Journal of Theoretical and Applied Information Technology
15th March 2022. Vol.100. No 5
2022 Little Lion Scientific

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

EXPLORATORY BIG DATA STATISTICAL ANALYSIS THE IMPACT OF PEOPLE LIFE'S CHARACTERISTICS ON THEIR EDUCATIONAL LEVEL

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

[1] Faculty of Science, Azhar University, Department of computer science, Cairo, Egypt
[2] Faculty of Computers and Information Technology, Future University in Egypt
[3] Faculty of Science, Azhar University, Department of computer science, Cairo, Egypt
[4] Faculty of Computers and Information Technology, Future University in Egypt
E-mail: [1] salwa.zaki@reefy.net , [2] neveen.ghali@fue.edu.eg
[3]afaf211@yahoo.com, [4] amira.mohamed@fue.edu.eg

ABSTRACT

Big data and cloud computing become the most important technologies in the most of fields over the world. That because their wide range of applicability in the main drivers of our life. Education is one of the most significant big data application area. Although there are many researches of analyzing big data in education sector as tools, performance and methods of teaching and how to enhancement them in all levels of education, there is absence of studies for applying big data analysis in census education data to study the impact of other people life's characteristics on their educational level. This research paper introduces exploratory big data analysis methods for categorical variables using python language to analyze the educational data in Egypt census (2017), and discuss the relations between educational data features as indicators of educational levels of Egyptians in the future. The implementation of exploratory big data analysis module (EBDA) displayed in this paper and the regression model used as a traditional statistical method to categorical data analysis Also, there is comparison between two results. This exploration of data analysis considered as the step for predictions, suggestions and recommendations of enhancement people educational level.

Keywords: *Big Data, Educational Data, Exploratory Analysis, Census Data, Python, Categorical Variables, Association Rules.*

1. INTRODUCTION

Big data refers to the large, diverse sets of information that grow at ever- increasing rates [1] [2]. It encompasses the volume of information, the velocity or speed at which it is created and collected, and the variety or scope of the data points being covered [3] [4]. Big data often comes from multiple sources and arrives in multiple formats. [5]. Data analytics technologies and techniques provide means to analyze data sets and provide conclusions about them, which help to make decisions. [6]. Big data analytics is a form of advanced analytics, which involves complex applications with elements as predictive models [7] [8], statistical algorithms and other high analytics systems that can bring new ways in decision-making, managing wide range of fields as health, finance and education [9]. The field of big data has been examined in different fields [10]. Specifically, in educational field, the use of big data analytics can construct a novel system for improving and discovering the prospect for valuable research, evaluation, and accountability [11]. This could be accomplished through technology integration like data analytics, web dashboards and data mining in education [12]. Therefore, this research makes exploratory for the Egypt census 2017 raw big data specified educational data using EBDA and analysis the relation between education level data and personal data using two analysis approaches, association rules applied as data mining analysis method and Ordinary Least Squares regression (OLS) as a traditional statistical method for describe the relationship between the gender and UR(Urban & Rural) as independent variables and Education level as a dependent variable, this showing the accuracy of numbers in the comparison of both analysis results . this step of analysis considers as a main stone of future implementation for machine learning prospective analysis.

Research scope explained in section 1, big data in education and related work discussed in section 2, the Proposed Model of (EBDA) for categorical variables elaborated in section3,the tools and methods listed in details in section 4, the experimental study displayed in section 5, the comparison between tow analysis approaches noted in section 6,and the last section makes concluding statements on the high and low points of the analysis module along with a statement on scope for future research in the same.

2. RESEARCH SCOPE

With The increase of disparity of educational levels geographically and qualitatively in Egypt and its impact on various fields such as industry, health care and teaching there has been a corresponding increase in its incorporation in the educational development [13]. In view of this situation, the present study analyzes the relation between life characteristics and education level of Egyptian citizen [14]. The study is restricted to big raw data of Egypt census 2017. The study in this research scope the educational data and the personalized data for each citizen in all sides of life as living place, insurance, work sectors, education levels and other characteristics. The study also involves an analysis of disparity education levels perspectives on the Gender and residential location (UR) by two ways, the first one is data mining association rules analysis, and the second approach is statistical sense of the data by processing them in a high analytical level to enable data-driven improvement of processes and procedures. [17]. It can find in analysis and discuss both results. Therefore, the scope of this study is limited to Egypt census, and more specifically to education field.

Industry	Use of Data Analytics with Big Data
Consulting\professional service	15%
Financial service	15%
Software\internet	10%
Healthcare	7%
Insurance	7%
Manufacturing(non-computers)	5%
telecommunications	5%
Government\federal	4%
Media\entertainment\publishing	4%
Advertising \ marketing	3%
Computer manufacturing	3%
Education	3%
Utilities	3%
Other	16%

3. BIG DATA IN EDUCATION AND RELATED WORK

The various topics of big data in education and previous related researches were discussed in this section. And explained the different sight of this paper.

3.1 Big Data in Education

Big Data refers to the process of combining enormous volumes of diversely outsourced data and analyzing them, using complex algorithms to inform decisions. [15]. Big data extensively used as a term today to describe and define the recent emergence and existence of data sets of high magnitude. In some cases, the data reach extremely big sizes such as in petabytes exceeding the hardware or human abilities to warehouse, manipulate and process them and therefore it characterized as big data. [16] .Data possessed in a system or a specific domain considered as big data when simultaneously the volume, the variety and the velocity are high irrespective of whether these three characteristics considered "small" to another domain. In this case, this is enough to challenge constrains in manipulating and analyzing the data so they can be used for different purposes. Depending on the domain, the size of data can vary from megabytes to petabytes. Thus, big data is context-specific and may refer to different sizes and types from domain to domain but the common challenge that all these domains must cope with is to being able to make
many sectors. The public, commercial, social and educational sectors receive and produce ceaselessly vast amounts of data from different sources and in different formats. As seen in the table [1] the use of Data Analytics with Big Data the education compared to other industries has only a 3%. In addition, other industries such as advertising, computer manufacturing and utilities have the same use and the rest of them are at 4% - 7%. Therefore, the use of Data Analytics with Big Data in education considered not bad compared to others.

