Milestone: Journal of Strategic Management Vol. 4, No. 2, September 2024
Faculty of Economics and Business
Pelita Harapan University

PREDICTION OF HEALTH INSURANCE PRODUCT PURCHASE ALLOCATION IN VARIOUS INDUSTRIES IN INDONESIA USING THE RANDOM FOREST METHOD

Hendra Achmadi[1)*], Eduard Ary Binsar Naibaho[2)], Sandra Sembel[3)], Herlina Lusmeida[4)]

[1,2,3,4)] *Fakultas Ekonomi dan Bisnis, Universitas Pelita Harapan, Indonesia*

e-mail: soleda2017@gmail.com
*(Corresponding Author indicated by an asterisk *)*

ABSTRACT

The objective of this research is identifying which industry can absorb the product of wealth management such as health insurance. Secondly is to identify what the most factors important to determine closing the health insurance premium. The life insurance penetration and density in Indonesia is the lowest level among the Asian country, so the data population in this research is from 38 different companies from different types of industries with 143 data sample, by using the purposive sampling. Most factors which influence the purchasing of health insurance are Listrik, Industry, domicile, age and position, whether the industry that the most contribution for the health insurance sales is banking and education industry. The methodology that is used in this research is called CRIPS-DM (Cross Industrial Standards Program Data Mining). The first steps what is the purpose of the organization, and the second is what data that needed, and continue to data preparation, after modeling, it will make an interpretation of the result, and the final steps is deployment, it will plan how it will be implemented in the real world, and the accuracy score from this model is 58 %. From the result of the projection closing health insurance from each industry, it can be concluded that the most industry that closed the health insurance is Banking Industry, the second is from insurance and the third is education and the next is education, retail, health, manufacturing and finance, hospitality, legal, publishing, technology and government and service industries.

Keywords: Premium Prediction; CRIPS-DM; Random Forest; The Most Contribution Industries

INTRODUCTION

Increasing effectiveness in sales is the main mission of insurance companies or financial industries in particular. In order to increase sales effectiveness, a targeting function is needed. Targeting here is a function where insurance companies target certain industries that will be the main target in marketing. So, the effectiveness of marketing depends on how long and how fast it takes to identify which customers are coming from which industry. Besides, the marketing also needs to know what factors will influence the customer closing the insurance sales. In the real world is very hard to know which factors will influence the sales, because different salesman, is different approach. So, the machine learning algorithm is needed to help marketing to focus on the target.

In operational activities, targeting makes the company focus on the target market, based on historical data it has. The insurance industry plays an important role as a provider of instruments used by the community for protection or risk management. The role of insurance is also the main characteristic of the insurance industry, namely as an industry that manages or bears the risks faced by individuals or business actors. This characteristic makes the aspect of insurance consumer protection very important, to ensure that insurance companies can fulfill their obligations to consumers when a risk occurs. Consumer protection can be maintained if all insurance industry players apply the principle of prudence effectively and carry out responsible business behavior.

Business competition on the one hand can encourage increased company efficiency and quality of service to consumers. Of course, this can benefit consumers because they can get

more affordable insurance products with good service quality. Increasing efficiency and service quality can also encourage industry growth and increase the competitiveness of the national insurance industry. However, on the other hand, tight competition can encourage companies to engage in market behavior that can harm consumers and industry stability, such as setting premium rates that are not commensurate with the risks borne, providing high marketing commissions to compete for marketing channels, and lack of transparency regarding the condition of the company and the products and services provided. Unhealthy business competition can harm other industry players and consumers.

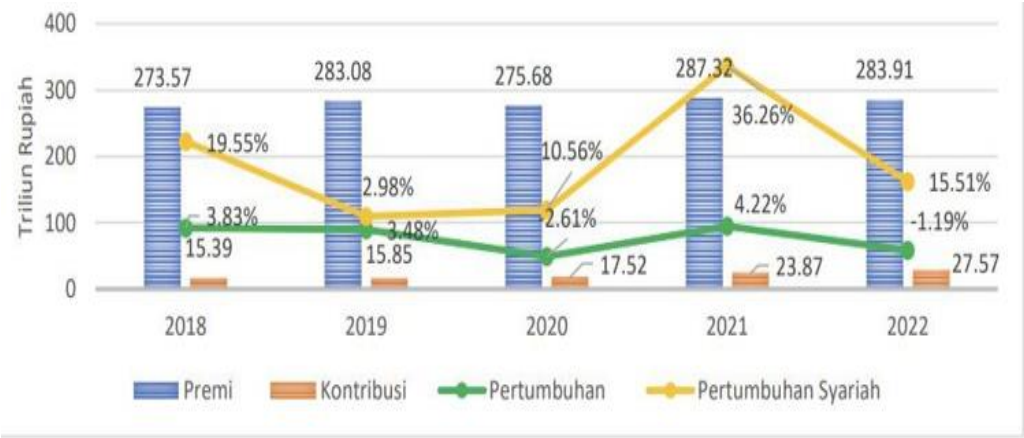


Figure 1. Premium Growth and Contribution in Indonesia
Souce: Otoritas Jasa Keuangan, 2023

