The Classification of Insurance Claim Risk Using the Multilayer Perceptron Method

Endang Wahyu Handamari[1)], Umu Sa'adah[2)], Maulana Muhamad Arifin[3]

[1-3)] Faculty of Science, Brawijaya University

Correspondence Author: ewahyu-math@ub.ac.id

Article Info :

Article History :
Received :
08-10-2024
Revised :
15-01-2025
Accepted :
18-01-2025
Available Online :
29-01-2025

Keyword :
Insurance, Risk Classes, Classification, Multilayer Perceptron.

ABSTRACT

Policyholders purchase insurance policies to protect themselves or their assets from potential financial risks in the future. Insurance guarantees that if an event covered by the policy occurs, the insurance company will provide compensation according to the agreed terms. Insurance companies conduct risk assessments for each policyholder to determine the premium that must be paid, making it essential to classify risk categories accurately. The Multilayer Perceptron (MLP) is one method used for classification problems. It is a machine learning algorithm belonging to the family of artificial neural networks. MLP is a flexible algorithm that can solve various classification problems, including those with complex features and non-linear relationships between input and output variables. The result of this research is the development and implementation of a Multilayer Perceptron method to classify risk categories. The evaluation of the Multilayer Perceptron model for risk classification shows satisfactory performance. Based on the classification report from training and test data, the model does not exhibit overfitting or underfitting.

1. INTRODUCTION

In the insurance industry, risk management is vital to ensure business continuity and provide appropriate services to policyholders. Risk classification is one of the methods used to understand and assess the risks faced by policyholders. This classification allows insurance companies to evaluate risks more accurately, determine appropriate premiums, and make better decisions about accepting or rejecting a particular risk. Research on machine learning modeling, using methods such as decision tree, random forest, and XGBoost, has been conducted to estimate insurance risks and make predictions (Sahai et al., 2023). A study also used the multinomial Naïve Bayes algorithm to classify premium payment status (Rinaldi et al., 2021). The results showed that the multinomial Naïve Bayes algorithm achieved a high level of accuracy with relatively low error rates.

In this research, insurance risk classification is performed using the Multilayer Perceptron (MLP) algorithm. MLP is a machine learning technique classified under neural networks. Several studies utilizing MLP have been conducted. For example, in research, a comparative analysis was performed on various classification algorithms, including Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Adaboost, K-Nearest Neighbors (KNN), Linear Regression (LR), Naive Bayes (NB), and Multilayer Perceptron (MLP) to detect insurance fraud (Rukhsar et al., 2022). Additionally, research has applied machine learning to address challenges in theoretical models and algorithms for advanced data analysis in the insurance industry (Shende et al., 2023). The dataset used in this research proposal is the Prudential Life Insurance assessment data, which includes several attributes from policyholders.

ISSN Print : 1979-7141
ISSN Online : 2541-1942

2. METHOD

2.1. Research data

This research utilizes secondary data from the Prudential Life Insurance Assessment dataset, available on Kaggle (**https://www.kaggle.com/competitions/prudential-life-insurance-assessment/data**). The dataset contains information related to life insurance policies, with a total of 128 attributes, including both features and metadata, such as an ID attribute and the target variable. For this analysis, 66 relevant features were selected after careful data preprocessing, as detailed in Table 1.

Table 1. Attribute Descriptions

Feature Name	Description
Ins_Age	Normalized age of the applicant
Ht	Normalized height of the applicant
Wt	Normalized weight of the applicant
BMI (Body Mass Index)	Normalized Body Mass Index (BMI) of the applicant
Employment_Info_1-6	6 attributes representing normalized employment history information of the applicant
InsuredInfo_1-6	6 attributes representing normalized personal information of the applicant
Insurance_History_1-9	9 attributes representing normalized insurance history of the applicant
Medical_History_1-41	41 attributes representing normalized medical history of the applicant
Response	Target variable representing an ordinal risk measure related to the final application decision

2.2. Research methodology

The research procedure consists of the following stages:

1. **Problem Identification and Formulation**: The problem is framed as a classification task where the goal is to predict the insurance risk category of applicants based on the available features. The aim is to develop a classification model that predicts risk more accurately.

