ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 14 (2024) No. 6, pp. 300-315
DOI:10.33168/JSMS.2024.0619

Classification and Recommendations for Insurance Products Using Naïve Bayes Classification Based on Customer Data

Denni Dwi Kurnianto, Sfenrianto

Information System Management Department, BINUS Graduate Program, Master of Information System Management, Bina Nusantara University, Jakarta 11480, Indonesia
dennidwikurnianto3.6@gmail.com

Abstract. The Digitalization Era was born with a series of benefits that were presented, such as being able to help companies to more easily reach their customers. Digitization is the increasing availability of digital data made possible by advances in creating, transferring, storing and analyzing digital data, and has the potential to structure, shape and influence the contemporary world. At the beginning of 2021 the insurance company had a huge impact due to the COVID-19 pandemic on the insurance claim system which required customers need to go to the office. This is due to social distancing regulations and requires working from home. Following the issue, insurance companies create applications that make it easier for customers such as insurance claims, payments, and application the insurance. There are 4 types of insurance available, Bebas Rencana Optimal, Sprint Link Smart, Critical Armor and BIPS. These are insurance can protect and cover all types of diseases and all age groups. The aim of this research is to classify customers using the Naïve Bayes method at PT XYZ and to find out the level of accuracy. The Naïve Bayes method is to classify probability by adding up the frequencies and combinations of values from a given dataset. Besides that, not only for insurance claims and payments, this application is a place of communication between insurance companies and customers. With this application, it is expected to strengthen customer relationships with insurance companies and make it easier for customers to make transactions and claim insurance. Regarding the accuracy level of product recommendations from 700 data, it was found that accuracy reached more than 92% in recommending types of insurance products to customers. Especially in this era of high technology, which must be utilized by companies in developing the company itself.

Keywords: Insurance, Smart Insurance System, Naïve Bayes Method, Customer Data, Classification.

1. Introduction

The Digitalization Era was born with a series of benefits that were presented, such as being able to help companies to more easily reach their customers. Various conveniences can be obtained by implementing digitalization both in the daily life of individuals and in the operations of an organization or company. Digitization is the process of transferring media from printed, audio, and video forms to digital forms (E. Sukmana, 2021). Digitization is the increasing availability of digital data made possible by advances in creating, transferring, storing and analyzing digital data, and has the potential to structure, shape and influence the contemporary world.

Companies in particular, have a lot to gain from investing in intelligence system-enabled technologies that cannot only automate executive- level task scheduling but can also enrich service quality by helping agents make informed decisions and unbiased judgments (Kumar et al., 2019). An automated system can help and replace the role of an agent that was previously done manually.

This automatic system will do the same role as the agent in conducting insurance analysis according to the data received. In the current process, agents are only served by a manual process in recording customer data using paper form. These are very ineffective where the same thing can be done in the system which will facilitate data processing and also save time. We know that the manual process can indeed minimize errors in filling out because there is direct confirmation from the agent to the customer.

In the application that was developed itself, there is the convenience of speed up insurance settlement because the approval and payment process can be done in the registration cycle. We added a signature process and payment through several types of payments such as VA, Credit Card or Offline at the end of the data input session. The developed application is also added with a digit system to perform tracking related to insurance applications, which makes customers aware of each verification process digitally. In this case, it will indirectly make the process more transparent and improve the company's services.

In the current process, agents are only served by a manual process in recording customer data using paper media. These are very ineffective where the same thing can be done in the system which can later facilitate data processing and also save time. We know that the manual process can indeed minimize errors in filling out because there is direct confirmation from the agent to the customer.

Table 1. Data Insurance Policy Submission

Insurance Product Type	Number of Policy at 2021	Insurance Policy Process Time	Accuracy of Insurance Product Recommendations	Insurance Policy Return to Review
Bebas Rencana Optimal	275	5 Days	256	**19**
Sprint Link Smart	180	6 Days	166	**14**
Critical Armor	308	5 Days	298	**9**
BIPS	140	4 Days	130	**10**

Fig. 1 : Accuracy of Insurance Product Recommendation

Table 1 and Fig. 1 show that human error still occurs which causes the allocation or recommendation of insurance products to be not on target. This could be due to lack of knowledge or comparison of data between client needs and the needs of each insurance product. By implementing the use of AI and classification systems, it is expected to increase the level of accuracy of insurance recommendations to XYZ company clients.

