DETERMINING GENERAL INSURANCE PREMIUM PRICING WITH ARTIFICIAL NEURAL NETWORK

FINAL PROJECT REPORT

DATA ANALYTICS

Yogyakarta, December 15, 2023

By Group B

1. Harindra Litsyachnaztyasia Berlian NIM: 21/473602/PA/20406
2. Maulida Nur Shabrina NIM: 21/473987/PA/20448
3. Gemma Praditya Pratama NIM: 21/474472/PA/20485
4. Evangeline Christine Feriardag Marpaung NIM: 21/475243/PA/20557
5. Brian Chang NIM: 21/479984/PA/20828
6. Kenya Azizah Weningjati Aryudi NIM: 21/483187/PA/21065

ACTUARIAL SCIENCE PROGRAM

FACULTY OF MATHEMATICS AND NATURAL SCIENCES

GADJAH MADA UNIVERSITY

2023

ABSTRAK

This paper investigates the application of Artificial Neural Networks (ANN) for determining general insurance premium pricing using data from motorcycle insurance. As Indonesia aims to achieve its Vision 2045, the mastery of science and technology, particularly in AI and machine learning, is crucial. The study aims to develop an ANN model for premium pricing and compare it with the Generalized Linear Model (GLM) method, traditionally used in insurance. The dataset consists of variables such as age, gender, geographic zone, vehicle age, and claim severity. The research involves data preprocessing, model training, and evaluation using Mean Square Error (MSE) and Root Mean Square Error (RMSE). The findings indicate that the ANN model has the potential to effectively manage non-linear patterns and provide accurate premium predictions, although the results for pure premium calculation showed significant differences compared to GLM due to data characteristics. The study concludes that with further research and computational resources, ANN models can significantly benefit the insurance industry.

Keywords: Artificial Neural Network, motorcycle insurance, premium calculation

INTRODUCTION

In order to achieve Indonesia Vision 2045 and based on the enforcement of one of the main pillars, mastery of science and technology, which will in turn improve the economic condition in Indonesia, there have been many efforts among Indonesian youths to compete with other neighboring and developed countries. In these last few decades, machine learning and data science, especially AI, have been playing a critical role in our society, starting from image recognition, fraud detection, pattern recognition, etc. However, to further develop the so-called AI and machine learning, we may need to focus more on the decision-making process as well. There are several methods in unsupervised-learning decision making, one of them is ANN (Artificial Neural Network).

In this paper, ANN is used for determining general insurance premium pricing using data from certain motorcycle insurance. As we all know insurance is a very integral part of everyday lives, certainly there will always be claims involved when certain incidents occur and in most cases, there have been some issues in insurance companies on deciding the right amount of reserves and premium has always been the main problem. Premium calculation is usually done by certain divisions in an insurance company and usually involves actuaries. As future actuaries ourselves, we must make sure that this can be done efficiently and make a huge step in achieving the Indonesia Vision 2045.

Premium itself depends on many factors, namely features or predictor variables. There are many (predictor) variables in the data set, such as age, gender, zone, mcklass, vehicle-age, bonus, duration, number of duration, and severity. Nevertheless, it is also expected that predictions might not be completely accurate, considering the varieties of data (sama characteristics might have different values of predictions) and there might be shortages of information on whether certain variables are considered significant or not. However, the analysis will certainly benefit us by giving us insights and recommendations on how the ANN model

should be developed in the future in insurance industries to anticipate the upcoming reality of the Indonesia Vision 2045.

OBJECTIVES

The objective of this paper is to develop an ANN model of general insurance premium pricing from certain motorcycle insurance. Previously in another similar project, we calculated the premium using the GLM method. Therefore, we want to compare the premium values using two different methods, namely ANN and GLM, because we know that ANN also involves features and may be useful in predicting values by methods of unsupervised learning. Thus, throughout the process, there will be comparisons of plots and values, and the final interpretation of the comparisons will be given in the conclusion part. Such interpretation will give us insights into whether the ANN model is adequately suitable for premium pricing, in which case it has rarely been done before, and whether the ANN model can be directly compared to GLM and is it actually better or not, through comparison of plots.

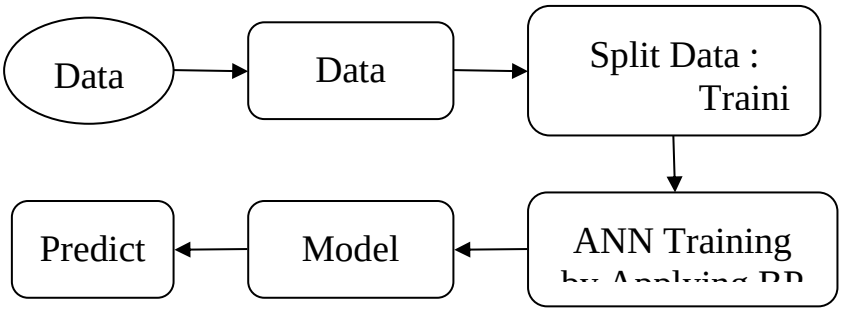
RESEARCH AND METHODS

The data used for this case study is based on motorcycle insurance data. It contains aggregate data on all insurance policies and claims during 1994-1998. The data set analyzed consisted of approximately 64,548 individual claims records, and contained the following variables:

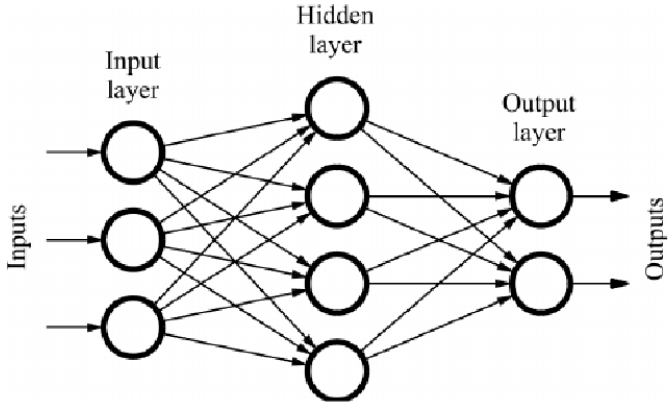
- Age : The owners' age is between 0 and 99.
- Gender : The owner's gender, M (male) or K (female).
- ZON : Geographic zone numbered from 1 to 7.
- MCKLASS : MC class is divided into seven classes from 1 to 7.
- Vec-age : Vehicle age, between 0 and 99.
- BONUS : Bonus class, taking values from 1 to 7.
- DURATION : The number of policy years.
- Nbclaim : The number of claims.

- Severity : The claim cost.

The frame of mind in this study can be observed in the following image.



The research data was analyzed with the Artificial Neural Network method. Artificial Neural Networks (ANN) are a type of computing system built based on the construction of the human brain. These networks consist of interconnected nodes or neurons that have weights associated with them. ANN can detect patterns within data and apply these patterns to predict new cases. ANN generally consists of three layers: an input layer, one or multiple hidden layers, and an output layer. The input layer receives the initial data that will be passed to the hidden layers to perform intermediate computations. It will later come out through the output layer producing the final results based on the input.



Artificial Neural Networks (ANN) have been used for general insurance pricing, including auto insurance pricing. In this context, ANNs are employed to predict pure premiums by analyzing various factors such as frequency and severity of claims. These are the advantages of using ANN:

- ANN can effectively manage non-linear and learn intricate representations by mapping inputs to outputs.

- The use of the ANN is flexible. It can be used for pattern recognition, time series, image processing, and so on.
- The result of the ANN is more successful than traditional statistical methods in the context of speed, simplicity, and capacity.
- By tuning the parameters through trial and error, the performance of the ANN can be enhanced.

The algorithmic structure of general insurance pricing with ANNs involves:

1. Data Collection

   The first step in determining general insurance with ANN is to collect relevant information and historical data on insurance policies, claims, and variables that can affect the price of general insurance premiums.

2. Data Preprocessing

   Before the analysis, the data underwent pre-processing to identify and address issues such as missing values, outliers, and inconsistencies in the data. Data normalization is also included in this step. Normalization is important to ensure the range becomes smaller. Since all the variables have distinct scales, it is advisable to bring them to a uniform scale. This aims to enhance efficiency and increase the accuracy of predictions. The normalization formula:

   $$x_i^{scaled} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

   where x_i is a vector representing the i^{th} row in the data.

3. Data Division

   The research data is further divided into two parts, training data and testing data. The training data is used for model development, and testing data is used for prediction. The division of data is essential for assessing the accuracy of models and ensuring their effectiveness in making predictions on new data.

4. Activation function

   The activation function is used to determine whether a neuron should be activated or not based on the weighted sum of inputs. There are

several types of activation functions used in ANN, the sigmoid function, ReLU function, hyperbolic tangent function (tanh), which introduces non-linearity in the network, and the identity function which introduces linearity in the network.

5. ANN Modeling

   a) Frequency Modeling

      Claim frequency refers to the number of claims during a specific period, usually a year. Claim frequency is calculated by dividing the number of claims filed by the duration of the claim. The formula is:

$$claim\ frequency\ = \frac{number\ of\ claims}{duration}$$

   b) Severity Modeling

      Claim severity refers to the average loss associated with a single claim. Claim severity is calculated by dividing the total amount of claims by claim count. The formula is:

$$severity\ = \frac{total\ claim\ amount}{claim\ count}$$

6. Model Assessment

   Model valuation involves the use of two statistical methods such as Mean Square Error (MSE) and Root Mean Square of Error (RMSE) to evaluate the accuracy and reliability of the models.

   a) MSE

      The Mean Squared Error (MSE) is calculated for each model trained on the data. Let Y_i be the observed value of the response variable and let $\widehat{Y}_i$ be the predicted value. Then, the MSE test is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (\widehat{Y}_i - Y_i)^2$$

      Model with the lower MSE values is considered as the most accurate and reliable, as it indicates a better fit to the data.

   b) RMSE

RMSE is the square root of the MSE and is a measure of the average magnitude of the errors in the predictions. RMSE is given by the formula

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}$$

A lower RMSE indicates better model performance, as it indicates that the predictions are more accurate and closer to the actual values.

7. Prediction

After comparing the models using RMSE, the best model will be selected. From that best model, relativities will be calculated. The relativities equations are generally used to determine the extent of the influence of a particular variable on claim frequency and claim severity. The formula is:

$$\mu_i = exp\left\{\sum_{j=1}^{r} x_{ij}\beta_j\right\}$$

In pricing models, the relativities equations can be applied to the base premium:

$$pure\ premium = \mu_{frequency} \times \mu_{severity}$$

ANALYSIS

From the motorcycle insurance data set, the analysis of premium calculations in general insurance will be conducted using artificial neural network methods.

- **Data Understanding**

	Age	gender	zon	mcklass	Vec-age	bonus	duration	nbclaim	Severity
0	0	M	1	4	12	1	0.175342	0	0
1	4	M	3	6	9	1	0.000000	0	0
2	5	K	3	3	18	1	0.454795	0	0
3	5	K	4	1	25	1	0.172603	0	0
4	6	K	2	1	26	1	0.180822	0	0
...	...	...	...	...	...	...	...	...	...
64543	86	M	4	5	16	3	0.413699	0	0
64544	86	M	4	6	9	7	1.057534	0	0
64545	87	M	4	6	10	7	0.323288	0	0
64546	91	M	1	5	17	1	0.000000	0	0
64547	92	M	2	6	13	7	0.386301	0	0

From Figure 3, it can be seen that there are 64548 rows and 9 columns in the dataset. There are 6 rating factors that will be used to determine the general insurance premium.

