IConEnt

The 5th International Conference on Entrepreneurship

Characteristics of Customer Who Makes Allocation for Pension Plan in Indonesia

Hendra Achmadi[a], Isana Sri Christina Meranga[b], Sylvia Samuel[c]

[a, b, c] Pelita Harapan University, M.H. Thamrin Boulevard 1100, Tangerang 15811, Indonesia

E-mail : hendra.achmadi@uph.edu[a]; isana.meranga@uph.edu[b]; sylvia.samuel@uph.edu[c]

ABSTRACT

This study sought to determine which characteristic from Indonesian people are most likely to make allocation for pension, and how to identify the key factors driving these purchases pension products. Recognizing Indonesia's relatively low allocation in pension plan, researchers analyzed a targeted sample of 199 data points from 38 companies across various sectors. Using the CRISP-DM data mining methodology, to determine what characteristics from customer with over 20 % allocation into the pension product, and in this study found that marriage status, income, industry, education, sex, domicile, electricity, position, age, status of property. The CRISP-DM process, which included defining objectives, data collection, preparation, modeling, and property ownership, as a characteristic of customer to invest in pension. This research uses data mining methodology with resulted in a model with 73.33 % accuracy. [1] The analysis revealed that the result demonstrated the highest stock market allocation by customers is based on industry (0,2), Listrik (0,13), Makan (0,12), transport (0,11), Pendidikan (0,10), income (0,06).

***Keywords* -** Pension Allocation, CRIPS-DM, Random Forest.

INTRODUCTION

In Indonesia, financial planning for retirement remains an underdeveloped practice among the general population. While the concept of retirement is widely acknowledged, proactive preparation particularly in the form of long-term savings or investment has not become a cultural norm. Many Indonesians rely heavily on government-provided programs such as the Social Security Administration (BPJS), which includes health and employment benefits. Although these schemes are helpful, they are often insufficient to fully support individuals during their retirement years, especially considering inflation, rising healthcare costs, and increasing life expectancy.

Several factors contribute to this lack of preparedness. First, financial literacy in Indonesia is relatively low. A significant portion of the population lacks a clear understanding of the importance of retirement planning and the tools available to secure financial stability in old age. Secondly, the majority of the workforce is employed in the informal sector, where pension schemes are rarely offered. Even among formal workers, many do not actively contribute to private retirement plans, relying instead on short term income and government support.

The consequence of this trend is a growing number of elderly citizens who face financial insecurity after retirement. This condition poses challenges not only for individuals and families but also for the broader social and economic systems. The government may be burdened with increased social assistance demands, and families often bear the responsibility of supporting older relatives, potentially impacting their own financial well-being.

Addressing this issue requires a multi-faceted approach: improving financial education, promoting awareness of personal retirement planning, and expanding access to inclusive and sustainable pension solutions. As Indonesia moves toward becoming a more developed economy, fostering a culture of financial independence in retirement is essential to ensuring social resilience and economic sustainability for future generations.

LITERATURE REVIEW

In the current digital age, data is considered a highly valuable resource. The ability to derive meaningful insights from vast datasets is essential. Data mining, which involves uncovering patterns and useful information from data, provides a powerful approach to this challenge. A widely adopted framework for this is CRISP-DM (Cross-Industry Standard Process for Data Mining).

Developed in the mid-1990s, CRISP-DM presents a systematic and adaptable approach to executing data mining projects. It consists of six key phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. By adhering to this process, professionals can maintain clear project direction and generate actionable outcomes. CRISP-DM has demonstrated its effectiveness across multiple industries, including finance, healthcare, retail, and manufacturing. Its flexibility in handling different data types and problem domains makes it a preferred methodology for organizations aiming to harness the power of their data.

Data Mining

The journey of Data Science and Artificial Intelligence (AI) traces back to Alan Turing's pioneering work on machine intelligence and the concept of the Turing Test (Winston, 2017).

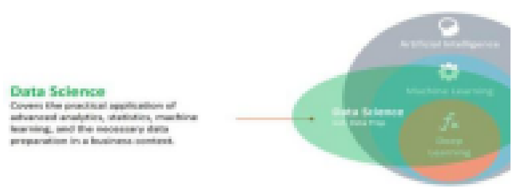


Fig. 1. Data Science

Source: (Academy, 2024)

This foundational work paved the way for AI's official emergence in 1956 and was followed by notable achievements, such as IBM's Deep Blue defeating chess champion Garry Kasparov in 1997, and the rise of Deep Learning in 2006, which brought major advancements in computer vision (Academy, 2024). These key developments are visually represented in Figure 3. Additionally, Singgalen (2024) outlines the CRISP-DM framework for data mining, which begins with meticulous data preparation—such as cleaning and transformation—before implementing suitable mining models and analyzing the generated insights. Figure 3 also captures the stages of this data mining approach.