Classification of insurance products is very important in meeting the needs or protection of insurance companies for customers. The level of accuracy and precision is needed from the agent's side in determining the right type of insurance product based on customer history data. At this time, XYZ company is still doing insurance classification

manually, which requires an agent who is trained and has the right level of decision. Decision making in determining the classification of this type of insurance, when done manually, definitely has shortcomings in terms of the level of accuracy in determining customer insurance needs. Human errors sometimes occur. This kind of case sometimes happens to customers who are registering for insurance. The agent always reads the data provided by the customer and performs data classification / matching related to what type of insurance is really needed. Recommendations will be given by agents by analyzing data such as medical history, age, occupation and other personal data. Indeed, in the case of determining/classifying the type of insurance by an agent, it is somewhat inaccurate, but in general the insurance can still provide protection to customers according to their needs. This is a priority for XYZ company because when a customer registers for an insurance, they must have a different premium amount. Therefore, company XYZ wants the amount paid by the customer to match what is needed by the customer.

The classification method used in designing the classification of insurance products is the Naive Bayes Classifier method. Naive Bayes is a probability and statistical method developed by the British scientist Thomas Bayes, which predicts future opportunities based on past experiences. Naive Bayes also has several advantages, namely it is easy to understand, requires only simple coding, is faster in calculation, handles quantitative and discrete data, only requires a small amount of training data to estimate the parameters (mean and variance of the variables) needed for classification.

2. Literature Review

2.1. Intelligence System

Intelligent systems (Intelligence System) are used to build automated systems that perform afferent synthesis of action programs that formulate goals (Pupkov, 2019). Regulatory control systems and management of various natural objects are also developed and acquired for practical use to obtain the desired useful effects. These problems offer in-depth study of human brain function challenges that must

be solved.

A company can view that the existing intelligence system is only a tool to improve company efficiency, but it can also be something that functions very strategically, in the sense that it can significantly provide customer satisfaction with the products and services provided by the company. There are several management views on the intelligence system in according to its function. First, it is something that is very familiar in companies to improve the efficiency of work processes or operational activities, especially for administrative matters and documentation, thus encouraging increasing efficiency, effectiveness in managing the company.

Company's need to survive and thrive in a global business environment are highly dependent on the company's competence in utilizing all the potential contained in information technology to break through various obstacles and turn this potential into increased speed, flexibility, integration, and continuous innovation (Turban et al., 2005). Breaking through various barriers requires a reliable enabler. One of the main enablers is information technology. Information technology itself consists of three components, namely telecommunications, electronic office equipment, and computers.

Intelligence Insurance System (IIS) is a system designed with the implementation of computer assistance that seeks to form an automated system that can be called an intelligent system. According to Kumar that states the intelligence level can be set to a threshold value to further categorize it into Weak, Strong, and Super Smart (Kumar et al., 2019).

Fig. 2: Sense Think and Act processes used by intelligence agents

Information technology is able to shorten response times to customers, thus enabling companies to increase customer value and cycle effectiveness. Information technology facilities enable companies to break through cost barriers by increasing productivity and improving the quality of decision making so as to achieve increased cost effectiveness (Pupkov, 2019).

2.2. Intelligence System in Decision Making

Decision making is the process of choosing among several actions for the purpose of achieving a goal or multiple goals. Managerial decision making is the same as the entire management process in the company (Lavendelis, 2013). Planning involves a series of decisions, what should be done? Like what? By whom? Other parts of the managerial process such as organizing and controlling also involve decision making.

In this decision making, there is also the "X" system which is adapted by a smart system known as the intelligence system. This intelligence system can help make the necessary decisions based on system analysis related to the data entered by the customer. This decision-making includes several things, namely determining the type of insurance, categorizing health based on a history of illness, work and family.