- **Data Preprocessing**

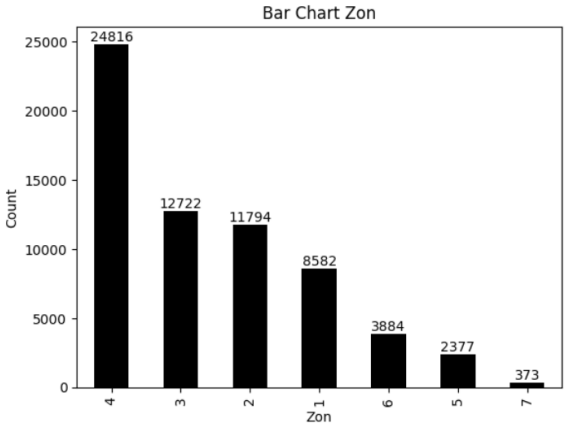
First, a check for missing values in the data will be conducted, and it was found that there are no missing values in each variable in the dataset used, as shown below

```
Age         0
gender      0
zon         0
mcklass     0
Vec-age     0
bonus       0
duration    0
nbclaim     0
Severity    0
dtype: int64
```

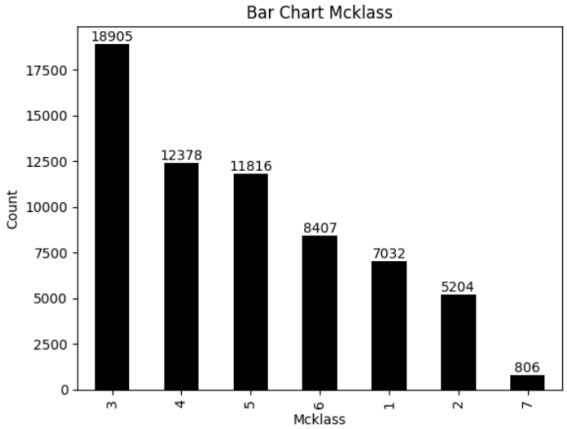
The next step in data preprocessing was to check the validation of data. Based on the check results, it is known that the data used is valid because the severity is set to 0 if no claim is made.

- **Data Visualization**

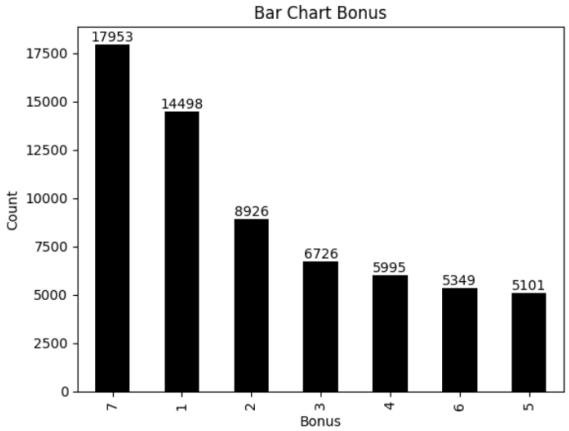
Exploratory Data Analysis (EDA) for Categorical Variables



From the zon barchart, we can conclude that zon 4 has the most motorcycle insurance (24816) and zon 7 has the least motorcycle insurance (373).

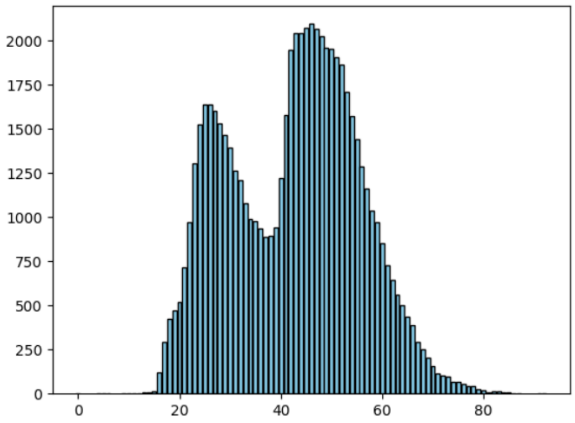


From the mcklass barchart, we can conclude that mcklass 3 has the most motorcycle insurance (18905) and mcklass 7 has the least motorcycle insurance (806).

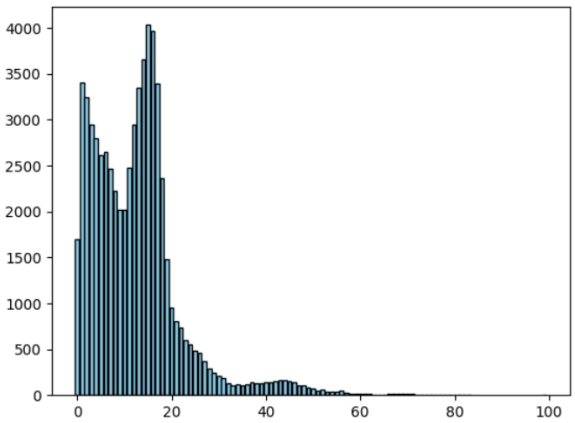


From the bonus barchart, we can conclude that bonus 7 has the most motorcycle insurance (17953) and bonus 5 has the least motorcycle insurance (5101).

Exploratory Data Analysis (EDA) for Continuous Variables



From the age histogram, we can see that the histogram has 2 peaks, so we can conclude that age has bimodal distribution.