The CRISP-DM methodology comprises six phases:

1. **Business Understanding**: Establishing the project's objectives and requirements from a business perspective.

2. **Data Understanding**: Gathering and examining initial data to identify issues and insights pertinent to the business objectives.

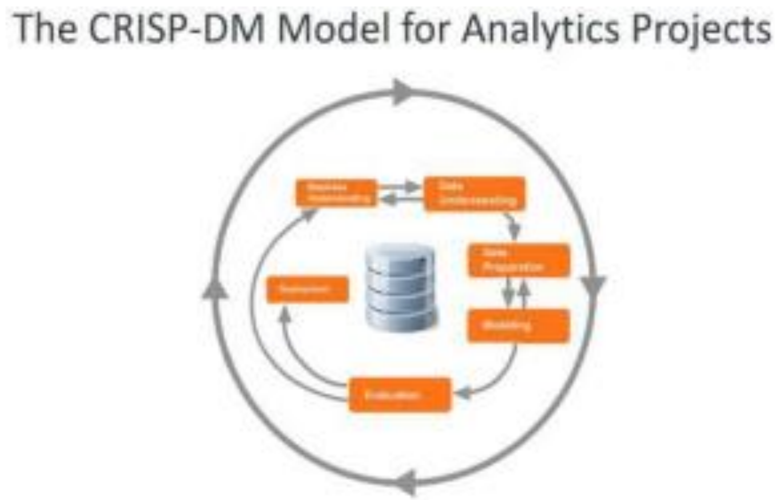


Fig. 2. CRIPS-DM

Source: (Academy, 2024)

3. **Data Preparation**: Processing raw data into a suitable format for modeling, which includes cleaning, transforming, and selecting relevant features.

4. **Modeling**: Selecting and applying appropriate modeling techniques to the prepared data. 5. **Evaluation**: Assessing the model's performance to ensure it meets the business objectives and is both reliable and valid.

6. **Deployment**: Implementing the model's insights into actionable strategies or systems within the business context.

This systematic methodology ensures that data mining initiatives remain aligned with organizational objectives and yield practical, actionable outcomes. The CRISP-DM framework starts with Business Understanding, where the goals of the analysis are clearly defined. It then moves to Data Understanding, which involves identifying and gathering the most relevant data to address the problem. In the Data Preparation phase, the data is cleaned—such as by addressing outliers validated for accuracy through tests like multicollinearity and heteroscedasticity checks, and duplicates are removed. Once the dataset is ready, a suitable modeling technique is chosen. This is followed by an evaluation of the model, interpretation of the findings, and the final step: deploying the insights into real-world applications.

Supervisor Learning

According to Duarte et al. (2019), classification methods involve teaching algorithms to recognize patterns in input data so they can categorize new instances accurately. Unlike classification, supervised regression is used to predict continuous numerical values. Charbuty and Abdulazeez (2021) point out that decision trees are commonly used in classification due to their ability to identify key features, making the results more interpretable. Additionally, Müller and Guido (2017) explain that supervised learning requires splitting the dataset into training and testing portions to evaluate how well the model can make predictions.

Random Forest

Random forests are a type of ensemble learning method that aggregates the results of several decision trees to enhance prediction accuracy (Ong et al., 2023). Each decision tree is built independently, using randomly selected subsets of features at each node split, which contributes to the model's strength and its ability to generalize well (Müller and Guido, 2017). One key advantage of random forests is their ability to evaluate the importance of features by measuring the reduction in node impurity—weighted by the likelihood of reaching each node across the entire ensemble (Muhajir and Widiastuti, 2022).

This technique is highly adaptable and has been successfully utilized in a range of fields, such as healthcare, finance, and image processing. Its capacity to manage large and complex datasets, evaluate feature relevance, and resist overfitting makes it a powerful approach for predictive analytics. Probst et al. (2019) delve into the statistical underpinnings and real-world uses of the Random Forest algorithm while Schonlau and Zou (2020) explore how various hyperparameters affect model performance and offer strategies for optimizing them effectively.

METHODOLOGY

3.1 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 criterion for purposive sampling is the respondent must have experience to close insurance policy.

3.2 Business Understanding

Data is collected from 199 respondents from 38 different industries. The question is which of these industries has the most adoption from the insurance products.

Table 1. Industrial Distribution

Banking	57
Insurance	17
Education	15

Retail	7
Agency	4
Food And Beverage	3
Bond	3
Manufacture	3
Telecommunication	3
Finance	2
Hospitality	2
Kesehatan	2
Legal	2
Publishing	2
Service	2
Technology	2
Audit	1
Construction	1
Dropship Kebutuhan Rumah Tangga	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 are 2 or 1.

