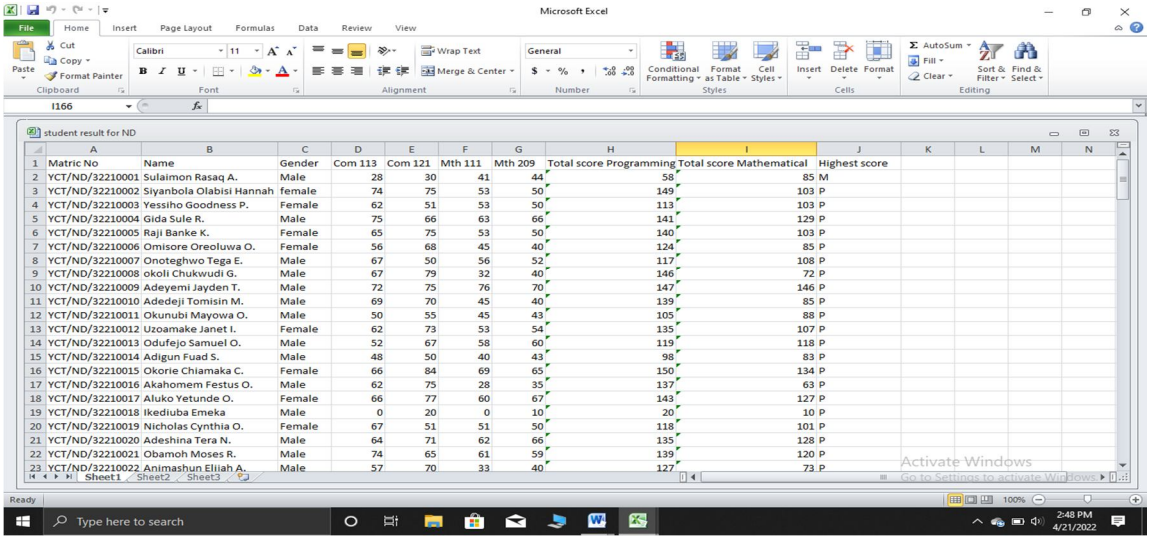
Data used for this work were obtained from Computer Science departmwnts of Yaba College of Technology (Yabatech) and Lagos State Polytechnic (Laspotech). The data covered National Diploma and Higher National Diploma students of Computer Science Department. The dataset consist of total of 161 National Diploma Students and 250 students for HND1.

This dataset were then entered into Microsoft Excel spreadsheet where various computations were done. The table below gives a detailed information of all the components that were entered into the Excel and description of each component and how it was gotten.

Table 4.1 Representation of the dataset

Component	Description
Instances in rows	ND students (161 instances) HND students (250 intsances)
Attributes in columns	The description of all the headings of each columns.
Matric No (Polynomial)	This consist of the matric numbers of students and this is a unique key that identifies each student
Name (String data type)	This gives personal details such as the surname and names of the students.
Gender (character data type)	This can be either male or female gender
Courses COM 113 COM121 COM 313 COM 325 MTH 112 MTH 209 MTH 311 MTH 312	 C Programming Java Programming C++ Programming Java Programming Algebra Calculus Advanced Algebra Advanced Calculus
Add Weight Grade	This is gotten by multiplying the WGP by the unit course for each individual courses and adding them together.
Grade Point Average	This is gotten by dividing the Add Weight Grade by the total number of each unit course
Grade	This is grading the students based on the results gotten from the grade point average.

On the Windows platform you will open Microsoft Excel and from there you input all the necessary details. Below is the outlook of how all the details will look like in Microsoft Excel. The Microsoft Excel spreadsheet is of two categories which comprises of student report for National Diploma and student report for Higher National Diploma.



Matric No	Name	Gender	Com 113	Com 121	Mth 111	Mth 209	Total score Programming	Total score Mathematical	Highest score
YCT/ND/32210001	Sulaimon Rasaq A.	Male	28	30	41	44	58	85	M
YCT/ND/32210002	Siyanbola Olabisi Hannah	female	74	75	53	50	149	103	P
YCT/ND/32210003	Yessiho Goodness P.	Female	62	51	53	50	113	103	P
YCT/ND/32210004	Gida Sule R.	Male	75	66	63	66	141	129	P
YCT/ND/32210005	Raji Banke K.	Female	65	75	53	50	140	103	P
YCT/ND/32210006	Omisore Oreoluwa O.	Female	56	68	45	40	124	85	P
YCT/ND/32210007	Onoteghwo Tega E.	Male	67	50	56	52	117	108	P
YCT/ND/32210008	okoli Chukwudi G.	Male	67	79	32	40	146	72	P
YCT/ND/32210009	Adeyemi Jayden T.	Male	72	75	76	70	147	146	P
YCT/ND/32210010	Adedeji Tomisin M.	Male	69	70	45	40	139	85	P
YCT/ND/32210011	Okunubi Mayowa O.	Male	50	55	45	43	105	88	P
YCT/ND/32210012	Uzoamake Janet I.	Female	62	73	53	54	135	107	P
YCT/ND/32210013	Odufejo Samuel O.	Male	52	67	58	60	119	118	P
YCT/ND/32210014	Adigun Fuad S.	Male	48	50	40	43	98	83	P
YCT/ND/32210015	Okorie Chiamaka C.	Female	66	84	69	65	150	134	P
YCT/ND/32210016	Akahomem Festus O.	Male	62	75	28	35	137	63	P
YCT/ND/32210017	Aluko Yetunde O.	Female	66	77	60	67	143	127	P
YCT/ND/32210018	Ikediuba Emeka	Male	0	20	0	10	20	10	P
YCT/ND/32210019	Nicholas Cynthia O.	Female	67	51	51	50	118	101	P
YCT/ND/32210020	Adeshina Tera N.	Male	64	71	62	66	135	128	P
YCT/ND/32210021	Obamoh Moses R.	Male	74	65	61	59	139	120	P
YCT/ND/32210022	Animashun Elliah A.	Male	57	70	33	40	127	73	P

Fig 4.1 Student report of National Diploma (ND1)

This is an Excel spreadsheet which contains student results and the various computations done inorder to achieve our aim. This spreadsheet is saved in Comma delimited format because it is format accepted by WEKA tool.