Table.1 Use of Data Analytics with Big Data in Industry. [18] [19]

2.1. Related Work

Various researches have focused on different analysis approaches [20] [21]. Recent literature related to the field includes [22] .in his case study describes the successful implementation of a Big Data analysis tool: “SAP's HANA”, in the University of Kentucky. Also, [23]in her research paper used different model, The model can be employed for learning analytics to move from generalized support to meaningful contextualized support for enhancing learning.

On the other hand, the study [24] develop a teaching outcome model (TOM) that can be used to inspire and inspect quality of teaching. The simulated approach reported in the research accomplished through Splunk. Splunk is a Big Data platform designed to collect and analyze high volumes of machine-generated data and render results on a dashboard in real-time. In addition, [25], in their paper describes a framework for modeling user's behavior. The proposed system learns individual policies from the movement of the players in the game and builds a cognitive model. He states that this type of modeling will help in understanding learning processes of the user who interacts with the system and in adapting the learning environment to the user.

In view of the relevance of data analytics to the education and research sectors, several applications are considered useful. Some of the applications that have well established reputation in these fields include quality assessment systems for higher education, research management systems, student performance analyzers and business intelligence applications for the education sector. Some prototypes have been proposed related to these application domains. However, research is still in its infancy and no commercially viable solutions are known to exist, which leaves immense scope for future research and development. [26]

In the education field, big data is relatively niche topic and all the previous studies of big data in education research about how to optimizing education intelligence by facilitating institutions management, educators, and learners improved quality of education. [9]. The most important study [27]I used as based last stone to my paper aims to review current research related to big data analytics in education and explain future research direction. Using Kitchenham's technique, that was selected and clustered the literature into the types of data, methods, type of data analytics and learning analytics application used. The results show that research of big data learning analytics generally aims to improve the learning process, analyze learner behavior for student profiling, improve student retention and evaluate student feedback in the context of MOOCs and Learning Management System. Several future directions for this topic are: 1) building a big open dataset including data pre-processing and addressing the problem of imbalanced dataset, 2) process mining for learning log activity to gain knowledge and insights from online behaviour, not only from the perspective of the learner but also from the activities of the teacher, 3) designing an automated framework which uses big data and allows descriptive, predictive, prescriptive analytical learning to be carried out. To summarize, embracing big data to learning analytics and educational data mining is an open research area that seems very powerful in education. So, this research act as a future work for the previous mentioned study by combination with other study [28] that aims to investigate how to improve online learning effectiveness during this special time. Through a mixed design, the study revealed the effect of educational levels, gender, and personality traits on online learning outcomes but, our paper dealt the topic with a new sight by studying the impact of people life's characteristics on their educational level using the main social features as gender, residential location, and work status from big data of Egypt census 2017. That can use this relational analysis result to enhancement the means and utilities of education.

3. THE MODEL DESCRIPTION

Once data has been processed and cleaned, the next phase is usually the exploration phase. In this phase, data scientists /analysts can apply a variety of techniques to understand and analyze the data [29]. The process of exploration can help the analyst to determine which features are important, the relationship among features as well as their interdependencies. Exploratory data analysis also helps the analyst to answer questions raised before the exploration begins and can greatly help in reporting and business intelligence. [30] This research implements the exploratory data analysis module, as a step for building more efficient machine learning and AI models.

3.3. The Proposed Module

The proposed module used as a basic in this research to explore the selected sample of census data displayed as the following

A. Data set description

The data of education section from the collected raw data of Egypt census 2017, which include the personalized data for each citizen in all sides of life as living place, insurance, work sectors, education levels and other characteristics.

B. Data processed

Raw data is must be transformed into a usable form by include converting data into structured form (tabular). To process this structured data, cleaning and removal of outliers or useless features, filling or removal of missing values and standardization/normalization of the data values.

C. Clean data set

Clean data set is often a part of the data processed stage. Clean data comes after data processed and organized. Data may contain duplicates, redundant features, errors, or be incomplete. The most common task done in this stage include data deduplication, data normalization, feature segmentation, record matching.

D. Association rules

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is an analytical tool for analyzing data. It allows users to analyze data, categorize it, and summarize the relationships among data. Technically, data mining is the process of finding correlations or patterns in large relational databases. Association rule learning searches for relationships among variables.

3.4. Enhanced Module

The following steps illustrate exploratory data analysis (EBDA) stages in this research:

Step 1. Raw data selection

The Sample of raw data collected for Egypt census2017 (education Part) as Big Data.

Step2. Data processed

Convert raw data to structured data as create relational database.

Step3. Data clean

Filling null values, cleared the duplicate and normalized data, the data collected by application on tablet, filling form on website and mobile app so there was a lot of null and duplicated values to procced, I apply more than method to cleansing data as dummy variables and replaced nulls by Nan using python libraries.

Step4. Exploratory data analysis

Implement the exploratory data analysis module that is standard exploratory data analysis module and configure it to apply on Big Data and add a new step of implement two approaches of analysis to design a new modified model called (EBDA).there were more difficulty to apply analysis methods, both of traditional statistical method or machine learning module that was for the all data variables are categorical variables which was needed more steps to coding before applying specific analysis methods for categorical data.

4. TOOLS AND METHODS

Firstly, a data warehousing solution is required, which can be used for data storage. Besides this, the data concerned needs to be accessed, queried, retrieved, manipulated and analyzed. This requires a processing language that can be interfaced with the storage solution. In accordance with the requirements of the system, the technologies chosen for the implementation of this exploration are SQL, exploratory data analysis methods as step before predictive analysis machine –learning model and python language. The reasons for chosen these tools and techniques explained in the following:

A. Structured Query Language (SQL)

is a proven winner that has dominated for several decades. And is chosen in this study for the following reasons:

SQL enables increased interaction with data and allows a broad set of questions to be asked against a single database design.

