COMPARISON OF FOUR ML PREDICTIVE MODELS PREDICTIVE ANALYSIS OF BIG DATA

SALWA ZAKI ABD ELHADY[1], NEVEEN I. GHALI[2] , AFAF ABO-ELFETOH[3] , AMIRA M. IDREES[4]

[1] Department of computer science, Faculty of Science, Azhar University, Cairo, Egypt
[2] Department of computer science, Faculty of Computers and Information Technology, Future University Cairo, Egypt.
[3] Department of computer science, Faculty of Science, Azhar University ,Cairo, Egypt
[4] Department of computer science, Faculty of Computers and Information Technology, Future University, Cairo, Egypt.

E-mail: [1]salwazaki@yahoo.com, [2]neveen.ghali@fue.edu.eg, [3a]faf211@yahoo.com, [4]amira.mohamed@fue.edu.eg

ABSTRACT

Big Data is the main factor in all fields of human existence be it medical, social networks, or research, it has also made inroads into education. The large size and complexity of datasets in Big Data need specialized statistical tools for analysis where python can come in handy. The Categorical component of any data set can be quantified using limited representations but evaluating it concerning the quantitative variables return a larger set of statistical inferences. This research explores the analysis of categorical and quantitative variables scalable to Big Data in education using a contemporary statistical tool. python provides multiple dimensions to statistical analysis of the dataset; this paper however explores the statistical inference rendered using the Box Plot feature through summary measures of the dataset. These statistical inferences can be used to train a Machine for predictions and classification under a certain category, one of these important analysis approaches is the Predictive analysis which is central to almost every research project. Whether the goal is to identify and describe trends and variation in populations, create new measures of key phenomena, or simply describe samples in studies aimed at identifying causal effects, predictive analyses are part of almost every empirical paper and report. many studies have shown the important directions that are extracted from predictive analysis in various sectors, and in this work, the predictive analytics applies to a big sample of educational data in Egypt census 2017 to produce estimates of the variation of educational level related to some other population features.

Keywords*:* *Big Data, Big Data Analysis, Machine Learning, Machine Learning Models, Predictive Analysis*

1. INTRODUCTION

Big data is currently a buzzword in both academia and industry, with the term being used to describe a broad domain of concepts, ranging from extracting data from outside sources, storing and managing it, to processing such data with analytical techniques and tools. [1]. Due to the rapid increase in interest in big data and its importance to academia, industry, and society, solutions to handling data and extracting knowledge from datasets need to be developed and provided with some urgency to allow decision-makers to gain valuable insights from the varied and rapidly changing data they now have access to. [2] [3] Such tasks involve recognition, diagnosis, planning, robot control, prediction, etc. [4] In This paper, many predictive models were applied for the same dataset and their results were compared and discussed to advise the most appropriate model for this data set. the various machine learning models were applied and compared the approximation of results to answer which model is appropriate for this sample and the data type. The main elements of the paper are the importance of Big Data, Big Data challenges, Big Data properties (V's Model), and Big Data's moving parts. The importance and requirements of big data analysis by the latest technologies, various Big Data Analysis approaches, and Tools Databases/warehouses, machines, Data Mining, and Programming languages were discussed under the big data process. Future research problems will promise the benefits of Big Data analytics and management.

2. RESEARCH SCOPE AND OBJECTIVES

Whilst education is the core of life welfare and due to the importance of studying the coefficients that impact the education level to aid in the enhancement of education's level. This study aims to report on the results of the various ML analysis models in a big sample of the big data of census Egypt2017 and compare it to define the most accurate result that predicts the future impact of the basic life characteristics on the people's educational level.

The scope of the study is limited to 10000000 records and 7 coefficients as a big sample of 100000000 records of Egypt census data and 30 fields related to education data. Each record has personal and educational data for Egyptian citizens. The long-term goal of the research is to enhance the most accurate and appropriate ML predictive Model after applying more of one model to our big sample education data can advise about what is the life coefficient must enhancement to increase the number of Egyptian citizens who gets high educational level the surrounding environment and educational level for constraint management is defined herein as the process of identifying, classifying, modeling, and resolving constraints. The objective of the current research is to study the high impact of the surrounding environment in Egypt on education status by analyzing the relationship between education data variables and personal data variables. Particularly, the study has the following sub-objectives:

1. To choose the Community representative big sample of all big data to make sure the accuracy of the analysis.
2. To apply various ML analysis models to this sample.
3. Compare the analysis results to nominate a suitable model.
4. To predict the future directions toward the enhancement of the factors affecting education.

