JURNAL CAHAYA MANDALIKA

P-ISSN: 2828-495X
E-ISSN: 2721-4796

PREDICTION OF DIABETES MELLITUS INSURANCE CLAIM MODELS USING MACHINE LEARNING METHODS

Livia Meristya Fitriani[1*], A. Arviansyah[2]
Master of Management, Faculty of Economics and Business, Universitas Indonesia, Jakarta, Indonesia,
Email: livia.meristya@ui.ac.id, arviansyah@ui.ac.id

Keywords:
Machine Learning; Classifications; Prediction Claims; Insurance

ABSTRACT

Diabetes mellitus is an increase in blood sugar levels accompanied by impaired metabolism of carbohydrates, lipids, and proteins as a result of insufficient insulin function. In 2021 the number of deaths due to diabetes mellitus in Indonesia reached 236,711 people, this is ranked sixth in the world and ranked first in Southeast Asia. Also in Indonesia, this disease increased by 8.5% in 2014 in people over 18 years of age. Many factors influence this disease, including age, gender, as well as the doctor's diagnosis of congenital diseases. The increasing number of cases of death from diabetes mellitus every year causes insurance companies to anticipate the situation, including calculating appropriate claim reserves. This paper aims to calculate the prediction of claims that can be generated using the variable limits of age, gender, and doctor's diagnosis of other congenital diseases by doing classification which carried out using the K-Modes clustering and the Heuristic Method. After classifying the data, we proceed with calculating claim predictions using Random Forest, Naïve Bayes, and Support Vector Machine algorithms. The results of this study indicate that the best model predictions are obtained using the Naive Bayes algorithm, while the best classification group uses the Heuristic model. This research will obtain the best accuracy if balanced with a large amount of data and more diverse variables. The results of this study are expected to be a guideline for insurance companies in determining the estimated amount of claims that may occur.

INTRODUCTION

Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves. WHO claimed about 422 million people worldwide have diabetes, the majority living in low-and middle-income countries, and 1.5 million deaths are directly attributed to diabetes each year. Both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades. In 2021, International Diabetes Federation provide that 537 million adults (20-79 years) are living with diabetes, also almost 1 in 2 (240 million) adults living with diabetes are undiagnosed (Gowriswari & Brindha, 2022).

According to WHO, the number of people with diabetes has increased from 108 million people in 1980 to 422 million in 2014. In 1980, less than 5% of people adults (over the age of 18 years) suffer from diabetes mellitus (Zhou et al., 2016). While in 2014 the level of adults suffering from diabetes mellitus increased 8.5%. The International Diabetes F ederation predicts that more adults will suffer from diabetes mellitus in middle or low income countries with predictions of up to 80% as a result of changing dietary habits fast. The effect of increasing disease diabetes mellitus, namely from changes in lifestyle patterns or changes in structure age as life expectancy increases (Popkin, 1994). Another thing that can affect the increase in diabetes mellitus is a person who are over 45 years old, are overweight (obese), have high blood pressure (hypertension), impaired fat metabolism, has a history of heredity with diabetes, a history of recurrent miscarriages, or giving birth to children with a body weight of more than 4 kg.

Analysis of claims in insurance is an important aspect because in the insurance industry 80% premium income is spent on insurance company claims (Mazviona et al., 2017). It is important to carry out a thorough analysis of claims to help increase the company's cash flow. Diabetes mellitus is one of the diseases that insurance claims most often make after heart disease and stroke. Various model predictions are carried out, one of which is by using machine learning methods. Machine learning methods, such as logistic regression, artificial neural network, and decision tree were used by Meng et al. to predict diabetes mellitus and pre-diabetes by risk factors. Other machine learning methods, such as Naïve Bayes, Decision Tree, and SVM used to detect diabetes mellitus and the results showed that Naïve Bayes algorithm works better than the other two algorithms (Lai et al., 2019).

In this article, we present the prediction model insurance claim for diabetes mellitus using Random Forest, Naïve Bayes, and Support Vector Machine techniques to predict the probability amount of claim based on historical claim insurance (Alghamdi et al., 2017). But before predict the model, we need to classify the risk with K-Modes and Heuristics method.

A. Rudi et al. conducted research in 2017 found that age factor can affect how a person can suffer from diabetes mellitus. The older a person then the function of the body will decrease, including the performance of the insulin hormone which results insulin not being able to work optimally so the resulting blood sugar increases. The other risk factor is gender, the results of Rudi A. and Kwureh research show that the percentage of patients female diabetes sufferers are greater than male patients due to women have a higher body fat composition than men, so women will get fat more easily which means women have greater potential for the risk of diabetes and obesity (Rudi & Kwureh, 2017).

Another research about diabetes mellitus conducted by Wilson et al. by developed the Framingham Diabetes Risk Scoring Model (FDRSM) to predict the risk for developing diabetes mellitus in middle-aged American adults (45 to 64 years of age) using Logistic Regression (Lai et al., 2019). The number of subjects in the sample was 3140 and the area under the receiver operating characteristic curve (AROC) was reported to be 85.0%. Data mining techniques have been widely used in diabetes mellitus studies to explore the risk factors. Meng et al. predict diabetes mellitus and pre-diabetes using Logistic Regression, Artificial Neural Network, and Decision Tree. The data included 735 patients who had diabetes mellitus or pre-diabetes and 752 who are healthy from Guangzhou, China. The accuracy was reported to be 77.87% using a Decision Tree model; 76.13% using a Logistic Regression model; and 73.23% using the Artificial Neural Network (ANN) procedure (Lai et al., 2019). However, in Meng et al. study there are limitations, including the sample data used only used data from the Guangzhou, China. If the sample space is taken from several regions in China, the results will be more representative.

Alaoui et al. also did some research in 2021 by classify soft tissue tumors using machine learning based approach which combines a new technique of preprocessing the data for features transformation, resampling techniques to eliminate the bias and the deviation of instability and performing classifier tests based on the Support Vector Machine and Decision Tree algorithms. These results confirm that machine learning methods could provide efficient and effective tools to reinforce the automatic decision making processes of soft tissue tumors diagnostics (Alaoui et al., 2021). However in this study, there are several anomalies related to the process of the construction of classification models (Liu et al., 2021).

In this study, we proposed a comparative analysis between Random Forest, Naïve Bayes, and Support Vector Machine for predict the insurance claim for diabetes mellitus by doing classification using the K-Modes and the Heuristic Method. We provide a comparison of data grouping with the aim the prediction results of the given model can be more accurate and reliable. The hypothesis for this prediction compares the performance of the modeling by looking at AUC, Accuracy, and F1-measure.

RESEARCH METHOD

In this study, 1157 sample data were used from an insurance company with a claim period from January 2019 to December 2022. This data provide the gender, age, approved claim numbers, incurred claim numbers, admission date, city, and the doctor’s diagnose whether there are other congenital diseases or not. The definition of diabetes for this paper is diabetes mellitus type 1 and type 2, controlled or uncontrolled, and excludes gestational diabetes, chemically induced (secondary) diabetes, neonatal diabetes, polycystic ovarian syndrome, hyperglycemia, prediabetes, or similar states or conditions (Association, 2021). There are no segregation or categorization of diabetes disease specifications was carried out. Table 1 shows that the means of age, the amount of claims submitted (incurred) and the amount of claims received (approved).

Table 1 The mean of variables

Gender	*N*	*Age*	*Incurred*	*Approved*
Male	780	47	3.165.300	3.612.583
Female	377	46	3.592.871	4.269.859

The other information, the gender variable will be coded 0 for female and 1 for male. The variable doctor's diagnosis or any other description will be given by “Remarks_Code” consists of several categories.

Code 1 : participants diagnosed only diabetes mellitus
Code 2 : participants diagnosed with other congenital diseases
Code 3 : participants who have excess fees
Code 4: participant with co-sharing difference

The analysis process begins with inputting the dataset into the tools (we used by Ms. Excel and R Studio). First, we needs to be done filling in data if there is any missing value, this process is called data cleansing. After the data goes through the data cleansing process, we need to explore the dataset to analyze the character of the variables which will be used. After the character of the variable is known, the process will continue data preparation followed by modeling the data.

Risk Classification

Two approaches to risk classification are now considered. The first is a data-driven clustering approach, using K-Modes method utilizing variables age, gender, and Remarks_Code to find groups of policy holders exhibiting similar characteristics. The second is the heuristic method of (Samson & Thomas, 1987) that classifies policy holders according to a set of pre-defined factors

K-Modes Clustering

The clustering model using the variables gender, age, and the variable Remarks_Code where the three variables are categories of data, then we used K-Modes clustering. The K-Modes clustering method is almost the same as K-Means clustering but the calculations are performed using the mode on the data or the data that appears most often.

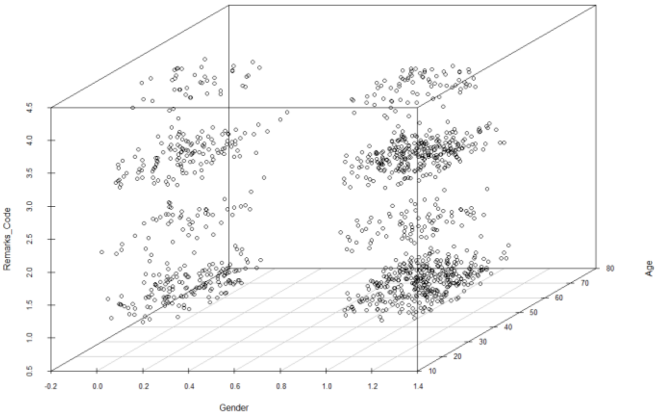

Figure 1 Grouping result K-Modes clustering

Based on the results of the grouping performed, it can be assumed that the cluster groups formed 4 clusters with ignoring the gender segregation that occur in the plot. As for the detail distribution of cluster groups based on the calculation of the K Modes clustering method can be seen in Table 2.

Table 2 The result of K-Modes clustering

Cluster	N	Age	Incurred	Approved
1	139	48	4.618.349	4.425.436
2	436	47	3.535.473	3.376.195
3	273	48	4.306.019	4.132.209
4	309	49	5.709.862	5.433.806

Based on the summary of the cluster groupings that occur, it can be seen that the characteristics of cluster 1 and cluster 4 have almost the same characteristics. As a result of the similarity of characteristics that are formed, the author will combine cluster 1 and cluster 4 into one group so that a new cluster group is obtained.

Table 3 The new clustering obtained

Cluster	N	Age	Incurred	Approved
1	448	46	3.998.764	3.623.173
2	436	47	3.535.473	3.376.195
3	273	48	4.306.019	4.132.209

Heuristic Method

The grouping performed on the Heusristic method will using the same variables as K-Modes clustering (insured's age, participant's gender, and Remarks_Code on the dataset). The data will be divided into 4 groups with grouping details as follows.

Cluster 1: women aged 18 – 40 years

Cluster 2: men aged 18-40 years

- Cluster 3: women aged 41 – 75 years
- Cluster 4: men aged 41-75 years

Table 4 The result of Heuristic method

Cluster	N	Age	Incurred	Approved
1	102	34	3.940.275	3.764.022
2	169	35	4.503.898	4.363.404
3	275	51	4.700.443	4.487.695
4	611	51	3.744.048	3.434.903

Based on the grouping results using the Heusristic method, the characteristics of the data cluster 1 and cluster 2 formed tend to be the same and the number of samples in the clusters formed tends to be small. As a result of this, the authors decided to combine the two clusters into one group to obtain a new cluster group.

Table 5 The new obtained Heuristic method

Cluster	N	Age	Incurred	Approved
1	271	35	4.291.759	4.137.807
2	275	51	4.700.443	4.487.695
3	611	51	3.744.048	3.434.903

Prediction Model

After the grouping are formed, the next step will be to predict the model based on the data grouping that has been formed to find the best modeling of each group. The dataset will be divided into training data and test data with the proportion of 80% distribution of data training and 20% for data test. The prediction model will be carried out using the Random Forest, Naïve Bayes, and Support Vector Machine (SVM) which is followed by evaluation model. The results of the evaluation will be followed by providing recommendations.

Random Forest

In a random forest, the first step that needs to be done is to calculate the entropy value that will be used as a determinant of the level of unauthenticity data. The entropy value obtained by using the following formula:

$$Entropy\ (Y) = \sum_i p(c|Y)\ \ log_2 p(c|Y)$$

where Y is the set of events and $p(c|Y)$ is the proportionvalue of Y against class c. After get the enttopy value then we need the value of the information gain obtained by using the following formula:

$\textit{Information Gain}\ (Y, a) = \textit{Entropy}\ (Y)\ - \sum_{v \epsilon Values\ (a)} \frac{|Y_v|}{|Y_a|} \textit{Entropy}\ (Y_v)$

where values (a) are values that have the possibility of belonging to in the set a, while Yv is a subclass of the value Y with v having relation to a, and Ya are values that are aligned with the set a.

Naïve Bayes

While in Naïve Bayes we predict the probability in a class group based on historical data. Assumed our variables are independent, the Naïve Bayes is formulated as follows.

$$P\ (C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

X : data with undefined class
C : specify class of X
$P(C|X)$: conditional probability of C based on condition of class X
$P(C)$: probability of class C
$P(X|C)$: conditional probability of X based on condition of class C
$P(X)$: probability of X
i : class

Support Vector Machine

In Support Vector Machine, it aims to maximize the hyperplane boundary which is the best boundary distance between the two groups obtained by calculate the hyperplane margin and calculate the maximum point of the data. Margins is the distance obtained between the hyperplane boundary and the group that has closest distance. The closest data referred to a support vector. In this Support Vector Machine calculation, the kernel function will be used as follows:

Kernel Linear

$$K(x, y) = (x^T y)$$

Kernel Radial

$$K(x, y) = exp - (\frac{|x-y|^2}{2y^2}).$$

After modeling is calculate in each group, a model evaluation will be carried out from the modeling that has been formed. In the evaluation model, we will use the matrix to see the performance of the model formed. After that, the data validation will be used to see the performance of the modeling generated on data that has never been studied before. The results of the model evaluation will be used in tuning parameters to improve the performance of the resulting model. Further evaluation will be carried out on models that have been tuned before. In finding the best modelling we need to be done tuning repeatedly. After finding the best modelling, a prediction test will be carried out to find out best of model performance for predicting future.

RESULT AND DISCUSSION

After grouping the data, the next step will be data modeling predictions are carried out for each grouping method. There are 3 types estimation modeling that will be carried out includes prediction modeling Random Forest, Naïve Bayes, and Support Vector Machine (SVM). Each method model predictions will be carried out on the grouping data clusters formed from grouping results using K-Modes clustering and Heuristic methods.

First, a summary of the information will be carried out on the results of grouping data from each classification method used. Each grouping result will be seen the average and the deviation to see the distribution of data that occurs in each group, this is done to anticipate by seeing how far the data deviation is. After seeing the summary of the data grouping, predictions can be made using Rstudio modeling.

Table 6 Summary statistics of dataset K-Modes clustering

Cluster	**Gender**	**N**	**Age**		**Incurred**		**Approved**	
			Mean	**Std. Deviasi**	**Mean**	**Std. Deviasi**	**Mean**	**Std. Deviasi**
1	Male	415	46	8	3.881.488	6.147.542	3.536.394	5.913.755
	Female	33	49	6	5.468.445	6.987.211	4.721.780	6.701.559
2	*Male*	358	47	9	4.008.476	6.853.185	3.748.052	6.147.785
	Female	78	47	9	3.062.471	4.484.783	3.004.338	4.490.993
3	*Male*	7	50	0	3.758.927	4.065.993	3.658.329	4.114.938
	Female	266	46	9	4.853.111	9.002.256	4.606.089	8.489.356

In the grouping of cluster 1, the number of male participants was formed 415 people and 33 female participants. For male participants the average age that makes claims for diabetes mellitus is 46 years with a standard deviation of 9 years. Also their claims submitted amounted to IDR 3,881,488 with a standard deviation of IDR 6,147,542. Whereas for female participants, the average age of the submission claim is 49 years with a standard deviation of 6 years, and the amount of the claim is submitted an average of IDR 5,468,445 with a standard deviation of IDR 6,987,211.

In cluster 2, there were 358 male participants and 78 female participants. In this cluster, male and female participants have an average age of 47 years with a standard deviation of 9 years. For male participants, the amount of claims submitted was IDR 4,008,476 with a standard deviation of IDR 6,853,185. Whereas for female participants, the average amount of claims submitted was IDR 3,062,471 with a standard deviation of IDR 4,484,783.

In the cluster 3 grouping, there were 7 male participants and 266 female participants. For male participants, the average age who made claims for diabetes mellitus was 50 years and the average amount of claims submitted was IDR 3,758,927 with a standard deviation of IDR 4,065,993. Whereas for female participants, the average age of filing claims is 46 years with a standard deviation of 9 years and the average amount of claims submitted is IDR 4,853,111 with a standard deviation of IDR 9,002,256.

Grouping the data using the Heuristic method also produces 3 clusters that will be modeled. In cluster 1, 169 male participants and 102 female participants are formed. In male participants, the average age of those who make claims for diabetes mellitus is 35 years with a standard deviation of 3 years. For male participants, the amount of claims submitted was IDR 4,504,159 with a standard deviation of IDR 7,662,631. Whereas for female participants, the average age of filing claims is 34 years with a standard deviation of 4 years, and the average amount of claims submitted is IDR 3,940,908 with a standard deviation of IDR 6,934,830.

In the cluster 2 grouping, there were 275 female participants who had an age range of 41-75 years, and the average age in this group was 51 years with a standard deviation of 6 years. Cluster 2 group, the average amount of claims submitted is IDR 4,700,443 with a standard deviation of IDR 8,499,914. In the cluster 3 grouping, there were 611 male participants who had an age range of 41-75 years, and the average age in this group was 51 years with a standard deviation of 6 years. Cluster 3 group has an average claim amount of IDR 3,744,048 with a standard deviation of IDR 6,081,329.

Table 7 Summary statistics of dataset Heuristic method

Cluster	Gender	N	Age		Incurred		Approved	
			Mean	Std. Deviasi	Mean	Std. Deviasi	Mean	Std. Deviasi
1	Male	169	35	3	4.504.159	7.662.631	4.363.415	7.647.262
	Female	102	34	4	3.940.908	6.934.830	3.764.930	6.887.094
2	Female	275	51	6	4.700.443	8.499.914	4.487.695	7.975.981
3	Male	611	51	6	3.744.048	6.081.329	3.434.903	5.452.911

Prediction Model

Based on the distribution of clusters that have been carried out, then the model predictions will be calculated by looking at the level of accuracy of the model formed. The method for model prediction will be carried out using algorithms from Random Forest, Naïve Bayes, and Support Vector Machine (SVM).

Table 8 Prediction model on K-Modes clustering

Cluster	RF	Naïve Bayes	SVM
1	76.79 %	77.90 %	34.35 %
2	70.41 %	74.77 %	31.01 %
3	67.03 %	73.99 %	30.95 %

Based on the results of the prediction modeling, the best model prediction results were obtained using the Naïve Bayes method with an accuracy rate in cluster 1 of 77.90%, cluster 2 prediction accuracy of 74.77%, prediction accuracy in cluster 3 of 73.99%. Modeling using Support Vector Machine (SVM) is not suitable for use in datasets because the prediction accuracy of the given model does not reach 50%.

Table 9 Prediction model on Heusristic method

Cluster	RF	Naïve Bayes	SVM
1	86.72 %	91.51 %	55.47 %
2	81.82 %	90.18 %	51.12 %
3	79.21 %	91.16 %	52.44 %

Based on the results of modeling predictions, the best model prediction results are the Naïve Bayes method with an accuracy rate in cluster 1 of 91.51%, cluster 2 prediction accuracy of 90.18%, prediction accuracy in cluster 3 of 91.16%. It is proven that the Naïve Bayes model prediction method using the K-Modes Clustering method and the Heuristic method is the best Machine Learning algorithm used. The model prediction method using the Support Vector Machine (SVM) algorithm is not suitable for use in datasets because the prediction accuracy of the given model tends to be low.

The grouping can be a consideration in classifying potential participants based on the type of risk and followed by model predictions. There are several other method options that can be used to predict the model to estimate the estimated amount of claims that may occur by group.

In this paper it has been proven that the clustering method using the Heuristic Method obtains the best model predictions compared to the K-Modes clustering method. The company will certainly tend to choose cluster groupings that can clearly define the description of the group. The results of grouping and predicting the size of these claims can assist insurance companies in preparing estimates of the amount of claims that need to be prepared. Other model prediction methods can also be applied to predictions

There are several things that need to be considered in classifying risks. The smaller the risk of a group, the more homogeneous the policyholders within the group and the amount of premium paid will be equivalent to the risks that may occur. The greater the risk in the group, the greater the prediction of losses that occur. Factors that can affect losses from the amount of claims that occur, among others.

a. Economic policy factors or natural disasters can affect costs, including an increase in claims costs (external factors).
b. There is a competition among insurance industry. An insurance company will certainly provide competitive rates which will lead to not too large profits.
c. The risk of inaccurate prediction of the amount of claims from an insurance company.

CONCLUSION

This paper shows that a grouping approach in risk classification can be used to predict the amount of claims that may arise. Clustering is advantageous because a wide variety of variables can be considered. In this study the approach using the Heuristic Method has a better predictive estimate. Claim cost prediction is carried out as a step aimed at determining the optimal premium. In this study the best model predictions were determined using Naïve Bayes. However, in future studies it is recommended to involve weight, blood pressure, or other supporting variables that can affect the size of the claim to obtain more accurate grouping results or model predictions.

Reference

Alaoui, E. A. A., Tekouabou, S. C. K., Hartini, S., Rustam, Z., Silkan, H., & Agoujil, S. (2021). Improvement in automated diagnosis of soft tissues tumors using machine learning. *Big Data Mining and Analytics*, *4*(1), 33–46. https://doi.org/10.26599/BDMA.2020.9020023

Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PloS One*, *12*(7), e0179805.

Association, A. D. (2021). 14. Management of diabetes in pregnancy: standards of medical care in diabetes—2021. *Diabetes Care*, *44*(Supplement_1), S200–S210.

Gowriswari, S., & Brindha, S. (2022). Hyperparameters Optimization using Gridsearch Cross Validation Method for machine learning models in Predicting Diabetes Mellitus Risk. *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, 1–4.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocrine Disorders*, *19*(1), 1–9. https://doi.org/10.1186/s12902-019-0436-6

Liu, X., Ding, Y., Tang, H., & Xiao, F. (2021). A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. *Energy and Buildings*, *231*, 110601.

Mazviona, B. W., Dube, M., & Sakahuhwa, T. (2017). An analysis of factors affecting the performance of insurance companies in Zimbabwe. *Journal of Finance and Investment Analysis*, *6*(1), 11–30.

Popkin, B. M. (1994). The nutrition transition in low-income countries: an emerging crisis. *Nutrition Reviews*, *52*(9), 285–298.

Rudi, A., & Kwureh, H. N. (2017). Faktor risiko yang mempengaruhi kadar gula darah puasa pada pengguna layanan laboratorium. *Wawasan Kesehatan*, *3*(2), 33–39.

Samson, D., & Thomas, H. (1987). Linear models as aids in insurance decision making: the estimation of automobile insurance claims. *Journal of Business Research*, *15*(3), 247–256.

Zhou, B., Lu, Y., Hajifathalian, K., Bentham, J., Di Cesare, M., Danaei, G., Bixby, H., Cowan, M. J., Ali, M. K., & Taddei, C. (2016). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants. *The Lancet*, *387*(10027), 1513–1530.