SQL is standardized, allowing users to apply their knowledge across systems and providing support for third-party add-ons and tools.

SQL scales, and is versatile and proven, solving problems ranging from fast write-oriented transactions, to scan-intensive deep analytics.

SQL is orthogonal to data representation and storage. Some SQL systems support JSON and other structured object formats with better performance and more features than NoSQL implementations. [31]

B. Exploratory Data Analysis methods

Exploratory Data Analysis is majorly performed using the descriptive statistics methods. Descriptive statistics provide detailed summaries about observations or sample of data. These statistics could be quantitative, summary statistics like mean, mode, medians, percentiles, max and min etc. or visual, such as graphs and plots. Descriptive statistics can divide into Uni-variate, Bi-variate and multi-variate analysis. [32].

1. Uni-variate analysis

Uni-variate analysis is one of the simplest ways for describing data. The prefix "Uni" means "one", meaning the analysis deals with one feature at a time. This means that when performing Uni-variate analysis, we do not consider ordinary relationships among features but instead the main purpose is to describe a single feature. The most popular descriptive statistics found in uni-variate analysis include central tendency (mean, mode and median) and dispersion (range, variance, maximum, minimum, quartiles and standard deviations. Using graphs and charts, there are several types of uni-variate analysis we can perform, some of which are Bar Charts, Histograms, Frequency Polygons and Pie Charts.

2. Bi-variate analysis

Bi-variate analysis is one of the simplest forms of descriptive statistics. It means two features compared side by side for possible relationships [33]. The result of bi-variate analysis used to answer the question of whether a feature "X" depends on another feature "Y", whether there is a linear dependence among these features and whether one can help predict another. Some popular types of bi-variate analysis include scatter plots, regression analysis and correlation coefficients.

3. Multi-variate analysis

Multi-variate analysis is the analysis of three or more features and the relationship among them. It is more complex than both uni-variate and bi-variate analysis. This type of analysis is mostly performed using special tools and software (Pandas, SAS, SPSS etc.), as working with three or more data features manually is infeasible. Multi-variate analysis is mostly preferred when the data set under consideration is diverse, and each feature or relation among features is important [34] [35].

A. The Association rules

The association rules used within a dataset to discover nontrivial hidden patterns between items in a set could utilize either descriptive or predictive models [36] [37]. In many cases, the algorithms generate large number of association rules, often in thousands or millions [38]. It is almost impossible for users to visualize or validate such a large number of complex association rules, which limits the usefulness of data mining results. Therefore, it is important to identify that the components of an association rule are two sets of items: Left Hand Side (LHS) and Right-Hand Side (RHS).

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

The LHS is the antecedent (an item found in the transactions) and the RHS is the consequent (an item that is found in combination with the former) [39]. There are three measurements for an association rule, the Support rule, the Confidence rule and lift measurement rule.

Support:

The support *rule* refers to the default popularity of an item and can be calculated by finding number of transactions containing a particular item (x, y) divided by total number of transactions (N). [40]

$$\textit{Support} = \frac{\boldsymbol{frequency}(\boldsymbol{x},\boldsymbol{y})}{\boldsymbol{N}} \qquad \textit{(1)}$$

Confidence:

The Confidence rule is considered a localization measure of correlation between X and Y denoted by confidence (x → y) the confidence rule is calculated as the ratio between the support of the union between X and Y subsets and the support of X [41]. The ratio that greater than a user-specified confidence is said to have minimum confidence [42].

$$\textit{Confidence} = \frac{\textbf{support }(\mathbf{x}\cup\mathbf{y})}{\textbf{support}(\mathbf{x})} \qquad \textit{(2)}$$

Lift:

A frequently utilized measure of association rule utility.

Lift interestingness measure defines the number of transactions that contain the items used to find interesting patterns. The Lift measure is denoted by lift (x → y) [43].

$$\textit{Lift} = \frac{\boldsymbol{support}(\boldsymbol{x}\cup\boldsymbol{y})}{\boldsymbol{support}(\boldsymbol{x})\times\boldsymbol{support}(\boldsymbol{y})} \qquad \textit{(3)}$$

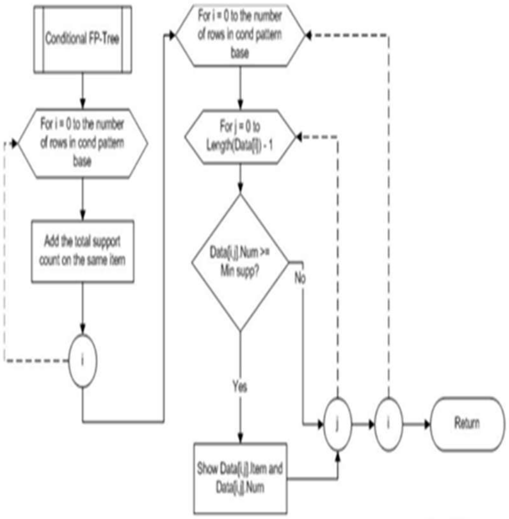
B. The Association rules implementation by FP –tree Algorithm

There are Different statistical algorithms have been developed to implement association rule mining. [44] Data mining is used to deal with the huge size of the data stored in the database to extract the desired information and knowledge. It has various techniques for the extraction of data; association rule mining is the most effective data mining technique among them. It discovers hidden or desired pattern from large amount of data. Among the existing techniques the frequent pattern growth (FP growth) algorithm is the most efficient algorithm in finding out the desired association rules. It scans the database only twice for the processing. [45]

Frequent Pattern Algorithm Steps

The frequent pattern growth method lets us find the frequent pattern without candidate generation.the following steps to mine the frequent pattern using frequent pattern growth algorithm:(Fptree algorithms flowchart shown in figure (1))

- The first step is to scan the database to find the occurrences of the itemset in the database. This step is the same as the first step of Apriori. The count of 1-itemsets in the database is called support count or frequency of 1-itemset.
- The second step is to construct the FP tree. For this, create the root of the tree. The root is represented by null.