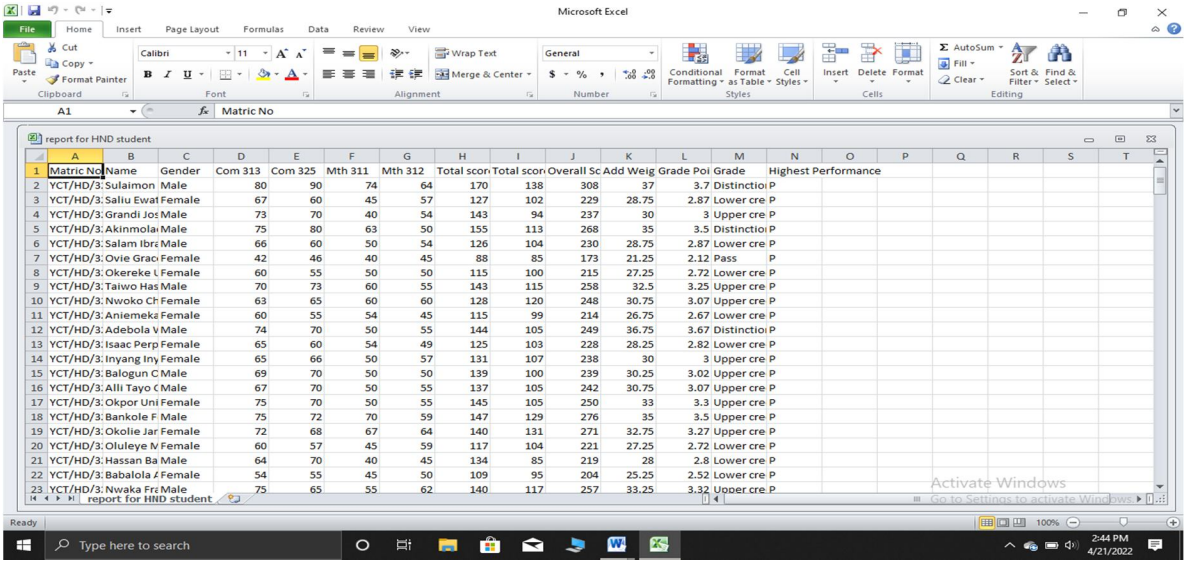


Fig 4.2 Student report of Higher National Diploma (HND1)

After getting this results and saving the excel file as a comma delimited file. This file is then opened in WEKA and the various machine learning algorithm are used inorder to discover the pattern and represent the knowledge gotten from the algorithm.

A. Result Interpretation Using Naives Bayes For Nd Result

From the diagram below, under the classifier we click on CHOOSE. This brings different options under WEKA we click on Bayes and select Naives Bayes. We click on Use training test from the test options presented and choose GRADE and start the interpretation. The results gotten shows 95% of correctly classified instances and 4% of incorrectly classified instances. It also shows detailed accurancy by class where the True Positive (TP) rate is higher than the False Positive (FP) Rate. Also it show the confusion matrix which is interpreted as thus: 12 students having Distinction, 33 students having Upper credit, 55 students having Lower Credit, a student having Pass and 15 students having Fail in their overall performance. This shows that mathematics and programming have an influence on the overall performance of students. From the statistics it can be deduced that there is an average overall performance of students and less failure. We can deduce that most students which perform excellent in programming course is intended to have a good performance in mathematics.

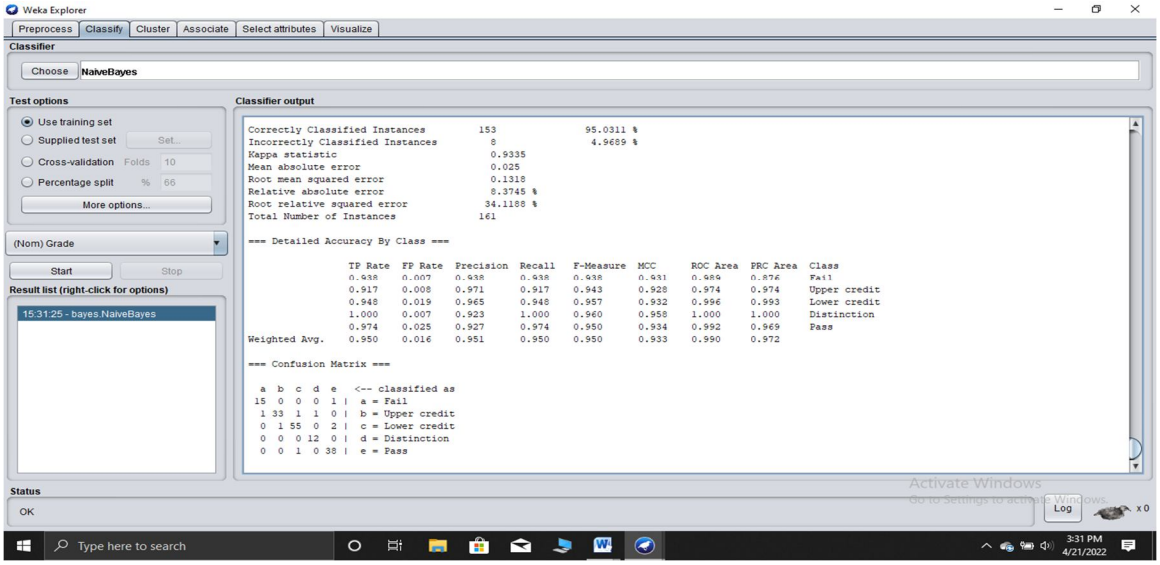


Fig 4.5 Result interpretation using Naives Bayes for ND students

1) Result Interpretation Using Logistic Regression

From the diagram below, under the classifier we click on CHOOSE. This brings different options under WEKA we click on Functions and select Logistic. We click on Use training test from the test options presented and choose GRADE and start the interpretation. The results gotten shows 100% of correctly classified instances and 0% of incorrectly classified instances. It also shows detailed accurancy by class where the True Positive (TP) rate is higher than the False Positive (FP) Rate. Also it show the confusion matrix which is interpreted as thus: 12 students having Distinction, 36 students having Upper credit, 58 students having Lower Credit, 39 student having Pass and 16 students having Fail in their overall performance. This shows that mathematics and programming have an influence on the overall performance of students. From the statistics it can be deduced that there is an average overall performance of students and less failure. We can deduce that most students which perform excellent in programming course is intended to have a good performance in mathematics.

