BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **20**, No 1
Sofia • 2020 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2020-0007

A Smart Social Insurance Big Data Analytics Framework Based on Machine Learning Algorithms

Youssef Senousy[1], *Abdulaziz Shehab*[2], *Wael K. Hanna*[1], *Alaa M. Riad*[1], *Hazem A. El-bakry*[1], *Nashaat Elkhamisy*[2]

[1]*Department of Computers and Information Systems, University of Mansoura, Mansoura, Egypt*
[2]*Department of Information Systems, Sadat Academy for Management Sciences, Cairo, Egypt*
E-mails: youssef_senousy@hotmail.com abdulaziz_shehab@mans.edu.eg wael_karam1@yahoo.com amriad2014@gmail.com elbakry@mans.edu.eg wessasalsol@gmail.com

***Abstract*:** *Social insurance is an individual's protection against risks such as retirement, death or disability. Big data mining and analytics are 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 need 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]. 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. Fig. 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 as wife, sons, and parents, who we call, pension beneficiaries if the pension owner dies.

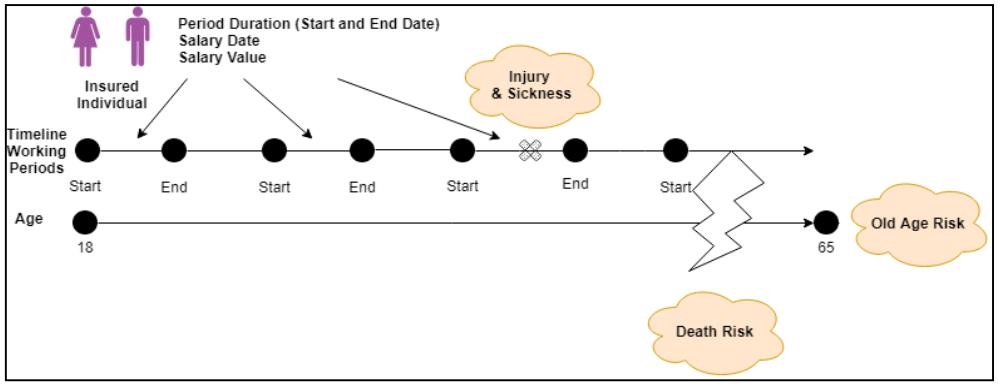


Fig. 1. Life cycle of insured individual

Big Data is a term used to describe a collection of data that is massive in size that 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 analyse 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 contain 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 contain rubbish data because of the bugs of the data entry applications since a long time ago.
- **Value.** Data does not have a fundamental value at the simplest level. Only when extracted the insight needed to solve a particular problem or meet a specific need, they become useful. NOSI's data are complex so they needs a lot of development to extract a meaningful value from them.

After discussing the 5 V's of Big Data, we may assign a score from 0 to 5 to determine if the social insurance data is considered as a big data or not. 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 (0-5)
Volume	> 1TB	5
Variety	Not Vary	0
Velocity	Needs Development	4
Veracity	Needs Development	3
Value	Needs Development	1

It is shown in the table above that the final score is 13 points which is more than half of the total score. So, we considered the social insurance data as big data.

This paper proposes a novel big data framework for Egyptian Social insurance. NOSI's data contains data, which need 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.

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 5 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 [3] 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. The study presented clear theoretical steps of collecting data. However a detailed implementation of these steps is not explained. 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 [4] present a research about big data in the finance and insurance sectors. The research discusses 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 illustrate 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 focus on are technical requirements like data extraction, quality, acquisition, integration/sharing, privacy, and security. Overall, the research presents a good background in the insurance and finance sectors. The research presents 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.

S o n g and R y u [5] present a big data analysis framework for healthcare and social sectors separately and assign them tentative names: 'health risk analysis centre' and 'integrated social welfare service. The authors face 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 need 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 analysing and processing large information on healthcare and welfare facilities needs to be developed. The study shows the obstacles that have occurred in related fields similar to social insurance such as healthcare and welfare. The study gives a starting point to the proposed framework to enhance the Egyptian social insurance scheme.

T s a i et al. [6] present a big data survey. The researchers analyse 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 concentrates on performance-oriented and results-oriented problems from the viewpoint of the big data analytics system and platform. Also, research offers 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. The research presents 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.

B h o o l a et al. [7] present big data challenges and possibilities using case studies that were implemented in the South African insurance industry and the technology and instruments required to analyse Big Data. They also discuss 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 is presented. The research reflects full information about big data insurance and supports 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.