From the vec-age histogram, we can see that the histogram has a tail that seems pulled to the right, so we can conclude that vec-age has right-skewed data.

- **Artificial Neural Network (ANN) Model**

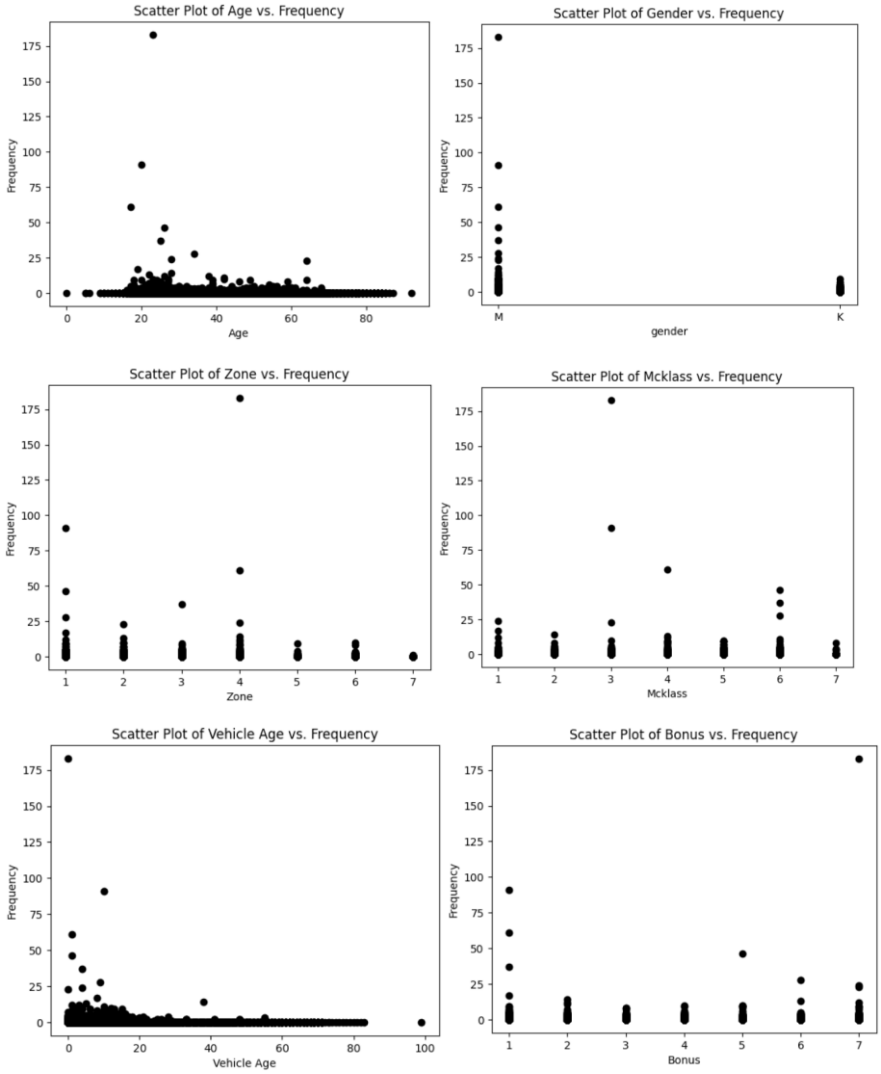
  Similar to the GLM model, we need both frequency and severity present (either in the dataset or through modifications), so that the final premium can be calculated through the multiplication of relativities of the best corresponding models.

Frequency Modeling

In frequency modeling, a new variable ‘frequency’ is created in the dataset as per the formula:

$$frequency = \frac{nbclaim}{duration}$$

For better understanding, a scatter plot visualization of the relationship between each rating factor and frequency was made.



Before modeling, we will first look at the data type and dimensions of the data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64548 entries, 0 to 64547
Data columns (total 10 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Age        64548 non-null  int64
 1   gender     64548 non-null  object
 2   zon        64548 non-null  int64
 3   mcklass    64548 non-null  int64
 4   Vec-age    64548 non-null  int64
 5   bonus      64548 non-null  int64
 6   duration   64548 non-null  float64
 7   nbclaim    64548 non-null  int64
 8   Severity   64548 non-null  int64
 9   frequency  62478 non-null  float64
dtypes: float64(2), int64(7), object(1)
memory usage: 4.9+ MB
```

It is observed that there are 8 variables of numeric type and 1 variable of object type. Unlike generalized linear models (GLMs), artificial neural network (ANN) models are unable to directly process categorical data. Therefore, the categorical variables needed to be transformed into the numerical format to be used in the model. Through the one-hot encoding process, the categorical variable 'gender' is transformed into a numerical form (M: 0, K: 1).

The value of frequency is set to '0' if the number of claims or duration is 0 to prevent 'NaN' errors. The frequency type is then changed to an integer because using a float type can cause errors in the training and testing processes that require data in integer form. This also improves testing accuracy.

In the training process, the severity, claim, and duration columns will not be used so these columns are omitted and the newest dataset is labeled ‘x’ while the frequency is labeled ‘y’. The result of modification as shown below:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64548 entries, 0 to 64547
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Age        64548 non-null  int64
 1   gender     64548 non-null  int64
 2   zon        64548 non-null  int64
 3   mcklass    64548 non-null  int64
 4   Vec-age    64548 non-null  int64
 5   bonus      64548 non-null  int64
 6   frequency  64548 non-null  int64
dtypes: int64(7)
memory usage: 3.4 MB
```

At this stage, the dataset will be divided into two parts (training and testing), with a proportion of 90:10, 95:5, and 99:1. The normalization process will also be performed to create a range (min and max) for each feature that has different values within the same interval. The identity activation function was chosen to train the neural networks. The reason for choosing the identity activation is that the weights in the network can be directly interpreted as the ordinary regression coefficients. This can simplify the interpretation of the model and the impact of input features on the output. The optimal proportion then will be determined based on the model with the smallest MSE and RMSE values.