The decision obtained is processing the input data which will be directed according to the needs of this "X" system. With the implementation of this intelligence system, it is hoped that it will be able to assist companies in classifying the insurance needs of XYZ company customers. Intelligence systems can increase the company's competitive advantage through the use of data, information, and knowledge owned by the company as a source of decision making. Analysis of company data is important in an effort to increase business competitiveness.

2.3. Intelligence System in Pattern Recognition

In designing a pattern recognition system, we need to pay attention to the model class definition,

application, pattern representation, feature extraction and selection, clustering analysis, classifier design and learning, training and testing sample selection, performance evaluation, etc. For application purposes, the content of each part of the pattern recognition system can vary greatly, especially in data processing and pattern classification.

To improve the reliability of the identification results, we need to add a knowledge base to correct errors that may occur, or by introducing constraints that greatly reduce the recognized patterns in the search space model library, to reduce computational matching. In certain applications, such as machine vision, in addition to identifying what object, the position and posture of the object must be determined to guide the work of the robot.

According to Pupkov statement, describing the general features of automated system programming allows us to draw the conclusion that it is a predetermined way based on data and habits (predictable) (Pupkov, 2019). If we can find similarities that humans often do and also their habits, then we can get data to be implemented into the system which will later help a decision quickly. We might notice that human actions can be characterized as actions with "high flexibility, changeability, and a kind of arbitrariness. "Therefore, we need a stable and neutral capability to be able to help determine and decide something based on the existing data processing".

2.4. Data Type Recognition Research Method

Data acquisition refers to the use of various sensors to convert various object information into a computer-acceptable set of values or symbols. We call this kind of numeric or symbolic (string) space a model space. The key to this step is sensor selection. To extract valid information from these numbers or symbols, data processing must be performed, including digital filtering and feature extraction.

2.5. Processing Data

According to (Wang & Wang, 2014), data processing is removing noise in the input data or information, and sorting out irrelevant data, leaving only the features and properties of the subject and identification methods that are closely related (such as object representation, perimeter, etc.). For example, in the introduction of data information, the system will read and process the searched data based on the data that has been entered into the system by utilizing the intelligence system and linkage with data storage.

Therefore, also said that it is necessary to adopt appropriate filtering algorithms, such as directional filtering, two-value filtering, and to filter out these unnecessary parts in the data processing of the system (Betrisandi, 2017).

2.6. Intelligence System and Human Being

Related to the definition of an intelligent system is an aggregation of technical and software tools that work together with a person (a group of people) or independently, which are united by information processing and are capable of synthesizing goals, decision making and discovery (Pupkov, 2019). Another advantage of intelligent systems (Intelligence System) over conventional systems is to recognize available data which may be incomplete, uncertain or fuzzy, where they can still make reasonable solutions, whereas conventional systems can only manipulate with complete or accurate data. In addition, automated decision support, artificial evolution, parallel execution, virtual integration, intelligent search and optimization bring unique capabilities of intelligent systems not available in conventional information systems (Pupkov, 2019).

2.7. Accuracy of Using Intelligence System in Insurance

The modern IIS (Intelligence Insurance System) is an innovation that is carried out due to several research reasons that want the company's processes to be carried out quickly, especially in the insurance sector. Some opinions say that this system should be able to perform tasks such as prediction, perform calculations and other tasks that require the ability to predict environmental developments, especially in the insurance sector. Intelligent models and mechanisms are needed to implement a proactive or goal-

oriented prognosis (Han & Kamber, 2006).

Second, the deals the company has to offer its clients must vary based on the actions taken (offers made) by other insurance companies that are direct competitors. Thus, IIS must monitor its reserves, market situation and react according to different changes by adapting the supply.

In this case, IIS (Intelligence Insurance System) is autonomous and represents its users in the market. Thus, the public will be helped to register for insurance quickly and instantly. Currently various intelligent mechanisms are used by autonomous intelligent computer systems to represent their users in the electronic market. Using such an approach in IIS simplifies the process of making various deals, for example issuing a policy can be simpler for both the company and its clients (Han & Kamber, 2006).