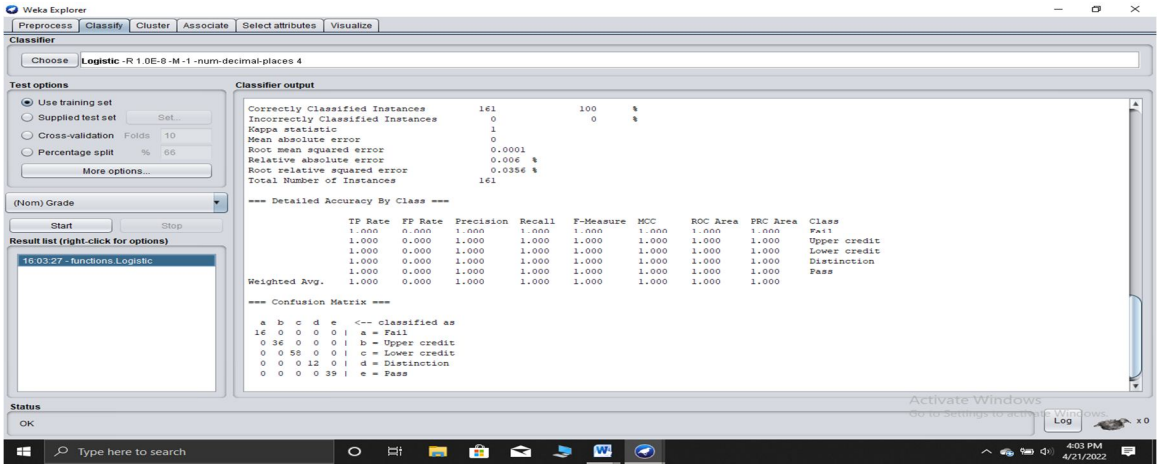


Fig 4.6 Result interpretation using Logistic regression for ND students.

2) Result Interpretation Using Decision TREE (J48)

From the diagram below, under the classifier we click on CHOOSE. This brings different options under WEKA we click on Trees and select J48. We click on Use training test from the test options presented and choose GRADE and start the interpretation. The results gotten shows 98% of correctly classified instances and 1% of incorrectly classified instances. It also shows detailed accurancy by class where the True Positive (TP) rate is higher than the False Positive (FP) Rate. Also it show the confusion matrix which is interpreted as thus: 12 students having Distinction, 35 students having Upper credit, 58 students having Lower Credit, 39 student having Pass and 15 students having Fail in their overall performance. This shows that mathematics and programming have an influence on the overall performance of students. From the statistics it can be deduced that there is an average overall performance of students and less failure. We can deduce that most students which perform excellent in programming course is intended to have a good performance in mathematics.

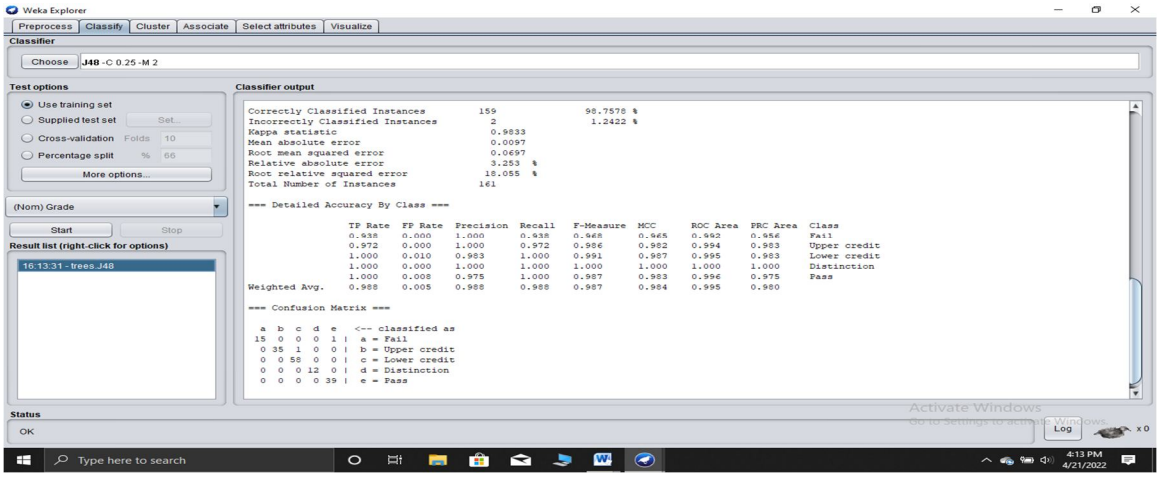


Fig 4.7 Result interpretation using logistic regression for ND students.

3) Result Interpretation Using Random Forest

From the diagram below, under the classifier we click on CHOOSE. This brings different options under WEKA we click on Trees and select Random forest. We click on Use training test from the test options presented and choose GRADE and start the interpretation. The results gotten shows 100% of correctly classified instances and 0% of incorrectly classified instances. It also shows detailed accurancy by class where the True Positive (TP) rate is higher than the False Positive (FP) Rate. Also it show the confusion matrix which is interpreted as thus: 12 students having Distinction, 36 students having Upper credit, 58 students having Lower Credit, 39 student having Pass and 16 students having Fail in their overall performance. This shows that mathematics and programming have an influence on the overall performance of students. From the statistics it can be deduced that there is an average overall performance of students and less failure. We can deduce that most students which perform excellent in programming course is intended to have a good performance in mathematics.