a) 90% Training

```
Age          58093
gender       58093
zon          58093
mcklass      58093
Vec-age      58093
bonus        58093
frequency    58093
dtype: int64
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	6404
1	0.00	0.00	0.00	19
2	0.53	1.00	0.69	17
3	0.00	0.00	0.00	8
4	0.00	0.00	0.00	1
5	0.00	0.00	0.00	3
7	0.00	0.00	0.00	1
10	0.00	0.00	0.00	1
12	0.00	0.00	0.00	1
accuracy			0.99	6455
macro avg	0.17	0.22	0.19	6455
weighted avg	0.99	0.99	0.99	6455

From the output above, we know that there are 58093 data for the 90% training process. After the normalization process, a smaller range is obtained (min: -3.270869453122334, max: 187.19607425920634). Using the identity activation function, a weighted average of over 90% was obtained (which is good enough), but the macro average is still very low due to too many '0' values in the data, which causes a slight bias.

b) 95% Training

```
Age          61320
gender       61320
zon          61320
mcklass      61320
Vec-age      61320
bonus        61320
frequency    61320
dtype: int64
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3204
1	0.00	0.00	0.00	9
2	0.57	1.00	0.73	8
3	0.00	0.00	0.00	4
5	0.00	0.00	0.00	1
7	0.00	0.00	0.00	1
10	0.00	0.00	0.00	1
61	0.00	0.00	0.00	0
accuracy			1.00	3228
macro avg	0.20	0.25	0.22	3228
weighted avg	0.99	1.00	0.99	3228

From the output above, we know that there are 61320 data for the 95% training process. After the normalization process, a smaller range is obtained (min: -3.269117448534325, max: 191.82357463171556).

Using the identity activation function, a weighted average of over 90% was obtained (which is good enough), but the macro average is still very low due to too many '0' values in the data and the disproportionality of the data.

c) 99% Training

```
Age          63902
gender       63902
zon          63902
mcklass      63902
Vec-age      63902
bonus        63902
frequency    63902
dtype: int64
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	636
1	0.00	0.00	0.00	2
2	0.50	1.00	0.67	4
3	0.00	0.00	0.00	3
10	0.00	0.00	0.00	1
accuracy			0.99	646
macro avg	0.30	0.40	0.33	646
weighted avg	0.98	0.99	0.99	646

From the output above, we know that there are 63920 data for the 99% training process. After the normalization process, a smaller range is obtained (min: -3.2679417500400083, max: 195.63493850415796). Using the identity activation function, a weighted average of over 90% was obtained (which is good enough). The macro average is still relatively low, but it shows significant improvement compared to the macro average in the previous models.

To determine the best frequency model, we compared the value of MSE and RMSE from the 3 models:

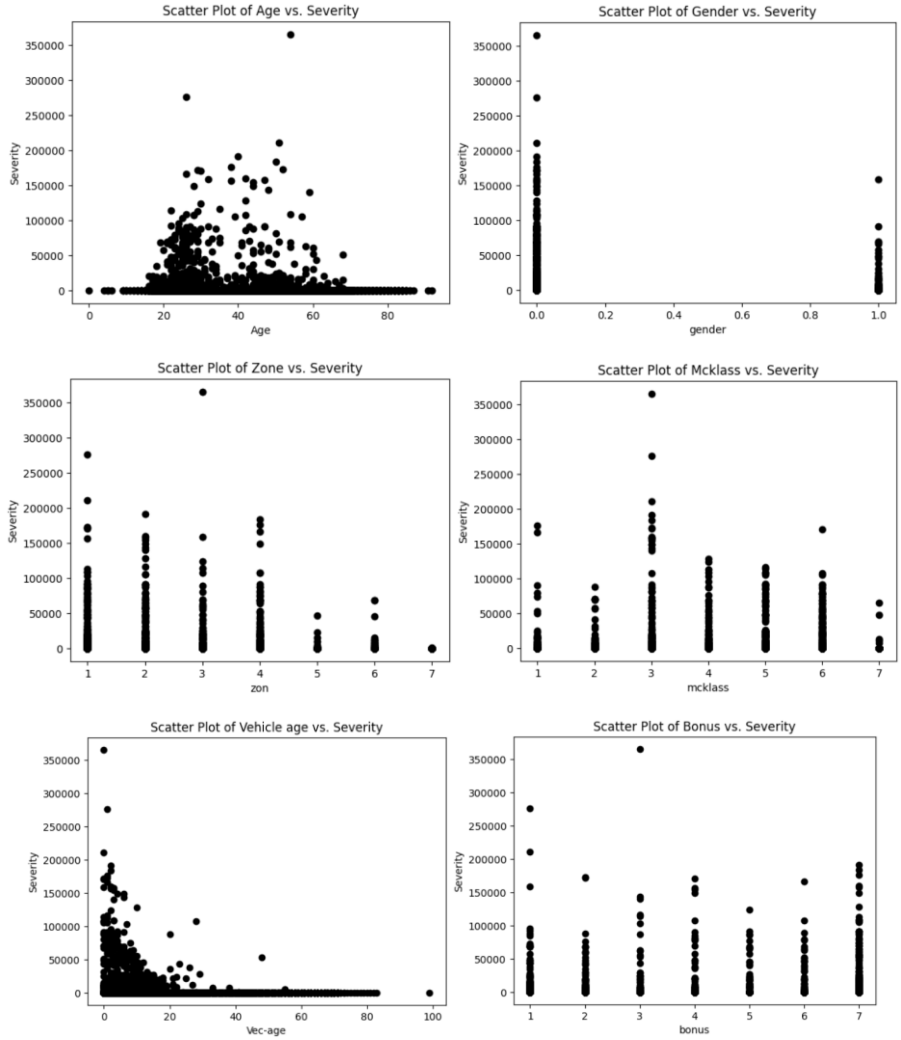
Model	MSE	RMSE
90% Training	0.03826491	0.19561418895
95% Training	0.82032218	0.90571639
99% Training	0.106811145	0.32681974

It can be concluded that the 90% training model is the best frequency model based on the smallest MSE and RMSE values.

Severity Modeling

Before the analysis, the severity type is changed to an integer since a float type can cause errors in the training and testing processes that require data in integer form. Through the one-hot encoding process, the categorical

variable 'gender' is transformed into a numerical form (M: 0, K: 1). For better understanding, a scatter plot visualization of the relationship between each rating factor and severity was made.



In the training process, the duration and nbclaim columns will not be used so these columns are omitted and the newest dataset is labeled ‘Xs’ while the severity is labeled ‘y’. The result of modification as shown below:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64548 entries, 0 to 64547
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Age       64548 non-null  int64
 1   gender    64548 non-null  int64
 2   zon       64548 non-null  int64
 3   mcklass   64548 non-null  int64
 4   Vec-age   64548 non-null  int64
 5   bonus     64548 non-null  int64
 6   Severity  64548 non-null  int64
dtypes: int64(7)
memory usage: 3.4 MB
```