2.8. Naïve Bayes Classifier

The Naive Bayes Classifier method is one of the methods contained in the classification technique (Tina R. Patil & S. S. Sherekar, 2013). Naive Bayes is a classification using probability and statistical methods proposed by British scientist Thomas Bayes, which predicts future opportunities based on previous experience, so it is known as Bayes' theorem (Tina R. Patil & S. S. Sherekar, 2013). The theorem is combined with Naive where it is assumed that the conditions between attributes are independent. The main characteristic of this naive Bayes classifier is a very strong (naive) assumption of the independence of each event condition (Tina R. Patil & S. S. Sherekar, 2013). Naive Bayes for each decision class, calculates the probability under the condition that the decision class is correct, given the object information vector (Ngai & Chau, 2009). This algorithm assumes that object attributes are independent.

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)}$$

Dengan $P(Y|X)$ adalah probabilitas data dengan vektor X pada ke $P(Y)$ adalah probabilitas awal kelas merupakan proba independen kelas Y dari semua fitur dalam vektor . Nilai selalu sehingga dalam perhitungan prediksi nantinya kita membagi memilih yang terbesar menjadi kelas dipilih sebagai hasil prediksi. $\prod_{i=1}^{q} P(X_i|Y)$ X $P(X)$ Y $P(Y)\prod_{i=1}^{q} P(X_i|Y)$

3. Research Methodology

3.1. Research Framework Using Crisp-Dm

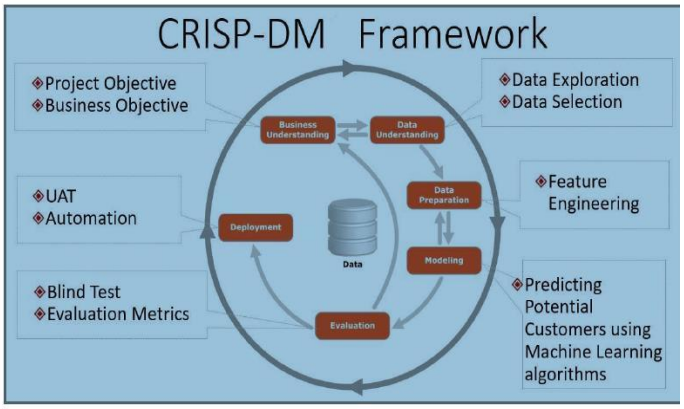


Fig. 3: Crisp-DM Framework (Chapman et al., 2014)

Many research results reveal that CRISP-DM is a data mining model that is still widely used in industry, partly because of its superiority in solving many problems in data mining projects. Mariscal, Marba and Fernandez (Mariscal, Marban, and Fernandez 2010) stated that CRISP-DM is the defacto standard for developing data mining and knowledge discovery projects because it is most widely used

in data mining development. This can be seen from the survey shown in Fig. 1 which was conducted on the use of the methodology in data mining projects. The CRISP-DM process model provides an overview of the life cycle of a data mining project. CRISP-DM has 6 stages, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment as shown in Fig. 2 (Chapman et al. 2000).