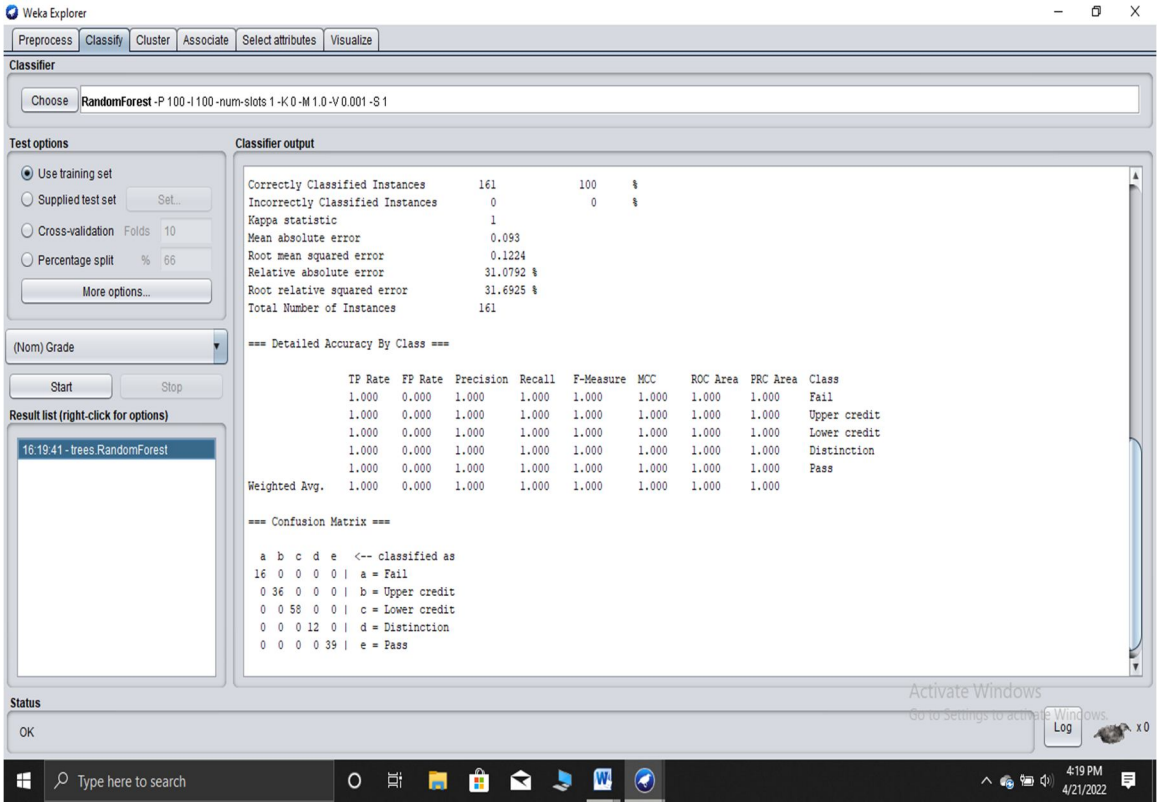


Fig 4.8 Result interpretation using Random Forest for ND students

B. Result Interpretation For HND Students

Having concluded the result interpretation for ND students we move forward to interpret the result for HND students to check if there is a relationship between mathematics and programming language courses. This results also help to determine how both courses affect the overall results for computer science students. From the diagram below, we could see the dataset being classified having relation as report for HND, 14 attributes and 249 instances. There is also a visualization area which breaks down each attribute showing their name, type, distinct, missing and unique features when an attribute is clicked upon.

After going through each attribute, we go over to the menu bar and click on Classify. This shows another interface which enables you to choose different machine learning algorithm and interpretation of the results gotten.

For this particular work we will be implementing four different machine learning algorithm namely:

- Naives Bayes:
- Logistic Regression
- Decision tree (J48)
- Random Forest.

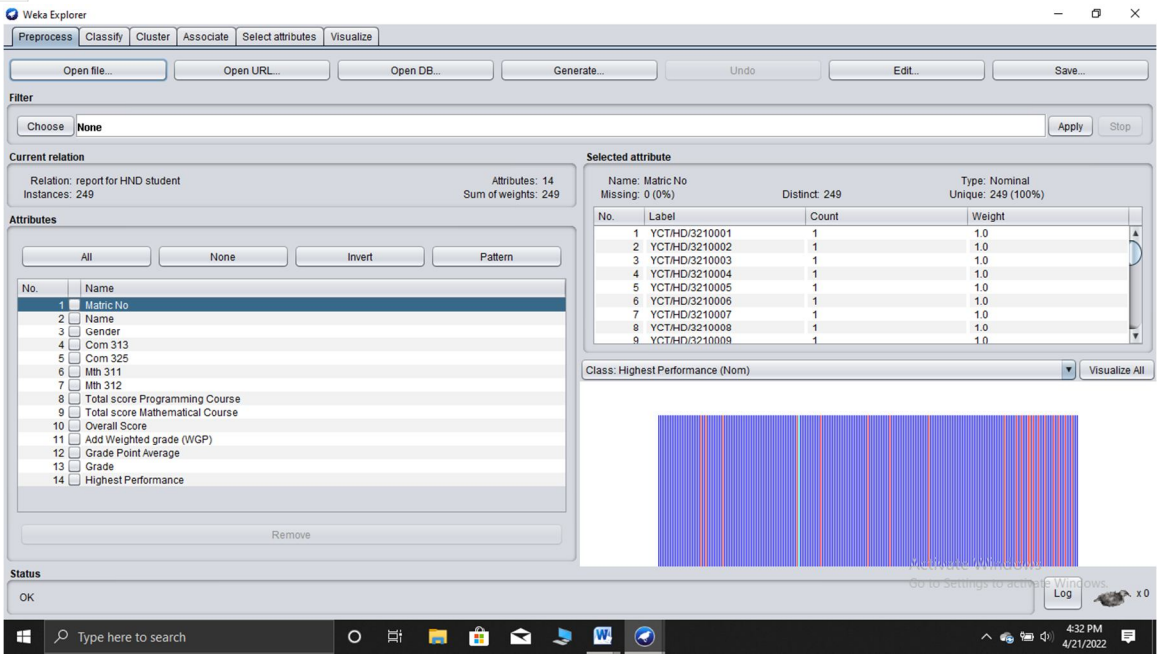


Fig 4.9. WEKA displaying Student result for HND.

1) Result Interpretation Using Naives Bayes

From the diagram below, under the classifier we click on CHOOSE. This brings different options under WEKA we click on Bayes and select Naive Bayes. We click on Use training test from the test options presented and choose GRADE and start the interpretation. The results gotten shows 95% of correctly classified instances and 4% of incorrectly classified instances. It also shows detailed accurancy by class where the True Positive (TP) rate is higher than the False Positive (FP) Rate. Also it show the confusion matrix which is interpreted as thus: 18 students having Distinction, 62 students having Upper credit, 97 students having Lower Credit, 51 student having Pass and 9 students having Fail in their overall performance. This shows that mathematics and programming have an influence on the overall performance of students. From the statistics it can be deduced that there is an average overall performance of students and less failure. We can deduce that most students which perform excellent in programming course is intended to have a good performance in mathematics.