Similar to the frequency modeling, the dataset will be divided into 2 parts (training and testing) with the proportions of 90:10, 95:5, and 99:1. The normalization process and selection of the best model using identity function activation will also be carried out at this stage.

a) 90% Training

```
Age         58093
gender      58093
zon         58093
mcklass     58093
Vec-age     58093
bonus       58093
Severity    58093
dtype: int64
```

accuracy			0.99	6455
macro avg	0.02	0.02	0.02	6455
weighted avg	0.99	0.99	0.99	6455

From the output above, we know that there are 58093 data for the 90% training process. After the normalization process, a smaller range is obtained (min: -3.27086945, max: 75.73556861). Using the identity activation function, a weighted average of over 90% was obtained (which is good enough), but the macro average is still very low due to too many '0' values in the data, which causes a slight bias.

d) 95% Training

```
Age         61320
gender      61320
zon         61320
mcklass     61320
Vec-age     61320
bonus       61320
Severity    61320
dtype: int64
```

accuracy			0.99	3228
macro avg	0.03	0.03	0.03	3228
weighted avg	0.99	0.99	0.99	3228

From the output above, we know that there are 61320 data for the 95% training process. After the normalization process, a smaller range is obtained (min: -3.2691174485, max: 76.27394412). Using the identity activation function, a weighted average of over 90% was obtained (which is good enough), but the macro average is still very low

due to too many ‘0’ values in the data and the disproportionality of the data.

e) 99% Training

```
Age         63902
gender      63902
zon         63902
mcklass     63902
Vec-age     63902
bonus       63902      accuracy                         0.98      646
Severity    63902     macro avg       0.07      0.07      0.07      646
dtype: int64        weighted avg      0.97      0.98      0.97      646
```

From the output above, we know that there are 63902 data for the 99% training process. After the normalization process, a smaller range is obtained (min: -3.2679417500400083, max: 77.50615867395385). Using the identity activation function, a weighted average of over 90% was obtained (which is good enough). The macro average is still relatively low, but it shows significant improvement compared to the macro average in the previous models.

To determine the best severity model, we compared the value of MSE and RMSE from the 3 models:

Model	MSE	RMSE
90% Training	8.840291e+06	2973.262754
95% Training	4.077772e+06	2019.349321
99% Training	4.856132e+06	2203.663205

It can be concluded that the 95% training model is the best frequency model based on the smallest MSE and RMSE values.

From the results of the models, the best model for frequency is the model with 90% training and 10% testing data, while for the severity model, the best model is the one with 95% training and 5% testing data. Moreover, from these ANN models, it is obtained that there are 3 layers: input, hidden, and output. The input consists of 7 inputs, namely age, gender, zone, mcklass, vehicle age, bonus, and frequency/severity. In the hidden layer, there are 8 nodes, which are the input nodes and an additional 1 bias node.

Lastly, there will only be 1 output, which is the frequency/severity. Thus, these models can be used for further analysis of the premium calculation

Premium Calculation

The premium value can be calculated by multiplying the frequency relativities and severity relativities. The relativities is the exponent value of the coefficients of both elements. The calculation below is the result of the multiplication of these two relative values.

- The frequency model expressed in relativities.

$$\mu_{frequency} = 0.26141272345008587 * e^{2.80908180e-03Age} * e^{2.30290171e-03Gender} * e^{6.16383251e-03Zone} * e^{3.54081897e-04MCKlass} * e^{2.89594896e-03Vec-age} * e^{-3.28961722e-03Bonus}$$

- The severity model expressed in relativities.

$$\mu_{severity} = 0.1273861154552569 * e^{5.92450581e-03Age} * e^{3.76093496e-03Gender} * e^{3.63904714e-03Zone} * e^{-5.96130490e-04MCKlass} * e^{1.23867872e-03Vec-age} * e^{-1.17060296e-03Bonus}$$

- The model to calculate pure premium

$$\mu_{frequency} * \mu_{severity} = 0.0333004 * e^{8.73358761e-03Age} * e^{6.06383667e-03Gender} * e^{9.80287965e-03Zone} * e^{-2.42048593e-04MCKlass} * e^{4.13462768e-03Vec-age} * e^{-4.46022019e-03Bonus}$$

Simulation of the premium calculation:

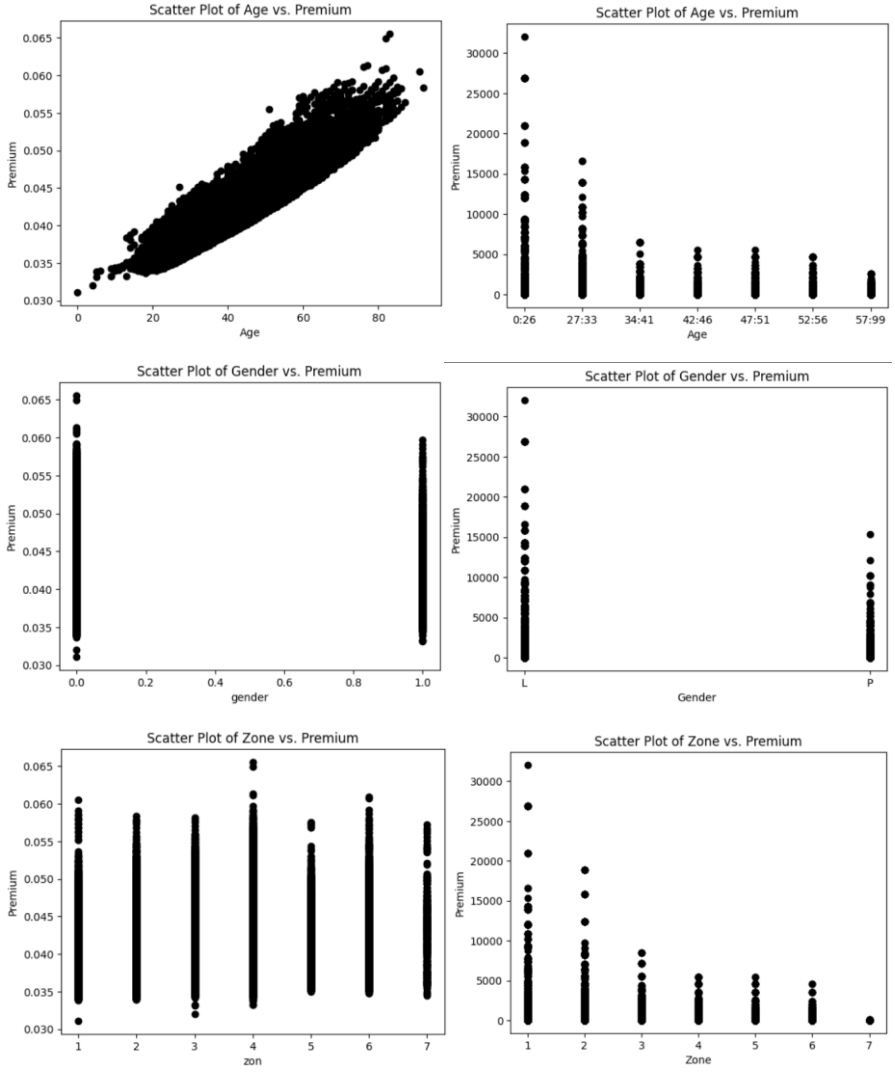
$$\mu_{frequency} * \mu_{severity} = 0.0333004 * e^{8.73358761e-03(30)} * e^{6.06383667e-03(0)} * e^{9.80287965e-03(2)} * e^{-2.42048593e-04(5)} * e^{4.13462768e-03(6)} * e^{-4.46022019e-03(7)}$$

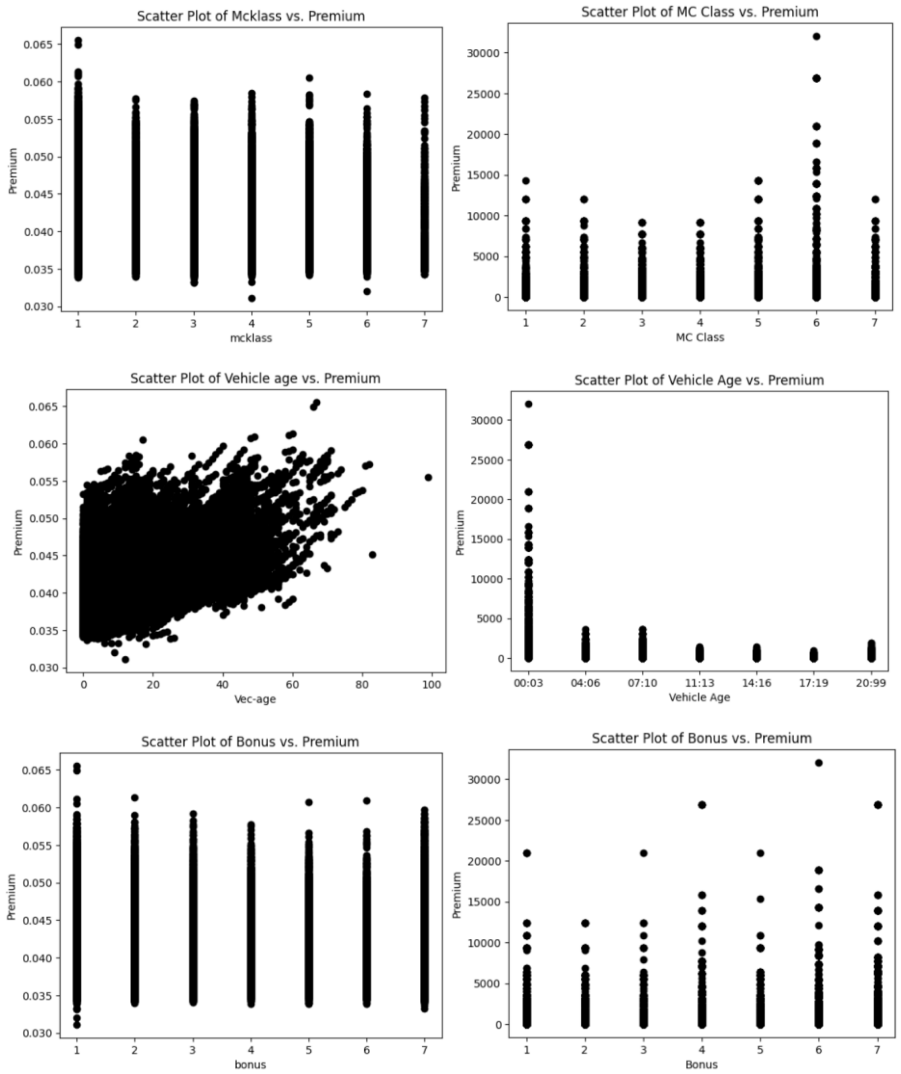
There is a case of a 30-year-old man who wants to insure his 6-year-old car with zone 2, class type 5, and bonus 7, with a premium value of