Conditional FP-Tree
For i = 0 to the number of rows in cond pattern base
Add the total support count on the same item
i
For i = 0 to the number of rows in cond pattern base
For j = 0 to Length(Data[i]) - 1
Data[i,j].Num >= Min supp?
No
Yes
Show Data[i,j].Item and Data[i,j].Num
j
i
Return

- The next step is to scan the database again and examine the transactions. Examine the first transaction and find out the itemset in it. The itemset with the max count is taken at the top, the next itemset with lower count and so on. It means that the branch of the tree is constructed with transaction itemsets in descending order of count.

- The next transaction in the database is examined. The itemsets are ordered in descending order of count. If any itemset of this transaction is already present in another branch (for example in the 1st transaction), then this transaction branch would share a common prefix to the root.
- This means that the common itemset is linked to the new node of another itemset in this transaction.
- Also, the count of the itemset is incremented as it occurs in the transactions. Both the common node and new node count is increased by 1 as they are created and linked according to transactions.
- The next step is to mine the created FP Tree. For this, the lowest node is examined first along with the links of the lowest nodes. The lowest node represents the frequency pattern length 1. From this, traverse the path in the FP Tree. This path or paths are called a conditional pattern base. Conditional pattern base is a sub-database consisting of prefix paths in the FP tree occurring with the lowest node (suffix).
- Construct a Conditional FP Tree, which is formed by a count of itemsets in the path. The itemsets meeting the threshold support are considered in the Conditional FP Tree.
- Frequent Patterns are generated from the Conditional FP Tree. [46]

Fig (1) the FP-tree algorithms flowchart

C. Analysis in python

Python is an increasingly popular tool for data analysis. In recent years, a number of libraries have reached maturity, allowing R and Stata users to take advantage of the flexibility, and performance of Python without sacrificing the functionality these older programs have accumulated over the years. [47]

The high level, dynamic and interactive nature of this language combined with the abundance of scientific libraries make it a preferred choice for analytical tasks. Python language has been on the increase since the early 2000s, in both industry applications and research [48].

The python ecosystem of libraries some of which used in this research explained as the following:

NumPy (Numerical Python) is a base data structure and fundamental package in Python and as such, numerous libraries are also built on top of it.

Pandas is another popular package built on top of NumPy that is used for data manipulation and analysis [49].Some of the features available in Pandas are Data Frame objects for data manipulation, tools for reading and writing data between memory, hierarchical axis indexing functions to work with high dimensional data, time series functionality, data filtration and data alignment for handling missing data.

Seaborn is built on top of Python’s core visualization library Matplotlib. However, Seaborn comes with some very important features as the following:

● built in themes for styling Matplotlib graphics.

● Visualizing univariate and bivariate data.

● Fitting in and visualizing linear regression m models.

● Plotting statistical time series data.

● Seaborn works well with NumPy and Pandas data structures. It comes with built in themes for styling Matplotlib graphics**.** [50]

Matplotlib is a multi-platform data visualization library built on NumPy arrays, and designed to work with the broader SciPy stack. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc. [51].

5. THE EXPERIMENTAL STUDY

Exploratory data analysis module (EBDA) implemented using the Python programming language, form, and finally, the visualization module holds numerous functions that aids easy visualization of (EBDA) builds by using many libraries like Panda's data frame for data structure, NumPy for data analysis, Seaborn and Matplotlib for data visualization. Figure (2) display the dependency modules of building EBDM.and its design based on three modules (feature engineering, struct data, and visualization). The feature-engineering module contains functions to handle tasks such as cleaning, filling missing, aggregating, counting features in a dataset, struct data module handles all tasks relating to structured data in tabular

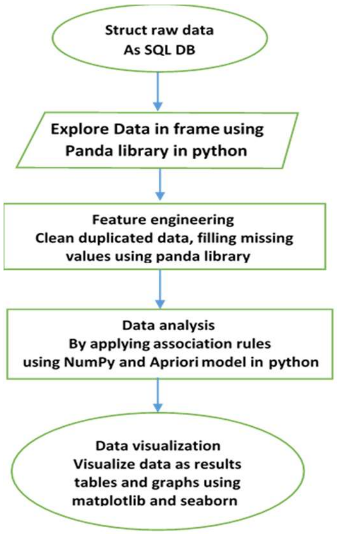

Struct raw data
As SQL DB
Explore Data in frame using
Panda library in python
Feature engineering
Clean duplicated data, filling missing
values using panda library
Data analysis
By applying association rules
using NumPy and Apriori model in python
Data visualization
Visualize data as results
tables and graphs using
matplotlib and seaborn

Fig (2) EBDA component tools diagram

Table (2) Panda's Data Frame (Data Before Cleansing)

EDU_CODE	EDU_JOIN_CODE	EDU_JOIN_LEVEL_CODE	GENDER_CODE	UR_CODE
2	5	NaN	1	2
2	4	2	1	2
1	3	1	2	2
2	5	NaN	1	2
1	3	2	2	2
9	2	NaN	2	1
6	3	3	1	2
9	2	NaN	1	1
2	5	NaN	2	2
11	2	NaN	2	1

Case study is a sample of education part in census Egypt 2017 raw data, the sample is about 10 million records and 30 fields that manipulate as big data. The census data in education sector include the educational level of each citizen beside other Personal Ingredients as the address, gender, living location (Urban or Reef), marriage status, age, work status.

The census data collected from various resources as social media, filled applications on tablets, phone application and paper application by more than one hundred million Egyptian citizen. Data includes all information for every citizen, but this paper focal of education department.