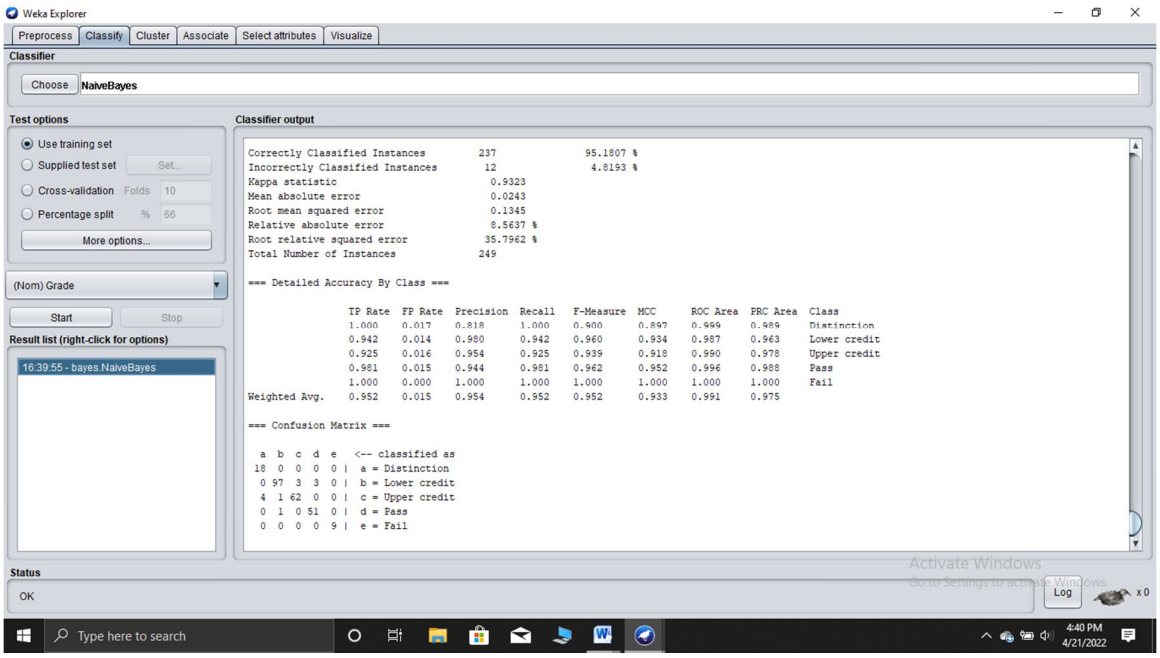


Fig 4.10 Result interpretation using Naive Bayes

2) Result Interpretation Using Logistic Regression

From the diagram below, under the classifier we click on CHOOSE. This brings different options under WEKA we click on Functions and select Logistic. We click on Use training test from the Test Options presented and choose GRADE and start the interpretation. The results gotten shows 100% of correctly classified instances and 0% of incorrectly classified instances. It also shows detailed accurancy by class where the True Positive (TP) rate is higher than the False Positive (FP) Rate. Also it show the confusion matrix which is interpreted as thus: 18 students having Distinction, 67 students having Upper credit, 103 students having Lower Credit, 52 student having Pass and 9 students having Fail in their overall performance. This shows that mathematics and programming have an influence on the overall performance of students. From the statistics it can be deduced that there is an average overall performance of students and less failure. We can deduce that most students which perform excellent in programming course is intended to have a good performance in mathematics.

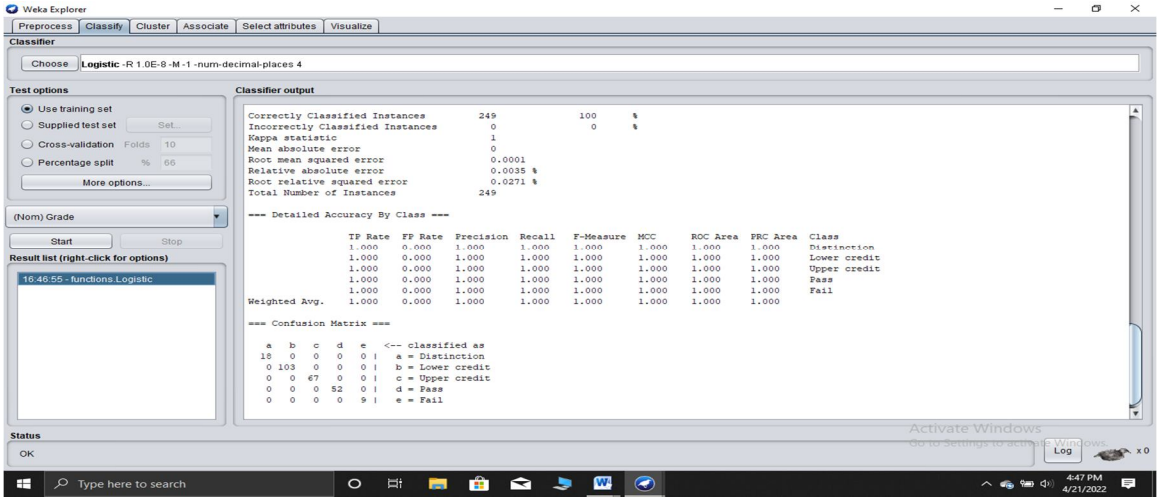


Fig 4.11 Result interpretation using Logistic Regression

3) Result Interpretation Using Decision TREE (J48)

From the diagram below, under the classifier we click on CHOOSE. This brings different options under WEKA we click on Trees and select J48. We click on Use training test from the test options presented and choose GRADE and start the interpretation. The results gotten shows 99% of correctly classified instances and 0.80% of incorrectly classified instances. It also shows detailed accurancy by class where the True Positive (TP) rate is higher than the False Positive (FP) Rate. Also it show the confusion matrix which is interpreted as thus: 18 students having Distinction, 66 students having Upper credit, 102 students having Lower Credit, 52 student having Pass and 9 students having Fail in their overall performance. This shows that mathematics and programming have an influence on the overall performance of students. From the statistics it can be deduced that there is an average overall performance of students and less failure. We can deduce that most students which perform excellent in programming course is intended to have a good performance in mathematics.

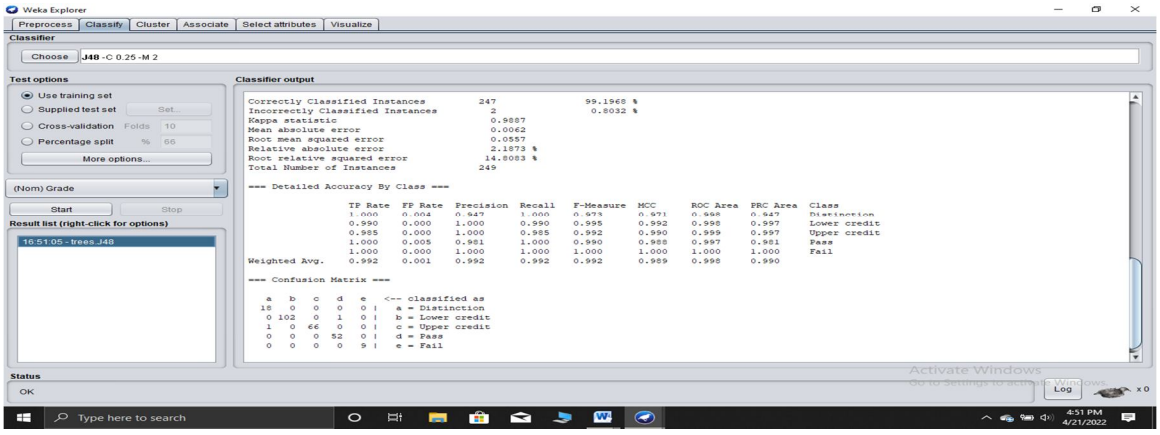


Fig 4.12 Result interpretation using Logistic regression.