Yenkar and Bartere [8] published a review on data mining with big data. The publishers implement heterogeneous combination learning in the big data revolution and also the data-driven model of data mining involves extracting and analysing large quantities of data to create large data models. The techniques comes 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. The research supports us to build our framework in social insurance, after the data extraction and integration the framework is 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 collecting of cases to be used in social insurance and how to apply it in Egypt. The framework as shown in Fig. 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 includes the data extraction and collection steps used. The second stage contains the data integration and selection of the extracted data and creating the dataset. The third stage is pre-processing of the dataset. The fourth stage consists of big data mining and analytics. The big data mining section is divided into classification and clustering. Big data analytics consist of statistical and actuarial analysis and lifetime value prediction.

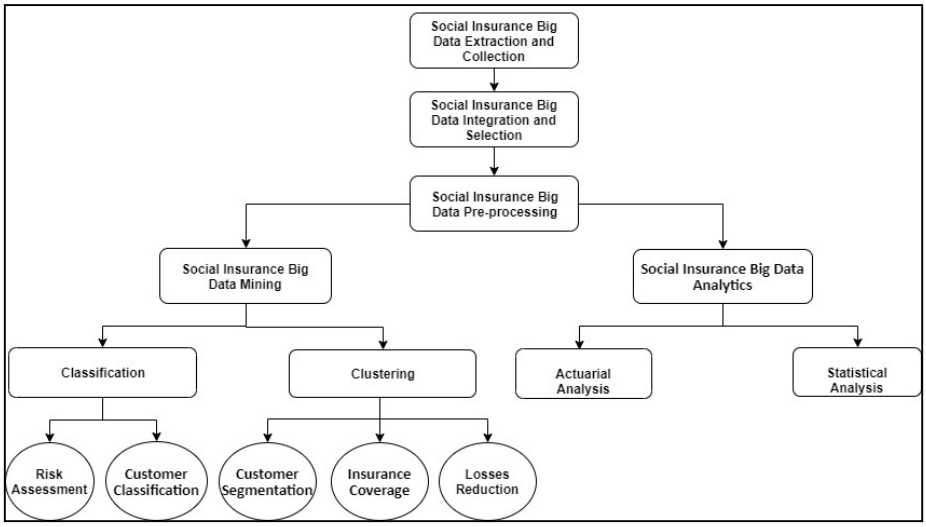


Fig. 2. Framework of Big Data in Egyptian Social Insurance

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 stores data on it is hierarchal. Data extraction and collection from mainframe is 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 of 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

The integration and selection of insurance data contain several important steps to produce the final dataset that will be used in clustering and classification. The first step is to create the same database table as the hierarchical mainframe database with the same attribute names. The second step is using integration services to insert data from text files into database tables. The third step is using some SQL queries to select important columns from database tables, and finally – collecting all data in NOSI dataset (Fig. 3).

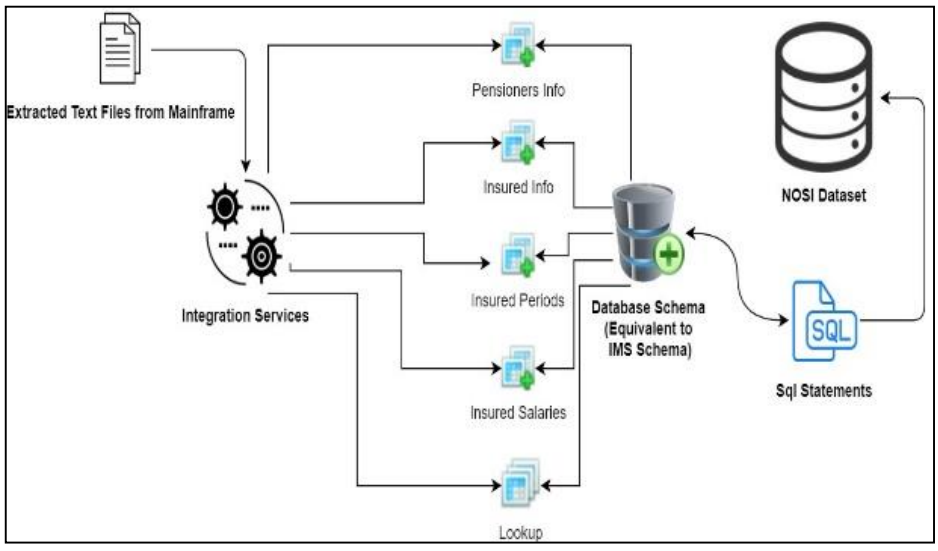


Fig. 3. Data integration and selection steps

NOSI's data consist of four basic tables: Insured Individuals Info, their Insured Periods, and Salaries. Also, Pensioners Individuals Info with some lookup tables like Cities, Job Categories, and Sectors. In the integration and selection processes, Microsoft SQL Server Management Studio is used to create database tables and NOSI dataset. Also, Microsoft SQL Server Integration Services is helpful in the insertion of the bulk of data in a short processing time. Every integration container consists of the File System Task, which is responsible for integrating the text file from the FTP server location to the database tables. Data Flow Task, which is used for handling every database table and datatypes for each table to be inserted successfully.