3. RELATED WORKS

Challenges to the use of technological tools in the study were carried out since the 1990s. However, studies on the use of big data in comprehensive education began 10 years ago.) [5] This study provides an in-depth review of Big Data Technology (BDT) advantages, implementations, and challenges in the education sector. BDT plays an essential role in optimizing education intelligence by facilitating institutions, management, educators, and learners improved quality of education, enhanced learning experience, predictive teaching and assessment strategy, effective decision-making, and better market analysis. Moreover, BDTs are used to analyze, detect and predict learners' behaviors, risk failures, and results to improve their learning outcomes and to ensure that the academic programmers undertaken are of high-quality standards. In addition, [6] This paper presents a detailed review of the latest developments in ML algorithms for Big Data Processing. ML-based Big Data Processing has gained popularity and new developments are on the rise for efficient data processing. This field is witnessing the unparalleled emergence of new methods and approaches for efficient data processing to discover interestingness for decision making. Thus, more and more ML-based data processing approaches are being used for Big Data Processing. [7] this paper focuses on the goals and purposes of Big Data in education, giving a clear picture of the value and effects of Big Data in education, how is the value potential of Big Data in recent years, and what will be the development shortly, and analyzing the learning benefits from Big Data and Open Data giving a brief description of how these technologies can contribute to a renowned education system. [8] This paper is a study on the use of Big Data in Education. Analyzed how Big Data and Open Data technology can involve education. Furthermore, how big amounts of unused data can benefit and improve education. Providing some new tools and methods bypassing the traditional difficulties and opening a new way of education. [9] this research explores the utility and applicability of deep learning for educational data mining and learning analytics. compare the predictive accuracy of popular deep learning frameworks/libraries, including, Keras, Theano, TensorFlow, fast.ai, and Pytorch. Experimental results reveal that performance, as assessed by predictive accuracy, varies depending on the optimizer used. Further, findings from additional experiments by tuning network parameters yield similar results. [10] This research compared the accuracy of machine learning algorithms that could be used for predictive analytics in higher education. The proposed experiment is based on a combination of classic machine learning algorithms such as Naive Bayes and Random Forest with various ensemble methods such as Stochastic, Linear Discriminant Analysis (LDA), Tree model (C5.0), Bagged CART (tree bag), and K Nearest Neighbors (KNN). applied traditional classification methods to classify the students' performance and to determine the independent variables that offer the highest accuracy. the results depict that the data with the 11 features using random forest generated the best accuracy value of 0.7333. However, revised the experiment with ensemble algorithms to reduce the variance (bagging), and bias (boosting) and to improve the prediction accuracy (stacking). Consequently, the bagging random forest outperformed other methods with an accuracy value of 0.7959. This research applied four different predictive models to choose more appropriate one to this data set, big categorical data, by comparison between their predictive values and accuracy.

4. BIG DATA

Big data refers to large data sets, collected by firms and governments, that are so large and complex that traditional data processing methods are inadequate to deal with the calculations needed to make sense of the data. These data sets are extremely valuable because of the vast information hidden within the data structures. When analyzed computationally, big data can provide more precise insights into hidden patterns, trends, and associations, especially in the context of human decision-making. [11]

The term big data was coined by Doug Laney in the early 2000s.1 Laney's definition includes three concepts:

- Volume: the type and detail of data being collected. Before the explosion in computing power, businesses and governments collected data but had a challenging time storing what was collected. Today, the volume of data collected from consumers and by agencies continues to grow, but because of computing capacity, storage is no longer an issue. This means that firms and agencies no longer have a data problem but instead have a computing puzzle.
- Velocity: the speed at which data are collected. Data are no longer lag. Instead, data are being collected in real-time at incredibly fast rates.
- *Variety: the types of data being collected. Whereas basic demographic data, attitudes and opinions, and possibly geographic information might have been collected in the past, today nearly anything and everything a consumer does online is being captured. Since Laney's original work, another concept has been added: veracity. This describes how much "noise" is in the data. Excessively large amounts of data can make it difficult to identify which data are important and which data are distractions.* [11]. “Fig .1”. [12]

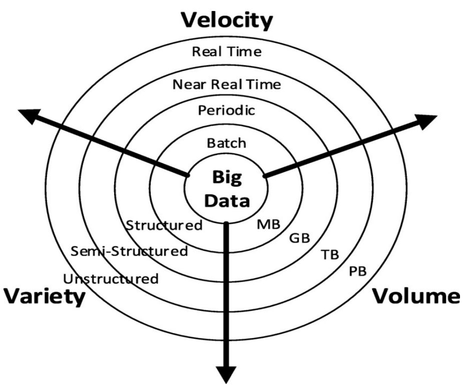

Fig .1 Big Data characteristics

4.1. Big Data Categories

Simply data is something that provides information about a particular thing and can be used for analysis. Data can have different sizes and formats. For example, all the information of a particular person in Resume or CV in pdf, Docx file format having a size in KBs. This is very small-sized data that can be easily retrieved and analyzed. However, with the advent of newer technologies in this digital era, there has been a tremendous rise in data size. data has grown from kilobytes (KB) to petabytes (PB). This huge amount of data is referred to as big data and requires advanced tools and software for processing, analyzing, and storing purposes. [13]

Big Data can be divided into the following three categories.