3.2 Data Understanding

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

3.3 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.

Table 2. Raw Data

```python
from sklearn.preprocessing import LabelEncoder
enc= LabelEncoder()
data['jk']=enc.fit_transform(data['jk'].values)
data['status']=enc.fit_transform(data['status'].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['rumah']=enc.fit_transform(data['rumah'].values)
data['pendidikan']=enc.fit_transform(data['pendidikan'].values)
data['pensiun']=enc.fit_transform(data['pensiun'].values)
```

jk	status	umur	bangsa	domisili	jabatan	industri	income	makan	transport	listrik	rumah	pendidikan	pensiun	nsiun_kategori
1	2	2	0	5	3	7	4	0	0	1	3	5	3	1
1	4	1	0	7	14	19	4	1	1	6	2	4	1	0
1	4	2	0	1	14	31	4	2	2	0	3	3	1	0
0	0	2	0	5	6	27	4	1	1	1	3	3	1	0
0	1	3	0	5	7	7	6	2	2	3	3	5	3	1
1	3	2	0	2	5	30	7	1	0	0	3	1	1	0
1	2	1	0	5	13	7	4	3	3	1	3	5	3	1
1	4	0	0	5	1	19	2	0	0	5	2	3	2	0
0	0	2	0	2	14	7	4	2	1	0	3	4	1	0
0	3	2	0	3	4	21	1	0	1	6	3	3	1	0
1	1	3	0	5	13	32	4	2	2	2	3	3	1	0
1	0	2	0	5	14	17	4	0	0	0	3	3	4	1
0	4	2	0	5	1	2	5	0	0	0	3	4	1	0
0	4	2	0	2	14	7	4	2	0	0	3	3	1	0
1	4	0	0	4	14	7	4	0	0	6	2	3	1	0
1	1	1	0	1	14	7	4	2	1	1	3	3	1	0
1	0	1	0	11	8	15	4	0	0	1	3	4	1	0
1	2	2	0	5	8	22	4	2	1	6	3	3	1	0
0	1	1	0	7	9	11	1	3	3	6	2	3	1	0
0	0	3	0	5	1	29	2	0	0	7	3	4	1	0
1	0	1	0	5	14	25	1	2	2	0	0	6	2	0
1	4	2	0	5	2	7	4	4	2	1	2	5	3	1

Table 3. Data Encoding

RESULTS

Modeling

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

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, cl

import sklearn.model_selection as ms
X=data[['status','umur','domisili','jabatan','industri','income'
X
y=data['pensiun']
y
```

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