Table 2 shows sample data selected from the integrated data, which contain the basic attributes that represent the life cycle of the insured individual. Note: the "Job End Date" column is empty which means that the insured individual is still working and does 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 pre-processing

There are some pre-processing tasks to improve and develop the accuracy of the data mining algorithms. There are some basic tasks in big data pre-processing 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 characteristics having values on a bigger scale. This will be illustrated in the experiment later.

3.4. SIBD mining

Clustering and classification are the most frequently used techniques 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 modelling method that 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 data divided into two nodes, are then

compared to all feasible splits for all values for all factors included in the assessment and an exhaustive search is conducted 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.

3.4.1.1. Risk assessment

In general, risk assessment is the process of estimating and evaluating 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, in 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 metallurgic 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.

3.4.1.2. 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 number, 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 at least 20 years 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 similar data into 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 losses reduction.

3.4.2.1. Customer segmentation

To offer the insurer the opportunity to obtain understanding about the customer from various perspectives, we can generate one or more dimensions such as demographics, behaviour, and 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 consists of young, adult and old men and women. The segmentation of behaviour should result in insured people groups sharing a common characteristic behaviour. This behaviour 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 compensated when risks occur.

3.4.2.2. Insurance coverage

The amount to which risks 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 to 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.

3.4.2.3. Losses reduction

As mentioned before, the idea of a social insurance scheme is to collect a contribution from insured individuals and benefit them if any risks happened to them. Clustering method, like partitioning by using K-means algorithm, can estimate the expenditure of Egyptian pensions by calculation of the pension rates, and what would happen if these rates are 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 fundamentals. 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 individuals. 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.

Actuarial models aim at demographic and financial projections of pension systems that 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 (behavioural) variables of workers. Actuaries need to recognize that developing an accurate and definitive formula of human behaviour is complicated and not always feasible. Accordingly, in relation to predictive analytics, actuarial methods need 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 implementing a year-by-year simulation technique to predict future expenses [17]. For instance, we can use, the following general equation that is used in their simulation technique to predict the future expenditure of pensions in Egypt:

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

The base of contributions is 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):

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

At this point, the research discusses the areas of the proposed framework. Next 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 pre-processing methods that can be applied to the dataset such as imputation, standardization, and outlier, and 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. 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? Table 4 describes the main characteristics of the dataset.

Table 4. Dataset characteristics

Dataset Characteristics	Multivariate
Attribute Characteristics	Numeric, Nominal and Date
Associated Tasks	Pre-processing, 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/her age, gender, his/her sector of work, etc.

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, etc.
Job Category	The category of work of the insured individual like Doctors, Engineers, Carpenters, 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 pre-processing

Before applying the pre-processing methods clarification of feature statistics is needed to decide which pre-processing method is suitable for the dataset. Feature statistics consists of centre (average), dispersion (median), minimum, maximum, and the missing values percentage. The following Table 6 expounds the dataset feature statistics.

Table 6. Feature statistics

Feature	Centre	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%
Takes A Pension	NO	0.22			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 [18]. 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, and 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 missing percentages is turned to 0%.

4.3.2. Z-score normalization

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

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

By applying this standardization on the feature values, present a mean equal to zero and a standard deviation of one [19]. After using the "Standardization" filer in WEKA, the feature "Age" is ranged from –3.342 to 7.09. Also, the "Full Insurance Period" feature is 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 [20]. In the dataset were found 19676 instances considered as outlier values and there are no extreme values. The instances were 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. 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 calculates 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 the next formula, we can explain a(i) as a measure of i and how it is fully assigned to its cluster; b(i) is the minimum average distance between b(i) and every data point in other clusters that are not included in K [21],

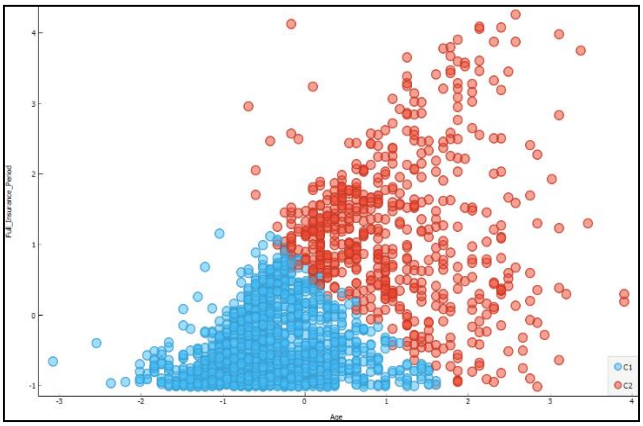
$$s\,(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad \text{if } |C_i| > 1. \tag{4}$$

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 neighbouring clusters. If most objects have a high value, then 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 are explained in Table 7 which contains the silhouette scores of two or more clusters on the dataset.

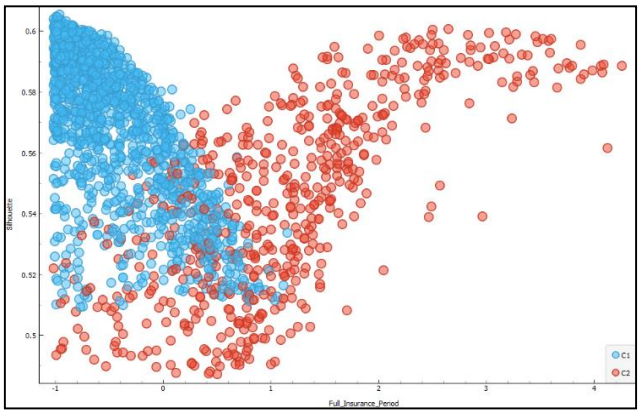
Table 7. Dataset silhouette scores

No of clusters	Silhouette score
2	0.138
3	0.001
4	–0.030
5	–0.062
6	–0.166
7	–0.178
8	–0.181

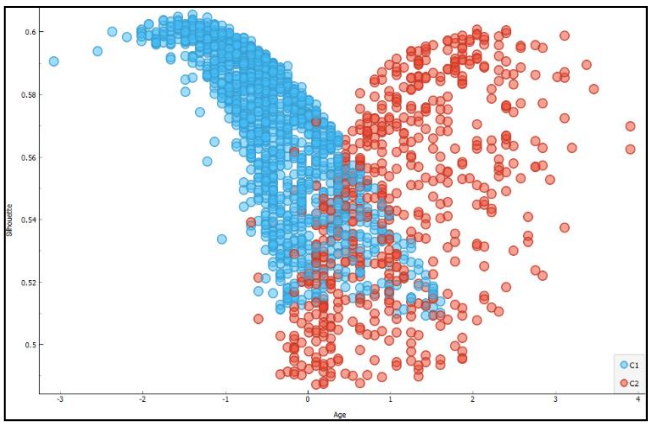
From the table above we found that two clusters would be suitable for the dataset because they have a higher score from other numbers of clusters. In cluster visualization of the K-means algorithm results, we used the scatter plot to describe the informative projection between clusters. The score plots detected three informative projections. The first projection is between “Age” and “Full Insured Period” features; the second projection – between “Age” and “Silhouette”; the third one – between “Full Insured Period” and “Silhouette”. Fig. 4 contains scatter plot of the three projections respectively.



(a)



(b)



(c)

Fig. 4. Scatter plot of age and full insured period (a); Scatter plot of full insured period and silhouette (b); Scatter plot of age and silhouette (c)

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 that 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 SVM divides all data objects into two categories in a feature space. The data object in SVM algorithm is defined, the features as $\{x_1,,,x_i\}$ and a class label as y_i. SVM treats every data object as a point in the space of the features so that the object belongs to any class. So, when the class label is $y_i = 1$ then the data object belongs to the class, or when $y_i = -1$ then the data object does not belong to the class. Therefore the general formula for the data [22]:

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

After applying the SVM algorithm on the dataset, the accuracy, precision, recall, and *f*-measure will be the evaluation measurements of the classification experiment, and accuracy, which is the ratio of correct predictions. Precision is the ratio of correct positive predictions. The recall is the ratio of positively labelled 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

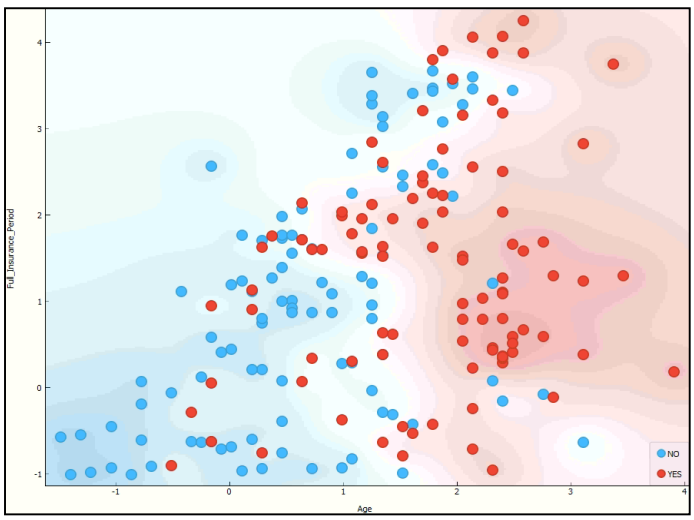


Fig. 5. Scatter plot for SVM Algorithm

5. Conclusion and future work

This paper proposes a new framework for big data in Egyptian Social insurance to collect all the basic steps and methods for better benefits to the insured people. The paper presents a literature review of some researches about big data mining and big data analytics in social insurance. The paper implements some parts of the presented framework through an experiment and explains the rest of it. Finally, the research overall tries 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 pre-processing 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, CN2 rule induction, and naïve Bayes algorithms, and compare measurement between them.

Also, the challenges that faces the researchers can be solved by the summarizing following points and this 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 is 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 of data selection in our framework needs 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., D. Reinsel. Extracting Value from Chaos, in IDC's Digital Universe Study. Sponsored by EMC, 2011.
3. Kim, H. Y., J. Cho. Data Governance Framework for Big Data Implementation with NPS Case Analysis in Korea. – Journal of Business and Retail Management Research, Vol. **12**, 2018, No 3.
4. Hussain, K., E. Prieto. Big Data in the Finance and Insurance Sectors. Book of New Horizons for a Data-Driven Economy, 2016, pp. 209-223.
5. Song, T., S. Ryu. Big Data Analysis Framework for Healthcare and Social Sectors in Korea. – Healthcare Informatics Research, Vol. **21**, 2015, No 1, pp. 3-9.
6. Tsai, C. W., C. F. Lai, H. C. Chao, A. V. Vasilakos. Big Data Analytics: A Survey. – Journal of Big Data, Vol. **2**, 2015, No 21, pp. 1-32.
7. Bhoola, K., T. Madzhadzhi, J. Narayan, S. Strydom, H. Heerden. Insurance Regulation in Africa: Impact on Insurance and Growth Strategies. Actuarial Society of South Africa's, Cape Town International Convention Centre, 2014, pp.145-196.

8. Y e n k a r, V., M. B a r t e r e. Review on Data Mining with Big Data. – International Journal of Computer Science and Mobile Computing, Vol. **3**, 2014, Issue 4, pp. 97-102.
9. G a r c í a, S., S. R a m í r e z-G a l l e g o, J. L u e n g o, J. M. B e n í t e z, F. H e r r e r a. Big Data Preprocessing: Methods and Prospects. BMC Big Data Analytics, 2016, pp. 1-22.
10. T i r u v e e d h u l a, S., C. M. S. R a n i, V. N a r a y a n a. A Survey on Clustering Techniques for Big Data Mining. – Indian Journal of Science and Technology, Vol **9**, 2016, No 3, pp. 1-12.
11. B ü c k e r, 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.
13. C a i, F., N. L e-K h a c, T. K e c h a d i. Clustering Approaches for Financial Data Analysis: A Survey. School of Computer Science & Informatics, 2016.
14. C h a h a l, H., P. G u l i a. Big Data Analytics. – Research Journal of Computer and Information Technology Sciences, Vol. **4**, 2016, pp. 1-4.
15. W a n g, C., M. C h e n, E. S c h i f a n o, J. W u, J. Y a n. 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.
18. J a d h a v, A., D. P r a m o d, K. R a m a n a t h a n. Comparison of Performance of Data Imputation Methods for Numeric Dataset. – International Journal in Applied Artificial Intelligence, 2019, pp. 913-933.
19. G a r c í a, S., J. L u e n g o, F. H e r r e r a. Preprocessing in Data Mining. Springer International Publishing, Switzerland, 2015.
20. A g g a r w a l, C. C. Outlier Analysis. Second Edition. Springer, Cham, 2016.
21. A m o r i m a, R. C., C. H e n n i g b. Recovering the Number of Clusters in Data Sets with Noise Features Using Feature Rescaling Factors. – Information Sciences Journal, 2016, pp. 1-34.
22. B r i d g e l a l l, R. Introduction to Support Vector Machines. Lecture Notes, 2017, pp. 1-18.

Received: 17.12.2019; Accepted: 24.02.2020