3.2. Stages of Classification Model Development

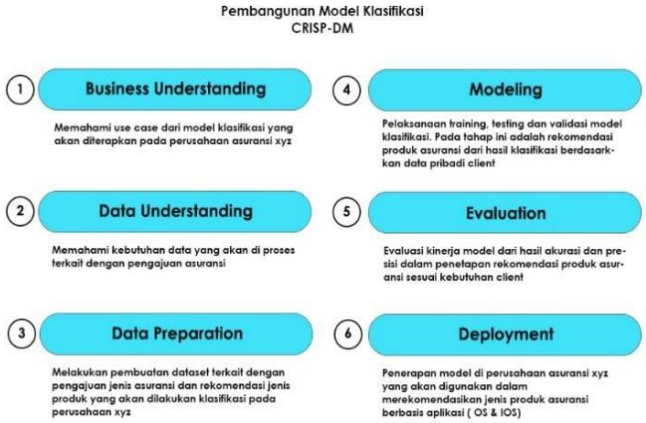


Fig. 4: Research Methodology Schematic

Intelligence Systems can be said more than anything else, because they can automate claim processing and document verification. This means agents with skills in data analysis and machine learning will be in high demand. Automation will only take over manual tasks in the future, allowing human personnel to be ethical on more complex tasks, such as managing portfolios and retaining clients. Intelligence System-based platforms such as Roadzen have added value to every phase of the insurance funnel and policy journey. By eliminating manual data extraction and data entry, an efficient, fast and error-free customer experience is provided.

3.2.1 Business Understanding

In this process, we can find processes that are considered important and necessary, including data selection, data processing, UW (UnderWritting) processes and final decisions related to registered policies. In final decision, the system will determine or recommend the client to take the type of insurance needed according to the data entered into the system.

Data Selection is the selection of data intended for clients in completing files or in the form of personal data that will be carried out online. The data is the main and important thing to continue processing the data afterwards. The data also serves as a reference for the system in directing the logic in selecting the type of insurance.

Data Processing can be called data processing carried out by the system based on the data selection step. In this data processing, data will be drawn according to system requirements which can later become a reference for classification which leads to a decision for the client.

UW Check (UnderWritting) is a risk identification and selection process imposed on prospective insured who wish to insure themselves at an insurance company. This process can detect risks that may be present and anticipate related claims that can be made.

When this system analyzes client demographics based on occupation, age, medical history, income and so on, this system will also estimate the possible claims that can be made against the client. When it does not find the expected claim based on data processing, the system will reject or hold a temporary hold regarding the client's insurance application.

3.2.2 Data Understanding

At this stage, the researcher explores and analyzes the data sources needed to apply the classification according to the specified use case. From the search results, the researcher decided to build a dataset using insurance client data which included personal data to the client's Medical History data. This dataset will be used to build a classification model based on the nave Bayes approach, both as a training dataset and as a testing dataset. At this stage, the researcher conducted a mapping of the data source profile from the dataset/data profiling to understand some general aspects of the data and collect statistical summaries of the data.

In analyzing the data collection, XYZ Company found that the client's activities in submitting policies and their general activities such as claims on websites or applications published by the company. With this in mind, we have drawn the associated sample required for a data set consisting of the following attributes:

1. **Gender**: The gender of the customer, as identified in their customer account. In this data set, 'M' is recorded for male and 'F' for Female.
2. **Occupation**: Unemployed, Student/Student, Employee, Manager, Housewife, TNI/Polri, Civil Servant, or Business Owner, Services/Finance, Government, Transportation, Trade, Construction, Manufacturing, Natural Resources, or Others.
3. **Region**: The area where the customer lives and the place of registers insurance
4. **Marital Status**: Married 'M', Not Married 'S' or Widowed 'D'/'J'.
5. **Smoker Status**: Y for smokers and T for non- smokers
6. **Medical History**: Selection of history according to the list of diseases that will be generated by the system for those who meet the Pre-Existing conditions in their health and "none" for those who do not have Pre-Existing conditions.
7. **Age**: This is the age when you first entered as an insurance customer
8. **Salary**: This is the average monthly salary of the premium payer.

3.2.3 Data Preparation

This study uses the Naïve Bayes Classifier method to classify insurance customer data into insurance products according to their demographic characteristics. Where, customers will be classified into one of 4 insurance products. The 4 insurance products used as classification classes are Bebas Rencana Optimal (BRO), Critical Armor (CA), Bebas Investa Prima Syariah (BIPS), and Sprint Link Smart (SLS). In this study, the dataset used 700 records and was taken from the central XYZ insurance company in Indonesia in 2021.