0.044. The result is significantly different from the calculation using GLM (approximately 317.45). The large difference in the modeling of frequency and severity, which does not consider the distribution and contains too many 0 values, is the cause of this discrepancy.

ANN Premium & GLM Premium

For better understanding, a scatter plot visualization of the relationship between each rating factor and premium using ANN (left) and GLM (right) was made.





We can break down the interpretations for each feature:

- Age: We can see from the ANN model that the older the policyholder is, the more expensive the premium gets. While for the GLM model, the older the policyholder is, the cheaper the premium gets. Realistically, the better model is the ANN model because it represents the real-world scenario where the premium is more expensive for older policyholders.
- Gender: We can see that males from both ANN and GLM models generally have higher premiums compared to females, although the difference is more significant in the GLM model.
- Zone: Based on the ANN graph, we can see that the premiums for all zones are almost similar. While for the GLM model, the premium

gets cheaper, starting from zone 2 to zone 7, and the decline is quite significant as well.

- Mcklass: Both ANN and GLM models have the same interpretations for Mcklass, except that the only difference is that zone 6 from the GLM model has a very high premium compared to the others.
- Vehicle Age: We can see from the ANN model that while it is not too visible, there has been an increasing trend. However, there is a sharp decline in premium value in the GLM model. This might be the same as the age interpretation that the ANN model might work best in real-life scenarios as the older the vehicle gets, the more maintenance it requires, therefore a higher premium is needed.
- Bonus: The premium in the ANN model is comparably similar for all bonus categories. Although there are no certain trends in the GLM model, notice that there are some upper-value outliers. The comparison between both models might not be significant, but there are clearly some differences.

CONCLUSIONS

This project explored artificial neural networks for pricing in general insurance. In the case of neural networks, nothing is assumed or specified about the distribution of the data. Furthermore, the model is allowed to determine the learning weights and optimize them without any restrictions placed on the possible distribution of claims.

The ANN model was trained and tested on the motorcycle insurance dataset. It was found that the 90% training model is the best model for frequency modeling based on the smallest values of MSE and RMSE. Meanwhile, for severity modeling, it is known that the 95% training model is the best model. Besides, it is recommended that additional hidden layers may be added for a more accurate analysis.

Based on the simulation in the case of a 30-year-old man who wants to insure his 6-year-old car with zone 2, class type 5, and bonus 7, it is known that the

pure premium result is 0.044 which is significantly different from the calculation using GLM that obtained a result of 317.45.

The final result of the pure premium calculation using an artificial neural network (ANN) indicates a value approaching 0. This occurrence is attributed to the abundance of zero values in both frequency and severity, resulting in the program capturing the value 0 too frequently. The artificial neural network (ANN) model for premium calculation will yield a significantly different value compared to the generalized linear model (GLM) when compared. In conclusion and based on the premium visualizations, neural networks will have the potential to be highly effective in the insurance industry with more time, computing power, and further research.

REFERENCES

Ibiwoye, A., Ajibola, O. O. E., & Sogunro, A. B. (2012, Winter). Research Gate. *Artificial Neural Network Model for Predicting Insurance Insolvency*, *2*(1), 59-68. https://www.researchgate.net/publication/259717034_Artificial_Neural_Network_Model_for_Predicting_Insurance_Insolvency

Jaszczyk, J. (2020, September 14). *Actuarial study: EDA,PCA,Cluster,Estimation (0.88)*. Kaggle. Retrieved December 15, 2023, from https://www.kaggle.com/code/jjmewtw/actuarial-study-eda-pca-cluster-estimation-0-88/input

König, D., & Loser, F. (2020, April 26). *GLM, Neural Nets and XGBoost for Insurance Pricing*. Kaggle. Retrieved December 15, 2023, from https://www.kaggle.com/code/floser/glm-neural-nets-and-xgboost-for-insurance-pricing/notebook#GLM,-Neural-Network-and-Gradient-Boosting-for-Insurance-Pricing,-Part-1:-Claim-Frequency

Shaw, B. S. (2022, December 20). *Let's Do: Neural Networks*. Medium. Retrieved December 15, 2023, from https://towardsdatascience.com/lets-do-neural-networks-d849d80fd012

Tukiyat, Suhaedi, A., & Sugiyanto. (2021, December). Humanis. *Forecasting The Stock Market Movements Of Unilever Companies Jakarta During The Covid-19 Pandemic Using Artificial Neural Network*, *2*(1), 329-335. http://www.openjournal.unpam.ac.id/index.p

APPENDIX

The notebook of this project: Final Project Syntax Andat Group B.ipynb

The data used in this paper:data.xlsx

The GLM premium: glm premium.xlsx