Journal of Theoretical and Applied Information Technology
31st January 2020. Vol.98. No 02

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

A SMART SOCIAL INSURANCE BIG DATA ANALYTICS FRAMEWORK BASED ON MACHINE LEARNING ALGORITHMS

YOUSSEF SENOUSY[1], ABDULAZIZ SHEHAB[1, 2], ALAA M. RIAD[1], NASHAAT ELKHAMISY[3]
[1]Department of Computers and Information Systems, Mansoura University, Mansoura, Egypt
[2]Department of Computer Science, College of Science and Arts, Jouf University, KSA
[3]Department of Information Systems, Sadat Academy for Management Sciences, Cairo, Egypt
E-Mail: youssef_senousy@hotmail.com, abdulaziz_shehab@mans.edu.eg , amriad2014@gmail.com, wessasalsol@gmail.com

ABSTRACT

Social insurance is an individual's protection against risks such as retirement, death or disability. Big data mining and analytics in a way that could help the insurers and the actuaries to get the optimal decision for the insured individuals. Dependently, this paper proposes a novel analytic framework for Egyptian Social insurance big data. NOSI's data contains data which needs some pre-processing methods after extraction like replacing missing values, standardization and outlier/extreme data. The paper also presents using some mining methods such as clustering and classification algorithms on the Egyptian social insurance dataset through an experiment. In clustering, we used K-means clustering and the result showed a silhouette score 0.138 with two clusters in the dataset features. In classification, we used the Support Vector Machine (SVM) classifier and classification results showed a high accuracy percentage of 94%.

Keywords: *Social Insurance, Data Integration, Big Data Mining and Big Data Analytics*

1. INTRODUCTION

Social insurance is one of the branches of the insurance sciences, its programs provide protection against wage loss resulting from retirement, prolonged disability, death, or unemployment, and protection against the cost of medical care during old age and disability [1] (USCB, 2004). The Social Insurance Authority in Egypt seeks to provide insured individuals, pensioners and their dependents with social protection as a replacement for revenue that is disrupted if one of the insured risks occurs to them. The following Figure 1 illustrates the life cycle of the insured individual. The timeline of working periods is consisting of durations (job start date and job end date) and every duration has its salary value and salary date. The Age of working individuals ranges from 18 to 60 for employees and 18 to 65 for employer owner. Pensions is the end of the individual life cycle. The good thing about the pension system is the income of the pensions is not given only to the pension owner but also his/her family like wife, sons, and parents which we call them pension beneficiaries if the pension owner dies.

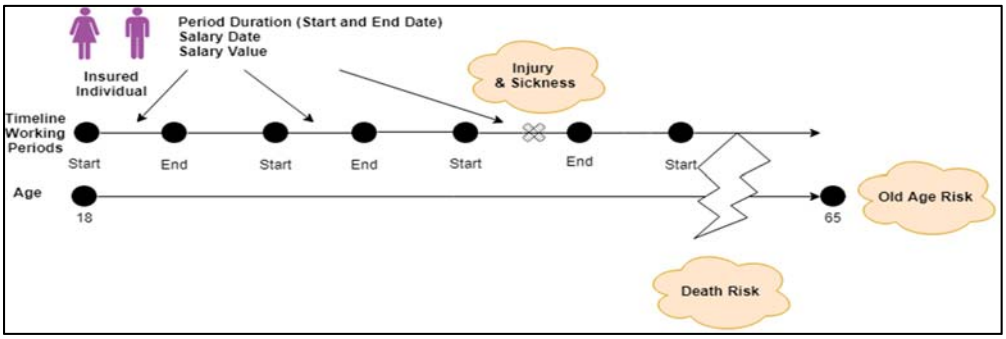

Figure 1: Life Cycle of Insured Individual

Journal of Theoretical and Applied Information Technology
31st January 2020. Vol.98. No 02

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

Big Data is a term used to describe a collection of data that is massive in size which grows exponentially over time like information from social insurance. The characteristics of Big Data usually include five V's: Volume, Velocity, Variety, Veracity, and Value [2].

Volume: Many data sets are too large to store or analyze using traditional database technologies and are being added or updated continuously as well. The National Organization for Social Insurance (NOSI) in Egypt has big data volumes which contains the full data of insured or non-insured people that are registered in the social insurance scheme. We determined all the data from their current datasets and it counted about 2,946,388,795 records. We found the size of approximately about 3.7 Terabytes.

Velocity: While data volumes are increasing, data creation and usage speeds are also increasing. NOSI's old systems are not effective to do fast analysis in real time data to make decisions we will discuss this point later in the implementation areas in the framework presentation.

Variety: From data to website logs, from tweets to visual data such as photos and videos, data comes in numerous shapes and forms. The nature of NOSI's data types does not vary. NOSI's data is mainly text data, numbers (integers/decimals) and dates. But in the future after developing NOSI's information systems it may contain pictures, scanned documents beside these data types.

Veracity: Veracity is about ensuring that the insights derived from data are reliable and valid. NOSI's data contains rubbish data because of the bugs of the data entry applications a longtime ago.

Value: Data does not have a fundamental value at the simplest level. Only when extract the insight needed to solve a particular problem or meet a specific need, it becomes useful. NOSI's data is complex so it needs a lot of development to extract a meaningful value from it.