Work started with transform this raw data to structured data by using SQL database and import in python as data set using panda's data frame. As shown in table (2)

Table (3) handling Null values

EDU_CODE	EDU_JOIN_CODE	EDU_JOIN_LEVEL_CODE	GENDER_CODE	UR_CODE
2	5	Null	1	2
2	4	2	1	2
1	3	1	2	2
2	5	Null	1	2
1	3	2	2	2
9	2	Null	2	1
6	3	3	1	2
9	2	Null	1	1
2	5	Null	2	2
11	2	Null	2	1

Second step is data filtering and replace missing values with NAN's values using Panda library's tools. As shown in table (3)

Final step is exploratory data analysis with three methods and using statistical and descriptive statistical libraries in python as Sklearn, Matplotlib, and Seaborn.as explained in the following section

5.1. Apply Univariate Analysis and Visualization Method

To explore variables one by one. For categorical variables, frequency table used to understand the distribution of each category. It is also used to highlight outlier values. The percentage of values can be calculated and represented under each category. It can be measured using two metrics, Count and Count Percentage against each category. A bar chart used as visualization. Create frequency tables (also known as crosstabs) in pandas using the pd. crosstab () function. The function takes one or more array-like objects as indexes or columns and then constructs a new Data Frame of variable counts based on the supplied arrays. The table below is one-way table of the EDU_CODE variable (education level coding from 1 to 14 as shown in table (4)), which get a sense of the distribution of records across the categories.

Table (4) One-Way Table Of The EDU_CODE

EDU_CODE	0	1	2	3	4	5	6	7	8
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	1	0	0	0
12	0	0	0	0	1	0	0	1	1
13	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0

The most useful aspects of frequency tables are that they allow extracting the proportion of the data that belongs to each category. The proportion of each education level display in the table (5).

Table (5) The Proportion Of Education Levels

EDU_CODE	Edu Level count%
1	0.64351
2	0.677238
3	0.612545
4	0.665534
5	0.497359
6	0.622168
7	0.596722
8	0.5613
9	0.594026
10	0.484619
11	0.473211
12	0.404597
13	0.416168
14	0.416501

5.2. Apply Bivariate Analysis and Visualization Method

To finds out the relationship between two variables and examines their correlation. By apply the Apply Bivariate Analysis and visualization, table (6) and graph (3) show the relation between EDU_CODE (Education level) and the GENDER_CODE (male 1, female 2).

Table (6) The Relation Between Education Levels And Gender

EDU_CODE	GENDER_CODE	
	1	2
1	144269	135499
2	93865	127122
3	45489	38206
4	2861	1848
5	1382	1079
6	44840	39600
7	39675	35351
8	31505	27577
9	106305	82020
10	14842	11520
11	55781	43977
12	1299	1007
13	811	528
14	682	324

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

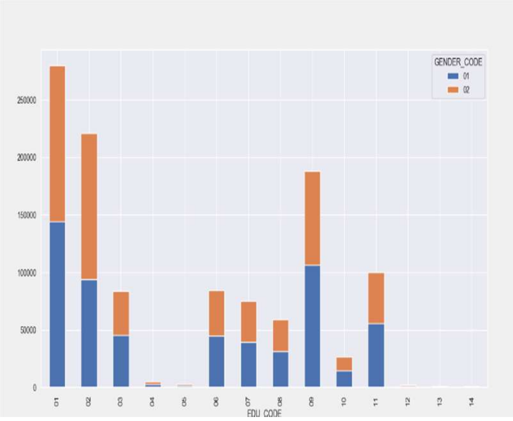


Fig (3) The Relation Between Education Levels And Gender

The Table (7) And Graph (4) Show The Relation Between EDU_CODE And UR.

EDU_CODE	UR_CODE	
	1	2
1	105657	174376
2	93865	67185
3	34528	49172
4	1786	2923
5	1737	724
6	33203	51242
7	32544	42483
8	32349	26737
9	81155	107182
10	17362	9001
11	68869	30892
12	1841	465
13	1064	272
14	822	184

Table (7) The Relation Between Education Levels And UR (Urban And Rural)

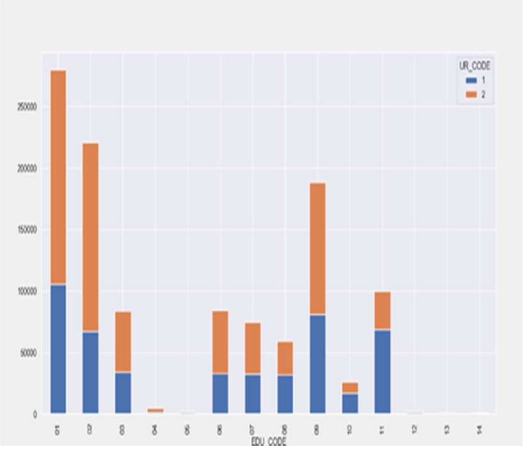


Fig (4) The Relation Between Education Levels And Ur (Urban And Rural)

5.3. Apply Multivariate Analysis and visualization method

To understand interactions between different fields in the dataset (or) finding interactions between variables more than two. The table (8) and graph (5) show the relation between EDU_CODE, UR and GENDER.

Table (8) The Relation Between EDU_CODE, UR And GENDER

UR_CODE	Urban		Rural	
GENDER_CODE	male	female	male	female
EDU_CODE				
1	54333	51226	89936	84273
2	29451	37731	64414	89391
3	18449	16076	27040	22130
4	1012	774	1849	1074
5	957	780	425	299
6	17120	16083	27720	23517
7	16862	15682	22813	19669
8	16481	15868	15024	11709
9	43564	37586	62741	44434
10	9480	7881	5362	3639
11	37578	31290	18203	12687
12	1015	826	284	181
13	628	435	183	89
14	549	273	133	51