• Structured Data • Unstructured Data • Semi-structured Data

- Structured Data

The data that has a structure and is well organized either in the form of tables or in some other way and can be easily operated is known as structured data. Searching and accessing information from such types of data is very easy. For example, data is stored in the relational database in the form of tables having multiple rows and columns. The spreadsheet is another good example of structured data.

- Unstructured Data

The data that is unstructured or unorganized operating such type of data becomes difficult and requires advanced tools and software to access information. For Example, images and graphics, pdf files, word documents, audio, video, emails, PowerPoint presentations, webpages and web contents, wikis, streaming data, and location coordinates.

- Semi-Structured Data

Semi-structured data is structured data that is unorganized. Web data such as JSON (JavaScript Object Notation) files, .csv files, tab-delimited text files, XML, and other markup languages are examples of Semi-structured data found on the web. Due to unorganized information, semi-structured is difficult to retrieve, analyze and store as compared to structured data. It requires a software framework like Apache Hadoop to perform all this. [14]

4.2. Big Data Analysis

Big data analytics refers to the complex process of analyzing big data for revealing information such as correlations, hidden patterns, market trends, and customer preferences. big data analytics is where advanced analytic techniques operate on big data sets. Hence, big data analytics is about two things—big data and analytics—plus how the two have teamed up to create one of the most profound trends in business intelligence (BI) today. [15] At the beginning of 2009, big data analytics entered the revolutionary stage. Not only had big-data computing become a breakthrough innovation for business intelligence, but also researchers were predicting that data management and its techniques were about to shift from structured data into unstructured data, and from a static terminal environment to a ubiquitous cloud-based environment. Big data analytics computing pioneer industries such as banks and e-commerce were beginning to have an impact on improving business processes and workforce effectiveness, reducing enterprise costs, and attracting new customers. [16] in education, big data involves a variety of data types about various levels of the educational systems, on complex and social interactions, stored at different places and in multiple systems, which need to be connected to be able to analyze processes taking place in education and to improve education. The potential of big data for education has been increasingly recognized and knowledge of patterns in data can be used for improving education [17] However, several issues related to big data require attention, such as privacy and ethical issues, responsibility, availability, and the quality of the data. [18]

4.3. Types of Big Data Analytics

Big Data analytics has four types as the following: [19]

a) Descriptive Analytics

This summarizes past data into a form that people can easily read. This helps in creating reports, like a company’s revenue, profit, sales, and so on. Also, it helps in the tabulation of social media metrics.

b) Diagnostic Analytics

This is done to understand what caused a problem in the first place. Techniques like drill-down, data mining, and data recovery are all examples. Organizations use diagnostic analytics because they provide an in-depth insight into a particular problem.

c) Predictive Analytics

This type of analytics looks into the historical and present data to make predictions of the future. Predictive analytics uses data mining, AI, and machine learning to analyze current data and make predictions. It works on predicting customer trends, market trends, and so on.

d) Prescriptive Analytics

This type of analytics prescribes the solution to a particular problem. Perspective analytics works with both descriptive and predictive analytics. Most of the time, it relies on AI and machine learning.

This paper is prospective for predictive analytics, that four models for predictive analysis are implemented on the same dataset, so the following section explains the predictive analysis and its various models in detail.

4.4. predictive analysis models

Predictive analytics is a broad term that refers to machine learning algorithms that make informed conclusions based on high volumes of data. In other words, it’s how we apply modern technology to take a look into the future and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

improve the way we work, live, and learn [20] as shown in the following .Fig .2.