After discussing the 5vs of big data, if we assign a score from 0 to 5 to determine if the social insurance data is considered as a big data or not? The following Table 1 shows each V and its score in NOSI.

Table 1: Big data Vs and its score in NOSI's Data.

Big Data Vs	Factor	Score (0 to 5)
Volume	> 1TB	5
Variety	Not Vary	0
Velocity	Needs Development	4
Veracity	Needs Development	3
Value	Needs Development	1

From the previous table showed that the final score is 13 points which is more than half of the total score. So, we considered the social insurance data as a big data.

This paper proposes a novel big data framework for Egyptian Social insurance. NOSI's data contains data which needs some preprocessing methods after extraction like replacing missing values, standardization and outlier/extreme data. The paper also presents using some mining methods such as clustering and classification algorithms on the Egyptian social insurance dataset.

The rest of the paper organized as follows: Section 2 presents a literature review of big data mining and big data analytics. Section 3 presents the social insurance big data framework, implements some parts of the framework and explains the rest with some insurance use cases. In Section 4, we will explore the Egyptian social insurance dataset and experiment results. Finally, the last section presents the conclusion and future work.

2. LITERATURE REVIEW

In the following section we will present some of the researches related to big data mining and analytics in insurance.

Kim and Cho presented a data governance framework for big data implementation with the national pension system. This research carried out a case analysis of South Korea's National Pension Service (NPS). They focused on public sector big data services to enhance people's quality of life. The procedures of NPS Big Data services data consist of four steps: extracting, transforming, cleaning, and loading. A data flow assessment scheme is built and used to handle information flow in a structured form that can be traced via a schematic diagram [3]. The study presented clear theoretical steps of collecting data. However not explained a

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

detailed implementation of these steps. The four procedures used in the author's paper will use some of these steps such as extracting, cleaning and loading in our proposed framework by using of Social Insurance Big Data (SIBD) in Egypt.

Hussain and Prieto presented research about big data in the finance and insurance sectors. The research discussed the benefits of analysis of industrial needs in the finance and insurance sectors. Benefits like enhancing the levels of customer insight, engagement, and experience. The authors illustrated the available data resources. Structured data: transaction data, data on account holdings and movements, market data from external providers, securities reference data, price information, and technical indicators. Unstructured data: daily stock, feeds, company announcements, online news media, articles, and customers' feedback. The most important point the authors focused on technical requirements like data extraction, quality, acquisition, integration/sharing, privacy, and security [4]. Overall, the research presented a good background in the insurance and finance sectors. The research presented a good methodology to use all insurance customer data; structured and unstructured to build a good vision to enhance their services. In our framework, we will use the structured data such as id, periods, and pensions of the insured individual. But the unstructured data is not applicable in the social insurance system in Egypt because of the lack of resources and technologies.

Song and Ryu presented a big data analysis framework for healthcare and social sectors separately and assigned them tentative names: 'health risk analysis center' and 'integrated social welfare service. The authors faced some obstacles in applying their framework. First, government ministries and agencies management committee is needed to correctly manage big data for healthcare and welfare services because big data needs to be managed in an integrated way. Second, a cooperative system with private organizations must be established that maintains unstructured big data associated with healthcare and welfare services. Most big data related to healthcare and welfare services are owned solely by the public sector. Third, technology for analyzing and processing large information on healthcare and welfare facilities needs to be developed. The study showed the obstacles that occurred in related fields similar to social insurance such as healthcare and welfare [5]. The study gave a starting point to the proposed framework to enhance the Egyptian social insurance scheme.

Tsai and et al. presented a big data survey. The researchers analyzed data analytics studies from the conventional data analysis to the new big data analysis. Three aspects of their framework are summarized: input, evaluation, and output. The paper concentrate on performance-oriented and results-oriented problems from the viewpoint of the big data analytics system and platform. Also, research offered a brief introduction to the big data mining algorithms consisting of clustering, classification, and regular pattern mining technologies from the perspective of the data mining problem [6]. The research presented a good mixture between big data analytics and mining which can be used as a good reference in the proposed framework. The framework will contain some use cases that can be applied in the social insurance data such as classification, clustering, statistical and actuarial analysis.

Bhoola and et al. presented big data challenges and possibilities using case studies that were implemented in the South African insurance industry and the technology and instruments required to analyze Big Data. They also discussed the roles that actuaries in Big Data Analytics and insurance room can play. Moreover, a brief introduction to data governance and laws as well as a possible perspective on what the future might hold. The research reflected full information about big data insurance and support this paper in the organization of our framework. The big data use cases in insurance will be presented in the framework like customer segmentation, risk assessment, and loss reduction [7].

Yenkar and Bartere published a review on data mining with big data. The publishers' implemented heterogeneous combination learning in the big data revolution and also in the data-driven model data mining involves extracting and analyzing large quantities of data to create large data models. The techniques came from artificial intelligence grounds and stats with a bit of database management. Normally, the data mining goal is either to predict or classify. The idea is to organize data into sets in classification. The plan is to predict a continuous variable's frequency in prediction [8]. The research supports us to build our framework in social insurance,

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

after the data extraction and integration the framework divided into two branches big data mining which contains classification and big data analytics which contains actuarial analytics that aims to predict the expenditure of pensions in the future.