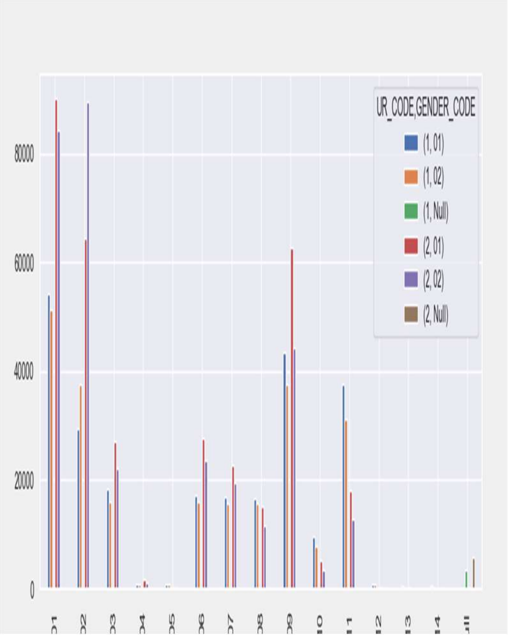


Fig (5) The Relation Between Education Levels, Gender And UR

5.4. Using FP_Tree To Implement Association Rules By Python

First: apply FP-tree to find the association between education levels (EDU_CODE) and gender (GENDER_CODE)

As seen in table (9), the support value for the first rule is 1.002. This number is calculated by dividing the number of transactions containing gender1 (male) divided by total number of transactions. The confidence level for the rule is 0.00069 which shows that out of all the transactions that contain gender1 (male), 0.07% of the transactions also contain EDU_Code01 (The literacy). Finally, the lift of 1.005 tells us that male is 1.005 times more likely to enroll the literacy compared to female.

Table (9) The Association Between Education Levels And Gender

antecedents	consequents	Support	Confidence	Lift
(EDU_CODE_01)	(UR_CODE_2)	0.153	0.622	1.1
(UR_CODE_2)	(EDU_CODE_01)	0.153	0.266	1.1
(EDU_CODE_02)	(UR_CODE_2)	0.134	0.696	1.2
(UR_CODE_2)	(EDU_CODE_02)	0.134	0.235	1.2

Second: apply FP-tree to find the association between education levels (EDU_CODE) And Urban & Rural (UR_CODE). As seen in table (10), the support value for the first rule is 0.153006. This number is calculated by dividing the number of transactions containing edu_code1 (The literacy) divided by total number of transactions. The confidence level for the rule is 0.622 which shows that out of all the transactions that contain ur_code2 (rurally), 60% of the transactions also contain EDU_Code01 (The literacy). Finally, the lift of 1.082377 tells us that the rurally Inhabitants is 1.08 times more likely to enroll the literacy compared to the urban Inhabitants.

Table (10) The Association Between Education Levels And UR

antecedents	Consequents	Support	Confidence	Lift
GENDER_CODE_01	EDU_CODE_01	1.02	0.0069	1.2
EDU_CODE_01	GENDER_CODE_01	1.05	0.0007	1.0
EDU_CODE_01	GENDER_CODE_02	1.09	0.0011	1.1
GENDER_CODE_02	EDU_CODE_01	1.03	0.0011	1.1
EDU_CODE_02	GENDER_CODE_01	0.85	0.017	0.8

A common statistical approach used by researchers to examine relationships between variables is the regression framework. Regression modeling allows researchers to examine the specific effects variables have on one another, while simultaneously controlling for the effects that other variables may also have. Although many types of regression frameworks exist, the most frequently used in categorical variables analysis are logistic regression techniques (e.g., binary logistic regression and ordinal logistic regression) and Ordinary Least Squares (OLS) regression, which was used in this paper.

5.5. Implement OLS Statistical Analysis Model in Python

The OLS model had implemented to find the relation between education level (EDU_CODE) as dependent variable and both gender (GENDER_CODE), Urban /Rural (UR_CODE) as independent variables. Unstandardized regression coefficients in an OLS linear regression represent the slope of the line between the independent variables and the dependent variable, while holding constant the effects of the other variables in the model. Thus, coefficients interpret as a one unit increase in the number of Gender (independent variable) corresponds to and expected increase/decrease in education level (dependent variable), independent of the effects of the remaining variables in the model.

Unstandardized regression coefficients should be used when examining results across studies.

The OLS model had implemented to find the relation between education level (EDU_CODE) as dependent variable and both gender (GENDER_CODE), Urban/Rural (UR_CODE) as independent variables. As seen in the following table (11), table (12) (OLS results tables), the coef value for the male (GENDER_CODE_01) is 0.0049 and the coef value of the female (GENDER_CODE_02) is 0.0027 (The coefficient of variation (CV) is a relative measure of variability that indicates the size of a standard deviation in relation to its mean. It is a standardized, unit less measure that allows you to compare variability between disparate groups and characteristics) which shows that male is 1.8 times more likely to enroll the literacy compared to female. The coef value for the urban (UR_CODE_1) is 0.0477 and the coef value of the rural (UR_CODE_2) is 0.0286 which shows that the rurally Inhabitants is 1.6 times more likely to enroll the literacy compared to the urban Inhabitants.

Table (11) The OLS Regression Results Of An Education Level (Dep Variable) And Male (GENDER_CODE_01) And Urban (UR_CODE_1) (Independent Variables)

Dep.Variable	EDU_CODE_01
Model	OLS
Method	Least Squares
No. Observations	1139670
inDep.Variables	Coef.
GENDER_CODE_02	0.0027
UR_CODE_02	0.0286

Table (12) The OLS Regression Results Of An Education Level (Dep Variable) And Female (GENDE_CODE_02) And Rural (UR_CODE_2) (Independent Variables)

Dep.Variable	EDU_CODE_01
Model	OLS
Method	Least Squares
No. Observations	1139670
inDep.Variables	Coef.
GENDER_CODE_01	0.0049
UR_CODE_1	0.0477

6. COMPARISON OF MODULES (FP-TREE MINING MODEL AND OLS STATISTICAL MODEL)

The comparison will be between FP-tree associations Module against OLS statistical Module. The comparison will attempt to compare the performance of the algorithms in terms of:

- Execution time.
- Results generated.

The number of items sets generated was included as an experimental result to determine whether there is a relationship between the number of item sets generated and execution time.

a. Execution time

In terms of execution time [52], the FP-tree model depends on the support value. The increase is at a rate of approximately 5% between 0.1% and 0.05% support, and 10% between 0.05% and 0.01% support, shows the execution time for the algorithm tends to increase as the minimum required support is lowered. While the execution time of OLS remains relatively constant and independent that is there is no specific value to implement model and get results. This recommends that at very low support levels, the FTP-tree algorithm would in fact be slower than the OLS algorithm. This reinforces the initial proposal that OLS performs well in extremely low support situations.

b. Results generated

As show in the above results tables of the both models there is big relativity between two models results with the differently of results' factors as in FTP-tree depends on support and confidence values as data mining analysis while OLS model depends on coef value as statistical analysis.

7. CONCLUSION

In conclusion, Education is important to everyone as it gives shape to people's life; it affects how we act, think, responds and gives path to life. Education is the process by which society deliberately transmits its cultural heritage and its accumulated knowledge, values and skills to each generation. In This paper we developed and implemented EBDA module to explore some effects of people life's characteristics on their education level- using census EGYPT big data- and analysis relations between them with two approaches that give big tight results which could be directors for predictions, suggestions and recommendations of enhancement people educational level. This concluding section consists of major findings, limitations of this study, and future research directions.

Major findings This study revealed the relation between the gender and geographical residence as a major sample of life characteristics and educational levels of Egypt citizen. This study could provide a meaningful reference for predictive analysis of the impact of the people life characteristic on their education levels. **Limitations** There are two limitations to this study. The first one, the data were limited to EGYPT only (EGYPT census Data). On the other hand, this study was implemented on education section of census data.

Future research directions. This research can lead to further analysis for more various life characteristics and predictive changes in education levels of people according to improvement of the surrounding coefficients.

REFERENCES

[1] M. Tamer, A. E. Khedr and S. Kholief, "A Proposed Framework for Reducing Electricity Consumption in Smart Homes using Big Data Analytics," *Journal of Computer Science,* vol. 15, no. 4, 2019.

[2] M. Othman, H. Hassan, R. Moawad and A. M. Idrees, "A Linguistic Approach for Opinionated Documents Summary," *Future Computing and Informatics Journal,* vol. 3, no. 2, pp. 152-158, 2018.

[3] A. M. Idrees, M. H. Ibrahim and A. I. El Seddawy, "Applying spatial intelligence for decision support systems," *Future Computing and Informatics Journal,* vol. 3, p. 384e390, 2018.

[4] A. M. Idrees, M. L. A. Khaled and A. H. A. Talkhan, "Spatial Data Mining, Spatial Data Warehousing, and Spatial OLAP," in *Emerging Trends in Open Source Geographic Information Systems*, IGI Global Publication, 2018, pp. 97-132.

[5] A. Oussous, "Big Data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences,* pp. Pages 431-448., 2018.

[6] S. Jeble and S. Kumari, "Role of Big Data in Decision Making," *Symbiosis International University, Pune, India,* p. Pages 36 – 44, 2018.

[7] A. M. Idrees, F. Gamal Eldin, A. M. Mohsen and H. Hassan, "Tasks, Approaches, and Avenues of Opinion Mining, Sentiment Analysis, and Emotion Analysis: Opinion Mining and Extents," in *E-Collaboration Technologies and Strategies for Competitive Advantage Amid Challenging Times*, IGI Global Publication, 2021, pp. 171-209.

[8] A. M. Idrees and E. Shaaban, "Reforming home energy consumption behavior based on mining techniques a collaborative home appliances approach," *Kuwait Journal of Science,* vol. 47, no. 4, 2020.

[9] S. Uthayasankar and M. ,. M. Kamal, "Critical analysis of Big Data challenges and analytical methods," *Journal of Business Research,* pp. Pages 263-286, 2017.

[10] A. Al Mazroi, A. E. Khedr and A. M. Idrees, "A Proposed Customer Relationship Framework based on Information Retrieval," *Expert Systems With Applications,* vol. 176, 2021.

[11] H. E. Elmasry, A. E. Khedr and M. M. Nasr, "An adaptive technique for cost reduction in cloud data centre environment," *International journal of Grid and Utility Computing,* vol. 10, no. 5, pp. 448-464, 2019.

[12] A. B. El Seddawy, T. Sultan and A. E. Khedr, "A Proposed Data Mining Technique to Improve Decision Support System in an Uncertain Situation," *International Journal of Engineering Research and Development,* vol. 8, no. 7, pp. 56-61, 2013.

[13] A. E. Khedr, A. M. Idrees and F. K. Alsheref, "A Proposed Framework to Explore Semantic Relations for Learning Process Management," *International Journal of e-Collaboration,* vol. 15, no. 4, 2019.

[14] A. E. Khedr, A. M. Idrees and A. I. El Seddawy, "Enhancing Iterative Dichotomiser 3 algorithm for classification decision tree," *WIREs Data Mining Knowledge Discovery,* vol. 6, p. 70–79, 2016.

[15] I. Yaqoob, . I. Abaker and T. Hashem, "Big Data: From Beginning to Future,," *International Journal of Information Management,,* pp. 1-10, 2016.

[16] C. Vaitsis, V. Hervatis and Z. Nabil, "Introduction to Big Data in Education and Its Contribution to the Quality Improvement Processes," in *Introduction to Big Data in Education and Its Contribution to the Quality Improvement Processes*, 2016, pp. 1-20..

[17] A. L'Heureux and K. Grolinger , "Machine Learning With Big Data: Challenges and Approaches," *IEEE Access,* pp. 1-23, 2017.

[18] Athanasios S. Drigas & Panagiotis LeliopoulosIJCSI , "The Use of Big Data in Education,," *International Journal of Computer Science Issues,* p. Page 61, 2014.

[19] Guires, "WHAT IS BIG DATA ANALYTICS, AND WHY IS IT IMPORTANT TO BUSINESS?," *Analytics, Big Data Analytics,* 2019.

[20] F. Abogabal, S. M. Ouf, A. M. Idrees and A. E. Khedr, "AN ARCHITECTURAL FRAMEWORK FOR GENERATING FOOD SAFETY KEY PERFORMANCE INDICATORS," *JOURNAL OF SOUTHWEST JIAOTONG UNIVERSITY,* vol. 55, no. 5, 2020.

[21] A. M. Idrees, M. H. Ibrahim and N. Hegazy, "A Proposed Model for Predicting Stock Market Behavior Based on Detecting Fake News," in *Science and Mathematics International Conference (SMIC) 2018*, 2018.

[22] V. Kellen, A. Recktenwald and S. Burr, "“Applying Big Data in Higher Education: A Case Study”," *Data Insight and Social BI Executive Report,* pp. 3-8, 2013..

[23] A. Shibani, " "contextualizable Learning Analytics Design: A Generic Model, and Writing Analytics Evaluations".," *University of Technology Sydney Sydney NSW Australia,* pp. 1-10, 2019.

[24] N. I. Glory, "Data Science Approach for Simulating Educational Data: Towards the Development of Teaching Outcome Model (TOM),," *Postgraduate Research,Big Data and Cognitive Computing 2(3):24,,* pp. 1-18, 2018.

[25] S. J. Lee, Y.-E. Liu and Z. Popovic, "Learning Individual Behavior in an Educational Game: A Data-Driven Approach, Certer for Game Science,," *Computer Science and Engineering University of Washington,* pp. 1-8, 2014.

[26] J. Vanthienen and K. De Witte, "Data Analytics Applications in Education,," *ISBN: 9781498769273,* pp. 1-7, 2017.

[27] A. Yunita, H. . B. Santoso and . Z. A. Hasibuan3, "Research Review on Big Data Usage for Learning Analytics," *Journal of Physics: Conference Series,* pp. 30-35, 2021.

[28] Z. Yu, "The efects of gender, educational level, and personality on online learning outcomes during the COVID-19 pandemic," *International J Educ Technol High Educ,* pp. 2,36-37, 2021.

[29] D. Cielen, A. D. B. Meysman and M. Ali, in *Introducing Data Science Big data, machine learning, and more, using Python tools*, 2016, p. chapter 2.

[30] R. Odegua, "DataSist: A Python-based library for easy data analysis, visualization and modeling, Department of Computer Science," *Department of Computer Science. Ambrose Alli University, Ekpoma. Nigeria,* pp. 1-16, 2019.

[31] G. . A. Schreiner , OrcID, D. Duarte and Ronald, "When Relational-Based Applications Go to NoSQL Databases: A Survey," *Federal University of Santa Catarina,* pp. 1-4, 2019.

[32] Shamami and A. Khademi, "Applied Univariate, Bivariate, and Multivariate Statistics," *Journal of Statistical Software 72(Book Review 2),* pp. 1-4, 2016.

[33] J. I. Hoffman, "Biostatistics for Medical and Biomedical Practitioners," *Exploratory Data Analysis, book chapter 27,,* pp. 451-500, 2015.

[34] S. Hossain and Collinearity, "A review of methods to deal with it and a simulation study evaluating their performance," 2019.

[35] 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.

[36] A. E. Khedr, A. M. Idrees and R. Salem, "Enhancing the e-learning system based on a novel tasks' classification load-balancing algorithm," *PeerJ Computer Science,* vol. 7:e669, 2021.

[37] Danh Bui-Thi1, Pieter Meysman1 and Kris Laukens1, "Clustering association rules to build beliefs and discover unexpected patterns," *Springer Science+Business Media, LLC, part of Springer Nature ,* pp. 1-12, 2020.

[38] A. E. Khedr and A. M. Idrees, "Enhanced e-Learning System for e-Courses Based on Cloud Computing," *Journal of Computers,* vol. 12, no. 1, 2017.

[39] H. Jabeen, "Market Basket Analysis using R," *datacamp Tutorials,,* 2018.

[40] U. Malik, ", Association Rule Mining via Apriori Algorithm in Python,," 2018.

[41] Nour E. Oweis, Vaclav Snaselm, "A Novel Mapreduce Lift Association Rule Mining Algorithm (MRLAR) for Big Data," *(IJACSA) International Journal of Advanced Computer Science and Applications,* pp. 1-7, 2016.

[42] A. M. Mostafa, Y. M. Helmy, A. E. Khedr and A. M. Idrees, "A PROPOSED ARCHITECTURAL FRAMEWORK FOR GENERATING PERSONALIZED USERS' QUERY RESPONSE," *JOURNAL OF SOUTHWEST JIAOTONG UNIVERSITY,* vol. 55, no. 5, 2020.

[43] Rajesh Natarajan and B Shekar, "Interestingness of association rules in data mining: Issues relevant to e-commerce," *Indian Institute of Management Lucknow (IIML),* pp. 1-19, 2015.

[44] D. Ai, H. Pan and X. Li, "Association rule mining algorithms on high-dimensional datasets," *Artificial Life and Robotics (2018),* pp. 1-8, 2018.

[45] M. Narvekara, F. Syed and Shafaque, "An optimized algorithm for association rule mining using FP tree," in *International Conference on Advanced Computing Technologies and Applications (ICACTA2015)2015*, 2015.

[46] https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/, "Frequent Pattern (FP) Growth Algorithm In Data Mining," *Software Testing Help,* pp. 3-4, 2021.

[47] W. McKinney, " Python for Data Analysis," *O'Reilly Media,* 2012.

[48] S. Raschka, J. Patterson and Corey, "Machine Learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *ResearchGate,* 2020.

[49] R. Odegua, "DataSist: A Python-based library for easy data analysis, visualization and modeling," *ResearchGate,,* pp. 1-16, November 2019.

[50] t. Seaborn, "Tutorials Point," 2017. [Online].

[51] G. El-Faloujy, "Data Visualization," *intrduction_to_python_for_data_science,* pp. 1-48, 2019.

[52] 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.

[53] M. A. Abdel-Fattah, A. E. Khedr and Y. Nagm Aldeen, "An Evaluation Framework for Business Process Modeling Techniques," *International Journal of Computer Science and Information Security (IJCSIS),* vol. 15, no. 5, pp. 382-392, 2017.