ISSN Print : 1979-7141
ISSN Online : 2541-1942

2. **Literature Review**: A review of relevant literature was conducted, covering the Multilayer Perceptron (MLP) classification algorithm, data preprocessing techniques, and model evaluation methods in Machine Learning.
3. **Data Collection**: The Prudential Life Insurance Assessment dataset was sourced from Kaggle, with 128 attributes, including 1 ID attribute for indexing, and the target variable 'Response', representing the risk class.
4. **Data Preprocessing**:
   - **Class Labeling**: The original target variable 'Response' contained 8 risk classes. In this study, the risk classes were consolidated into 3 categories:
     - Class 1: Combining risk classes 1 to 4.
     - Class 2: Combining risk classes 5 to 7.
     - Class 3: Representing risk class 8.
   - **Data Cleaning**: Missing values were handled based on their proportion:
     - Attributes with missing values ≤ 50% were filled using the mean value.
     - Attributes with missing values > 50% were dropped to maintain data integrity.
   - **Data Transformation**: The data was transformed through two key processes:
     - **Encoding**: The feature, a categorical variable, was converted to a numerical format using Label Encoding since it had no inherent ordering.
     - **Normalization**: Min-Max Normalization was applied to scale the features into a range of [0, 1] to ensure uniformity across all features. This process was performed on both the training and testing datasets for consistency.
   - **Data Split**: The dataset was split into two sets:
     - 75% for training data (44,535 samples).
     - 25% for testing data (14,846 samples). This step ensures the model is tested on unseen data to avoid overfitting.
5. **Model Building**: The classification model was built using the Multilayer Perceptron (MLP) method, implemented via the MLPClassifier function from the Scikit-learn library. Key hyperparameters tuned include:
   - Number of hidden layers: The model used 7 hidden layers.
   - Number of nodes per layer: Varying nodes were used, starting with 16 nodes in the first layer and gradually increasing, based on optimal performance metrics.
6. **Model Evaluation**: Model performance was evaluated using several metrics, including accuracy, precision, recall, and F1-score, based on the confusion matrix. Evaluation on the testing data is critical to ensure generalization and predictive power.

3. RESULTS AND ANALYSIS

3.1. Implementation of the Multilayer Perceptron (MLP) Method for Risk Class Classification

The implementation of the Multilayer Perceptron (MLP) method for classifying risk classes is divided into three classes. They are Class 1, Class 2, and Class 3. The initial stages involve data input and data labeling. The data input process into Google Colab uses the Pandas package. This research processes a dataset contains information related to life insurance policies, with a total of 128

ISSN Print : 1979-7141
ISSN Online : 2541-1942

attributes, including both features and metadata, such as an ID attribute and the target variable. For this analysis, 66 relevant features were selected after careful data preprocessing. Next, the labeling of the response attribute will be conducted in this research. Initially, the original data from applicants for Prudential life insurance includes a response variable (target) where the risk level is multiclass (comprising 8 categories), as shown in Figure 1. However, in this research, the risk levels have been modified into three classes, named Class 1, Class 2, and Class 3, to assist the insurance company in making accurate and appropriate decisions.