```
=== EVALUASI MODEL ===
Akurasi: 0.7333
```

```
print(met.classification_report(y_test,y_prediksi))
```

precision recall f1-score support

0 0,91 0,77 0,83 13,00

1 0,25 0,50 0,33 2,00

accuracy 0,73 0,73 0,73 0,73

macro avg 0,58 0,63 0,58 15,00

Weighted avg 0,82 0,73 0,77 15,00

Fitur	Importance
industry	0,20
Listrik	0,13
Makan	0,12
transport	0,11
Pendidikan	0,10
income	0,06

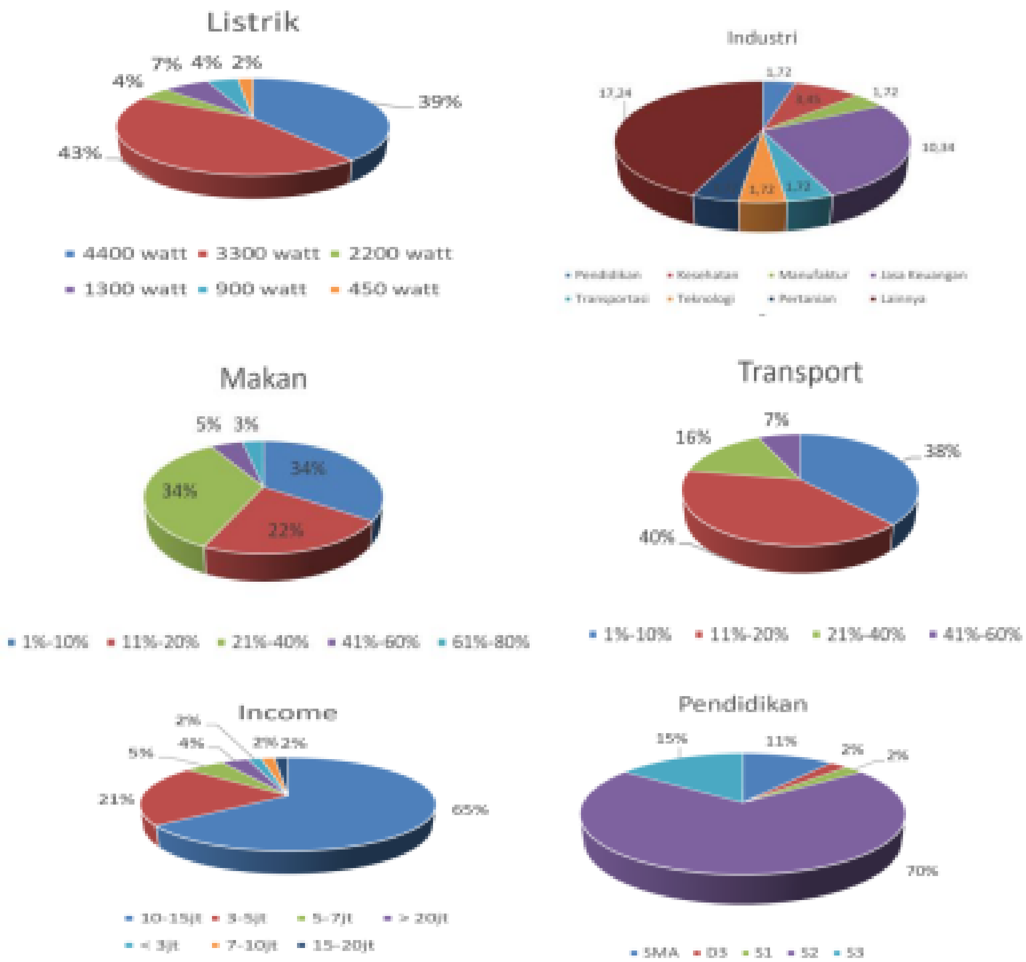
Table 4. Feature Important with Pension Allocation > 20 %.

The feature importance results indicate that ‘industry’ holds the highest influence (0.20) in the model, suggesting that the industry sector in which an individual works significantly affects the target prediction—potentially related to retirement readiness or financial behavior. This makes sense, as different industries offer varying levels of income stability, benefits, and retirement plans.

Following closely are ‘Listrik’ (0.13), Makan’ (0.12), and ‘transport’ (0.11), reflecting monthly expenditure categories that correlate with financial lifestyle and spending capacity. These features provide strong signals about an individual's financial commitments and cost of living, which can influence savings or pension decisions.

‘Pendidikan’ (0.10) and ‘Jabatan’ (0.08) are also important, indicating that educational background and job position play a crucial role in financial outcomes. Higher education and higher-ranking positions typically correlate with better income and financial literacy.

Interestingly, income’ ranks low (0.06), due to redundancy with other variables like ‘jabatan’ and spending features. Demographic factors like ‘domisili,’ rumah’, ‘umur’, ‘status’, ‘jk’, and ‘bangsa’ show even lower importance, suggesting they have less predictive power in this context.



EVALUATION

The feature importance analysis reveals that spending behavior , and there are six important features that determine the behavior to make investment to pension more than 20 % , the first is financial industry and health industry , second the electricity has 3300 watt, third is the behavior to tread a friend is only 1%-10 % , the fourth is the expense for transportation is 11 % - 20 % and the fifth is the most education is a master degree and finally the most income is 15 – 20 million rupiah.

CONCLUSION

The analysis shows that people who invest over 20% of their income for pension typically work in the financial or health industry, have high electricity usage (3300W), spend only 1–10% behavior to treat on friends, allocate 11–20% for transport, hold a Master's degree, and earn 15–20 million Rupiah monthly. To prepare for retirement, individuals should adopt a disciplined financial mindset, limit non-essential spending, and consistently allocate a portion of their income ideally 20% or more toward pension savings. Education, balanced expenses, and long-term planning are key to building a secure future. Start early, stay consistent, and adjust as income increases.

REFERENCES

Academy, R. (2024). *Data Science*. RapidMiner Academy.

https://academy.rapidminer.com/learning-paths/data-science-professional-with rapidminer

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Duarte, V., Zuñiga-Jara, S., & Contreras, S. (2019). *Machine Learning and Marketing: a literature review*. https://ssrn.com/abstract=4006436

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). https://doi.org/10.3390/informatics9020045

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*(May), 100168. 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 Introduction to Machine Learning with Python*.

OJK, O. J. K. (2023). Draft Roadmap Pengembangan Perasuransian Indonesia. *Departemen Pengaturan Dan Pengembangan IKNB, OJK*, 1–63. www.ojk.go.id.

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).

https://doi.org/10.3390/wevj14080227

Singgalen, Y. A. (2024). Implementation of CRISP-DM for Social Network Analysis (SNA) of Tourism and Travel Vlog Content Reviews. *Jurnal Media Informatika Budidarma, 2024*, *8*, 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