3. SIBD FRAMEWORK

The proposed Social Insurance Big Data (SIBD) framework comprises social insurance big data extraction and collection to use cases in social insurance and how to apply it in Egypt. The framework as shown in Figure 2 aims to enhance social insurance services, help the insurers and actuaries to take a good decision in the fast and right way and give us more insights into the social data. The first stage including the data extraction and collection steps used. The second stage containing the data integration and selection of the extracted data and creating the dataset. The third stage pre-processing of the dataset. The fourth stage consists of big data mining and analytics. The big data mining section, divided into classification and clustering. Big data analytics consist of statistical and actuarial analysis and lifetime value prediction.

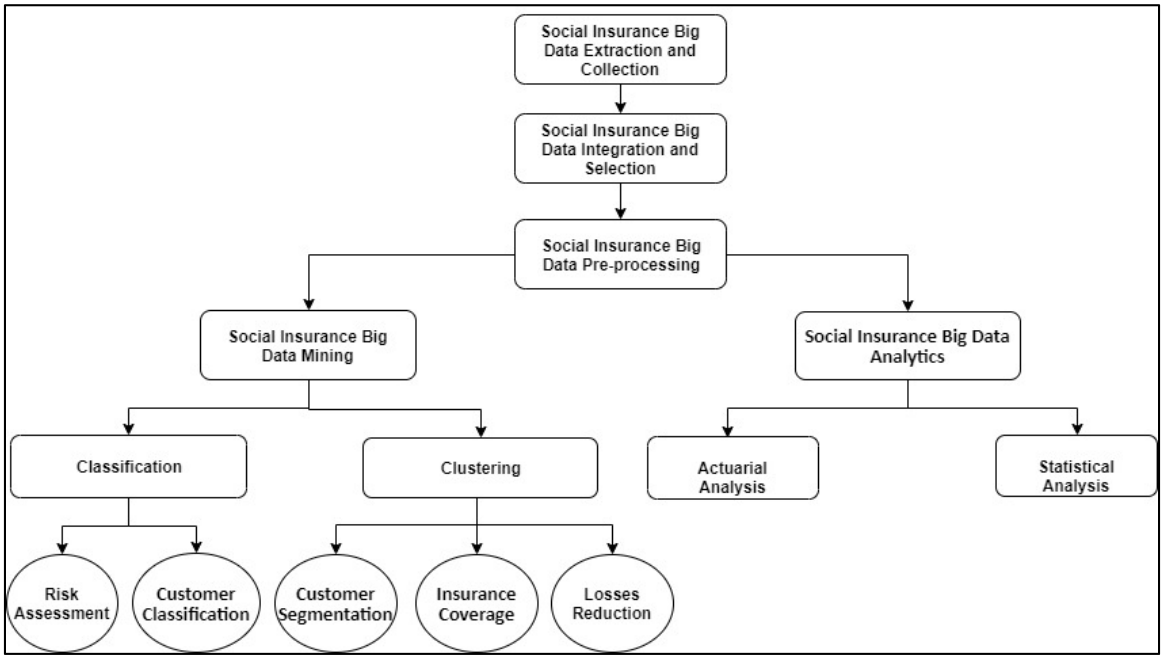

Figure 2: Framework of big data in Egyptian Social Insurance

3.1 SIBD Extraction and Collection

National Organization for Social Insurance (NOSI) is the only authority responsible for Egyptian Social Insurance data. The architecture of a NOSI information system mainly consists of a centralized IBM mainframe. The type of database that stored data on it is hierarchal. Data extraction and collection from mainframe going through the following stages:

- Choosing the hierarchical mainframe schema that will be extracted.
- Creating batch programs that are responsible to read the data and write on data files.
- Initiating batch processing which compilation the batch program from production control to computer operation.
- Provide the necessary spaces, which presented by mainframe tapes. Data is written in tape blocks.
- Data is integrated from tapes to the sequential dataset.
- Transfer data from sequential dataset to FTP server as flat text files.

3.2 SIBD Integration and Selection

Integration and selection of insurance data contain several important steps that will support in dataset creation (see Figure 3). The first step is to create a database schema equivalent to the IMS schema with the same attribute names. For instance, if we want to integrate a file have data the same line format described above. So, we will create a table with the same description that

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

extracted files from the mainframe. The second step using integration services to insert data from text files into database tables. The third step is using SQL to select important columns from tables. Finally, collecting all data in NOSI dataset.