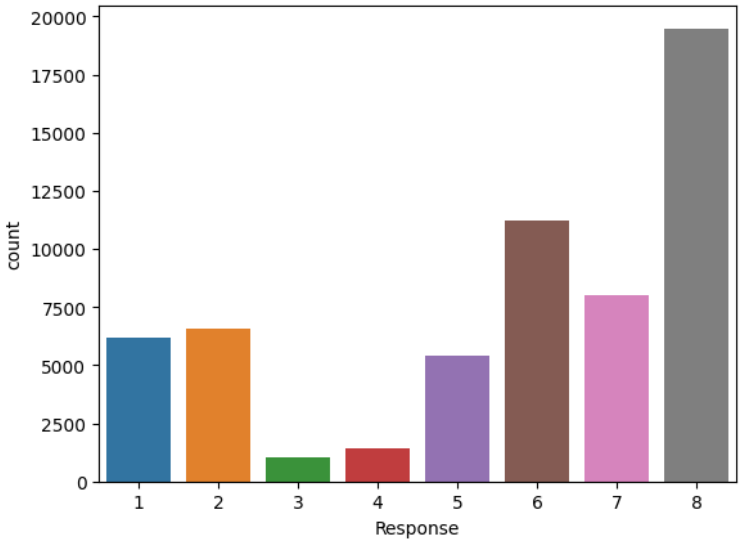

Figure 1. Eight response attributes before modification

The modification of the response attribute combines Classes 1 to 4 into Class 1, Classes 5 to 7 into Class 2, and Class 8 into Class 3. Figure 2 illustrates the modification of the response attribute, which consists of three classes.

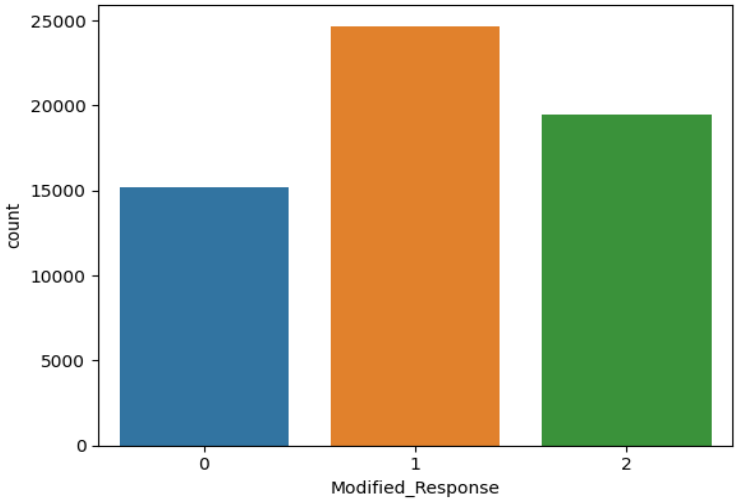

Figure 2. Modification of response attributes

The next process after data input and labeling is data cleaning. Data cleaning is performed to detect missing values and resolve related issues. Based on the programming results, attributes with missing values less than or equal to 0.5 are identified, as shown in Table 2 below.

ISSN Print : 1979-7141
ISSN Online : 2541-1942

Table 2. Attributes with Missing Values ≤ 0.5

Attribute	Percentage of Missing Values
Family_Hist_2	0.482579
Insurance_History_5	0.427679
Family_Hist_4	0.323066
Employment_Info_6	0.182786
Medical_History_1	0.149694
Employment_Info_4	0.114161
Employment_Info_1	0.000320
Family_Hist_2	0.482579

Attributes with missing values less than or equal to 0.5 will be filled with the mean value. Attributes with missing values greater than 0.5 will be removed to preserve data integrity. The attributes scheduled for removal are shown in Table 3 below.

Table 3. Attributes with Missing Values > 0.5

Attribute	Proportion of Missing Values
Medical_History_10	0.990620
Medical_History_32	0.981358
Medical_History_24	0.935990
Medical_History_15	0.751015
Family_Hist_5	0.704114
Family_Hist_3	0.576632

Before implementing the MLP method, data splitting is performed to divide the data into training, validation, and testing sets. The aim is to ensure that the developed model can be properly evaluated and performs well when faced with new data. This process is crucial for making accurate predictions on the test data. The data is split into 75% for training and 25% for testing, with the expectation that the resulting model will avoid overfitting. As a result, 44,535 training data and 14,846 test data are obtained, forming new variables: x_train, y_train, x_test, and y_test.