4) Result Interpretation Using Random Forest

From the diagram below, under the classifier we click on CHOOSE. This brings different options under WEKA we click on Trees and select Random forest. We click on Use training test from the test options presented and choose GRADE and start the interpretation. The results gotten shows 100% of correctly classified instances and 0% of incorrectly classified instances. It also shows detailed accurancy by class where the True Positive (TP) rate is higher than the False Positive (FP) Rate. Also it show the confusion matrix which is interpreted as thus: 18 students having Distinction, 67 students having Upper credit, 103 students having Lower Credit, 52 student having Pass and 9 students having Fail in their overall performance. This shows that mathematics and programming have an influence on the overall performance of students. From the statistics it can be deduced that there is an average overall performance of students and less failure. We can deduce that most students which perform excellent in programming course is intended to have a good performance in mathematics.

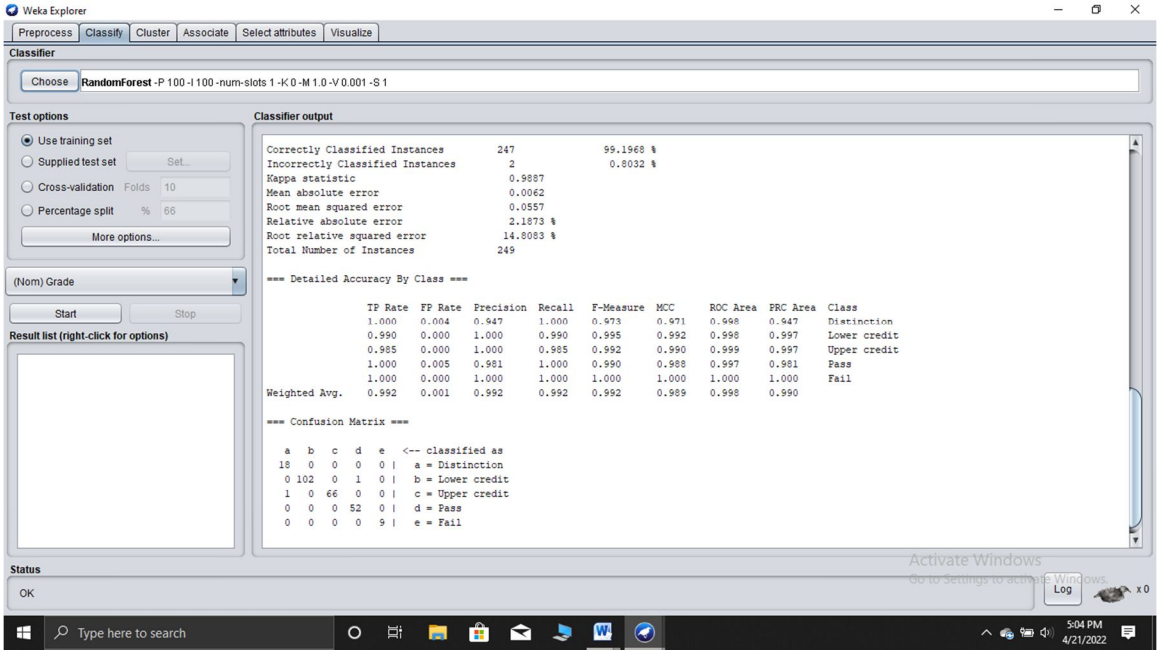


Fig 4.13 Result interpretation using Random Forest.

VI. DISCUSSION OF RESULT

The results gotten from the machine learning algorithm will be discussed further with the aid of tables to compare the algorithm based on different criterions. The analysis will help in drawing conclusion on the work.

This tables will be categorised into 2, Table 1 will highlight classification of algorithm using their efficiency while Table 2 will highlight classfier accurancy evaluation measure by class and this will also include a break down of the confusion matrix.

Table 4.2 Classification Of Algorithm Using Their Efficiency For Nd Students.

Classifier (total Instance)	Algorithm Implemented	Correctly Classified Instances	Incorrectly Classified Instances	Time Taken (seconds)	Kappa Statistics	Mean Absolute error	Root Mean Error	Relative absolute error (%)	Relative square error (%)
Bayes	Naive Bayes	153	8	0.01	0.8677	0.025	0.1318	8.37	34.11
Functions	Logistic	161	0	0	1	0	0.0001	0.006	0.0356
Trees	J48	159	2	0	0.8677	0.0419	0.1733	13.98	44.83
Trees	Random Forest	161	0	0.01	1	0.093	0.1224	31.07	31.69

TABLE 4.3 Classifier Accurancy Evaluation Measure For Nd Students.

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Naive Bayes	0.950	0.016	0.951	0.950	0.950	0.990
Logistic Regression	1.000	0.000	1.000	1.000	1.000	1.000
J48	0.901	0.033	0.902	0.901	0.900	0.984
Random Forest	1.000	0.000	1.000	1.000	1.000	1.000

Table 4.4 Classification Of Algorithm Using Their Efficiency For Hnd Students.

Classifier (total Instance)	Algorithm Implemented	Correctly Classified Instances	Incorrectly Classified Instances	Time Taken (seconds)	Kappa Statistics	Mean Absolute error	Root Mean Error	Relative absolute error (%)	Relative square error (%)
Bayes	Naive Bayes	237	12	0	0.9435	0.0206	0.1149	7.26	30.56
Functions	Logistic	249	0	0.01	0.9829	0.004	0.0446	1.42	12.27
Trees	J48	247	2	0	0.9887	0.0062	0.0557	2.1873	14.8083
Trees	Random Forest	249	0	0	1	0.091	0.225	32.11	32.60

TABLE 4.3 Classifier Accurancy Evaluation Measure For Hnd Students.