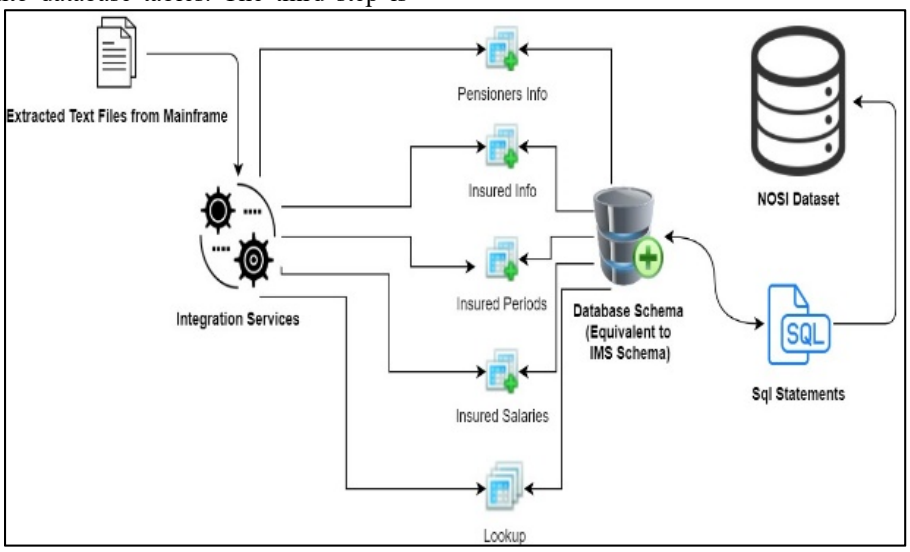

Figure 3: Data Integration and Selection Steps

NOSI's data consists of four basic tables: Insured Info, Insured Periods, Insured Salaries, Pensioners Info, and some lookup tables like Cities, Job Categories, and Sectors. Microsoft SQL Server Management Studio is used to do all the creation of a database, tables, and NOSI dataset. Also, Microsoft SQL Server Integration Services helped us to insert the bulk of data in short processing time. Every integration container consists of the File System Task which is responsible for reading the text file from the FTP server location and inserting in the database. Data Flow Task which is used for handling every database table and datatypes for each table to be inserted successfully.

The following Table 2 shows sample data selected from the integrated data which contains the basic attributes that represent the life cycle of the insured individual. A remarkable note the "Job End Date" column is empty which means that the insured individual is still working and not have a job end date.

Table 2: Sample Data Selected

Gender	Age	Sector	Total_Period	Job_Start_Date	Job_End_Date
Male	64	Private	21.3	1989-02-08	2010-05-27
Male	40	Private	16.5	2015-09-01	
Female	61	Public	11.1	2002-01-01	2012-07-31
Male	35	Private	10	2017-08-17	
Male	77	Private	31.1	1976-01-30	2007-03-06
Male	53	Private	12.2	2012-02-08	
Male	68	Private	40.1	2000-04-01	2011-09-11

3.3 SIBD Preprocessing

There are some preprocessing tasks to improve and develop the accuracy of the data mining algorithms. There are some basic tasks in big data preprocessing such as data cleaning, complete the missing values, and data normalization [9]. Data cleaning is the biggest challenge in SIBD, because of the age of this data. Sometimes you may find errors in data type's conversion because some rows have illogical data. The unknown numeric missing values will be replaced by mean values. Data normalization is used to handle characteristics on a different scale, otherwise it may reduce the efficiency of an equally significant attribute due to other

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

characteristics having values on a bigger scale. This will be illustrated in the experiment later.

3.4 SIBD Mining

Clustering and classification is the most techniques used in big data mining. These techniques were chosen based on the description of the issue and our interest in experimenting with two separate methodologies, whose fundamental features are summarized next.

3.4.1 Classification

Classification is a binary modeling method includes dividing the set of data into precisely two subgroups (or "nodes") that are more homogeneous to the reaction variable than the original set of information. It is recursive because for each of the resulting nodes the process is repeated. The divided data into two nodes, then compare all feasible splits for all values for all factors included in the assessment and conduct an exhaustive search through them all, choosing the split that separates information into two nodes with the greatest degree of homogeneity. There are some of case studies in Egyptian social insurance using the classification techniques such as risk assessment and customer classification. They are explained below:

Risk Assessment

In general, risk assessment is the process of estimate and evaluate the risk. In addition, it detects the possibility of the occurrence of something beneficial or harmful to the individual at a certain time. There are two types of risk assessment: Risk-Taking Model which is focused on normal things like rights, choices, and participation of the individual and Risk Minimization Model focused on danger and health. Social insurance is typically considered as the second type of risk assessment. We can divide the insured individuals in Egypt into three categories: Insured individuals from 18 to 24 years, from 25 to 45 and, from 46 to 65. Every category has its social risk assessment from death, retirement, and disability. For example, the first category the death, retirement, and disability rates will be lower. In the second category, the retirement will be lower than death and disability because in this category there are some dangerous jobs such as military, drilling, and steel jobs. In the last category, the retirement rate is higher than death and disability. The classification in risk assessment can support us in the estimation of expenditure of pensions of the insured individuals.

Customer Classification

The classification algorithm's task is to find out how that set of characteristics can lead us to take a certain decision. Individuals in Egyptian Social Insurance can be classified by using some of the dataset attributes such as insurance no., age, total periods. For example, if the analyst wants to predict the number of pensioners that have an early retirement and there is a rule of early retirement which is the insured individual must have a 20 years minimum or more of total working periods as shown in Table 3.

Table 3: Selected Data for Customer Classification

Gender	Age	Total_Period	Deserve Pension
1	64	21.3	YES
1	40	8.5	NO
2	61	11.1	NO
1	35	15	NO
1	77	31.1	YES
1	53	12.2	NO
1	68	40.1	YES

The Gender, Age, and Total Period predictor columns determine the value of the Deserve Pension (predictor attribute). The predictor attribute is recognized in a training set. The classification algorithm then attempts to determine how the predictor attribute value was achieved.