Before implementing the MLP method for risk class classification, a data transformation step is performed as part of data cleaning. The data transformation includes encoding and normalization processes to ensure uniformity among features within the same range for optimal model performance. The data transformation is applied to both the training and test data, specifically on the features x_train_resampled and x_test, as it will not alter the authenticity of the test data.

ISSN Print : 1979-7141
ISSN Online : 2541-1942

3.2. Encoding

The dataset contains one object-type feature, Product_Info_2, which requires encoding to convert the feature type to a non-object type. In this study, label encoding is utilized because the feature is a label (predictor) without any inherent order. The data type of the Product_Info_2 feature is changed to integer, making it uniform with other non-text features.

Table 4. Comparison of Product_Info_2 Feature Before and After Label Encoding

Before Encoding	After Encoding
A1	0
A2	1
A3	2
A4	3
A5	4
A6	5
A7	6
A8	7
B1	8
B2	9
C1	10
C2	11
C3	12
C4	13
D1	14
D2	15
D3	16
D4	17
E1	18

3.3. Normalization

Normalization is performed to ensure that all features are within the same range, specifically from 0 to 1. Min-max normalization is applied. After undergoing data preprocessing, the next step is to build a classification model using the Multilayer Perceptron (MLP) method and to train the model. The algorithm implemented in this research utilizes the MLPClassifier function from scikit-learn. Hyperparameter tuning for the MLP method involves setting the number of hidden layers and nodes.

ISSN Print : 1979-7141
ISSN Online : 2541-1942

In this study, the MLP model has 13 hidden layers, with the number of nodes in the first hidden layer set to 16, the second hidden layer to 22, and so on. The research is limited to 13 hidden layers because the accuracy tends to decline after the seventh hidden layer, as shown in Table 5 below:

Table 5. Maximum Accuracy for Each Hidden Layer

Hidden Layer	Accuracy Value
1	0.4421
2	0.4427
3	0.4454
4	0.4661
5	0.4865
6	0.4998
7	0.5103
8	0.4932
9	0.4846
10	0.4844
11	0.4832
12	0.4819
13	0.4798

The accuracy for the test data, according to Table 5, is 0.5103. The next step involves obtaining the confusion matrix. The confusion matrix is used to determine how well the model predicts each target class on the test data. Figure 3 shows the confusion matrix from the MLP method implemented on the data. Based on the confusion matrix in Figure 3, there are 1,900 true positive values for Class 0, 2,400 true positive values for Class 1, and 3,300 true positive values for Class 2.

Various metrics can be derived to assess the performance of the MLP method, including accuracy, precision, recall, and f1-score, as shown in Figure 3.

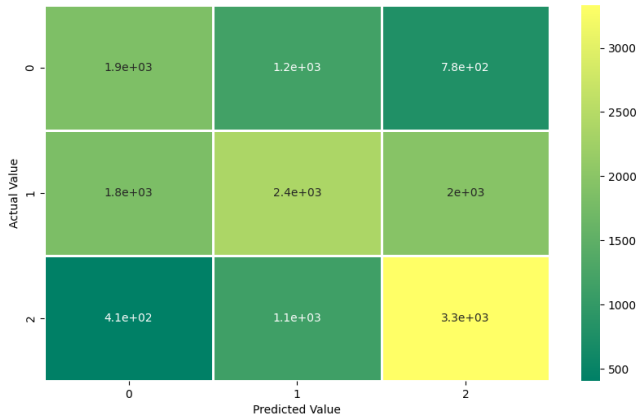

Figure 3. Confusion matrix for the multilayer perceptron model

ISSN Print : 1979-7141
ISSN Online : 2541-1942

Then, based on the classification report for the test data in Table 6, the precision value is 0.51, the recall value is 0.51, and the f1-score is 0.5. The accuracy of the MLP model on the test data is 0.51. In addition, for comparison, the precision, recall, f1-score, and accuracy values of the MLP model for the training data were also calculated. The results of the classification report for the training data are shown in Table 7, where the precision is 0.53, recall is 0.54, f1-score is 0.53, and the accuracy of the MLP model on the training data is 0.54. The classification report results for both the test and training data do not show significant differences. Therefore, it can be concluded that the MLP model used in this study does not experience overfitting or underfitting.

Table 6. Classification report from test data

	Precision	Recall	F1-Score	Support
0	0.45	0.49	0.47	3800
1	0.51	0.39	0.44	6173
2	0.55	0.68	0.61	4873
Accuracy			0.51	14846
Macro avg	0.50	0.52	0.51	14846
Weighted avg	0.51	0.51	0.50	14846

Table 7. Classification report from training data

	Precision	Recall	F1-Score	Support
0	0.57	0.50	0.53	18519
1	0.45	0.40	0.42	18519
2	0.57	0.72	0.64	18519
Accuracy			0.54	55557
Macro avg	0.53	0.54	0.53	55557
Weighted avg	0.53	0.54	0.53	55557

4. CONCLUSION

The conclusion drawn from this research is:

1. The MLP model demonstrated moderate success, achieving a precision of 0.51, recall of 0.51, F1-score of 0.50, and accuracy of 0.51 on the test data. For the training data, the precision was 0.53, recall 0.54, F1-score 0.53, and accuracy 0.54. These results indicate that the model is neither overfitting nor underfitting.
2. However, the performance remains limited, highlighting the need for improvement. Future work should focus on refining features, optimizing the model, or exploring other algorithms to enhance prediction reliability. While the model serves as a foundation, additional efforts are necessary for better outcomes.

ISSN Print : 1979-7141
ISSN Online : 2541-1942

5. ACKNOWLEDGMENT

We declare that we have no conflict of interest.

6. REFERENCES

Dewi I.A.M.S., 2019. *Manajemen Risiko*. UNHI Press. Denpasar.

Guntara D., 2016. Asuransi dan Ketentuan-ketentuan Hukum yang Mengaturnya. *Jurnal Justisi Ilmu Hukum 1(1)*. 29-46.

Muller A.C. & Guido S., 2017, Introduction to Machine Learning with Python. O'Reilly Media, Inc. Sebastopol.

Nawaz N., Harun S., Ali R., & Heryansyah A., 2016. Rainfall Runoff Modeling by Multilayer Perceptron Neural Network LUI River Catchment. *Jurnal Teknologi 78*. 37-42.

Rinaldi R., Goejantoro R., & Syaripuddin, 2021, Penerapan Metode Klasifikasi Multinomial Naive Bayes (Studi Kasus: PT Prudential Life Samarinda Tahun 2019). *Jurnal Informatika 12(2)*. 111-118.

Rukhsar L., Bangyal W.H., Nisar K., & Nisar S., 2022, Prediction of Insurance Fraud Detection using Machine Learning Algorithms. *Mehran University Research Journal of Engineering and Technology 41*. 33-40.

Sahai R., Al-Ataby A., Assi S., Jayabalan M., Liatsis P., Loy C.K., Al-Hamid A., Al-Sudani S., Alamran M., & Kolivand H., 2023, Insurance Risk Prediction Using Machine Learning. In *Lecture Notes on Data Engineering and Communications Technologies*. Lecture Notes on Data Engineering and Communications Technologies, vol. 165, Springer Science and Business Media Deutschland GmbH, 419-433.

Shende M.Y., Bhandekar R., & Kannake R., 2023, Insurance Claim Prediction using Machine Learning. *International Research Journal of Modernization in Engineering Technology and Science 05*. 7008-7011.

ISSN Print : 1979-7141
ISSN Online : 2541-1942