Despite the growth, the role of the insurance industry in the national economy is relatively stagnant. This condition can be seen, among others, from the development of the insurance penetration rate which only grew from 2.81% in 2019 to 2.82% in 2022 (including social/mandatory insurance). Insurance penetration in Indonesia is also relatively low compared to other ASEAN countries. Based on data in the ASEAN insurance surveillance report 2022 (excluding mandatory/social insurance), in 2021 Indonesia's insurance penetration was 1.4%, Vietnam 2.2%, the Philippines 2.5%, Malaysia 3.8%, Thailand 4.6%, and Singapore 12.5%.

Figure 2. Penetration in ASEAN Country
Source: Otoritas Jasa Keuangan, 2023

Insurance density increased from IDR1,551,026 in 2018 to IDR2,006,214. Although nominally insurance density has increased, the figure is relatively low. Based on data in the ASEAN Insurance Surveillance Report 2022 (excluding mandatory/social insurance), in 2021 Indonesia's insurance density was IDR1,882,636, Brunei IDR6,115,960, the Philippines IDR1,354,763, Malaysia IDR6,575,558, Thailand IDR6,115,960, and Singapore IDR136,314,431.

LITERATURE REVIEW

Data Mining Steps

In 1950 Alan Turing published a paper entitled “Computing Machinery and Intelligent”, Winston (2017) which tells about computers that can think. In this paper he proposed a question which he called the Turing Test. The idea of the Turing test itself is that in order for a computer to pass the test, it must convince humans that the computer is human, the computer must be able to have a real conversation with humans. 1956: John Mc Carthy & Martin Misnky introduced the term artificial intelligence. So that 1956 was the year of the rise of the artificial intelligence era. 1997: For the first time, an IBM computer named IBM Deep Blue defeated world chess player Kasparov. 2006: Geoffrey Hinton introduced the term "Deep Learning" to explain a new algorithm that can make computers "see". With this deep learning, computers can distinguish between objects and text in images and videos as seen in this figure (Rapidminer Academy, 2024).

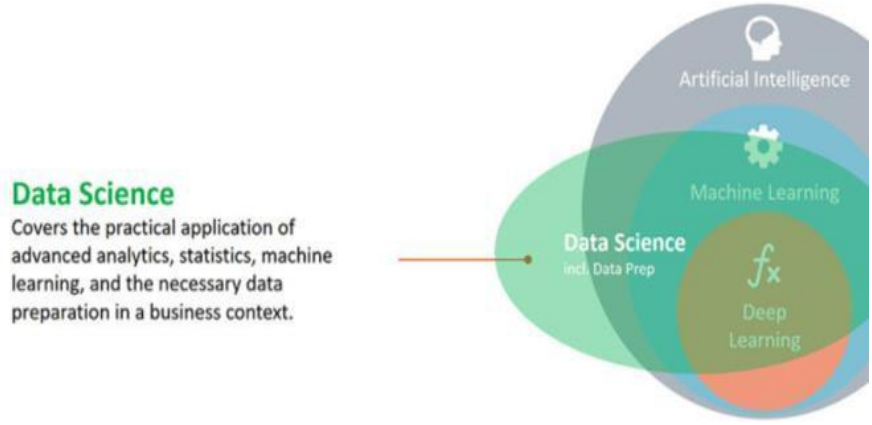


Figure 3. Data Science
Source: Rapidminer Academy, 2024

The data mining process is carried out with data preparation and continued with data processing or data cleaning. At this stage, data preparation begins for further processing, for example whether the data has a type of number or factor or date, and then the data in the cleaning data is also done by removing special characters, then after that the transformation is carried out, namely changing the data from cleansing data to target data, namely the next process is to carry out data mining or data models based on methods that are suitable for the data. The last stage is the process of interpreting knowledge obtained from data processing The development method using data mining is called CRIPS-DM (Cross Industrial Standards Program Data Mining) (Singgalen, 2024), as figure 3.