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Naive Bayes	0.960	0.013	0.961	0.960	0.960	0.995
Logistic Regression	1.000	0.000	1.000	1.000	1.000	1.000
J48	0.901	0.033	0.902	0.901	0.900	0.984
Random Forest	1.000	0.000	1.000	1.000	1.000	1.000

TABLE 4.5 Confusion Matrix Of The Algorithm For Nd Students

Algorithm	Class	Correctly Classified Instance	Incorrectly Classified Instance
Naive Bayes	Distinction	12	0
	Upper Credit	33	3
	Lower Credit	55	3
	Pass	38	1
	Fail	15	1
Logistic Regression	Distinction	12	0
	UpperCredit	36	0
	Lower Credit	58	0
	Pass	39	0
	Fail	16	0
J48	Distinction	12	0
	Upper Credit	35	1
	Lower Credit	58	0
	Pass	35	1
	Fail	15	1
Random Forest	Distinction	12	0
	Upper Credit	36	0
	Lower Credit	58	0
	Pass	39	0
	Fail	16	0

TABLE 4.6 Confusion Matrix Of The Algorithm For Hnd Students

Algorithm	Class	Correctly Classified Instance	Incorrectly Classified Instance
Naive Bayes	Distinction	18	0
	Upper Credit	63	4
	Lower Credit	98	5
	Pass	51	1
	Fail	9	0
Logistic Regression	Distinction	18	0
	UpperCredit	67	0
	Lower Credit	102	1
	Pass	50	2
	Fail	9	0
J48	Distinction	18	0
	Upper Credit	66	1
	Lower Credit	102	1
	Pass	52	0
	Fail	9	0
Random Forest	Distinction	18	0
	Upper Credit	67	0
	Lower Credit	103	0
	Pass	52	0
	Fail	9	0

The dataset was analyzed using WEKA and Random Forest, J48, Naïve Bayes and Logistic Regression algorithms. The algorithms were applied for the HND and ND dataset and there was comparison based on their accurancy, learning time and percentage of correctly classified instances. This comparison shows that there is direct relation between the execution time in building models and volume of data records. From the analysis, the Kappa Statistics which is a metric that compares an observed accurancy with an expected accurancy. The Kappa statistics of the algorithms implemented show that the value is less than 1 or equal to 1. This shows that the accurancy level was high and also it was used to evaluate classifiers among themselves due to the varying degree of the Kappa statistics. The mean absolute error measures the average of the difference between the actual value and the predicted value. The values gotten from all statistics were closer to 0. This signifies that the model built was a better model. The machine learning algorithms Logistic Regression and Random forest gave the best accurancy in both database. Each having an accurancy of 100% in ND result dataset. 100% in Random forest for HND result dataset and 99% for Logistic Regression in HND results.

The confusion matrix show less incorrect instances which means that mainly all the instances were correctly classified. This is as a result of the data cleansing done while inputing the data. Dirty and missing dataset were removed from the dataset.

This shows that the predictor will not only predict the number of students that are likely to be in distinction, upper credit, lower credit, pass and fail but also show the relationship between having the knowledge of mathematics and programming language for an overall performance in computer science. The knowledge pattern represented further satisfies our aim that it is imperative for students to have a standard knowledge of mathematics as this will help in being the best in their chosen profession.

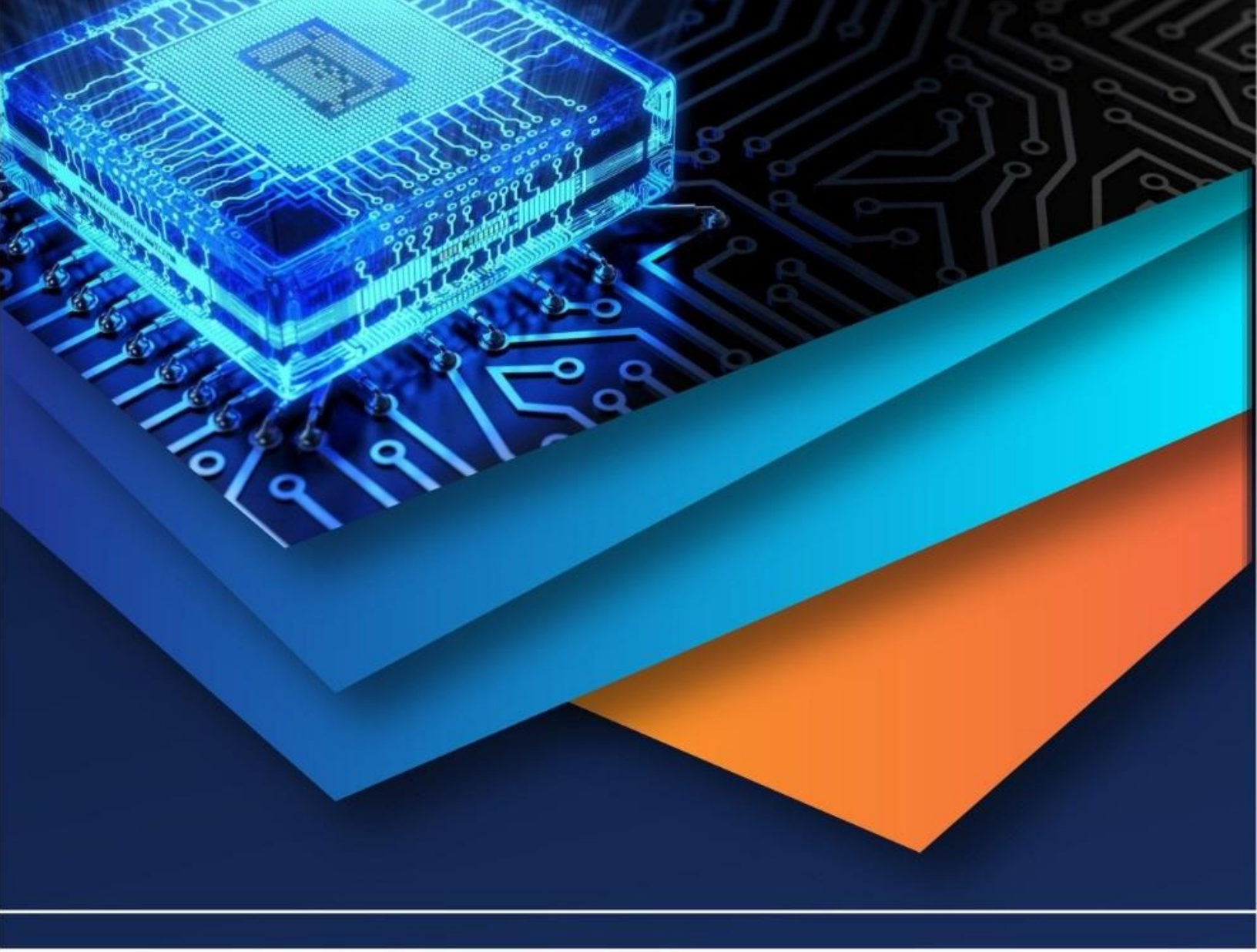
VII. CONCLUSION

This work shows that it is imperative for students to have a standard knowledge of mathematics for the study of programming language courses as this will help them in their reasoning and understanding of programming logic. Understanding Programming logic is the key to understanding and writing computer programs. Random Forest algorithm emerged the best algorithm. Random forest has the best precision and recall accurancy which is 1.000 for ND and HND predictive model. Also it gave the best classification accurancy of classifying all the correct instances. This shows that Random forest gave an accurancy of 100% respectively and learning time of 0 seconds and 0.1 seconds for HND and ND predictive model.