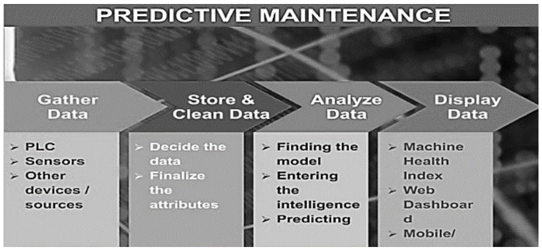

Fig .2. Predictive Model Maintenance

The most widely used predictive models are:

4.4.1. Decision trees

Decision trees are a simple, but powerful form of multiple variable analysis. They are produced by algorithms that identify various ways of splitting data into branch-like segments. Decision trees partition data into subsets based on categories of input variables, helping you to understand someone's path of decisions. [21] Decision trees are non-linear, which means there's a lot more flexibility to explore, plan and predict several possible outcomes of decisions, regardless of when they occur. Decision trees are extremely useful for data analytics and machine learning because they break down complex data into more manageable parts. They're often used in these fields for prediction analysis, data classification, and regression.

Decision trees can deal with complex data, which is part of what makes them useful. However, this doesn't mean that they are difficult to understand. At their core, all decision trees ultimately consist of just three key parts, or 'nodes':

- Decision nodes: Representing a decision (typically shown with a square)
- Chance nodes: Representing probability or uncertainty (typically denoted by a circle)
- End nodes: Representing an outcome (typically shown with a triangle).as shown in "Fig .3". [22]

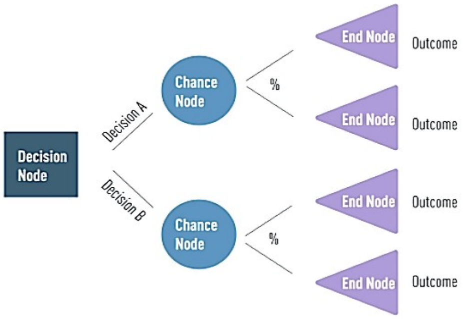

Fig .3. Decision tree diagram

4.4.2. Regression (linear and logistic)

Regression is one of the most popular methods in statistics. Regression analysis estimates relationships among variables, finding key patterns in large and diverse data sets, and how they relate to each other. Regression analysis is a statistical approach for modeling the connection between one or more independent variables and a dependent (target) variable. Regression analysis, in particular, allows us to see how the value of the dependent variable changes in relation to an independent variable while the other independent variables are maintained constant. Temperature, age, salary, price, and other continuous/real data are predicted. Regression is simply the "best guess" method for generating a forecast from a set of data. Fitting a set of points to a graph is what it is called. The most common regression models used as the following:

- Linear Regression - The most basic regression model, works best when data is linearly separable and there is little or no multicollinearity.
- Lasso Regression - Linear regression with L2 regularization.
- Ridge Regression - Linear regression with L1 regularization.
- Support vector regression (SVR) - based on the same concepts as the Support Vector Machine (SVM) for classification, with a few small modifications.

Ensemble regression- It aims to increase prediction accuracy in learning situations with a numerical target variable by combining many models. [23]

4.4.3. Neural networks

Patterned after the operation of neurons in the human brain, neural networks (also called artificial neural networks) are a variety of deep learning technologies. They're typically used to solve complex pattern recognition problems – and are incredibly useful for analyzing large data sets. They are great at handling nonlinear relationships in data – and work well when certain variables are unknown. [24] deep neural network examines data using learned representations that are like how people think about problems. The algorithm is given a collection of relevant characteristics to examine in typical machine learning, but in deep learning, the algorithm is given raw data and derives the features itself.

Deep learning is a subset of machine learning that employs multilayer neural networks. Deep learning has progressed in sync with the digital era, which has resulted in an avalanche of data in all formats and from all corners of the globe.

Big data is gathered from a variety of sources, including social media, internet search engines, e-commerce platforms, and online theatres.

However, since the data is typically unstructured, it might take decades for humans to analyze and extract useful information.

Companies are increasingly using AI systems for automated support as they see the enormous potential that can be realized by unlocking this wealth of data. Let's look at some of the most important deep learning models based on neural network architecture:

- Multi-Layer perceptron
- Convolution Neural Networks (CNN)
- Recurrent Neural Networks (RNN)
- Boltzmann machine
- Auto encoders etc. [25]

4.4.4. Random forest

Random Forest is a powerful and versatile supervised machine-learning algorithm that grows and combines multiple decision trees to create a "forest." It can be used for both classification and regression problems in R and Python. Random Forest's ensemble of trees outputs either the mode or mean of the individual trees. This method allows for more accurate and stable results by relying on a multitude of trees rather than a single decision tree. It's kind of like the difference between a unicycle and a four-wheeler [26] .Random Forest is a technique used in modeling predictions and behavior analysis and is built on decision trees. It contains many decision trees representing a distinct instance of the classification of data input into the random forest. The random forest technique considers the instances individually, taking the one with the majority of votes as the selected prediction. as shown in "Fig.4". [27]

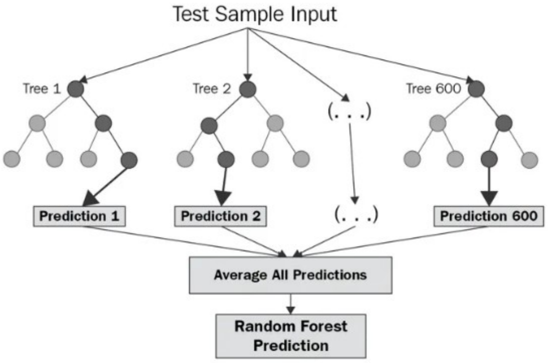

Fig .4. Random Forest technique

5. METHODOLOGY

The data was obtained from the census data 2017, and education level department data. It contains 10 million records in rows and 14 features in columns. The features are classified into two major groups Demographic Features (DF), and Academic level Features (AF), The demographic features are resident place, and gender. The academic level features consist of the educational stage and grade level. We have utilized these features using traditional classification methods such as a random forest. Moreover, we applied ensemble algorithms to choose the correct features and to predict the education level has a high number of citizens with high accuracy. The data provider is raw data. There were a lot of missing fields. The data was pre-processed again before the analysis using python software. The data was split into 80% for training and 20% for testing.

A combination of ensemble techniques were used to discuss the accuracy of machine learning algorithms results and to improve the accuracy of the prediction. The comparison of the results using four-machine learning methods random forest, decision tree, neural network, and logistic regression. the features such as residence place, military status, and gender were selected to determine whether these features are contributing factors to citizen education level.

Firstly, we checked the importance of the features of the dataset. After applied the logistic regression to select features the result was that the demographic features are the most important features. We used traditional classification methods to classify the citizen's education level and to determine the independent variables that offer the highest accuracy.

6. IMPLEMENTATION AND DISCUSSION

Generally, about 80% of the time spent in data analysis is cleaning and retrieving data, The data set consists of Egyptian citizens' data. This data is classified by different Education levels, having a certain category assigned to the review with a rating numerical value of EDU_Level= {Edu1 , Edu2 ,Edu3 … Edu13 }using 10 features and 10000000 rows.

First: the feature selection using featurewiz

Featurewiz is a new open-source python package for automatically creating and selecting important features in the dataset that will create the best model with higher performance. It also uses advanced feature engineering strategies to create new features before selecting the best set of features with a single line of code. It's suitable for all kinds of variables and targets. Featurewiz uses the SULOV algorithm and Recursive XGBoost to reduce features to select the best features for the model. [28] [29].

- **SULOV**

means Searching for an Uncorrelated List of Variables. The algorithm works in the following steps.

First step: find all the pairs of highly correlated variables exceeding a correlation threshold (say absolute (0.7)).

Second step: find their Mutual Information Score to the target variable. Mutual Information Score is a non-parametric scoring method.

Third step: take each pair of correlated variables, then knock off the one with the lower Mutual Information Score.

Final step: Collect the ones with the highest Information scores and least correlation with each other.

- **Recursive XGBoost**

After selecting the features with less correlation and high mutual information score, the Recursive XGBoost is used to find the best features among the remaining features. Here is how it works.

First step: Select all features in the dataset and split the dataset into train and valid sets.

Second step: Find top X features on the train using valid for early stopping (to prevent overfitting).

Third step: Take the next set of features and find top X.

Final step: Repeat this 5 times and finally combine all selected features and de-duplicate them.

The results extracted from the program for features selected using featurewiz shown in the following tables as:

Table (1). Loading random sample of 100000 records to panda frame

FAST FEATUREENGG AND SELECTION! Be judicious with featurewiz. Don't use it to create too many un-interpretable features!
Skipping feature engineering since no feature_engg input... **INFO: featurewiz can now read feather-formatted files. Loading train data...
Shape of your Data Set loaded: (1139670, 8) Loaded train data. Shape = (1139670, 8)
Setting a hard limit of 900K samples for train since some it is huge and breaks pandas... No test data filename given...
Classifying features using a random sample of 10000000 rows from dataset...
Single_Label Multi_Classification problem loading a random sample of 10000000 rows into pandas for EDA

Table (2) show the result of classifying variables by XGBoost into train & test

CLASSIFYING VARIABLES
Classifying variables in data set... 7 Predictors classified...
No variables were removed since no ID or low-information variables found in data set
No GPU active on this device
Tuning XGBoost using CPU hyper-parameters. This will take time...
After removing redundant variables from further processing, features left = 7 No interactions created for categorical vars since feature engg does not specify it
Single_Label Multi_Classification problem Starting feature engineering...Since no test data is given, splitting train into two... Source X_train shape: (720000, 7) | Source X_test shape: (180000, 7)