3.4.2 Clustering

Clustering is a process of dividing data into similar objects called clusters. There are many types of clustering algorithms in big data mining such as partitioning, hierarchical, density, grid, model, and constraint based clustering algorithms [10]. The use cases in social insurance using clustering techniques are customer segmentation, insurance coverage and loses reduction. They are explained below:

Customer Segmentation

To offer the insurance the opportunity to obtain customer understanding from various perspectives, we can generate one or more dimensions such as demographics, behavior, and

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

value. In each dimension, the clustering method must follow the same schema. The variables used for segmentation should be chosen based on a set of criteria [11]. The demographic segmentation in Egyptian social insurance can be consists of young, adult and old men and women. The segmentation of behavior should result in insured people groups sharing a common characteristic behavior. This behavior should identify insured individuals that can be managed in the same manner, but this should be managed differently from other segments. The segmentation's third and last dimension is value. This dimension concludes the insured is going to be financially rewarded when risks occur.

Insurance Coverage

The amount of risks that are covered for an insured individual is called insurance coverage. Insurance coverage represents the level of effectiveness of implementing social insurance policy, measured by the percentage of the insured population in the total working-age population under social insurance policies [12]. The working population in Egypt can be divided into four clusters. The first cluster is about the insured youth & adults with low income. The second one is about insured employees with high income. The third one collects old insured employees before retirement. The fourth cluster can be about the pensioners.

Losses Reduction

As mentioned before, the idea of a social insurance scheme is to collect a contribution from insured individuals and benefit them if happened to them any risks. Clustering method like partitioning by using K-means algorithm can estimate the expenditure of Egyptian pensions by calculation of the pension rates and what will happen if these rates increased or decreased. The pension rates can affect the expenditure by a profit or loss to the government. Clustering helps simplify the problem of classification of financial data based on their characteristics rather than on labels such as customer gender, living place, income or last transaction success, etc. [13].

3.5 SIBD Analytics

Big data analytics mainly encompasses big data analytical methods, systematic big data architecture, and analytical software. Data analysis in big data is the most important step to explore meaningful values, make suggestions and make decisions. Analysis of data can explore potential values. However, data analysis is a wide, dynamic and very complex area [14]. Social insurance with big data analytics will be useful in some type of analysis like statistical analysis, and actuarial analysis.

3.5.1 Statistical Analysis

The statistical analysis for big data can be grouped into two main types: resampling and divide-conquer. Resampling is a straightforward method created to serve two fundamental. First, resampling offers a deviation from the fixed hypotheses underlying many statistical processes. Second, resampling offers a structure for estimating the distribution of statistics that are very complicated. Therefore, because of the second fundamental, it will be useful to use resampling to estimate the distribution of pensions over the insured individual. The technique of dividing and conquering usually has three steps: (1) partitioning a large data set into K blocks; (2) processing each block individually (potentially in parallel), and (3) aggregating the alternatives from each block to create a final solution to the complete information [15]. A divide and conquer strategy can cause efficiency and enhancement of social insurance.

3.5.2 Actuarial Analysis

Actuarial models aim to demographic and financial projections of pension systems were generally derived from models that had been applied to occupational pension schemes covering groups of workers based on demographic variables, economic variables, and social (behavioral) variables of workers. Actuaries need to recognize that developing an accurate and definitive formula of human behavior is complicated and not always feasible. Accordingly, in relation to predictive analytics, actuarial need methods to considerably improve their knowledge of expected conduct or occurrences and support their policies and decisions [16]. Therefore, big data analytics will support actuaries in their work. One of the most important goals for actuaries is implement a year-by-year simulation technique to predict future expenses [17]. For instance, the following general equations is used in their simulation technique we can use it to predict the future expenditure of pensions in Egypt:

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

Next Year's Expenditure = Current Year's Expenditure * Survival rate * Adjustment Factor + Cost of New Pensions Projected for Award Next Year (1)

The base of contributions are calculated by multiplying the assumed amount of active insured individuals by the projected average insurable income and the contribution rate of the system (contribution factor)

Contribution Base = No. of insured * Average Insurable Earnings * Contribution Factor (2)

At this point, the research discussed the areas of the proposed framework. The following section presents an experiment with the implementation of some parts in the framework.

4. EXPERIMENT

The experiment section contains a description of the Egyptian social insurance dataset and its features. The preprocessing methods that can be applied to the dataset such as imputation, standardization, and outlier. Clustering between data nodes. Finally, applying a classification algorithm and measure accuracy, F-measure, recall, and precision. In the experiment, we used two miner tools WEKA and Orange.

4.1 Dataset

Due to the hugeness of Egyptian Social insurance datasets, the experiment dataset was carefully constructed based on the fundamentals of social insurance and its basic features [18]. Also, we decided to get the data of the individuals from 20 years only, not the whole data. In short, the dataset includes the insured individuals who have actual work periods and the target feature determines whether they are beneficiaries of the insurance system or not? In other words, are they taking pensions or not? The following Table 4 describes the main characteristics of the dataset.

Table 4: Dataset Characteristics

Dataset Characteristics	Multivariate
Attribute Characteristics	Numeric, Nominal and Date
Associated Tasks	Preprocessing, Clustering, and Classification
Number of Instances	13,800,427
Number of features	9
Extraction Date	2019-11-03

4.2 Dataset Features Description

Dataset features (attributes) represent the basic information about the insured individual such as his age, gender, his sector of work, and etc. The following Table 5 illustrates the description of each dataset features and the target.

Table 5. Dataset Features

Feature	Description
Age	The age of the insured individual
Gender	The gender Male or Female
City	It contains the last Egyptian city that the insured individual has/had a job in it.
Sector	The insured individual work sector such as public, private and etc.
Job Category	The category of work of the insured individual like Doctors, Engineers, Carpenters and etc.
Last Job Start	The last starting date of his/her work.
Last Job End	The last ending date of his/her work. As mentioned before, the job end date may contain empty values and this means that the insured individual is still working.
Full Insurance Periods	Calculated by subtracting insured working duration dates the start date and end date and sum the result durations.
Target	**Description**
Takes A Pension	It contains YES or NO values. 'YES' means the insured individual takes a pension, 'NO' means the insured not taking a pension.

4.3 Dataset Preprocessing

Before applying the preprocessing methods. Clarification of feature statistics needed to decide which preprocessing method is suitable for the dataset. Feature statistics consists of center (average), dispersion (median), minimum, maximum, and the missing values percentage. The following Table 6 expound the dataset feature statistics.

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

Table 6. Feature Statistics

Feature	Center	Dispersion	Min	Max	Missing
Age	39.23	0.29	18	84	0%
Gender	Male	0.54			0%
City	East Cairo	3.35			16%
Sector	Private	0.70			0%
Job Category	Maintenance Engineers	3.12			28%
Last Job Start	2011-07-01	6.87			0%
Last Job End	2011-02-28	6.62			55%
Full Insurance Periods	7.93	0.97	0	41.90	0%

4.3.1 Replacing Missing Data with Mean Imputation

This replaces the missing value with the mean or average sample or model depending on the data being distributed [19]. The feature that has more than 50% will be removed from the dataset because when missing values are large in number, all of these values will be replaced by the same imputation value, that is mean, contributes to a shift in the distribution form. Therefore, the "Last Job End" feature will be removed. By using WEKA "replace missing values" filter all the missing percentages turned to 0%.

4.3.2 Z-Score Normalization

This method also called "Standardization". The method aim to rescaling the characteristics of data between zero and one. If A is represent the mean of the values of feature A and σ_A is the standard deviation, original value v of A is normalized to V' using the following equation:

$$V' = \frac{\mathrm{v} - \bar{\mathrm{A}}}{\sigma_A} \qquad (3)$$

By applying this standardization on the feature values present a mean equal to zero and a standard deviation of one [20]. After using the "Standardization" filer in WEKA, the feature "Age" be ranged from -3.342 to 7.09. Also, the "Full Insurance Period" feature be ranged from -1.017 to 4.48.

4.3.3 Outlier/ Extreme Values

An outlier is an occurrence in dataset values that have distance from other values. Extreme values are either too large or too small values [21]. In the dataset founded 19676 instances considered as outlier values and there is no extreme values. The instances will be removed by "remove with value" filer. So, the number of instances reduced from 13,800,427 to 13,780,751.

4.4 Clustering

In clustering on the Egyptian social insurance dataset, we decided to use the K-means clustering. K-means is a method by which observations are grouped into a specific number of disjoint clusters. The "K" refers to the specified number of clusters. There are different distance measures that are used to determine which observation is to be appended to which cluster.

To detect the number of suitable clusters on the dataset the K-means calculate the silhouette score of each cluster. The silhouette value is a similarity measure of an object is to its own cluster compared to other clusters. In formula (5), we can explain $a(i)$ as a measure of i and how it fully assigned to its cluster. The $b(i)$ is the minimum average distance between $b(i)$ and every data point in other clusters that are not included in K [22].

$$s\,(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, if\ |C_i| > 1 \qquad (4)$$

The silhouette score range from −1 to +1. The high value indicates that the object fits well with its own cluster and is negatively aligned with neighboring clusters. If most objects have a high value, the configuration for clustering is suitable. If many points have a low or negative value, then there may be too many or too few clusters in the cluster configuration. The K-means implementation presents the following results that explained in Table 7 which contains the silhouette scores of 2 or more clusters on the dataset.

Table 7. Dataset Silhouette Scores

No. of Clusters	*Silhouette Score*
2	0.138

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

3	0.001
4	-0.030
5	-0.062
6	-0.166
7	-0.178
8	-0.181

From the table above we found that two clusters will be suitable for the dataset because it has a higher score from other numbers of clusters.

In cluster visualization of the K-means algorithm results, we used the scatter plot to describe the informative projection between clusters. The score plots detected three informative projection. The first projection between "Age" and "Full Insured Period" features. The second projection between "Age" and "Silhouette". The third one between "Full Insured Period" and "Silhouette". The following Figures 4-a, 4-b, and 4-c contain the scatter plot of the three projections.

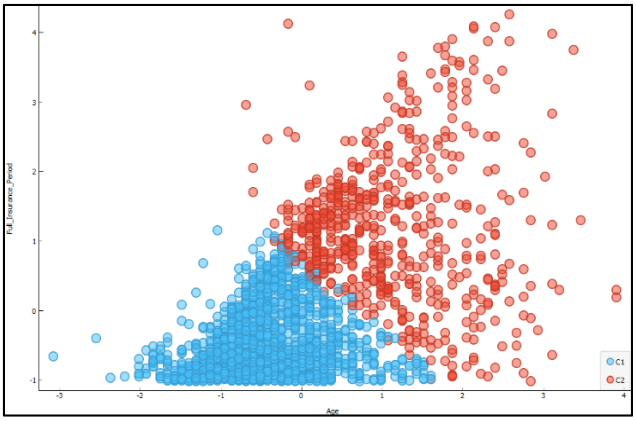

Figure 4-a: Scatter Plot of Age and Full Insured Period

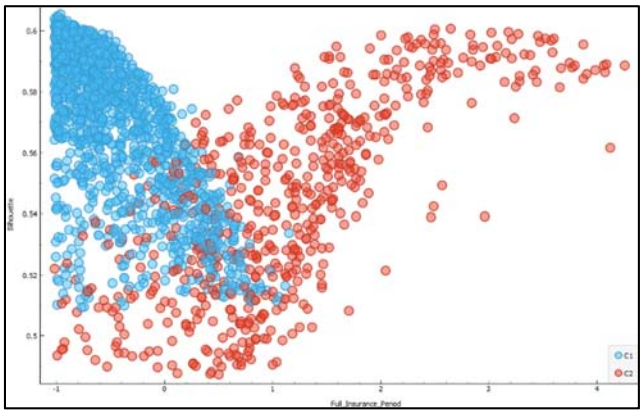

Figure 4-b: Scatter Plot of Full Insured Period and Silhouette

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

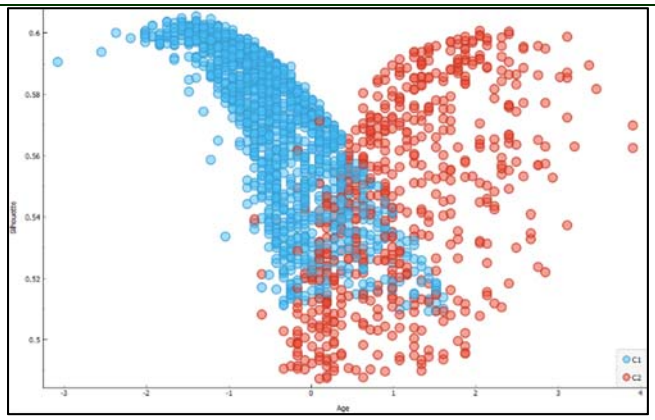

Figure 4-c: Scatter Plot of Age and Silhouette

4.5 Classification

Data classification consists of two main stages. The first stage is to create a classification model that involves the learning process, pick the algorithm to construct a classification model, and use the training set to construct a classification model. The second stage is to use the classification system which involves the classification method analysis and the classification model can be applied to the new test data if the reliability is appropriate. The dataset is divided into a training set and a test set; 80% of the data has been used for training, and 20% of the data has been used for testing.

The Support Vector Machine (SVM) is the chosen algorithm in the classification experiment. The data object in SVM algorithm is defined the features as $\{x_1,,,x_2\}$and a class label as y_i. SVM treats every data object as a point in the space of the feature so that the object belongs to any class. So, when the class label y_i= 1 then the data object belongs to the class or when y_i= -1 then the data object doesn't belongs to the class. Therefore the general formula for the data:

$$\{(x_i, y_i)|x_i \in \Re^p, y_i \in (-1,+1)\}_{i=1}^n$$

After applying the SVM algorithm on the dataset, the accuracy, precision, recall, and F-measure will be the evaluation measurements of the classification experiment. Accuracy which is the ratio of correct predictions. Precision is the ratio of correct positive predictions. The recall is the ratio of positively labeled instances, also predicted as positive. Finally, F-measure combines precision and recall in the harmonic mean of precision and recall. Table 8 shows the classification measures result of the dataset.

Table 8: Classification Measures

Accuracy	F-measure	Precision	Recall
0.94	0.91	0.88	0.94

The following Figure 9 represent the scatter plot of the SVM algorithm after classification.

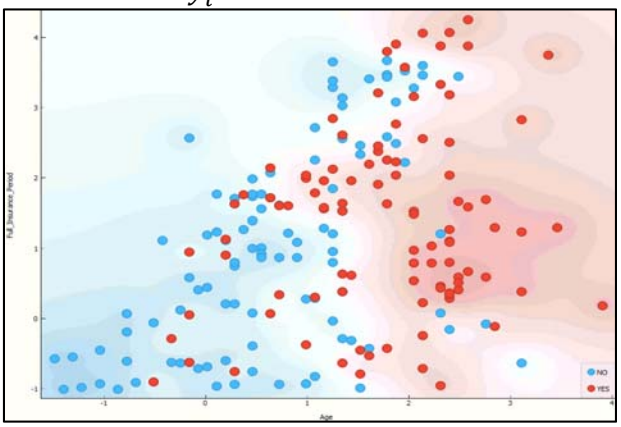

Figure 9: A Scatter Plot for SVM algorithm

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

5. CONCLUSION AND FUTURE WORK

This paper proposed a new framework for big data in Egyptian Social insurance to collect all the basic steps and methods for better benefits to the insured. The paper presented a literature review of some researches about big data mining and big data analytics in social insurance were presented. The paper implemented some parts of the presented framework through an experiment and explained the rest of it. Finally, the research overall tried to spotlight on social insurance field and give a mixture of using big data mining and analytics to help the insurers and the actuaries to get the best decision for the insured individuals.

For future work, we will extend this work with more preprocessing methods on the selected social insurance dataset. Furthermore, we will use some big data mining techniques by focusing on classification algorithms and apply some of the supervised learning algorithms such as decision tree J48, JRip, and Naïve Bayes algorithms and compare measurement between them.

Also, the challenges that faced the researchers it can be solved by the summarized following points and it can consider as our future framework updates:

- Structuring of big data sampling methods is important to reduce the amount of big data to a manageable size.
- A higher level of data science expertise needed to implement big data strategies to determine how long such data need to be kept, as some data are useful in making long-term decisions, while other data are not applicable. Therefore, the importance data selection in our framework need experience in business rules of social insurance.
- Data mining algorithms need data to be loaded into the main memory even if having super-large main memory to store all data for computing. So, the development of data collection and integration is important for big data mining in the framework.
- The importance of creating knowledge indexing framework to ensure real-time data monitoring and classification for big data applications.

REFERENCES

[1] U.S. Census Bureau [USCB], *Social Insurance and Human Services*, Section 11, 2004, pp. 355-381.

[2] Gantz, J. and Reinsel, D., *Extracting Value from Chaos, in IDC's Digital Universe Study*, sponsored by EMC, 2011.

[3] Kim, H. Y. and Cho, J., *Data Governance Framework for Big Data Implementation with NPS Case Analysis in Korea*. In Journal of Business and Retail Management Research, 12(03), 2018.

[4] Hussain, K. and Prieto, E., *Big Data in the Finance and Insurance Sectors*, In book of New Horizons for a Data-Driven Economy, 2016, pp.209-223.

[5] Song, T. and Ryu, S., *Big Data Analysis Framework for Healthcare and Social Sectors in Korea*, Healthcare Informatics Research, 21(1), 2015, pp.3-9.

[6] Tsai, C. W., Lai, C. F., Chao, H. C., and Vasilakos, A. V., *Big Data Analytics: A Survey. Journal of Big Data*, 2:21, 2015, pp.1-32.

[7] Bhoola, K., Madzhadzhi, T., Narayan, J., Strydom, S., and, Heerden, H., *Insurance Regulation in Africa: Impact on Insurance and Growth Strategies.* Presented at the Actuarial Society of South Africa's, Cape Town International Convention Centre, 2014, pp.145-196.

[8] Yenkar, V. and Bartere, M. *Review on Data Mining with Big Data*, In International Journal of Computer Science and Mobile Computing, Vol.3 Issue 4, 2014, pp. 97-102.

[9] García, S., Ramírez-Gallego S., Luengo, J., Benítez, J. M. and Herrera, F., *Big Data Preprocessing: Methods and Prospects*. BMC Big Data Analytics, 2016, pp.1-22.

[10] Tiruveedhula, S., Rani, C.M. S., Narayana, V., *A Survey on Clustering Techniques for Big Data Mining*, In Indian Journal of Science and Technology, Vol 9(3), 2016, pp.1-12.

[11] Bücker, T., *Customer Clustering in the Insurance Sector by Means of Unsupervised Machine Learning*, Internship Report, 2016, pp. 1-112.

[12] International Labour Organization [ILO], *Social Insurance: Enhancing Social Security Right for Everyone,* Policy Brief, Vol. 3, 2014.

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

[13] Cai, F., Le-Khac, N., and Kechadi, T., *Clustering Approaches for Financial Data Analysis: a Survey,* School of Computer Science & Informatics, 2016.

[14] Chahal, H., and Gulia, P., *Big Data Analytics.* In Research Journal of Computer and Information Technology Sciences, Vol. 4, 2016, pp.1-4.

[15] Wang, C., Chen, M., Schifano, E. Wu, J., and Yan, J., *Statistical Methods and Computing for Big Data*. Statistics and Its Interface, 2016, pp.399-414.

[16] American Academy of Actuaries [AAA]. Big Data and the Role of the Actuary. Big Data Task Force, 2018.

[17] International Labor Office [ILO], *ILO Pension Model Technical Guide,* 2018, Switzerland.

[18] Senousy, Y., Hanna, W.K., Shehab, A., Riad, A.M., El-Bakry, H.M., Elkhamisy, N., Egyptian Social Insurance Big Data Mining Using Supervised Learning Algorithms, *In Revue d'Intelligence Artificielle*, Vol. 33, No. 5, 2019, pp. 349-357.

[19] Jadhav, A., Pramod, D., and Ramanathan, K., Comparison of Performance of Data Imputation Methods for Numeric Dataset. *In Applied Artificial Intelligence, an International Journal*, 2019, pp. 913-933.

[20] García, S., Luengo, J., and Herrera, F. *Preprocessing in Data Mining*. Springer International Publishing Switzerland, 2015.

[21] Aggarwal, C. C., *Outlier Analysis Second Edition*. Springer, Cham, 2016.

[22] Amorima, R. C., Hennigb, C., *Recovering the Number of Clusters in Data Sets with Noise Features using Feature Rescaling Factors*. In Information Sciences Journal, 2016, pp. 1-34.