WEKA tool was mainly used in carrying out the data analysis and classification of this dataset. Random forest serves as the best algorithm in generating the result predictor application.

REFERENCES

[1] Aderibigbe, I. A. and Odunayo S. (2019). Towards an improved learning process: the relevance of ethnicity to data mining prediction of students' performance. Applied Sciences 2(8).
[2] Aderibigbe, I. A. and Odunayo, S. (2019). The impact of Engineering students' performance in the first three years on their graduation results using educational using educational data mining. Heliyon, 5(2).
[3] Adekitan, I. A. and Noma-Osaghie, E. (2018). Data mining approach to predicting the performance of first year student in a university using the admission requirement. Education Information Technology.
[4] Almarabeh, H. (2017). Analysis of students' performance by using different data mining classifier. International Journal of Modern Educational Computer Science, 9(8) (pp. 9).
[5] Ahmed, A. B. E. D. and Elaraby, I. S. (2014). Data mining: A prediction for student's performance using classification method. World Journal of Computer Application and Technology, 2(2) (pp. 43-47)
[6] Belford, Geneva, G. and Allen T. (2019). Computer Science. The Encyclopedia Britannica. Retrieved from http://www.britannica.com/science/computer science.
[7] Bocconi, S., Chioccariello, A. and Earp J. (2018). The Nordic approach to introductory computational thinking and programming in compulsory education. Retrieved from http://www.doi.org/10.17411
[8] Cristobal, R. and Sebastian V. (2010). Educational data mining: A review of the start of the art. IEEE Transactions on Systems man and Cybernetics part (Applications and Review), 40(6) (pp. 601-618).
[9] Dr. Sudhir, B. J., and Dr. Kodge B. G. (2013). Census Data mining and Analysis using WEKA. International Conference in Emerging trends in Science, Technology and Management.
[10] Evaristus, D. M., David, F., David, J. M. S. and Johanes, A. (2021). Big data in Educational Institutions using Rapid Miner to predict learning effectiveness. Journal of Computer Science, 17(4) (pp. 403-413).
[11] Fayyad, U., Platesky-shapiro, G., and Smyth P. (1996). From data mining to knowledge discovery in database. AI magazine, America Association for Artificial Intelligen
[12] Fadhilah, A., Nur, H. I. and Azwa A. A. (2015). The prediction of students' academic performance using Classification Data mining techniques. Applied mathematical science, 9(129) (pp. 6415-6426).
[13] Grover, S. and Pea, R. (2013) computational thinking in K12: A review of the state of field. Educational research, 42(1).
[14] Hafez, M. and Maghari, A. Y. A. (2017). Students' performance using Data mining Classification. International Journal of Advanced Research in Computer and Communication Engineering 6(8).
[15] Ke, F. (2014). An Implementation of design bared learning through creating educational computer games. A case study on mathematics learning during design and computing, computers and education, 73(26)
[16] Khasanah, A. U. and Haiwati, (20180. A comparative study to predict student performance using Educational data mining Techniques. IOP Conference Seriies. Materials Science and Engineering 215.
[17] La paglia, F., La cascio C., Francomano, M. and La barbera, D. (2017). Educational robotics to improve mathematics and metacognitive skills: Annual review of cypertherapy and telemedicine, 15(70).
[18] Lambic, D. (2011). Presenting practical application of mathematics by use of programming software with easily available visual components: Teaching mathematics and its applications. International journal of the IMA, 30(1).
[19] Maqsud, S. K., and Dr. Nidhi H. D. (2017). Analysis of data using Data mining tool ORANGE. International Journal of Engineering development and research, 5(2).
[20] Michael, R., Nicolas, C., Fabian, O., Tobias, K., Thirsten, M., Peter, O., Thomas R. G. and Beind, W. (2009). KNIME: Konstanz Information : version 2.0 and beyond. ACM Slakod explorations newsletter, 11(1) (pp. 26-31).
[21] Mohammed, A. (2021). Intro to Rapidminer: a No-code development platform for data mining (with case study). 2021 Data Science Blogathon.
[22] Papert, S. (1980). Mindstorms: children, computers and powerful ideas. New York, Badio Books Incorporation.
[23] Rashmi, K. (2021). Data mining functionalities-an overview. Naukri learning Retrievd from http://www.naukri.com.
[24] Rohit, S. (2020). KDD Process in data mining: what you need to know. Upgrad Education Retrieved from http://www.upgrad.com.
[25] Sadiq, H., Neama, A. D., Fadl, M. B. and Ribata, N. (2018). Educational Data mining and analysis of students' academic performance using WEKA. Indonesian Journal of Electrical engineering and Computer Science, 7(2) (pp. 447-459).
[26] Saran, E. F. and Odd, T. K. (2018). A literature review exploring the use of programming in Mathematics Education. International Journal of Learning, Technology and Educational Research, 17(12) (pp. 18-32)
[27] Shivangi, G, and Neeta, U. (20160. Comparative analysis of Classification Algorithm using WEKA tool. International Journal of Scientific and Engineering Research 7(8).
[28] Strecht, P., Luis, C., Carlos, S. and Joao, M. M. (2015). A comparative study of classification and regression algorithms for modellings students' academic performance. International conference on Educational Data mining.
[29] Vladan, D. (2001). Knowledge discovery and data mining in database. World Scientific Publishing.
[30] Yelland, N. (1995). Mindstorms or a storm in a teacup? A review of research with Loga, International Journal of Mathematical Education in Science and Technology, 26(6).
[31] IntelliPaat, (2023), Top 10 Data Mining Applications, https://intellipaat.com/blog/top-data-mining-applications/?US



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY