applied sciences

MDPI

*Article*

# Need-Based and Optimized Health Insurance Package Using Clustering Algorithm

**Irum Matloob *, Shoab Ahmad Khan, Farhan Hussain, Wasi Haider Butt, Rukaiya Rukaiya and Fatima Khalique**

Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering (CEME), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan; shoabak@ceme.nust.edu.pk (S.A.K.); farhan.hussain@ceme.nust.edu.pk (F.H.); wasi@ce.ceme.edu.pk (W.H.B.); rjavaidsh@gmail.com (R.R.); fatima.khalique@ceme.nust.edu.pk (F.K.)

* Correspondence: irum.matloob@ceme.nust.edu.pk

**Abstract:** The paper presents a novel methodology based on machine learning to optimize medical benefits in healthcare settings, i.e., corporate, private, public or statutory. The optimization is applied to design healthcare insurance packages based on the employee healthcare record. Moreover, with the advancement in the insurance industry, it is rapidly adapting mathematical and machine learning models to enhance insurance services like funds prediction, customer management and get better revenue from their businesses. However, conventional computing insurance packages and premium methods are time-consuming, designation specific, and not cost-effective. During the design of insurance packages, an employee's needs should be given more importance than his/her designation or position in an organization. The design of insurance packages in healthcare is a non-trivial task due to the employees' changing healthcare needs; therefore, using the proposed technique employees can be moved from their existing package to another depending upon his/her need. This provides the motivation to propose a methodology in which we applied machine learning concepts for designing need-based health insurance packages rather than professional tagging. By the design of need-based packages, medical benefit optimization which is the core goal of our proposed methodology is effectively achieved. Our proposed methodology derives insurance packages that are need-based and optimal based on our defined criteria. We achieved this by first applying the clustering technique to historical medical records. Subsequently, medical benefit optimization is achieved from these packages by applying a probability distribution model on five years employees' insurance records. The designed technique is validated on real employees' insurance records from a large enterprise.The proposed design provides 25% optimization on medical benefit amount compared to current medical benefits amount therefore, gives better healthcare to all the employees.

**Keywords:** clustering; fraudsters; health insurance; healthcare; medical benefit; premium amount; probability distributions



**Citation:** Matloob, I.; Khan, S.A.; Hussian, F.; Butt, W.H.; Rukaiya, R.; Khalique, F. Need-Based and Optimized Health Insurance Package Using Clustering Algorithm. *Appl. Sci.* **2021**, *11*, 8478. https://doi.org/10.3390/app11188478

Academic Editor: Albert Smalcerz

Received: 26 July 2021
Accepted: 8 September 2021
Published: 13 September 2021



## 1. Introduction

In developing countries, the income persons/employees cannot afford the high cost medical services which are exponentially rising with time. Under such conditions, good insurance policies can prevent the employees from financial troubles when required [1]. Healthcare provision is the main challenge in health care industry. The cooperate organizations/enterprises help their employees with healthcare insurance coverage to increase their productivity. When insurance policy amount is not enough to fulfil the healthcare needs of the employees in that case employees tries to overcome their burden by mis-utilizing medical benefit amount. The main cause of these misutilizations is the inappropriate design of medical benefit packages. Moreover, the enterprises normally provide one package with a fixed premium amount and a medical benefit amount for each type of category in

their organizational hierarchy, without considering that one employee need for healthcare services may vary as compared to others in the same category.

There are three main elements in the health insurance industry namely client, healthcare providers, and insurer [2]. Some of the important concepts related to the health insurance industry are listed below in Table 1.

**Table 1.** Important Terminologies.

| Terminology | Description |
| --- | --- |
| Insured | The employee and his family members, who are availing the insurance policy. |
| Sum amount | A maximum amount paid by the insurance company in case the insured person gets hospitalized. For example, the sum insurance of an employee is PKR 3 Lacs and got hospitalized three times in a year. On his first treatment, the billed amount is 50 thousands PKR, on his second and third treatments, the amount is of PKR 1 Lac and 2 Lacs, respectively. The total bill amount for the year is PKR 3.5 Lacs which exceeds the sum insured of the employee hence, paid by the insured person. |
| Premium amount | The fixed annual amount paid by the insured. In the case of organizations or employers, insured are employees who are availing insurance policy, and their employers are the payer of premium amounts to insurance companies. |
| Payer | It is the employer/ organization paying the premium. |
| Insurance policy | It is a legal document that includes details regarding particular insurance coverage for an insured. |
| Contract | It is the legal agreement between the insurer and insured. |

### *1.1. Problem Description*

In healthcare insurance, the sum insured and premium are directly proportional to each other. Whenever the sum insured amount is increased premium amount also increases accordingly.

The four main problems in health insurance are: moral hazard, adverse selection, healthcare fraud [2] and excess sum insured amounts as shown in Figure 1. Moral hazard is the term used for the situation, once an employee gets insurance coverage, he/she tries to get treatment for a long-standing health condition like hernia, etc. The adverse selection is the case when groups with poor health conditions, got insurance coverage. Pre-existing member groups with a fewer number of healthy members, join an insurance scheme results in monetary losses for these schemes because the insurance claims are more than the total premium collected. The third main problem is healthcare fraud which is the misutilization of available funds. The fourth important problem is that organizations are paying more premium amounts for their employee's health insurance. If there are 10,000 employees in any organization, there is a possibility that only 5000 employees are visiting the hospital for availing health services whereas the remaining never avail services. In some cases, the premium amount is greater than the total computed medical benefit utilization. Therefore, there is a need to optimize the premium amount according to the medical benefit utilization amount.

The main goal is to keep a balance between two important factors: Employee need (sum insured amount) and employer budget (budget limits (premium)). The imbalance between these two factors results in overburdened employers or stressful employees as shown in Figure 1.

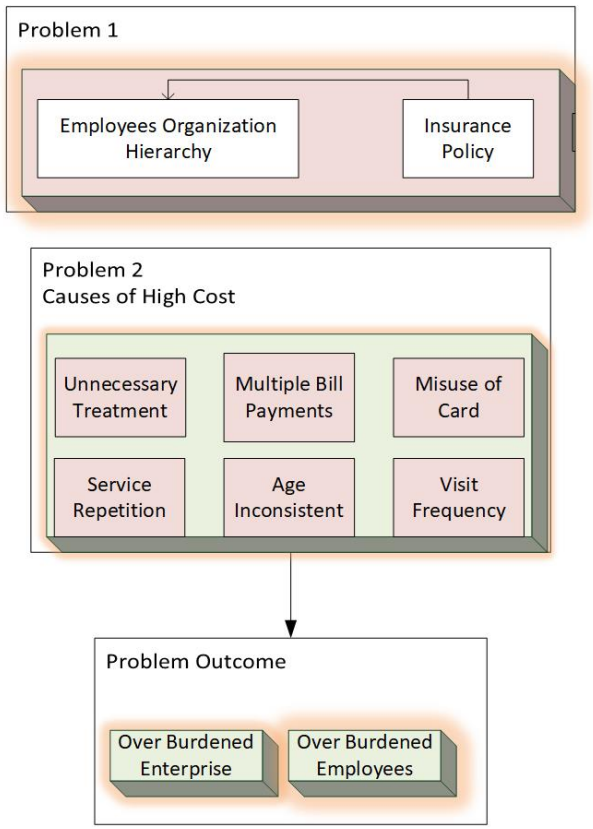


**Figure 1.** Description of problems to be addressed.

The objective is to develop a dynamic premium amount calculation technique for different groups within each category of employees in an organization using machine learning. Employees can be transferred from one group to another depending upon their needs. These packages are useful for employers as well as for insurance companies or employers.

Solution based on rigorous and systematic methodology is required, which can minimize the financial crisis of employees and enterprises. To the best of our knowledge, none of the existing designs used machine learning techniques to generate need-based insured packages as well as the data driven analysis on real enterprise record using probability distribution model. In the proposed methodology, optimization refers to the amount of medical benefit performed. The existing medical benefits amount is optimized by using the need-based packages evaluated using machine learning techniques.

### *1.2. Research Contributions*

The proposed methodology focuses on medical benefit optimization. The designed methodology provides following research contributions to the research area:

1. Incorporates medical benefit optimization using kmeans clustering and probability distribution model.
2. Generates employee need-based health insurance packages for enterprises and insurance companies. The historical medical records of employees are analyzed for generating these packages.
3. Provides detailed data-driven analysis for medical benefit optimization by defining lower and upper bounds for each package amount and estimation of out of pocket employees in each category.

## 2. Related Work

Most of the research is conducted on highlighting different issues in healthcare but none of them is focused on the cause of the incidence. The prevalence of the issues lies in

addressing the lack of available funds to low-income employees for healthcare services. There are researches that employed machine learning techniques to address many problems in the insurance industry [3]. Recent research is conducted to optimize benefit amounts in the automobile insurance industry but still lacks the focus on the optimization of the amounts using need-based insurance packages.

The genetic algorithm-based methodology for optimization of technical benefits over car insurance amount is proposed in [4]. The design uses a combination of heuristic and predictive techniques to perform optimization for moving targets. An automated framework for insurance companies is suggested using an extreme gradient boosting algorithm on blockchain-based transactional records [5]. The results showed that XGboost produces high performance as compared to other existing learning algorithms. An online learning solution is proposed in [6] that can automatically handle real-time updates of the insurance network. Machine learning-based techniques are applied to predict fake claims in automobile insurance and simplifies the calculation of premium amounts based on previous financial details for different customers. The importance of machine learning models in the insurance domain is sssed, and a comparison of household and motor insurance is provided in [7]. The detailed evaluation of machine learning algorithms in the insurance industry is performed in [8] and proved the effectiveness of the random forest algorithm. A study on the health insurance scheme implementation constraints of an organization is discussed in [9]. It uses the qualitative research design over the employee data to explore the challenges counter in the implementation process. A strategy to model the renewal price adjustment problem as a sequential decision problem and single and constrained Markov chain process is proposed in [10]. The model analyzes revenue maximization and its effects on the customers' retention levels by applying a model-free reinforcement learning algorithm. The results are validated on employee's data of a Spanish company named BBVA. A hybrid methodology is proposed in [11] that analyzes the data imbalance problem in the automobile insurance industry by applying K Reverse Nearest Neighborhood and one-class support vector machine (OCSVM) algorithms. A comparative study is performed in [12] that compares various classifiers and an application of genetic algorithm-based fuzzy C-Means clustering in the automobile insurance industry.

Few proposed schemes utilized probability distributions and other statistical models for funds prediction and annuity valuation. A new risk-function premium principle is proposed in [13]. An investigation is performed in [14] that proves the fourth-order statistics as an accurate approximation of the expected loss. The research conducted in [15] utilizes extreme trees and neural networks (NNs) models for bond return predictability. A methodology based on a machine-learning technique for pricing arithmetic and geometric average options is proposed in [16]. It provides a model-free scheme for efficient and quick results. The advanced ensembles such as random forests and boosted trees are used for insurance pricing, presented in [17]. The proposed design uses probability distribution models such as Poisson and gamma deviation for the purpose. A framework based on machine learning models is proposed in [18] that elaborates the impact of models on individual insurance consumers. The evaluation of the risk and severity of an insurance claim is performed using the telematics data and prior knowledge in [19]. The scheme computes the premium amount using variables, e.g., age, postal code and car model and driving patterns are used for generating premium. The methods for variable annuity valuation based on machine learning and data clustering are proposed in [20]. The tree boosted models of machine learning are used to optimize the proposed premium, presented in [21]. A system to adjust the insurance rates in real-time according to the change in character traits of the user is designed in [22]. The cost optimization issue is addressed in [23], that provides a comprehensive review of existing studies and analyzed research studies based on healthcare cost optimization problems. The impact of artificial intelligence on the insurance industry is presented in [24]. The study illustrates that with the focus on cost efficiency and new revenue streams, the insurance business model is transformed from a loss compensation model to prediction and prevention.

## 3. Materials and Methods

The proposed methodology is divided into three cascaded steps, including generation of need-based package using mapping and clustering, computation of medical benefits using the generated packages and data driven analysis using probability distribution to estimate the amount. The notations used in the algorithms are depicted in Table 2.

**Table 2.** Notation with description.

| Notation | Description |
|---|---|
| $\varphi$ | depicts transactional data of employees. |
| $k$ | There are seven categories of employees in this organization and each category is availing separate insurance plan. Categories are represented by k where k= (A, B,C,D,E,F,G) |
| $\phi$ | denotes total computed medical benefit. |
| $\alpha_j$ | denotes total amount in each category. |
| $\beta_i$ | depicts packages for $k$ categories. |
| $\alpha_{G_i}$ | depicts amount of each group in each category |
| $O$ | is the Premium amount which the hospital/organization is paying. The organization is giving total premium $T$ and for each category there is a separate value of premium amount |
| $\epsilon$ | Number of total employees are denoted by $E$ and number of employees in each group $G_x$ of each category are denoted by $\epsilon$ |
| $V^i$ | is the out of pocket employees in each category i = k where k = A,B,C,D,E,F,G. |

Figure 2 depicts the overall methodology used for the medical benefit optimization. The concept of clustering is incorporated to identify natural groups in existing categories of employees.The kmeans clustering is applied on each category and the clusters centroid are mapped on the selected category records.

### 3.1. Need Based Package Generation Using Clustering

The machine learning concept of clustering is used for classification purposes. Clustering is the natural grouping of records or data, in which similar records are in the same group whereas dissimilar records are in the different group. For generating need-based packages, we must perform category based analysis of transactional data. One category of employees is considered at one time and then clustering and mapping are applied as shown in Figure 3, amount is the money used by the employees of the selected category and on the x-axis number of records are the number of transactions in the selected category. Each category can be divided into groups, as depicted in the Figure 3. As a result of clustering, clusters are obtained in each category. Once a centroid for each cluster is obtained, we compute the distance of each computed centroid from all records of the considered category. The distance between centroids and all records of each category are computed and based on these calculated distances, patients are assigned to each cluster. The values of amount and visits are kept same as that of the centroid for all records. These steps are repeated for all the centroids.

Algorithm 1, explains the procedure for the need-based package generation. Inputs for the algorithm 1 are: number of clusters, max_iterations, $\varphi_1$,$\varphi_2$,$\varphi_3$,$\varphi_4$,$\varphi_5$,$\varphi_6$,$\varphi_7$ along with all considered attributes. First of all, the Kmeans algorithm is executed for one category. The second for loop is on x, where x denotes the number of clusters and $C_x$ denotes centroid for cluster $x$. Mapping of $C_x$ on specific category records is performed. The meaning of term mapping is different in our scenario, it is the computation of distance of $C_x$ from all the records in the considered category. Fourth, for loop iterates for the number of records in a specific category. $D_u$ denotes the distance of record from centroid where subscript $k$ denotes category. The mean of all the $D_u$ is computed. All those records whose $D_u$ is less than the Mean are added to the $G_x$. All remaining are discarded. At line no 18, the amount of considered record is $Am_u$, and $Am_x$ is the amount of centroid. $Am_u$ is replaced with $Am_x$. In the end, the sum of all amounts $Am_x$ in one group is computed.

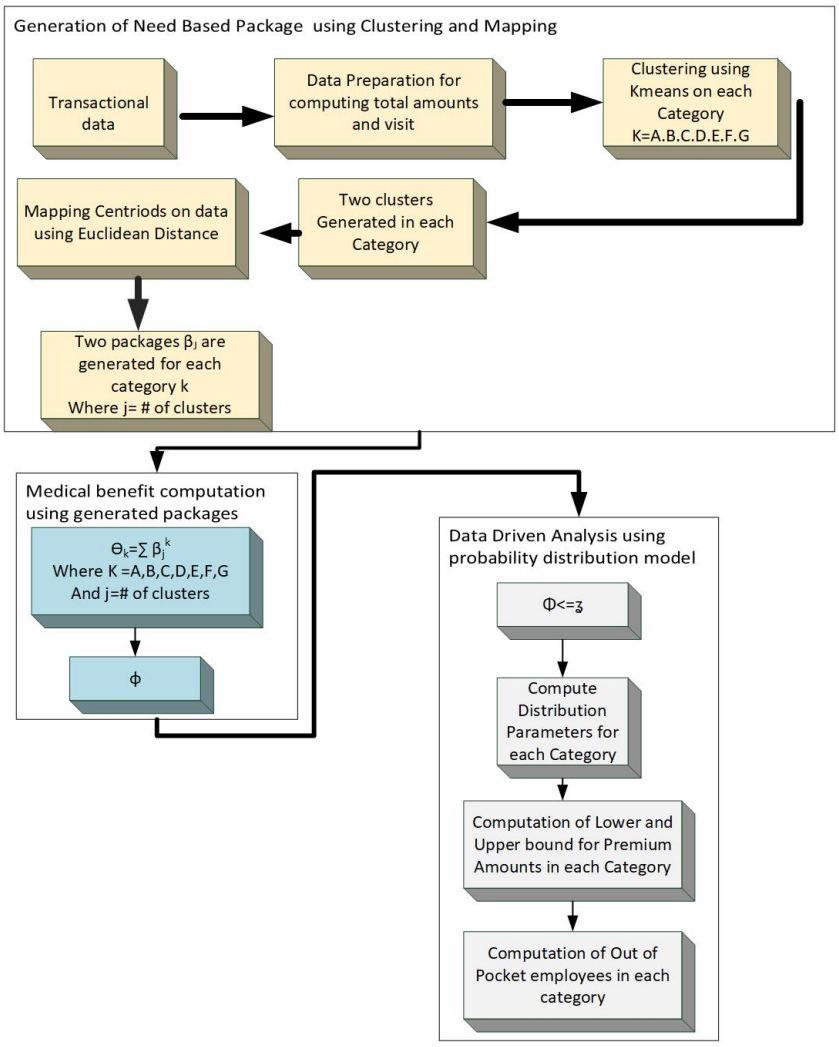


**Figure 2.** Graphical Summary of Proposed Methodology.

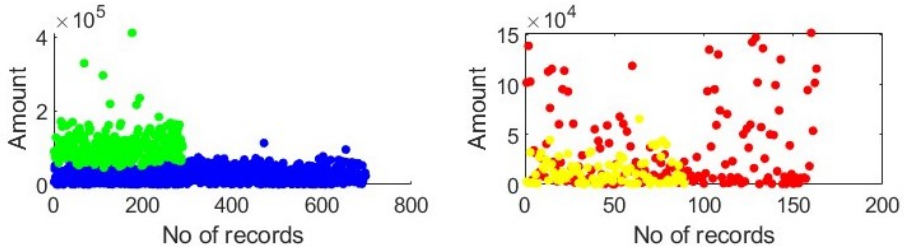


**Figure 3.** Clusters within categories.

For generating packages, there is a need to sum up the total amount of all members in one group. Within every category $u$, total amount is computed for each group $\beta_i$. Total amount $\alpha_j$ for each category is computed in the next step.

**Algorithm 1:** Need based Package generation using clustering and mapping.

**Input:** $\varphi_1, \varphi_2 \ldots . \varphi_7$ , number of clusters,*max_iterations* and *Considered attributes*$\{age_u,\ gender_u,\ status_u,\ Amount_u,\ visits_u\}$
**Output:** $\omega$;

**for** $u \leftarrow 1$ *to* $7$ **do**
  Call Kmeans ($\varphi_u^k$)
  **foreach** $\varphi_k$ **do**
    **for** $x \leftarrow 1$ *to* $C$ **do**
      Find instances closet to $C_x$
      $C_x$=$\{agec_x, genderc_x, statusc_x, Amountc_x, Visitc_x\}$
      **foreach** $C_x$ **do**
        Map $C_x$ to $\varphi_k$
        **for** $u \leftarrow 1$ *to size*$(\varphi_k)$ **do**
          $D_x$=$\sqrt{(ac_x - a_u)^2 + (gc_x - g_u)^2 + (sc_x - su)^2 \ + (Amc_x - Am_u)^2 + (Vc_x - V_u}$
        **end**
        Calculate Mean $D_u$
        Set Mean as $M_x$
        // Perform matching for grouping of employees with same attributes like centriod
        **if** $(D_u <= M_x)$ **then**
          Add $D_u \rightarrow G_x$
          **else**
            Discard
          **end**
          Replace $Am_u$ with $Amc_x$ in all Members in $C_x$
          Sum_Amount+=$Amc_x$
        **end**
      **end**
    **end**
  **end**
  Generate Packages for each Category $u$.
**end**

### *3.2. Medical Benefit Computation Using Generated Packages*

We generate different insurance packages for each category in the first step. Once all insurance packages are generated, the next step is to calculate the total medical benefit utilization is computed using the following equations:

$$\epsilon_{G_i} = \alpha_{G_i} / amount_{G_i} \tag{1}$$

where $amount_{G_i}$ is the amount value of centroid of considered group in each category and $i$ is the number of groups in each category. In Equation (2), we are computing number of employees from generated packages.

$$\epsilon_k = \sum_{i \in G} \epsilon_{G_i} \tag{2}$$

Once we get $\epsilon$, Equation (3) is used, for computing total medical benefit utilization.

$$\phi = \sum_{j=1}^{j=k} \sum_{i=1}^{i=k} \epsilon_i * \alpha_j \tag{3}$$

$\alpha_j$ is the total computed amount for each category where $j$ = A,B,C,D,E,F,G. The step by step implementation is represented in Algorithm 2.

**Algorithm 2:** Medical benefit computation using generated packages.

**Input:** Packages: $\{\omega_1, \omega_2 ... \omega_7\}$
$\epsilon = 0$;
**for** $(u \leftarrow 1$ *to* $7)$ **do**
  **for** $(r \leftarrow 1$ *to* $M)$ **do**
    $\epsilon_{G_r} = \alpha_{G_r} / amount_{G_r}$
  **end**
  $\epsilon = \epsilon_{G_r}$
**end**
$sum = 0$;
**for** $(j \leftarrow 1$ *to* $7)$ **do**
  **for** $(i \leftarrow 1$ *to* $M)$ **do**
    $sum = \epsilon_i * \alpha_j$
  **end**
**end**
$\phi = sum$;

### 3.3. Data Driven Analysis Using Probability Distribution Model

Once we get $\phi$, then we will move to the third step. In this step, data-driven analysis is performed for computing:

1. Lower and upper bound for each group of each category and
2. Out of pocket employees before and after optimization
3. Estimation of total medical benefits amount by using a probability distribution model

A probability distribution is a function that depicts the possibility of getting the values that a random variable can have within the distribution. In our case, this random variable is the *amount*. We evaluate the results of the above two steps, with the help of probability distribution models. The probability distribution model that fits employee insurance records is the Gamma distribution model as shown in Figure 4. On the X-axis, there number of employees and on the Y-axis is the probability density function.

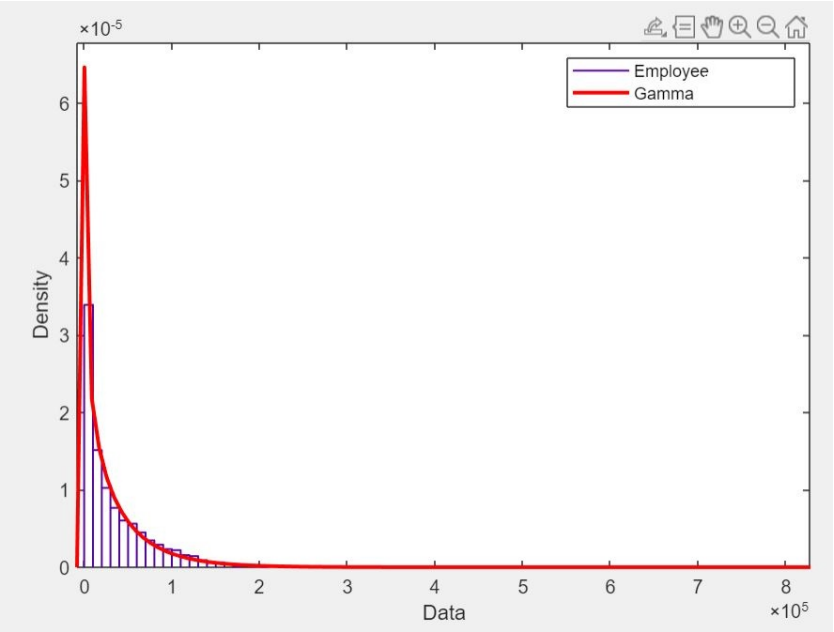


**Figure 4.** Probability distribution model for employees insurance records.

The shape and scale parameters of gamma distributions are a and b respectively. The formula for Mean $\mu$ and standard deviation $\sigma$ in terms of gamma distribution parameters are represented in Equations (4) and (5).

$$\mu_{G_h} = a_{G_h} b_{G_h} \tag{4}$$

$$\sigma_{G_h} = \frac{\sqrt{a_{G_h}}}{b_{G_h}} \tag{5}$$

where h = {1,2,3,4..}, depending upon the number of groups within the category. Once we get $\mu_{G_h}$ and $\sigma_{G_h}$ for each group of all categories, we can adjust the premium amount for each group of every category, by using a gamma distribution plot. Algorithm 3 explains the computation of lower and upper bound for premium amounts for each category using the gamma probability distribution model. $V^i$ is the out of pocket employees where i represents a category.

**Algorithm 3:** Data driven analysis using probability distribution Model.

**Input:**: $\{\phi_1, \phi_2 ... \_7\}$
**for** $(i \leftarrow 1$ *to* $7)$ **do**
  $\mu^i = a^i b^i$
  $\sigma^i = \frac{\sqrt{a^i}}{b^i}$
  **for** $(h \leftarrow 1$ *to* $G$ **do**
    $\mu^i_{G_h} = a^i_{G_h} b^i_{G_h}$
    $\sigma^i_{G_h} = \frac{\sqrt{a^i_{G_h}}}{b^i_{G_h}}$
  **end**
  **for** $(f \leftarrow 1$ *to* $3)$ **do**
    **if** $(T < \phi)$ **then**
      $\mu^i + f * \sigma^i$
    **end**
  **end**
  Set lower bound= $\mu^i$
  Set Upper bound=$\mu^i + 1 * \sigma^i$
  **for** $(s \leftarrow 2$ *to* $3)$ **do**
    $V^i$+=[P($\mu^i + s * \sigma^i$)*($\mu^i + s * \sigma^i$ )]*100
  **end**
**end**

The main topic of discussion nowadays is that most employers or organizations are paying the excess premium amount to the insurance companies while their employees are not frequently consuming the amount of the medical benefit. The objective is to align the premium amount with actual medical utilization made by each category of employees. The historical medical records enable us to generate need-based packages, with the help of these packages we can estimate actual utilization of amount using the probability distribution concept. After data visualization we observed that the gamma distribution fits the data as shown in Figure 4. The process of optimization depends on the difference between $T$ real medical benefits amount and $\phi$ computed medical benefits amount. The data-driven analysis using a probability distribution plot is performed for each group in the specified category. For each group, we evaluate the mean $\mu$ and standard deviation $\sigma$. It is observed that as the amount increases the number of employees availing that amount is reduced. This fact can be observed from the shape of the gamma distribution curve. When we get a value for a standard deviation for each group, we know this fact that standard deviation is used as a constant in the probability distribution. Within the gamma distribution, we can move towards the right side of the mean and can get estimates of how many employees

are getting out of pocket while availing healthcare services. All medical expenses equal and greater than $\mu + 2\sigma$ are not affordable for employees. These estimates let us find out the number of out of pocket employees in each category before and after optimization.The same concept is used for finding lower and upper bound for the premium amount for each group of all categories. We set the bracket from $\mu$ till $\mu + \sigma$ as the safe range for estimating the premium amount for the specified group. The members in one group can be moved to another group depending upon the amount that the employee has utilized. This optimization methodology introduces a model for employers and insurance companies. The complexity of these three algorithms is also not a limiting factor as the planning is only done once a year for deciding the packages for any enterprise.

## 4. Results and Analysis

The database used for medical benefit optimization methodology is depicted in Figure 5.

For medical benefit optimization methodology, we need to prepare data first. The table named facttable_data contains all the prepared columns for each employee. The table named as 'centers' contain centroids after the kmeans algorithm. Euclidean_distance table contains results after computing distance from all centroids. Member_center contains members of each center.

### *4.1. Data Preparation*

The data is prepared before the implementation of the Kmeans algorithm. The set of attributes providing details about the availed and provided services are shown in Table 3.

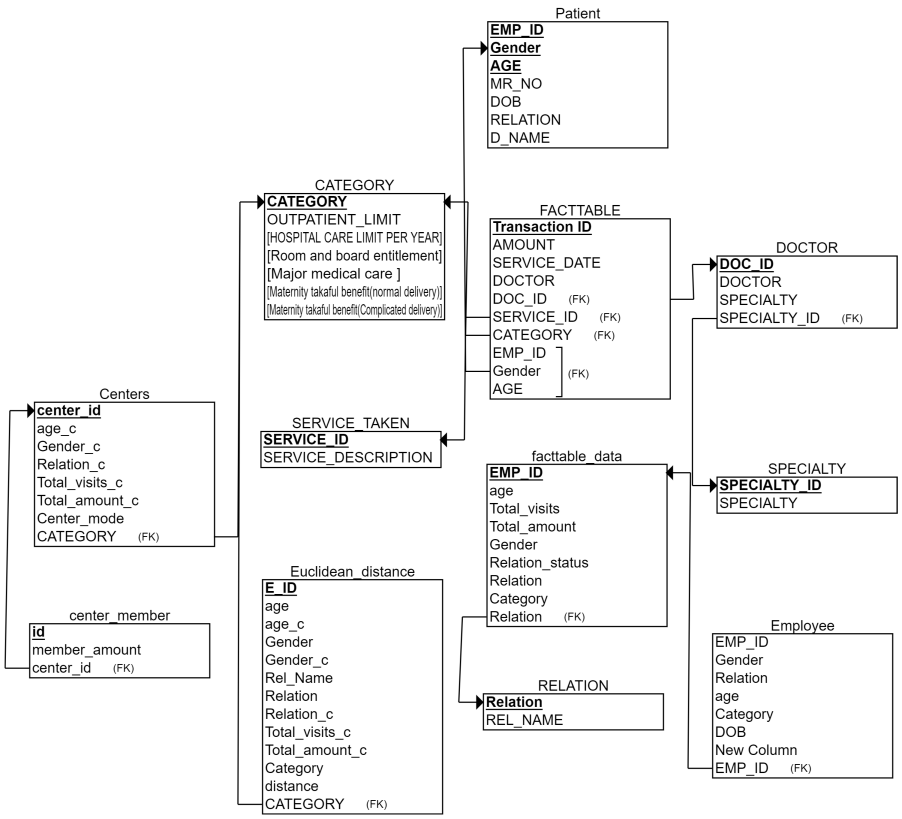


**Figure 5.** Relational database for medical benefit optimization methodology.

**Table 3.** Attributes.

| Name | Data Type |
|---|---|
| EMP_ID | varchar(255) |
| Age | Float |
| Total_visits | Float |
| Total_amount | Float |
| Marital_status | Float |
| Relation | Float |
| Gender | Float |
| Category | nvarchar(255) |

The queries which are applied on transactional data are provided below.

```
CREATE TABLE facttable_data (
EMP_ID varchar(255),
age float,
total_visits float,
total_amount float,
relation_status varchar(255),
relation float,
gender float,
CATEGORY nvarchar(255))
```

After creating table named facttable_data, we are going to insert values in this table by executing stored procedure 'prepareDataForKmeans' Appendix A and following queries. As it is already mentioned that FACTTABLE is the main table which contains all transactions of last five years.

```
INSERT INTO facttable_data (EMP_ID,
total_visits,
total_amount, relation_status)
Select EMP_ID, COUNT(EMP_ID)
as total_visits,
SUM(amount) as total_amount,
STRING_AGG(RELATION_ ,',')
as relation_status
from FACTTABLE GROUP BY EMP_ID
```

In the original data set, relation and gender both are of string data type, but we define them as floats for further processing. This is done by using the following query.

```
UPDATE facttable_data
SET facttable_data.age
= employee.AGE,
facttable_data.CATEGORY
= employee.CATEGORY,
facttable_data.gender =
(Case when employee.gender = 'M'
then 1 when employee.gender
= 'F' then 0 else 0 end)
FROM facttable_data
INNER JOIN employee
on facttable_data.EMP_ID
= employee.EMP_ID
```

For relation(status) we used 1 for married and 0 for single.

```
UPDATE facttable_data
SET facttable_data.relation =
(Case when
facttable_data.relation_status
like '%SON%'or
facttable_data.relation_status
like '%WIFE%'or
facttable_data.relation_status
like '%DAUGHTER%'or
facttable_data.relation_status
like '%HUSBAND%'  then 1
else 0 end )
```

The proposed methodology is implemented in eclipse using JAVA programming language. The screenshots of the computed values are provided in Figures 5–11.

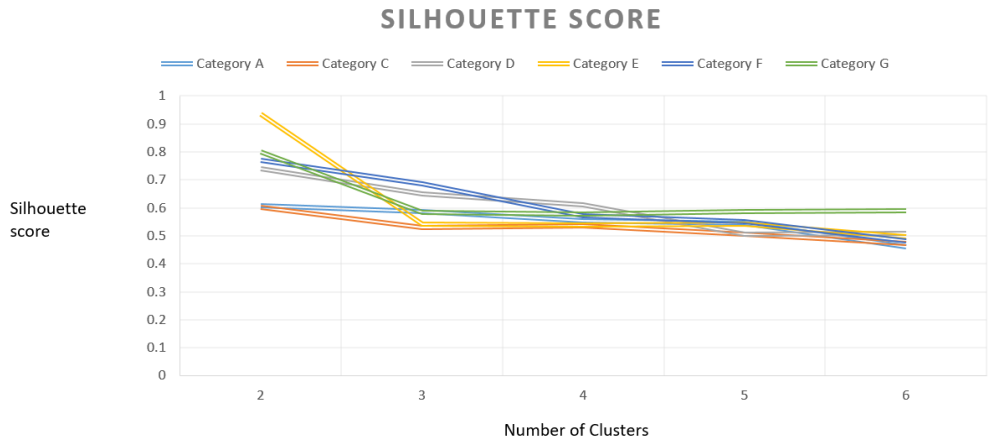


**Figure 6.** Determining Number of clusters.

Kmeans clustering algorithm is used for performing clustering. For Kmeans we need to enter the number of clusters and number of iterations as an input to the kmeans algorithm. We select the number of clusters by using the Silhouette score method. In Figure 6, it can be seen that peaks are at 2, which identifies the optimum number of cluster for all the categories in the provided data. Based on the Silhouette score, we kept the number of clusters as 2. The number of iterations is decided by observing the change in centroids values. Once the number of clusters and maximum iterations is decided. Then, need-based packages are generated for each category of employees. Let us first analyze the already in use packages for each category and then we discuss in detail the generated need-based packages for each category.

From Table 4, it can be seen that there is a single premium amount for each type of category. It seems that while designing such packages insurance companies did not consider the need of employees. The need of any person can be analyzed by different parameters like age, marital status, gender, amount and number of visits, etc. By considering above mentioned parameters, it can be seen that based on any person's need each category can be divided into two groups. The proposed methodology generates packages by dividing each category into further two classes based on the above-mentioned parameters.

**Table 4.** Category wise Annual Premium amount and number of employees.

| Category Type | Premium Amount (PKR) | Number of Employees |
|---|---|---|
| A | 39,036 | 170 |
| B | 17,466.6 | 5 |
| C | 31,922 | 985 |
| D | 13,075.57 | 252 |
| E | 26,309 | 1991 |
| F | 8862 | 1619 |
| G | 68,811 | 120 |

Table 5 depicts a need-based package for category A. The methodology further divides category A into two groups. It can be seen that five attributes are considered for defining the need of any employee age, gender, relation status(relationship status), amount, number of visits. The Sum amount is the sum of all employees amount in the first group. In category A, age is not the distinguishing attribute, all-male employees whose ages are greater than and equal to 30 and are married come in the first group. The premium amount for each employee in the first group is Rupees 55,205. All the female employees whose age is greater than and equal to 30 and are married come in the second group and the premium amount for each employee in this group is rupees 13,189. The total premium amount for category A $\alpha_A$ is six million six hundred thirty-six thousand fifty-six hundred and seventy rupees. The total number of employees in category A is 170. The main purpose of this analysis is to define the premium amount according to the need of employees. Out of 170 employees, not all the employees visit the hospital for taking healthcare services every year but still employer is paying the huge and same amount of premium for all the employees of category A. From Table 5, it can be seen that the sum amount is the sum of amounts of all employees in one group. We use equation 1 for computing the number of employees for both groups. The number of employees in generated packages is 64, which is less than 170. This analysis on five years data proves that 40% of total employees in category A visit hospital. So the premium amount is computed accordingly. Equation (3) is used for computing the total sum amount for all groups of one category.

**Table 5.** Need based Package for Category A.

| Age | Gender | Relation Status | Sum Amount | Amount | Visit |
|---|---|---|---|---|---|
| 30.0 | 0.0 | 1.0 | 2,484,225 | 55,205.0 | 2700 |
| 30.0 | 1.0 | 1.0 | 250,591 | 13,189.0 | 342 |

In Table 6, it can be seen that all employees whose ages are less than and equal to 26, male and married are in group I. For all employees whose ages are greater than and equal to 31, female and married, there is a total of 10 employees in this category and only 4 of them visited the hospital for availing health services. Again, 40% of employees are using health care services. Premium amount for each group in category B is computed using Equation (3).

**Table 6.** Need based Package for Category B.

| Age | Gender | Relation Status | Sum Amount | Amount | Visit |
|---|---|---|---|---|---|
| 26.0 | 0.0 | 1.0 | 66,212 | 33,106.0 | 116.0 |
| 31.0 | 1.0 | 1.0 | 40,518 | 20,259.0 | 34.0 |

Table 7 depicts the package for category C. Group I in category C consists of all employees whose ages are less than and equal to 28, female and married. The annual premium for the group I is rupees 17,208. In group II, all employees whose ages are

greater and equal to 29, male and married. The premium amount for each employee in this group of category C is 58,784 rupees. The number of employees in group I are 318 and in group 2 are 150. The original number of employees in category C is 985. The number of employees who avail health services is 468. In category C, almost 50% employees are availing healthcare services.

**Table 7.** Need based Package for Category C.

| Age | Gender | Relation Status | Sum Amount | Amount | Visit |
|---|---|---|---|---|---|
| 28.0 | 1.0 | 1.0 | 2,460,744 | 17,208.0 | 3003.0 |
| 29.0 | 0.0 | 1.0 | 18,693,312 | 58,784.0 | 21,306.0 |

Table 8 depicts the package for category D, employees whose ages are less than and equal to 28, male, married are in a group I and all employees whose ages are greater than and equal to 29, male and single are in group II. The original number of employees in category D is 252. Using equation1, the numbers of employees in both groups are computed for category D. Forty-nine employees are in a group I and one hundred and fifty-seven employees are in group II. In category D, 85% employees are availing healthcare services.

**Table 8.** Need based Package for Category D.

| Age | Gender | Relation Status | Sum Amount | Amount | Visit |
|---|---|---|---|---|---|
| 28.0 | 0.0 | 1.0 | 2,488,048 | 54,088.0 | 3036.0 |
| 29.0 | 0.0 | 0.0 | 1,697,484 | 10,812.0 | 1727.0 |

Table 9 depicts the package for category E, all employees whose ages are less than and equal to 28, male and married are in group I and all employees whose ages are greater than and equal to 29, female and married are in group II. The premium amount for each employee in group I is 54,665 rupees and the premium amount for each employee in group II is 13,802. The original number of employees in category E is 1991. In category E, 50% of employees are availing health care services.

**Table 9.** Need based Package for Category E.

| Age | Gender | Relation Status | Sum Amount | Amount | Visit |
|---|---|---|---|---|---|
| 28.0 | 0.0 | 1.0 | 30,557,735 | 54,665.0 | 36,894.0 |
| 29.0 | 1.0 | 1.0 | 4,182,006 | 13,802.0 | 5151.0 |

In Table 10 package for the category, F is shown, all employees whose ages are 28, and are male and married are in one group and all employees whose ages are 26, and are female and single are in group II. In this category, only 40% are using healthcare services. The premium amount for each employee in group I is 41,191 and for each employee in group II is 7,768.

**Table 10.** Need based Package for Category F.

| Age | Gender | Relation Status | Sum Amount | Amount | Visit |
|---|---|---|---|---|---|
| 28.0 | 0.0 | 1.0 | 10,750,851 | 41,191.0 | 13,572.0 |
| 26.0 | 1.0 | 0.0 | 8,171,936 | 7768.0 | 10,520.0 |

Table 11 depicts the package for category G. Employees who are married, male and aged less than and equal to 30 are in group I and all-female employees with age greater than and equal to 30, married are in group II. The premium amount for each employee in group I is 96,912 rupees and the premium amount for each employee in group II is 19,656.

The total number of employees in category G is 1619. About 82% employees are using health services.

**Table 11.** Need based Package for Categoy G.

| Age | Gender | Relation Status | Sum Amount | Amount | Visit |
|---|---|---|---|---|---|
| 30.0 | 0.0 | 1.0 | 3,779,568 | 96,912.0 | 4046.0 |
| 30.0 | 1.0 | 1.0 | 1,631,448 | 19,656.0 | 1826.0 |

The current medical expense for 5143 employees and the total computed medical expense in rupees are depicted in Figure 7, which is the output of Algorithm 2.

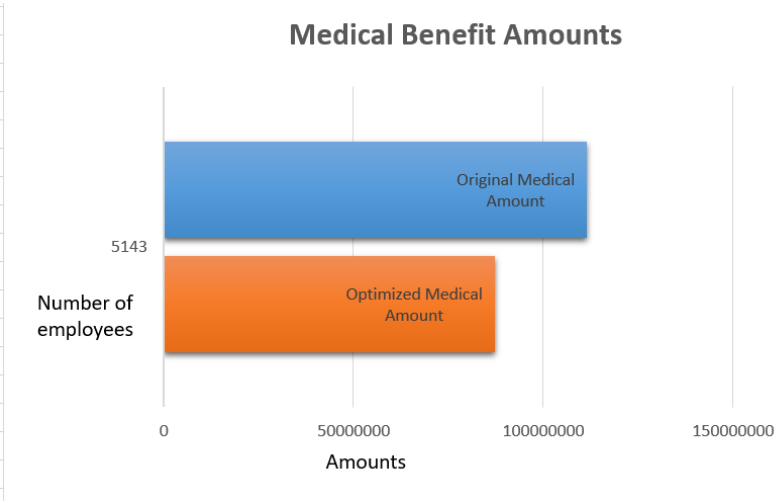


**Figure 7.** Total Medical Amount Comparison.

Now there is a need for data-driven analysis. The probability distribution which fits our data is 'Gamma'. Let us first consider Category A. The probability distribution for each group in category A is shown in Figures 8 and 9.

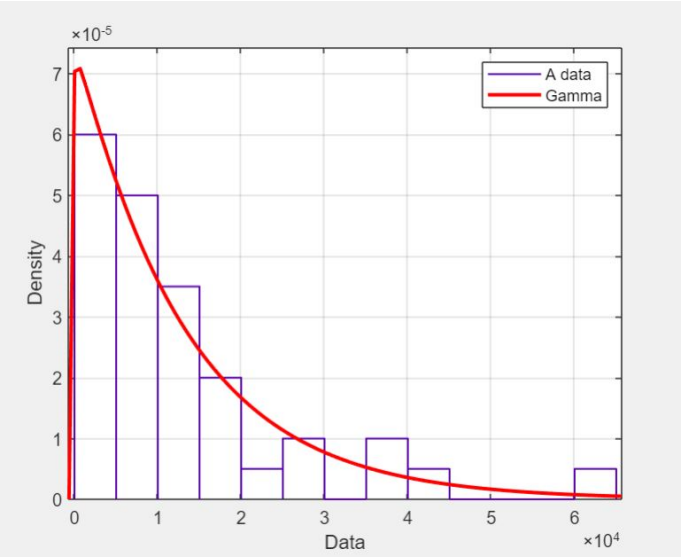


**Figure 8.** HISTFIT for group 1 in Category A.

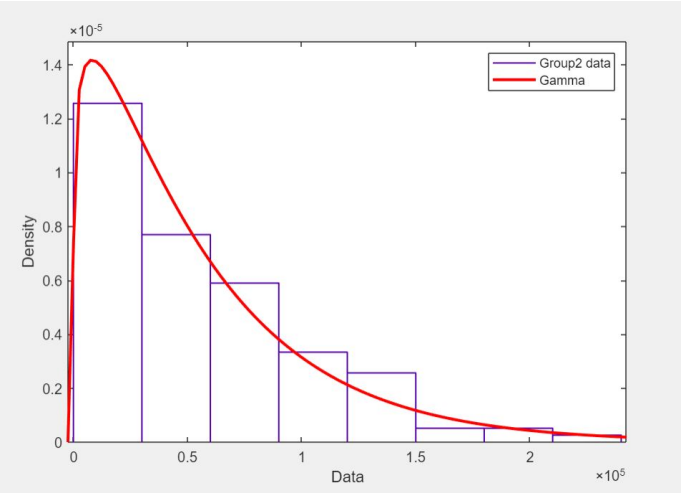


**Figure 9.** HISTFIT for group 2 in Category A

We combine both groups in category A and compute the total amount for this category. $\mu$ is mean, $\phi$ is the original medical amount and $\sigma$ is the standard deviation. For data-driven analysis, we used a gamma probability distribution model. The parameters of gamma distribution for both groups of category A are shown in Table 12.

**Table 12.** Parameter Computation for Category A.

| Group | a(Shape) | b(Scale) | $\mu$ | $\sigma$ |
|---|---|---|---|---|
| 1 | 1.0165 | 12,868.7 | 13,081 | 12,974 |
| 2 | 1.17172 | 47,418.5 | 55,561.2 | 51,328 |

In category B there are four employees in group I and only one employee is in group II. In category C there are two groups.

In Figures 10 and 11, a histogram with a distribution fit(histfit) is plotted for group I and group II in category C. In group I, the mean value is around 10,000 and standard deviation $\sigma$ is 40,000. In group II, it can be seen that mean value $\mu$ is around 25,000 and standard deviation $\sigma$ is 19,690.

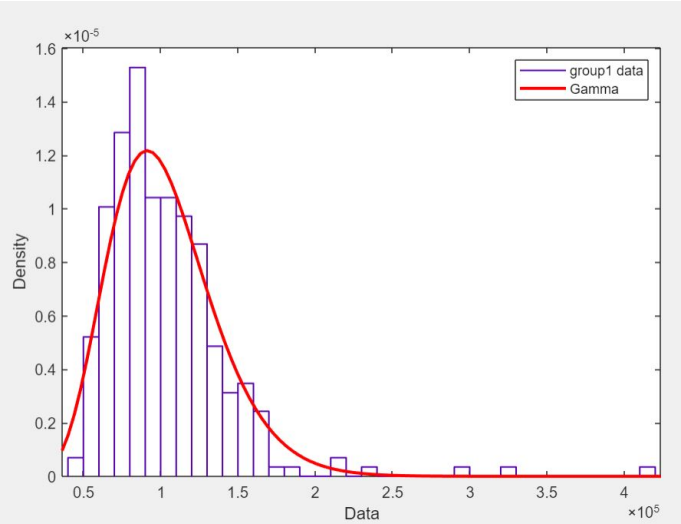


**Figure 10.** HISTFIT for group 1 in Category C.

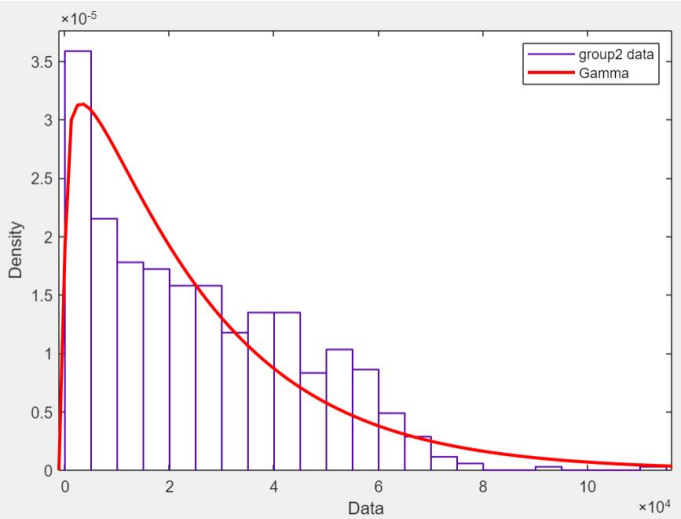


**Figure 11.** HISTFIT for group 2 in Category C.

The parameters $a$ and $b$ for gamma distribution are computed for both groups of category C as shown in Table 13.

**Table 13.** Parameter Computation for Category C.

| Group | a(Shape) | b(Scale) | $\mu$ | $\sigma$ |
|---|---|---|---|---|
| 1 | 8.91083 | 11,528.9 | 10,2732 | 34,414.82 |
| 2 | 1.1381 | 22,539.6 | 25,652.4 | 24,045.73 |

In group I, of category D, the mean $\mu$ is around 54,087, and $\sigma$ is around 42,686. In group II, the mean $\mu$ is 10,812 and the standard deviation $\sigma$ is 11,105.67 as shown in Table 14. The probability distribution model for groups of category D is depicted in Figures 12 and 13.

**Table 14.** Parameter Computation for Category D.

| Group | a(Shape) | b(Scale) | $\mu$ | $\sigma$ |
|---|---|---|---|---|
| 1 | 1.26108 | 42,889.9 | 54,087.8 | 48,164.51 |
| 2 | 0.947823 | 1407.2 | 10,812 | 11,105.67 |

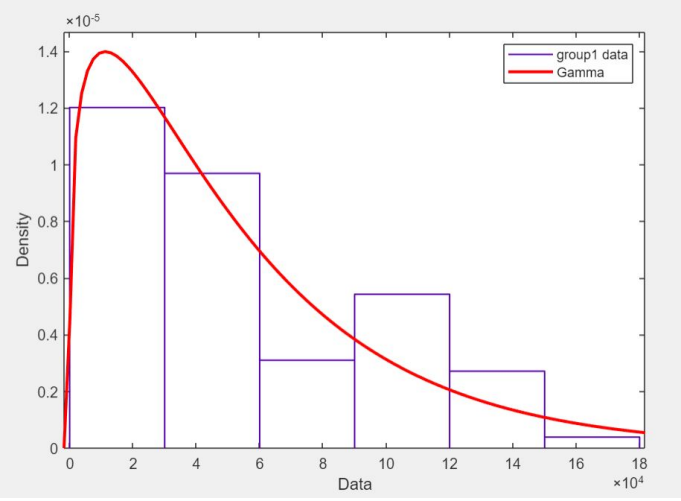


**Figure 12.** HISTFIT for group 1 in Category D.

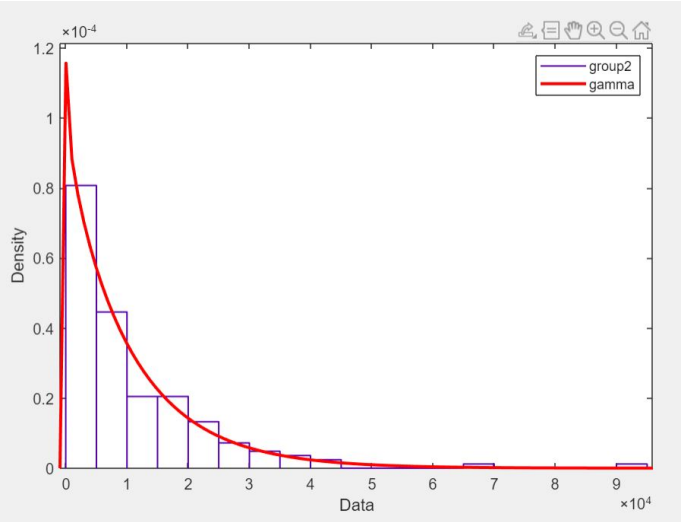


**Figure 13.** HISTFIT for group 2 in Category D.

Similarly, gamma parameters for all categories E, F and G are computed as shown in Table 15.

**Table 15.** Parameters estimation for Categories E, F and G.

| Category | Group | a(Shape) | b(Scale) | $\mu$ | $\sigma$ |
|---|---|---|---|---|---|
| E | 1 | 1.26108 | 42,889.9 | 54,087.8 | 48,164.51 |
| | 2 | 0.761365 | 18,175 | 13,837.8 | 15,858.85 |
| F | 1 | 0.944072 | 8228.13 | 7767.95 | 7994.73 |
| | 2 | 1.12469 | 36,624.5 | 41,191.3 | 38,840.83 |
| G | 1 | 4.60034 | 21,066.4 | 96,912 | 45,183 |
| | 2 | 1.31399 | 14,959 | 19,656.1 | 17,147.48 |

Table 16 depicts the number of out of pocket employees in the current medical benefits amount and out of pocket employees in optimized benefit amount using Gamma distribution. The term out of pocket is used to explain how many employees are spending additional amounts for availing healthcare services and getting financially overburdened.

**Table 16.** Out of Pocket Employees in Current Medical Benefit and in Optimized Amount.

| Category | Out of Pocket Employees in Current Amount | Out of Pocket Employees in Optimized Amount |
|---|---|---|
| A | 45 | 33 |
| B | 30 | 10 |
| C | 54 | 33 |
| D | 47 | 15 |
| E | 57 | 31 |
| F | 52 | 12 |
| G | 30 | 17 |

In Figure 14, outlier are detected and these outliers are basically depicting the out of pocket employees in each category.

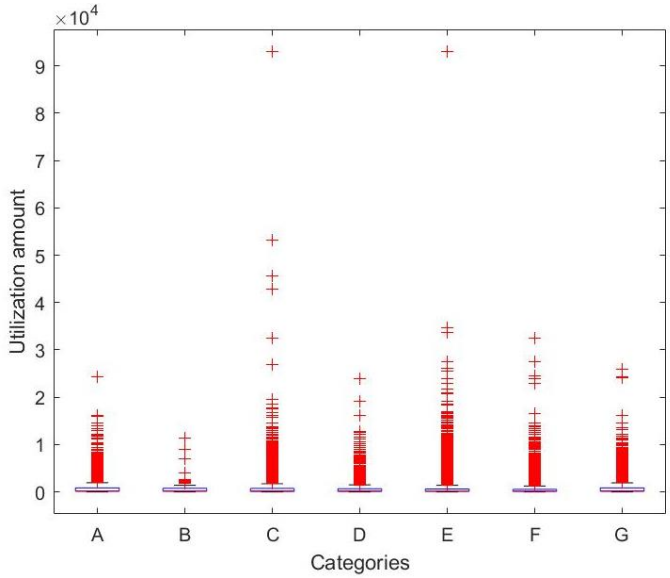


**Figure 14.** Out of pocket employees in original data.

In Figure 15, out of pocket employees before and after medical benefit optimization are shown. The y-axis on the left side is depicting the number of out-of-pocket employees and the y-axis on the right side is depicting the number of out of pocket employees after optimization. The x-axis is depicting categories used for insurance coverage.

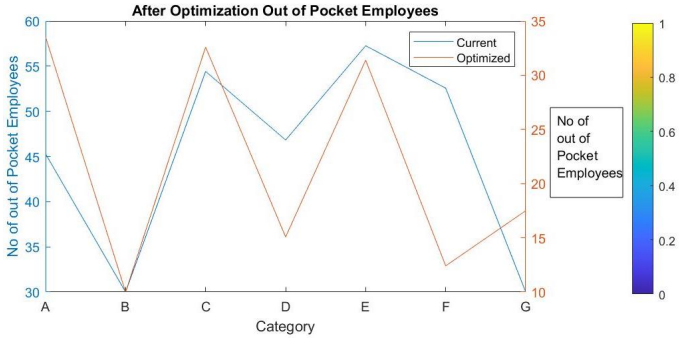


**Figure 15.** Number of Out of Pocket employees before and after Optimization.

In Table 17, the gamma distribution is depicted for each group in all categories. If we move to the right of gamma distribution each time by adding $\mu$ to $\sigma$. As we know the concept of scaling in distributions, we multiply constant with standard deviation $\sigma$ as depicted in Table 9. It can be observed that the value of the probability density function decreases as we move right to the $\mu$. By using these observations we defined lower and upper bound using gamma distribution, for already generated need-based packages. From $\mu$ till $\mu + \sigma$, we can set affordable premium amounts for each package in each category. However, if we move to or greater than $\mu + 2\sigma$, then there is the chance of monetary losses. From generated gamma distribution, we can define lower bound and upper bound for each group of every category. Deciding when to move from lower to upper bound is based on attributes like age, gender and marital status which are defining needs of the employees.

Table 18 is showing lower and upper bounds for setting premium amounts for each group of all categories. Employers/organizations can decide on the affordable premium amount within defined bounds exceeding these amounts can result in financial losses.

**Table 17.** Optimization of total medical benefit for each Category A,B,C,D,E,F and G.

| Category | Group No | $\mu$ | $\sigma$ | $\mu+\sigma$ | $\mu+2\sigma$ | $\mu+3\sigma$ | $P(\mu+\sigma)$ | $P(\mu+2\sigma)$ | $P(\mu+3\sigma)$ |
|---|---|---|---|---|---|---|---|---|---|
| A | 1 | 13,081 | 12,974 | 26,055 | 39,029 | 52,004 | $1.05\times10^{-5}$ | $3.84\times10^{-6}$ | $1.41\times10^{-6}$ |
| | 2 | 55,561 | 51,328 | 106,889 | 158,218 | 209,547 | $2.75\times10^{-6}$ | $9.9\times10^{-7}$ | $3.5\times10^{-7}$ |
| C | 1 | 102,732 | 34,414 | 137,146 | 171,560 | 205,974 | $5.70\times10^{-6}$ | $1.60\times10^{-6}$ | $3.60\times10^{-7}$ |
| | 2 | 25,652 | 24,045 | 49,698 | 73,743 | 97,789.59 | $5.80\times10^{-6}$ | $2.11\times10^{-6}$ | $7.50\times10^{-7}$ |
| D | 1 | 54,088 | 48,164 | 102,252 | 150,416 | 198,581 | $2.90\times10^{-6}$ | $1.07\times10^{-6}$ | $3.70\times10^{-7}$ |
| | 2 | 10,812 | 11,105 | 21,917 | 33,023 | 44,129 | $1.20\times10^{-5}$ | $4.44\times10^{-6}$ | $1.65\times10^{-6}$ |
| E | 1 | 54,087 | 48,164 | 102,252 | 1,504,16 | 198,581 | $3.10\times10^{-6}$ | $1.03\times10^{-6}$ | $3.26\times10^{-7}$ |
| | 2 | 13,837 | 15,859 | 29,697 | 45,555 | 61,414 | $7.80\times10^{-6}$ | $2.97\times10^{-6}$ | $1.15\times10^{-7}$ |
| F | 1 | 7768 | 7995 | 15,762 | 23,757 | 31,752 | $1.66\times10^{-5}$ | $6.16\times10^{-6}$ | $2.29\times10^{-6}$ |
| | 2 | 41,191 | 38,841 | 80,032 | 118,873 | 157,713 | $3.59\times10^{-6}$ | $1.30\times10^{-6}$ | $4.60\times10^{-6}$ |
| G | 1 | 96,912 | 45,184 | 142,096 | 187,279 | 232,463 | $4.03\times10^{-6}$ | $1.27\times10^{-6}$ | $3.20\times10^{-7}$ |
| | 2 | 19,656 | 17,147 | 36,803 | 53,950 | 71,097 | $8.45\times10^{-6}$ | $3.03\times10^{-6}$ | $1.05\times10^{-6}$ |

**Table 18.** Optimization of total medical benefit for each Category A,B,C,D,E,F and G.

| Category | Group # | Lower Bound | Upper Bound |
|---|---|---|---|
| A | 1 | 13,081 | 26055.4 |
| | 2 | 55,561 | 1,06,889 |
| C | 1 | 1,02,732 | 137146 |
| | 2 | 25,652 | 49,698 |
| D | 1 | 54,087 | 1,02,252 |
| | 2 | 10,812 | 21,917 |
| E | 1 | 54,087 | 1,02,252 |
| | 2 | 13,838 | 29,696 |
| F | 1 | 7768 | 15,763 |
| | 2 | 41,191 | 80,032 |
| G | 1 | 96,912 | 142,096 |
| | 2 | 19,656 | 36,803 |

The total amount is computed by first using the mean $\mu$ as the premium amount for each group in each category. Then $\mu+\sigma$ as the premium amount for each group of each category. After that $\mu+2\sigma$ is computed for each group of all categories and used as a premium amount. Table 19 depicts the total amount computed using gamma distribution for categories A, B, C, D, E, F and G. The last three columns in Table 19, are probability density functions for $\mu+\sigma$, $\mu+2\sigma$ and $\mu+3\sigma$. Category B, D and F, we can see currently allocated medical amount is less than the need of the employees who are in these categories. The proposed methodology solves the real-life problem of insurance package design. The insurance packages are designed by learning from historical medical records. This data-driven analysis provides an easy way to compute out of pocket employees and to define lower/upper bound for the premium amount for all groups of every category.

**Table 19.** Optimization of total medical benefit for each Category A,B,C,D,E,F and G.

| Category | $\phi$ | $\beta$ | $\mu + \sigma$ | $\mu + 2\sigma$ | $\mu + 3\sigma$ |
|---|---|---|---|---|---|
| A | 2,722,499 | 6,636,057 | 14,224,155 | 20,743,556 | 27,262,958 |
| B | 126,989 | 87,333 | 231,184 | 335,380 | 439,575 |
| C | 21,274,512 | 31,442,546 | 72,638,057 | 97,809,596 | 122,981,136 |
| D | 4,347,796 | 3,295,044.144 | 12,159,063 | 17,871,775 | 23,584,488 |
| E | 35,067,873 | 52,380,635 | 161,637,982 | 237,239,625 | 312,841,268 |
| F | 19,504,878 | 14,347,412 | 60,059,019 | 89,927,607 | 119,796,195 |
| G | 4,113,456 | 8,257,405 | 11,036,095 | 15,387,108 | 19,738,122 |

After analyzing each category, we compute the overall amount by moving to the right on the probability distribution curve as depicted in Table 19. The organization can adjust the amount of their medical benefit by using our proposed methodology. The need of the employees can be changed, so our model enables employers to move an employee from one group to another.

#### 4.1.1. Case 1: $T < \phi$

In this case no optimization is required.

#### 4.1.2. Case 2: $T > \phi$

There can be a case in which $T$ is greater than the $\phi$ optimized amount, then there is a need of optimization. The scale parameter $\sigma$ is multiplied two or three times and then added to the $\mu$, according to the scenario.

#### 4.1.3. Case 3: Need of Employees Changes with Time

There can be a case in which employee healthcare needs change due to multiple reasons for example, he/her gets married, requirements with ageing increase, etc. Thus, the employee group changes and his insurance package is accordingly updated. In such cases, probability distributions will be generated again for each category.

#### 4.1.4. Case 4: $T = \phi$

The ideal case is the case where $T$ the current medical benefit is equal to the optimized medical benefit $\phi$. In such cases, no more optimization is required.

#### 4.1.5. Observation

Figure 16, depicts the real number of employees before the optimization methodology is applied in the left y-axis, and the number of employees in each category after medical benefit optimization methodology is applied is depicted by the right y-axis. The number of employees in each category is reduced after the optimization methodology is applied. This proves that not all insured employees are using healthcare services. The observation is that organization is paying extra premium amounts as not all employees are using healthcare services frequently.

It can be seen from Figure 17 that the allocated benefit amount is not fully utilized.The x-axis is representing categories and the y-axis is depicting the percentage of amount utilization in each category.

In Figure 18, it can be seen that after optimization total amount for each category is reduced. It is observed that in the three categories B, D and F, there are out of pocket employees. They paid an additional amount from their pocket for using healthcare services.

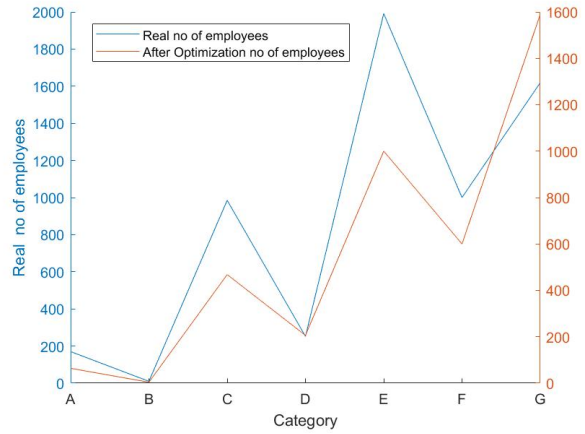


**Figure 16.** Number of employees in each category.

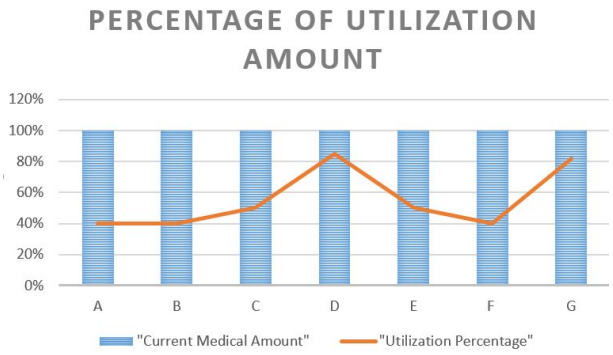


**Figure 17.** Percentage of Utilization Amount in each category.

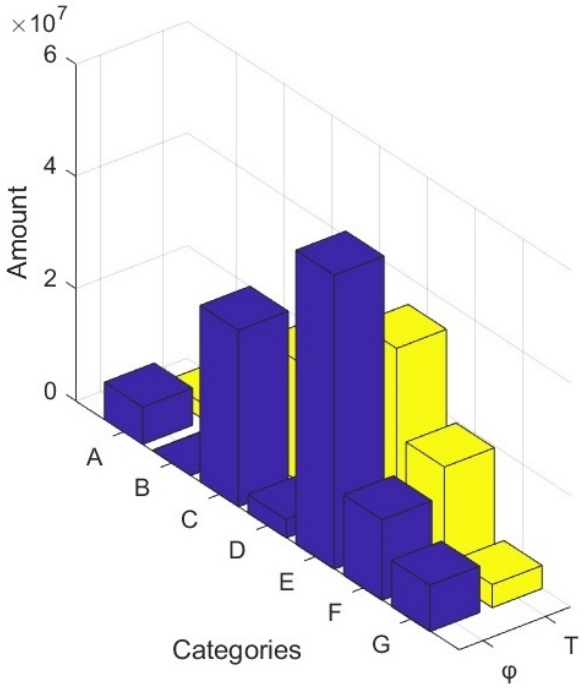


**Figure 18.** Comparison of Utilization Amount in each category.

Figure 19 depicts the data-driven analysis of amounts for each category using a gamma distribution. After the optimization, it is necessary to compute the percentage of optimization achieved. Optimization percentage is computed by using Equation (6):

$$P = \frac{(\phi - T)}{T} * 100 = \frac{29288429.57}{116446432.6} * 100 = 25\% \quad (6)$$

$\phi$ is computed optimized amount and $T$ is current total amount, both amounts are shown in Figure 12. The $T$ is optimized to $\phi$ by 25% with the help of proposed methodology. This optimization is achieved as we have divided each category into two groups, each group has different sum amount and different premium amounts. By using these dynamic amounts for each category, need based packages are derived.

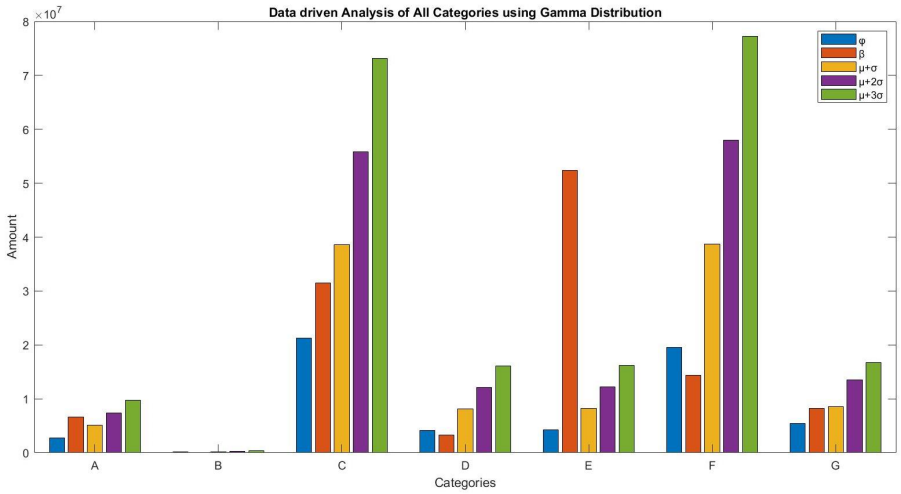


**Figure 19.** Data Driven Analysis using Gamma Distribution.

## 5. Discussion

The techniques and studies which are discussed in Section 2, shows a trend of applying machine learning techniques and gaining gradual recognition and acceptance for solving insurance-related issues in the healthcare industry. The focus of most of the research is more on cost efficiency and fraud detection. The designs are based on prediction, analysis and evaluation of healthcare services, operations and resources. It is observed that the data-driven analysis using probability distributions is the most helpful platform for analyzing existing loopholes. Optimization is an interdisciplinary term, which combines mathematics, computer science, economics and engineering. There is a pervasive requirement of the applicability of optimization approaches to healthcare-related problems. The optimization approaches improve the provision of healthcare services by making them more cost-effective and efficient. It is a good practice to perform experimentation, using different values for the parameters considered in the optimization problem, to verify the robustness of the optimization results. For this purpose, intensive experimentation is performed on medical benefits amount in a considered case study using the concepts of data clustering and a probability distribution model. After the detailed literature review in Section 2, we observed that there is a requirement of an efficient methodology to generate insurance packages that must not be designation specific rather evenly distributed fairly in all employees of an organization. None of the existing methodologies are focused on generating insurance packages and optimization of medical benefits using machine learning techniques. Our proposed methodology use machine learning concepts and perform data-driven analysis of original transactional data using probability distribution concepts. Table 20 shows the novelty of our proposed methodology.

**Table 20.** Comparison with Existing Techniques.

| ML Related Researches in Insurance Industry | Type of Research | Comparison |
| --- | --- | --- |
| [8,12,25–28] | Fraud Detection | All of these researches are proposing methodologies for detecting fraud in insurance industry. |
| [3,6,13,19,21,22,29] | Premium amount in auto insurance | These researches focus on risk functions premium calculation, automobile premium computation based on user driving pattern etc |
| [30–33] | Customer related | All these researches are related to customer management. |
| Proposed Methodology | Medical benefit optimization | We observe that in the last decade focus of the researcher is more on fraud detection, risk prediction and customer management. None of the research focus on the medical benefit optimization. The main goal of our research is the medical benefit optimization from need based packages. |

This data-driven analysis is helpful in ensuring longterm sustainability of the insurance programs. The main purpose of using Kmeans is to get a group of all similar employees in each category. Further processing on centroids is performed to get better premium amounts for packages. Subsequently, the gamma probability distribution is applied to optimize the amount according to the requirement. There is a chance that any employee's group can be changed, and accordingly, the amount can be optimized. The probability distribution for each category is computed again.

## 6. Practical Implications

Healthcare expenditure is continuously rising, especially in developing and low-income countries. According to the World Health Organization's report published in 2016, the annual rise in healthcare costs in developing countries is around 6% greater than 4% of the developed countries.

In Pakistan, unluckily, the healthcare sector is among the most neglected sectors and is spending only 3% of its gross domestic product (GDP) on the education sector, health, and nutrition. Since the last decade, the government allocates 0.5pc to 0.8pc of its GDP for the health sector, less than the WHO benchmark that is 6pc of GDP. According to the World Bank report (April 2017), this figure is significantly less than other countries.

In most countries, including Pakistan, the government has just initiated medical support programs through several national-level initiatives. One of these initiatives is the establishment of the Prime Minister Task Force on IT and Telecom (in 2018) to lay down the foundation of data standards and annotations for incorporating the improved plans in healthcare service delivery to the common person. The Sehat card scheme was recently introduced in Pakistan to provide need-based insurance packages. The proposed methodology can be adopted to improve the provision of health insurance in healthcare institutions to control their overall expenditures. The designed methodology is beneficial for insurance companies as well as for enterprises.

## 7. Limitations

With the increase in data size, cloud computing will be required for processing. Although, the complexity of proposed three algorithms is not a limiting factor as the planning is only done once a year for deciding the packages for any enterprise. We have used age, gender, marital status, relation, amount and number of visits during the design of need based insurance packages. As more attributes are added analysis can produce

more cost effective packages. So the proposed methodology can be extended by adding more features that can contribute to the evaluation of employee needs. It would reflect better and improved performance if applied for a wider perspective. We received data in a raw form and proper preprocessing of data was required before implementation. The data acquisition from well-reputed hospitals is a troublesome task and is the main hurdle in the implementation of this type of methodology.

## 8. Conclusions and Future Work

In the healthcare industry, there is a dire need to focus on health insurance issues and to introduce efficient cost-effective techniques for fair distribution of insurance facilities. The linkage of healthcare benefits with employee roles in the organization is one of the critical problems. There is a need to replace the current strategies with methodologies that can ensure need-based healthcare benefits to the employees. This will not only minimize the chances of fraud/misutilization of healthcare benefits but will also enhance the sense of health security among the employees irrespective of their grades or designations. The existing studies are based on practices and strategies, emphasized funds prediction, customer management and estimation of insurance premium amounts. Our proposed methodology generated need-based packages using a machine learning model based on K_means clustering. With the help of this model, we have computed the optimum premium amount. The data-driven analysis based on the probability distribution model is used for analyzing the adjustment of premium amounts within each package depending upon the available financial resources. The proposed methodology is validated via five years' transactional data of employees of a large enterprise. The results indicate that the medical premium amount is optimized by 25% of the current benefit amounts. Therefore, if adopted, it will not only allow employers and insurance companies to design suitable insurance schemes for the provision of healthcare benefits but will also prevent financial losses in the long run. By such extensive analysis, we can understand the power of insurance industry policies which can save millions of lives but at the same time can result in a disaster. In this research, we have used only data from one enterprise, but shall be applicable to other similar enterprises, as the algorithm easily scales for several employees and the number of clusters.

**Author Contributions:** Conceptualization, I.M. and S.A.K.; methodology, I.M. and S.A.K.; software, I.M.; validation, I.M. and S.A.K.; formal analysis, I.M.; investigation, S.A.; resources, I.M.; data curation, S.A.K.; writing—original draft preparation, I.M.; writing—review and editing, F.H., W.H.B., R.R., F.K.; visualization, I.M.; supervision, S.A.K.; project administration, I.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Acknowledgments:** The authors thank Shifa International Hospital, Islamabad, Pakistan for providing the employee's insurance data to validate our proposed methodology. This work is part of PM Task Force on IT and Telecom initiative that genuine people can easily get medical benefits from the government level initiative and is presented as an initial proposed methodology to the taskforce for fraud prevention and package generation.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

### Stored Procedure

```
CREATE PROCEDURE prepareDataForKmeans
(
@category nvarchar(255)
```

```
)
AS
BEGIN
DECLARE @amountAvg float
DECLARE @ageAvg float
DECLARE @visitAvg float
DECLARE @relationAvg float
DECLARE @genderAvg float
DECLARE @amountStdev float
DECLARE @ageStdev float
DECLARE @visitStdev float
DECLARE @relationStdev float
DECLARE @genderStdev float
SET @amountAvg = (
SELECT AVG(total_amount)
FROM facttable_data
);
SET @visitAvg = (
SELECT AVG(total_visits)
FROM facttable_data
);
SET @ageAvg = (
SELECT AVG(age)
FROM facttable_data
);
SET @relationAvg = (
SELECT AVG(relation)
FROM facttable_data
);
SET @genderAvg = (
SELECT AVG(gender)
FROM facttable_data
);
SET @amountStdev = (
SELECT STDEV(total_amount)
FROM facttable_data
);
SET @visitStdev = (
SELECT STDEV(total_visits)
FROM facttable_data
);
SET @ageStdev = (
SELECT STDEV(age)
FROM facttable_data
);
SET @relationStdev = (
SELECT STDEV(relation)
FROM facttable_data
);
SET @genderStdev = (
SELECT STDEV(gender)
FROM facttable_data
);
select EMP_ID, (age - @ageAvg) / @ageStdev as age,
```

```
(total_visits - @visitAvg) / @visitStdev as total_visits,
(total_amount - @amountAvg) / @amountStdev as total_amount,
(relation - @relationAvg) / @relationStdev as relation,
(gender - @genderAvg) / @genderStdev as
gender from facttable_data where CATEGORY = 'A'
END
```

## References


1. Rao, S. Health insurance: Concepts, issues and challenges. *Econ. Political Wkly.* **2004**, *39*, 3835–3844
2. Radermacher, R.; Dror, I.; Noble, G. Challenges and strategies to extend health insurance to the poor. In *Protecting the Poor: A Microinsurance Compendium*; ILO: Geneva, Switzerland, 2006.
3. Ding, K.; Lev, B.; Peng, X.; Sun, T.; Vasarhelyi, M.A. Machine learning improves accounting estimates: Evidence from insurance payments. *Rev. Account. Stud.* **2020**, *25*, 1098–1134. [CrossRef]
4. Groba, C.; Sartal, A.; Vázquez, X.H. Solving the dynamic traveling salesman problem using a genetic algorithm with trajectory prediction: An application to fish aggregating devices. *Comput. Oper.* **2015**, *56*, 22–32. [CrossRef]
5. Dhieb, N.; Ghazzai, H.; Besbes, H.; Massoud, Y. A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement. *IEEE Access* **2020**, *8*, 58546–58558. [CrossRef]
6. Kowshalya, G.; Nandhini, M. Predicting fraudulent claims in automobile insurance. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*; IEEE: Piscataway, NJ, USA, 2018; pp. 1338–1343.
7. Grize, Y.-L.; Fischer, W.; Lützelschwab, C. Machine learning applications in nonlife insurance. *Appl. Stoch. Model. In Business Ind.* **2020**, *36*, 523–537. [CrossRef]
8. Itri, B.; Mohamed, Y.; Mohammed, Q.; Omar, B. Performance comparative study of machine learning algorithms for automobile insurance fraud detection. In *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–4.
9. Hossain, S.S.M.R.; Salman, S.M. Implementation challenges of the mandatory health insurance scheme. *Bull. Natl. Res. Centre* **2019**, *43*, 151. [CrossRef]
10. Krasheninnikova, E.; García, J.; Maestre, R.; Fernández, F. Reinforcement learning for pricing strategy optimization in the insurance industry. *Eng. Appl. Artif. Intell.* **2019**, *80*, 8–19. [CrossRef]
11. Sundarkumar, G.G.; Ravi, V. A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. *ENgineering Appl. Artif. Intell.* **2015**, *37*, 368–377. [CrossRef]
12. Subudhi, S.; Panigrahi, S. Use of optimized fuzzy c-means clustering and supervised classifiers for automobile insurance fraud detection. *J. King Saud-Univ.-Comput. Inf. Sci.* **2020**, *32*, 568–575. [CrossRef]
13. Challa, A. Insurance Models and Risk-Function Premium Principle. 2012. Available online: https://www.semanticscholar.org/paper/Insurance-models-and-risk-function-premium-Challa/1f19f01beafdb3c451861c9275901e68ab3a0377 (accessed on 20 May 2021).
14. Mazzoccoli, A.; Naldi, M. The expected utility insurance premium principle with fourth-order statistics: Does it make a difference? *Algorithms* **2020**, *13*, 116. [CrossRef]
15. Bianchi, D.; Büchner, M.; Tamoni, A. Bond risk premiums with machine learning. *Rev. Financ. Stud.* **2020**, *34*, 1046–1089. [CrossRef]
16. Gan, L.; Wang, H.; Yang, Z. Machine learning solutions to challenges in finance: An application to the pricing of financial products. *Technol. Forecast. Soc. Chang.* **2020**, *153*, 119928. [CrossRef]
17. Henckaerts, R.; Côté, M.-P.; Antonio, K.; Verbelen, R. Boosting insights in insurance tariff plans with tree-based machine learning methods. *North Am. Actuar. J.* **2020**, *25*, 255–285. [CrossRef]
18. Kuo, K.; Lupton, D. Towards explainability of machine learning models in insurance pricing. *arXiv* **2020**, arXiv:2003.10674.
19. Kröger, V.; Nordström, R. *Expected Individual Insurance Cost Based on Driving Pattern: Machine Learning Methods Using Telemetric Data*; Digitala Vetenskapliga Arkivet: Uppsala, Sweden, 2020.
20. Gan, G. Application of data clustering and machine learning in variable annuity valuation. *Insur. Math. Econ.* **2013**, *53*, 795–801. [CrossRef]
21. Spedicato, G.A.; Dutang, C.; Petrini, L. Machine Learning Methods to Perform Pricing Optimization. A Comparison with Standard Glms. 2018. Available online: https://www.semanticscholar.org/paper/Machine-Learning-Methods-to-Perform-Pricing-A-with-Spedicato-Dutang/6a51b2c8557acde21389193ea86f3d00482036c3 (accessed on 20 May 2021).
22. Collopy, F.; Nard, C.A.; Amin, H.S.; Turocy, G.; Takieh, S.V.S.; Krosky, R.C.; Noonan, D.; Narvaez, G.A.; Asquith, B. Dynamic Insurance Rates. US Patent App. 12/536,999, 27 May 2010.
23. Abdalkareem, Z.A.; Amir, A.; Al-Betar, M.A.; Ekhan, P.; Hammouri, A.I. Healthcare scheduling in optimization context: A review. *Health Technol.* **2021**, *11*, 445–469. [CrossRef]
24. Eling, M.; Nuessle, D.; Staubli, J. The impact of artificial intelligence along the insurance value chain and on the insurability of risks. In *The Geneva Papers on Risk and Insurance-Issues and Practice*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 1–37.

25. Hassan, A.K.I.; Abraham, A. Modeling insurance fraud detection using imbalanced data classification. In *Advances in Nature and Biologically Inspired Computing*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 117–127.
26. Jha, B.K.; Sivasankari, G.; Venugopal, K. Fraud detection and prevention by using big data analytics. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*; IEEE: Piscataway, NJ, USA, 2020; pp. 267–274.
27. Matloob, I.; Khan, S.A.; Rahman, H.U. Sequence mining and prediction-based healthcare fraud detection methodology. *IEEE Access* **2020**, *8*, 143256–143273. [CrossRef]
28. Matloob, I.; Khan, S.; Hussain, F.; Rahman, H. Medical health benefit management system for real-time notification of fraud using historical medical records. *Appl. Sci.* **2020**, *10*, 5144. [CrossRef]
29. Singh, D.; Kumar, P. Conceptual mapping of insurance risk management to data mining. *Int. J. Comput. Appl.* **2012**, *975*, 8887. [CrossRef]
30. Soeini, R.A.; Rodpysh, K.V. Applying data mining to insurance customer churn management. *Int. Proc. Comput. Sci. And Information Technol.* **2012**, *30*, 82–92.
31. Bhatnagar, V.; Ranjan, J.; Singh, R. Analytical customer relationship management in insurance industry using data mining: A case study of indian insurance company. *Int. J. Netw. Virtual Organ.* **2011**, *9*, 331–366. [CrossRef]
32. Zhikun, X.; Yanwen, W.; Zhaohui, L. Optional insurance compensation rate selection and evaluation in financial institutions. *Int. e-Serv. Sci. Technol.* **2014**, *7*, 233–242.
33. Goonetilleke, T.O.; Caldera, H. Mining life insurance data for customer attrition analysis. *J. Ind. Intell. Inf.* **2013**. [CrossRef]