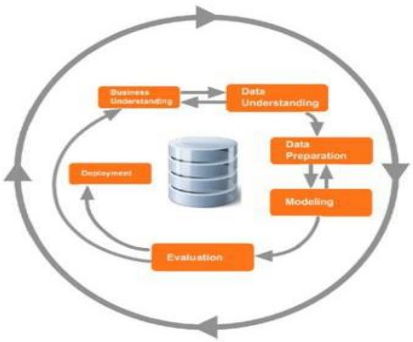


Figure 4. CRIPS-DM
Source: Rapidminer Academy, 2024

In the CRIPS-DM method, the first step is business understanding, where the purpose is to analyze using data mining for what, after that data that supports or is suitable for solving problems in business called data understanding, then data preparation is carried out, where data preparation is carried out data cleansing from the outlier, then reliability and validity and multicollinearity are tested, and finally in the heteroskedasticity test, and finally in the data duplication check. Then a model that matches the problem at hand is selected, after which the results are evaluated and interpreted, and the implementation plan is implemented.

Supervisor Learning

According to Duarte et al. (2019) classification techniques are computer programs that learn from given input data and use this training data with the aim of learning to classify based on observation patterns in the data. On the other hand, supervised learning for regression is a set of algorithms used to predict continuous values.

Strengthened by the statement Jijo & Abdulazeez (2021) decision tree has many used for classification case, and with using decision tree, there are which characteristics more important and can be used as a guideline for interpretation. Müller & Guido (2017) also stated that supervised learning is a type of machine learning in which the data will be divided into two different parts of data, first is training data dan second is testing, data dan it can be calculate how much of the acceleration from this data.

Random Forest

According to Ong et al. (2023) forests are an ensemble learning method that combines multiple decision trees to improve the accuracy of predictions. The basic idea behind random forests is to build many decision trees and then use the average of their predictions as the final prediction.

Besides that, according to Gkikas et al. (2022) a random forest is an ensemble-based classification algorithm that constructs a collection of decision trees and then aggregates their outputs to make a final decision. Each decision tree in the forest is grown independently, with a random subset of the available features used to determine the best split at each node. Moreover, according to Müller & Guido (2017) random forest has the unique feature, it's called feature important, whether this feature important, which feature are most influence to the dependent variable.

From Muhajir & Widiastuti (2022) random forests are a popular machine learning algorithm that can be used for classification and regression tasks—another feature importance of Random forests. The Feature importance from random forests can be used to identify the essential features for the prediction task. An essential feature because the algorithm considers

multiple features when constructing the decision trees, and the importance of each feature can be calculated based on its contribution to the performance of the forest.

RESEARCH METHOD

Population

Population from this research from many 38 industries, and with 148 sample data, with purposive sampling, has been collected by google form method. The criteria for purposive sampling is the respondent must have experience to closed insurance police.

Business Understanding

Data is collected from 148 respondents from 38 different industries. The question is which of these industries has the most adoption from insurance products.

Table 1. Industrial Distribution

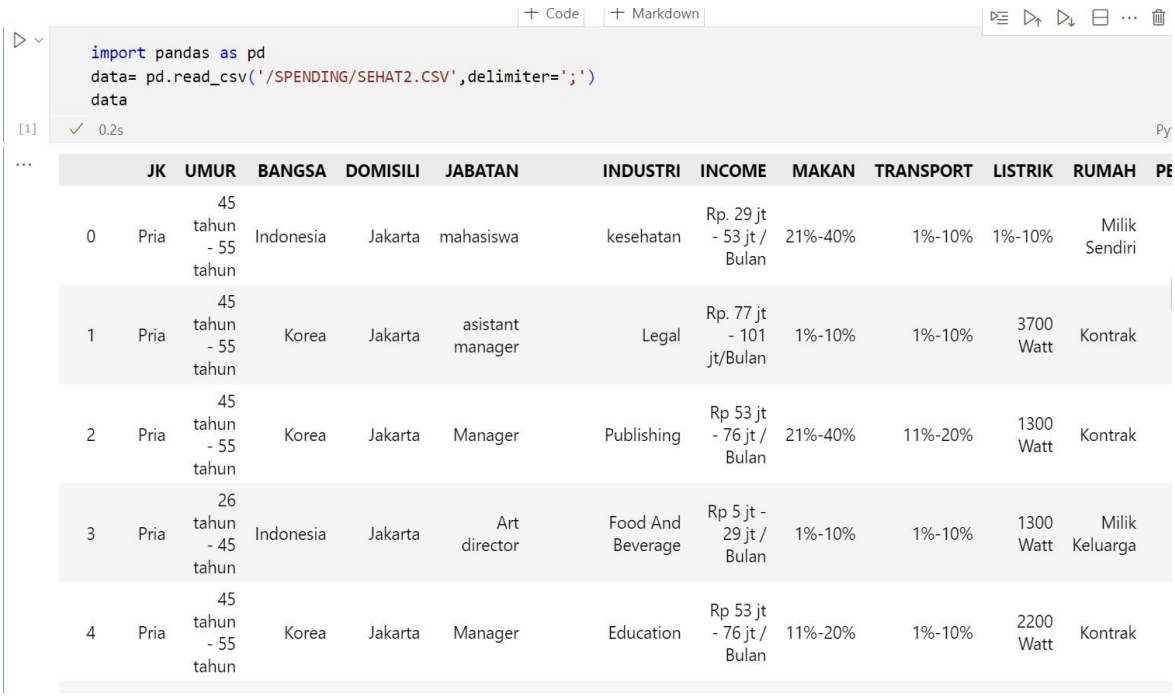
Banking	57
Insurance	17
Education	15
Retail	7
Agency	4
Food And Beverage	3
Health	3
Manufacture	3
Telecommunication	3
Finance	2
Hospitality	2
Health	2
Legal	2
Publishing	2
Service	2
Technology	2
Audit	1
Construction	1
Dropship Stairs Household	1
Ecommerce	1
Government	1
Government office	1
Healthcare	1
Home	1
IT	1
Financial Services	1
Leasing	1
Medical and Technology	1
Online business	1
Outsourcing	1