Table (3) show the results of preprocessing of data variables

preprocessing
Cleaned NaNs in numeric features test and train have similar NaN columns
Start preprocessing with 7 variables
Final Number of Features: 7
New X_train rows: 720000, X_test rows: 180000
New X_train columns: 7, X_test columns: 7
Completed feature engineering. Shape of Train (with target) = (900000, 8)

Table (4) shows the results for the last stage of feature selection by featurewiz model as selected five features from the main seven feature

Searching for Uncorrelated List Of Variables (SULOV) in 7 features
There are no null values in dataset...
Removing (2) highly correlated variables: ['EDU_LEAVE_REASON_CODE', 'CEN_YEAR']
Following (5) vars selected: ['EDU_JOIN_LEVEL_CODE', 'GENDER_CODE', 'UR_CODE', 'EDU_JOIN_CODE', 'GOV_CODE']

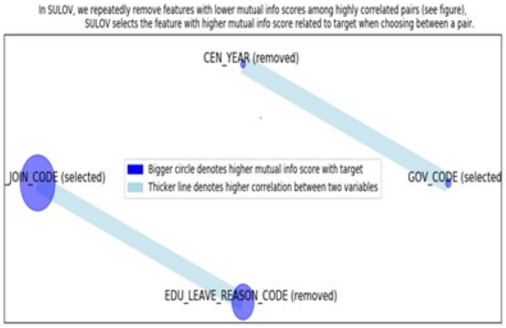
In SULOV, we repeatedly remove features with lower mutual info scores among highly correlated pairs (see figure),
SULOV selects the feature with higher mutual info score related to target when choosing between a pair.
CEN_YEAR (removed)
_JOIN_CODE (selected)
Bigger circle denotes higher mutual info score with target
Thicker line denotes higher correlation between two variables
GOV_CODE (selected
EDU_LEAVE_REASON_CODE (removed)

Fig .5. explore and visualize the result of the feature selection by featurewiz

6.1. Classification And Prediction

This stage was carried out as follows: - Data training and testing were performed by the selected classification method using 4-fold cross-validation. - Calculating the average classification accuracy for the test data. The infrastructure of the data-processing cluster consists of the master with 2 vCPU and 16 GB of memory and two workers with 2 vCPU, each of them having 13 GB of memory. The experiments were done using SQL 2016, Python v 3.5. The data is classified by different Education levels, having a certain category assigned to the review with a rating numerical value of EDU_Level= {Edu1 , Edu2 ,Edu3 … Edu7 } . The most common four classification ML models used in this research are Decision Tree Classifier model, Random Forest Classifier model, logistic regression model and ANN (artificial Neural Network) classifier. The models fitted first before calculate accuracy and predict the y (EDU_Level) values as the number of citizen in each level or education stage. The following figures explain the relation between the True values of y to the prediction values resulted by applied the four ML models.

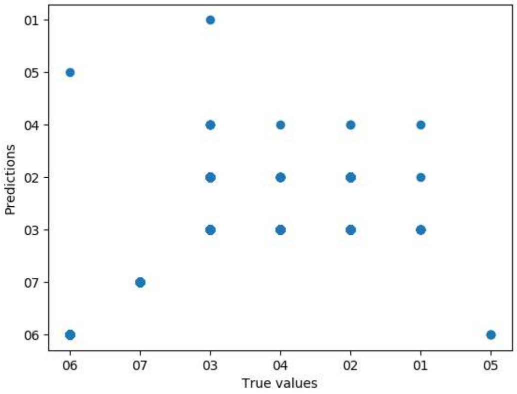

Fig.6. the relation between true values of (EDU_Level) and predictions Appling the Random fore

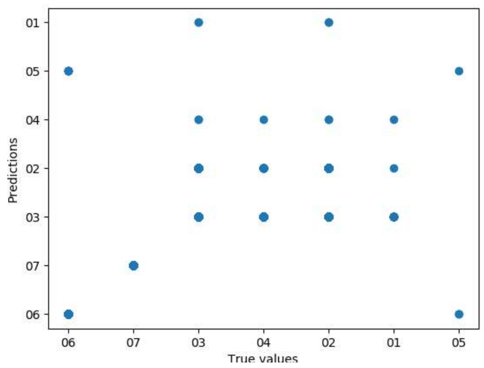

Fig.7. the relation between true values of (EDU_Level) and predictions Appling the Decision Tree model

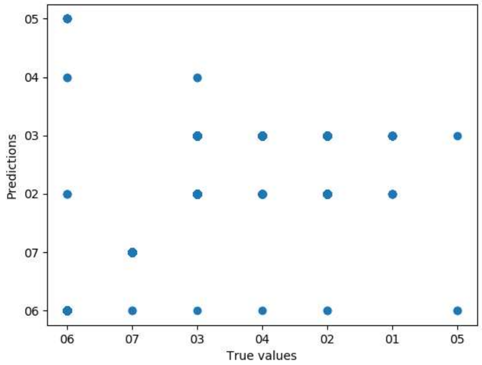

Fig.8. the relation between true values of (EDU_Level) and predictions Appling the Logistic Regression model

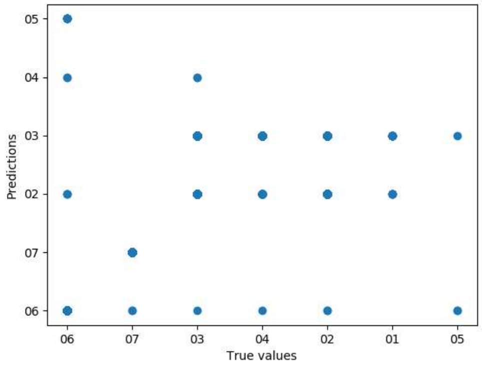

Fig.9. the relation between true values of (EDU_Level) and predictions Appling the ANN model

7. EVALUATION OF THE EXPERIMENT

The ML models namely Logistic Regression, Neural network, random forest and Decision Tree were used and the results can be seen in Table (5). It illustrates that the average values of predictive accuracy, Random Forest and decision tree are similar, and Logistic Regression has achieved 1 – 2% higher average classification accuracy results in comparison to random forest and Decision Tree, but the difference is not statistically significant. Except for Logistic Regression, the performance of analyzed classification methods contains more stability, and the values of the average classification accuracy are less distributed.

The accuracy of bagging Neural network outperformed the other methods with the accuracy of 0.769. In this study, the predictive Ml Logistic Regression had given the best result. We need to take more combination of features to increase the accuracy of the predictive percentage for Edu_level. We can also improve the result by including more features such as the disability in the module and in the module to classify the Edu_level in the future study.

Table (5) Comparison of predictive Accuracy between four ML models

ML model	Accuracy	LB(lower bound)	UB(upper bound)
Logistic Regression	0.895	0.682	0.955
Neural network	0.769	0.678	0.885
Random Forest	0.878	0.652	0.838
Decision Tree	0.879	0.669	0.895

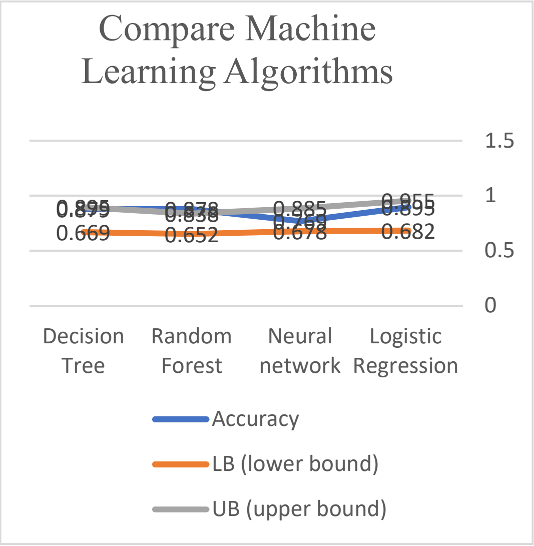

Fig.10. The Accuracy Comparison Between Applied ML Models

8. CONCLUSION AND FUTURE DIRECTION

A predictive model should be generated by trying a combination of various machine-learning algorithms, and the model should be validated to obtain optimum accuracy. In this research the comparison of neural network, Random Forest, Decision Tree and Logistic Regression methods for multi-class text classification is presented. The findings indicate that the Logistic Regression multi-class classification method has achieved the highest classification accuracy in comparison with neural network, Random Forest, Decision Tree Machines classification methods. On the contrary, Decision Tree has the lowest average accuracy values. The investigation indicates that increasing the training data set size per class leads to insignificant growth in classification accuracy. Based on the data analytics experiment, we found that random forest and logistic regression gave accurate results, then neural network and decision tree. In the future direction of this research, use advanced deep learning algorithms to build a predictive analytics model to predict the education level enhancement in Egypt.

REFERENCES

[1] B. Raja, "Predictive Analytics: A Study, Inclinations, Applications & Challenges," *IJETIE VOL. 5, ISSUE 12, DECEMBER ,* 2019.

[2] D. B. K. Kamesh , "A Study on Big Data and its Importance," *research gate,* p. 10, 2020.

[3] P. Pedamkar, "Machine learning life cycle," *www.educba.com,* 2022.

[4] a. bameda, "Machine learning usually refers to the changes in," *Johar Institute of Professional Studies, Lahore,* pp. 56-57, 2022.

[5] M. A. Bamiah, S. N. Brohi and B. B. Rad, "Big data technology in education: Advantages, implementations, and challenges," *Journal of Engineering Science and Technology,* 2018.

[6] R. Bhatnagar, "Machine Learning and Big Data Processing: A Technological Perspective and Review," *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018) (pp.468-478),* 2018.

[7] A. Drigas and P. Leliopoulos, "The Use of Big Data in Education," *IJCSI International Journal of Computer Science Issues,* 2016.

[8] C. Selvarani and P. Vani, "The Use Of Big Data In Education," *National Conference on Recent Advances in Commerce, Management and Computer Science (NRCACMC-2020),* 2020.

[9] T. Doleck, D. Lemay, R. Basnet and P. Bazelais, "Predictive analytics in education: a comparison of deep learning frameworks," *Education and Information Technologies ,* 2020.

[10] S. N. Brohi, . T. R. Pillai, . S. Kaur, H. Kaur and . S. Sukumaran, "Accuracy Comparison of Machine Learning Algorithms for Predictive Analytics in Higher Education," *International Conference for Emerging Technologies in Computing,* 2019.

[11] A. C. Lyons and J. Grable, "An Introduction to Big Data," *JOURNAL OF FINANCIAL SERVICE PROFESSIONALS ,* 2018.

[12] S. Choubey, R. Benton and T. Johnsten, "Towards Big data Governance in Cybersecurity," *Data-Enabled Discovery and Applications,* 2019.

[13] jitender, "Structured, Semi-Structured And Unstructured Data," 2020. [Online].

[14] . R. Allen, "What are the Types of Big Data?," 2022. [Online].

[15] Simplilearn, "What is Big Data Analytics and Why It is Important?," 2022. [Online].

[16] Y. WangaLee, A. Terry and A. Byrda, "Big data analytics: Understanding its capabilities and potential benefits," *Technological Forecasting and Social Change,* 2018.

[17] B. Veldkamp, K. Schildkamp and M. Keijsers, "Big Data Analytics in Education:: Big Challenges and Big Opportunities," *International Perspectives on School Settings, Education Policy and Digital Strategies,* 2021.

[18] N. Gruschka, V. Mavroeidis and K. Vishi, "Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR," *arXiv:1811.08531v1 [cs.CR] ,* 2018.

[19] R. Pathak, "What is Big Data Analytics? Definition, Advantages, and Types," *www.analyticssteps.com,* 2021.

[20] T. BUSH, "Predictive Analysis: Definition, Tools, and Examples," *PESTLE ANALYSIS,* 2020.

[21] A. Gupta and s. singh, "Decision Tree Introduction with example," *ide.geeksforgeeks.org,* 2022.

[22] I. H. Sarker, A. Colman and J. Han, "BehavDT: A Behavioral Decision Tree Learning to Build User-Centric Context-Aware Predictive Model," *Springer Science+Business Media, LLC, part of Springer Nature 2019,* 2019.

[23] A. . S. Rawat, "Top 8 Machine learning Models," *https://analyticssteps.com/,* 2021.

[24] M. Gh-Farahani, "Machine Learning models for prediction," *https://simulatoran.com/,* 2020.

[25] P. Balamuthukumar, "Detecting Attacks to Computer Networks Using a Multi- Layer Perceptron Artificial Neural Network," *IJIRT Journal,* 2020.

[26] N. Donges, "Random Forest Algorithm: A Complete Guide," *Expert Contributor,* 2022.

[27] A. PAVLOV, "Random Forest - A Machine Learning Technique," *www.c-sharpcorner.com/article,* 2020.

[28] D. David, "Automatic Feature Selection in Python: An Essential Guide," *https://hackernoon.com/,* 20 july 2021.

[29] N. A. Bayomy, L. A. Abd-Elmegid, A. E. Khedr and A. M. Idrees, "A Literature Review for Contributing Mining Approaches for Business Process Reengineering," *Future Computing and Informatics Journal,* vol. 5, no. 2, 2021.

[30] A. E. Khedr, A. M. Idrees and A. Elseddawy, "Adaptive Classification Method Based on Data Decomposition," *Journal of Computer Science,* vol. 12, no. 1, pp. 31-38, 2016.