Procurement of goods and services	1
PNS	1
Property	1
Psychology Practice	1
Service Office	1
Tax and Audit	1
Think tank	1
Transportation	1

From table 1, most of data belongs to banking (57), insurance (17) and education (15) and retail (7) and the agency (4), and the rest is 2 or 1.

Data Understanding.

These data were taken from the primary data, whose taken from the questioner, and after that it will be put into the data set using python.



```
import pandas as pd
data= pd.read_csv('/SPENDING/SEHAT2.CSV',delimiter=';')
data
```

[1] ✓ 0.2s

	JK	UMUR	BANGSA	DOMISILI	JABATAN	INDUSTRI	INCOME	MAKAN	TRANSPORT	LISTRIK	RUMAH	PE
0	Pria	45 tahun - 55 tahun	Indonesia	Jakarta	mahasiswa	kesehatan	Rp. 29 jt - 53 jt / Bulan	21%-40%	1%-10%	1%-10%	Milik Sendiri	
1	Pria	45 tahun - 55 tahun	Korea	Jakarta	asistant manager	Legal	Rp. 77 jt - 101 jt/Bulan	1%-10%	1%-10%	3700 Watt	Kontrak	
2	Pria	45 tahun - 55 tahun	Korea	Jakarta	Manager	Publishing	Rp 53 jt - 76 jt / Bulan	21%-40%	11%-20%	1300 Watt	Kontrak	
3	Pria	26 tahun - 45 tahun	Indonesia	Jakarta	Art director	Food And Beverage	Rp 5 jt - 29 jt / Bulan	1%-10%	1%-10%	1300 Watt	Milik Keluarga	
4	Pria	45 tahun - 55 tahun	Korea	Jakarta	Manager	Education	Rp 53 jt - 76 jt / Bulan	11%-20%	1%-10%	2200 Watt	Kontrak	

Figure 5. Primary Data

Data Preparation

After the data set is input, data must decode first, because the data has much contained the category types. The data must decode into numeric types.

JK	UMUR	BANGS	DOMISILI	JABATAN	INDUSTR	INCOME	MAKAN	TRANSPORT	LISTRIK	RUMAH	PEND	SOCIAL_MA	SJJKELUARGA	SMOBIL2	KESEHATAN	STATUS
Pria	45 tahun - 55 tahun	Korea	Jakarta	asistant manage	Legal	Rp. 77 jt - 101 jt/Bulan	1%-10%	1%-10%	3700 Watt	Kontrak	S2	1%-10%	1%-10%	1%-10%	1%-10%	Menikah anak 2 orang
Pria	45 tahun - 55 tahun	Korea	Jakarta	Manager	Publishing	Rp 53 jt - 76 jt / Bulan	21%-40%	11%-20%	1300 Watt	Kontrak	S1	11%-20%	11%-20%	1%-10%	1%-10%	Menikah anak 2 orang
Pria	45 tahun - 55 tahun	Korea	Jakarta	Manager	Education	Rp 53 jt - 76 jt / Bulan	11%-20%	1%-10%	2200 Watt	Kontrak	S2	1%-10%	1%-10%	1%-10%	1%-10%	Menikath anak 3 orang
Pria	17 tahun - 25 tahun	Indonesia	Jakarta	Staff	Education	Rp 5 jt - 29 jt / Bulan	21%-40%	11%-20%	1300 Watt	Milik Keluarga	S1	1%-10%	1%-10%	1%-10%	11%-20%	Single
Wanita	17 tahun - 25 tahun	Indonesia	Jakarta	Staff	Technology	Rp 5 jt - 29 jt / Bulan	1%-10%	11%-20%	3700 Watt	Milik Keluarga	S1	21%-40%	21%-40%	11%-20%	11%-20%	Single
Pria	17 tahun - 25 tahun	Indonesia	Jakarta	Staff	Education	Rp 5 jt - 29 jt / Bulan	21%-40%	1%-10%	2200 Watt	Milik Keluarga	S1	1%-10%	11%-20%	1%-10%	1%-10%	Single
Wanita	17 tahun - 25 tahun	Indonesia	Manado	Staff	Banking	Rp 53 jt - 76 jt / Bulan	21%-40%	11%-20%	3700 Watt	Milik Keluarga	S1	11%-20%	21%-40%	21%-40%	41%-60%	Single
Pria	26 tahun - 45 tahun	Indonesia	Jakarta	Manager	Banking	Rp 5 jt - 29 jt / Bulan	11%-20%	1%-10%	900 Watt	Kontrak	S1	1%-10%	1%-10%	1%-10%	1%-10%	Menikah anak 1 orang
Pria	26 tahun - 45 tahun	Indonesia	Depok	Senior Manager	Banking	Rp 5 jt - 29 jt / Bulan	1%-10%	11%-20%	900 Watt	Kontrak	S2	1%-10%	11%-20%	1%-10%	1%-10%	Menikah anak 1 orang
Wanita	17 tahun - 25 tahun	Indonesia	Depok	Staff	Banking	< Rp 5 jt / Bulan	41%-60%	11%-20%	450 Watt	Kontrak	S1	11%-20%	11%-20%	1%-10%	11%-20%	Single
Wanita	26 tahun - 45 tahun	Indonesia	Jakarta	Staff	Banking	Rp 5 jt - 29 jt / Bulan	11%-20%	1%-10%	1300 Watt	Kontrak	S1	11%-20%	11%-20%	1%-10%	1%-10%	Menikah anak 1 orang
Pria	lebih dari 55 tahun	Indonesia	Jakarta	Senior Manager	Banking	Rp 53 jt - 76 jt / Bulan	1%-10%	1%-10%	900 Watt	Milik Sendiri	S1	1%-10%	1%-10%	1%-10%	1%-10%	Menikah anak 2 orang
Pria	lebih dari 55 tahun	Indonesia	depok	Staff	Banking	Rp 53 jt - 76 jt / Bulan	1%-10%	1%-10%	> 4400 Watt	Milik Sendiri	S3	1%-10%	1%-10%	1%-10%	500 ribu - 1 juta	Menikath anak 3 orang
Pria	45 tahun - 55 tahun	Indonesia	Jakarta	Staff	Banking	Rp. 77 jt - 101 jt/Bulan	11%-20%	1%-10%	> 4400 Watt	Milik Sendiri	S1	11%-20%	11%-20%	11%-20%	2 juta - 3 juta	Menikah anak 2 orang
Pria	17 tahun - 25 tahun	Indonesia	Tangerang	Staff	Banking	Rp 5 jt - 29 jt / Bulan	11%-20%	41%-60%	2200 Watt	Kos Kosan	S1	11%-20%	11%-20%	1%-10%	1 juta - 2 juta	Single
Wanita	26 tahun - 45 tahun	Indonesia	Jakarta	Manager	Banking	Rp 5 jt - 29 jt / Bulan	21%-40%	1%-10%	3700 Watt	Milik Sendiri	S1	1%-10%	1%-10%	1%-10%	500 ribu - 1 juta	Menikah anak 2 orang
Wanita	17 tahun - 25 tahun	Indonesia	Jakarta	Staff	Banking	Rp 5 jt - 29 jt / Bulan	1%-10%	1%-10%	900 Watt	Milik Keluarga	S1	21%-40%	21%-40%	1%-10%	1 juta - 2 juta	Single
Wanita	17 tahun - 25 tahun	Indonesia	Jakarta	Senior Manager	Banking	< Rp 5 jt / Bulan	1%-10%	11%-20%	2200 Watt	Milik Keluarga	S1	1%-10%	21%-40%	1%-10%	500 ribu - 1 juta	Single
Wanita	17 tahun - 25 tahun	Indonesia	Jakarta	Staff	Insurance	Rp 5 jt - 29 jt / Bulan	11%-20%	1%-10%	2200 Watt	Milik Sendiri	S1	1%-10%	1%-10%	21%-40%	500 ribu - 1 juta	Single
Wanita	26 tahun - 45 tahun	Indonesia	Tangerang Se	Senior Manager	Banking	Rp 5 jt - 29 jt / Bulan	21%-40%	21%-40%	2200 Watt	Milik Sendiri	S1	11%-20%	11%-20%	11%-20%	1 juta - 2 juta	Menikah anak 1 orang
Pria	26 tahun - 45 tahun	Indonesia	Tangerang	Manager	Banking	Rp 5 jt - 29 jt / Bulan	11%-20%	11%-20%	1300 Watt	Milik Sendiri	S1	1%-10%	1%-10%	1%-10%	500 ribu - 1 juta	Menikah anak 2 orang
Wanita	17 tahun - 25 tahun	Indonesia	Jakarta	Senior Manager	Banking	Rp 5 jt - 29 jt / Bulan	11%-20%	1%-10%	2200 Watt	Milik Keluarga	S1	11%-20%	11%-20%	1%-10%	500 ribu - 1 juta	Single
Wanita	17 tahun - 25 tahun	Indonesia	Jakarta	Manager	Banking	Rp 5 jt - 29 jt / Bulan	21%-40%	11%-20%	1300 Watt	Kontrak	S1	11%-20%	11%-20%	1%-10%	500 ribu - 1 juta	Single
Wanita	17 tahun - 25 tahun	Indonesia	Jakarta	Manager	Banking	Rp 5 jt - 29 jt / Bulan	21%-40%	1%-10%	2200 Watt	Milik Keluarga	S2	1%-10%	1%-10%	1%-10%	500 ribu - 1 juta	Single
Wanita	26 tahun - 45 tahun	Indonesia	Jakarta	Manager	Banking	Rp. 30 jt - 52 jt / Bulan	21%-40%	1%-10%	4400 Watt	Milik Keluarga	S1	21%-40%	21%-40%	1%-10%	1 juta - 2 juta	Single
Wanita	26 tahun - 45 tahun	Indonesia	Jakarta	Manager	Banking	Rp. 30 jt - 52 jt / Bulan	1%-10%	1%-10%	4400 Watt	Milik Sendiri	S1	1%-10%	1%-10%	11%-20%	3 juta - 5 juta	Menikath anak 3 orang
Pria	26 tahun - 45 tahun	Indonesia	Jakarta	Manager	Insurance	Rp. 30 jt - 52 jt / Bulan	1%-10%	1%-10%	2200 Watt	Milik Sendiri	S1	1%-10%	1%-10%	1%-10%	2 juta - 3 juta	Menikah anak 1 orang
Pria	26 tahun - 45 tahun	Indonesia	Jakarta	Senior Manager	Insurance	Rp. 30 jt - 52 jt / Bulan	21%-40%	11%-20%	2200 Watt	Milik Sendiri	S1	1%-10%	11%-20%	1%-10%	2 juta - 3 juta	Menikah anak 1 orang

Figure 6. Raw Data

```
from sklearn.preprocessing import LabelEncoder
enc= LabelEncoder()
data['JK']=enc.fit_transform(data['JK'].values)
data['UMUR']=enc.fit_transform(data['UMUR'].values)
data['BANGSA']=enc.fit_transform(data['BANGSA'].values)
data['DOMISILI']=enc.fit_transform(data['DOMISILI'].values)
data['JABATAN']=enc.fit_transform(data['JABATAN'].values)
data['INDUSTRI']=enc.fit_transform(data['INDUSTRI'].values)
data['INCOME']=enc.fit_transform(data['INCOME'].values)
data['MAKAN']=enc.fit_transform(data['MAKAN'].values)
data['TRANSPORT']=enc.fit_transform(data['TRANSPORT'].values)
data['LISTRIK']=enc.fit_transform(data['LISTRIK'].values)
data['PENDIDIKAN']=enc.fit_transform(data['PENDIDIKAN'].values)
data['SOCIAL_MAKANTEMAN']=enc.fit_transform(data['SOCIAL_MAKANTEMAN'].values)
data['SJJKELUARGA']=enc.fit_transform(data['SJJKELUARGA'].values)
data[' SMOBIL2 ']=enc.fit_transform(data[' SMOBIL2 '].values)
data[' STATUS ']=enc.fit_transform(data[' STATUS '].values)
data[' KESEHATAN ']=enc.fit_transform(data[' KESEHATAN '].values)
data['RUMAH']=enc.fit_transform(data['RUMAH'].values)
```

```
data
```

✓ 0.0s

	JK	UMUR	BANGSA	DOMISILI	JABATAN	INDUSTRI	INCOME	MAKAN	TRANSPORT	LISTRIK	RUMAH	PENDIDIKAN
0	0	2	0	5	20	39	6	2	0	0	5	4
1	0	2	1	5	19	21	8	0	0	3	1	5
2	0	2	1	5	10	30	4	2	1	1	1	4
3	0	1	0	5	1	9	3	0	0	1	3	4
4	0	2	1	5	10	7	4	1	0	2	1	5
...	...	...	...	...	...	...	...	...	...	...	...	...
143	1	0	0	5	16	31	3	0	3	1	3	4
144	1	0	0	5	16	36	3	1	2	1	3	4
145	1	0	0	5	10	1	3	0	0	1	1	4
146	0	1	0	0	10	31	7	1	0	2	5	3
147	1	1	0	5	15	3	3	1	1	2	5	4

JK	UMUR	BANGSA	DOMISILI	JABATAN	INDUSTRI	INCOME	MAKAN	TRANSPOR	LISTRIK	RUMAH	PENDIDIKA	SOCIAL_M	SJKELUAR	SMOBIL2	STATUS
0	2	1	5	18	21	6	0	0	2	1	4	0	0	0	1
0	2	1	5	11	30	3	2	1	0	1	3	1	1	0	1
0	2	1	5	11	7	3	1	0	1	1	4	0	0	0	2
0	0	0	5	15	7	2	2	1	0	3	3	0	0	0	3
1	0	0	5	15	35	2	0	1	2	3	3	2	2	1	3
0	0	0	5	15	7	2	2	0	1	3	3	0	1	0	3
1	0	0	8	15	3	3	2	1	2	3	3	1	2	2	3
0	1	0	5	11	3	2	1	0	6	1	3	0	0	0	0
0	1	0	4	14	3	2	0	1	6	1	4	0	1	0	0
1	0	0	4	15	3	0	3	1	4	1	3	1	1	0	3
1	1	0	5	15	3	2	1	0	0	1	3	1	1	0	0
0	3	0	5	14	3	3	0	0	6	5	3	0	0	0	1
0	3	0	27	15	3	3	0	0	7	5	5	0	0	0	2
0	2	0	5	15	3	6	1	0	7	5	3	1	1	1	1
0	0	0	24	15	3	2	1	3	1	2	3	1	1	0	3
1	1	0	5	11	3	2	2	0	2	5	3	0	0	0	1
1	0	0	5	15	3	2	0	0	6	3	3	2	2	0	3
1	0	0	5	14	3	0	0	1	1	3	3	0	2	0	3
1	0	0	5	15	18	2	1	0	1	5	3	0	0	2	3
1	1	0	25	14	3	2	2	2	1	5	3	1	1	1	0
0	1	0	24	11	3	2	1	1	0	5	3	0	0	0	1
1	0	0	5	14	3	2	1	0	1	3	3	1	1	0	3
1	0	0	5	11	3	2	2	1	0	1	3	1	1	0	3
1	0	0	5	11	3	2	2	0	1	3	4	0	0	0	3
1	1	0	5	11	3	5	2	0	3	3	3	2	2	0	3
1	1	0	5	11	3	5	0	0	3	5	3	0	0	1	2
0	1	0	5	11	18	5	0	0	1	5	3	0	0	0	0
0	1	0	5	14	18	5	2	1	1	5	3	0	1	0	0
1	0	0	5	15	12	2	2	3	0	3	3	1	2	2	3

Figure 7. Encoding Data

RESULTS AND DISCUSSION

Modeling

After the decoding process, it will be ready to make a model. The model which chooses based on the data types was the random forest.

```
Create and train model
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.tree import DecisionTreeClassifier
import sklearn.model_selection as ms
X=data[['JK','UMUR','BANGSA','DOMISILI','JABATAN','INDUSTRI','INCOME','MAKAN','TRANSPORT','LISTRIK','RUMAH',
X
y=data[' KESEHATAN ']
y
```

```
    accuracy=met.accuracy_score(y_test,y_prediksi)
    print('Accuracy= ',accuracy)
 ✓ 0.0s

Accuracy=  0.5813953488372093
```

Figure 8. Modeling

After the modeling run, it will result the accuracy 58 %, so the model can predict 58 % how much of the health insurance will be taken from each industry. from table 3. The random forest algorithm has function, is called a feature important. The feature important is a formula that indicates, where factors that will influence the closing with health product from insurance. The highest factor is electrical (0,102), this factor is reflected in style of life, the bigger the electricity the wealthier. The second factor is industry, the most closing is become from banking (57), insurance (17), education (15), retail (7), from table 4. Projection closing health insurance from each industry, and the third is domicile, and the most recent from Jakarta (95), Depok (6), Bogor (6), Tangerang (5), Surabaya (4), so the distribution of insurance is dominated

from JABODETABEK. Next from the age, the most closing is become from age 26 to 46 (62), the second from age 17–25 (48). So, in this case, the marketing must approach the candidate customer from the age of 26 to 46 years old. And the final analysis, the position the most closed the health insurance is from staff (68), manager (34) and senior manager (19).

Table 2. Feature Important

Factor	Amount
Electricity	0.10270245066995426
Industry	0.09599428308565679
Domicile	0.08415363155356075
Age	0.08233508935446782
Position	0.07853930095280212

Table 3. Projection Closing Health Insurance from Each Industry

Industry	Industry Code	Total
Banking	3	57
Insurance	18	17
Education	7	15
Retail	31	7
Agency	1	4
Food And Beverage	9	3
Health	12	3
Manufacture	22	3
Telecommunication	36	3
Finance	8	2
Hospitality	15	2
Health	39	2
Legal	21	2
Publishing	30	2

Table 4. The Most Contribution from Domicile

Jakarta	95
Depok	6
Bogor	6
Tangerang	5
Surabaya	4
Bekasi	4
Semarang	2
Medan	2

Table 5. Distribution of Age

Age	Total
17 years - 25 years	48
26 ears - 45 years	62
45 years - 55 years	24
>55 years	9

Table 6. Distribution of Position

Position	Total
Staff	68
Manager	34
Senior Manager	19
Director	5
assistant manager	2
CEO	2
Art director	1
Lecturer	1

Evaluation

From the result of the projection closing health insurance from each industry, it can be concluded that the most industry that closed the health insurance is Banking Industry, the second is from insurance and the third is education and the next is retail, health, manufacturing and finance, hospitality, legal, publishing, technology and government and service industries. This results because there are from the financial literacy (Lopus et al., 2019).

CONCLUSION

Nowadays many insurance agents have been taught to know about the needs of customers, but today is more accurate to use the behavior of the customers. So, by using machine learning algorithms is more accurate to make a prediction of health insurance premium by using the random forest algorithm. From the result of the projection closing health insurance from each industry, it can be concluded that the most industry that closed the health insurance is banking industry, the second is from insurance and the third is education and the next is education, retail, health, manufacturing and finance, hospitality, legal, publishing, technology and government and service industries, with 58 accuracy score.

REFERENCES

Duarte, V., Zuñiga-Jara, S., & Contreras, S. (2022). Machine learning and marketing: A systematic literature review. *IEEE Access, 10*, 93273–93288. https://doi.org/10.1109/ACCESS.2022.3202896

Gkikas, D. C., Theodoridis, P. K., & Beligiannis, G. N. (2022). Enhanced marketing decision making for consumer behaviour classification using binary decision trees and a genetic algorithm wrapper. *Informatics, 9*(2), 1–29. https://doi.org/10.3390/informatics9020045

Jijo, B. T., & Abdulazeez, A. M. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(1), 20–28. https://doi.org/10.38094/jastt20165

Lopus, J. S., Amidjono, D. S., & Grimes, P. W. (2019). Improving financial literacy of the poor and vulnerable in Indonesia: An empirical analysis. *International Review of Economics Education*, *32*(11), 196–202. https://doi.org/10.1016/j.iree.2019.100168

Muhajir, M., & Widiastuti, J. (2022). Random forest method approach to customer classification based on non-performing loan in micro business. *Jurnal Online Informatika*, *7*(2), 177–183. https://doi.org/10.15575/join.v7i2.842

Müller, A. C., & Guido, S. (2017). *Introduction to machine learning with python: A guide for data scientists.* O'Reilly Media.

Ong, A. K. S., Cordova, L. N. Z., Longanilla, F. A. B., Caprecho, N. L., Javier, R. A. V., Borres, R. D., & German, J. D. (2023). Purchasing intentions analysis of hybrid cars using random forest classifier and deep learning. *World Electric Vehicle Journal*, *14*(8), 1–26. https://doi.org/10.3390/wevj14080227

Otoritas Jasa Keuangan. (2023). *Roadmap pengembangan perasuransian Indonesia 2023–2027*. Departemen Pengaturan dan Pengembangan IKNB, OJK. www.go.id

Rapidminer, Academy. (2024). *Data science professional*. Rapidminer Academy. https://academy.rapidminer.com/learning-paths/data-science-professional-with-rapidminer

Singgalen, Y. A. (2024). Implementation of CRISP-DM for social network analysis (SNA) of tourism and travel vlog content reviews. *Jurnal Media Informatika Budidarma*, *8*(1), 572–583. https://doi.org/10.30865/mib.v8i1.7323

Winston, P. H. (2017). On computing machinery and intelligence. *Boston Studies in the Philosophy and History of Science*, *324*, 265–278. https://doi.org/10.1007/978-3-319-53280-6_11