Table 2. Medical Type and Risk Level

Medical Record	Risk Level
Cancer	High
Hyperthyroidism, HIV, Kidney Disease	High
Repeated consultations or there are still follow-up controls with specialist doctors	High
Hepatitis B atau Hepatitis C	Medium
Epilepsy, Ovarian Cyst (Female), Chronic skin disorders	Medium

Hearing Loss, Leukemia, Asthma, Epilepsy	Medium
Family History and Lifestyle	Medium
Waiting for Laboratory Results	Medium
None of it	Low

Insurance products taken by customers, namely BRO, CA, BIPS, SLS with the following Fig. below:

Product Coverage and Claims									
Produk Asuransi	Age Coverage			Primary Insured 21>?	Medical Record Coverage			Cancer? Repeating or Daily Checkup? Hipertiroid? Kidney Disease?	Policy Holder Job Class
	0 - 18	19 - 55	56>		Low	Medium	High		
BRO	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	1 - 4
SLS	Yes	Yes	No	No	No	Yes	Yes	No	1 - 4
BIPS	Yes	Yes	Yes	Yes	Yes	Yes	No	No	1 - 3
CA	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1 - 2

Fig. 5: Insurance Specification Product

The Naïve Bayes approach is used to estimate the maximum posterior probability of the i customer for the insurance product class BRO, CA, BIPS, or SLS based on vector x . The final objective of this research is to produce a recommendation system for insurance products that are suitable for prospective insurance customers based on their demographic characteristics.

Data Preprocessing

In general, the raw data obtained still contains some data that cannot be used in the data mining process due to errors or anomalies in the data. So, it is necessary to pre-process the data or pre-processing before extracting information on the data. Where the data that has been obtained will be normalized, in other words the data will be removed from the noise. In this study, pre-processing begins with the elimination of duplication of data, if there are 2 or more data with the same policy number in one dataset, one of the data must be removed. The next step is to handle the missing value, this missing value can happen due to several things, it could be because the customer doesn't have it, the customer forgets to enter it or the input error in the company's database system.

For some data where one attribute is not filled in, the average value will be immediately replaced for the numeric attribute (continuous), while for the categorical attribute, the missing value will be replaced with the mode in the attribute. The attributes of data processing are gender, occupation, region, marital status, smoking status, medical records, age and salary. These data will be collected and store to the system to determine the product recommendation.

Underwriting Check

After there is no noise in the dataset, the next step is to carry out the training process. This stage consists of several steps:

1. *Seek Prior Opportunities.*

To find the value of the prior probability $P(Y)$ of each class, it can be searched by calculating the fraction of each data l owned by each class.

2. *Finding the likelihood value*

a. Likelihood attribute Categorical

On the category attribute, conditional probability $P(X_i = x_i|Y = y)$,

But when $P(X_i = x_i|Y = y) = 0$ $\bar{x}$ then

it $P(X_i = x_i|Y = y)$ will be calculated using the Laplacian formula smoothing.

Decision Of Application (Classification)

The third stage is the classification / decision stage. After the prediction model has been built on the training data, it is time to classify the data whose class label is not known. model has been built on the training data, it is time to classify the data whose class label is not known.

Conditions for classifying records,

a. If P(BRO|X) > P(CA|X), P(BRO|X) > P(SLS|X), P(BRO|X) > P(BIPS|X)

then the record is classified into BRO class

b. If P(CA|X) > P(BRO|X), P(CA|X) > P(SLS|X), P(CA|X) > P(CA|X)

then the record will be classified into CA class.

c. If P(BIPS|X) > P(CA|X), P(BRO|X) > P(SLS|X), P(CA|X) > P(BRO|X)

then the record will be classified into the BIPS class with a record of the absence of chronic disease at HIGH Level.

d. If P(SLS|X) > P(CA|X), P(SLS|X) > P(BIPS|X), P(SLS|X) > P(BRO|X)

then he belongs to the SLS class. However, if the maximum posterior opportunity is in the SLS class but the age of the additional insured customer is more than 21 years old or already working, then the customer is not eligible to take the SLS insurance product, so it will be recommended for the second maximum opportunity class.

3.2.4 Modeling

Based on the data listed in the "Preparation Data", we will use the data to process and carry out designs related to the classification of submissions and product recommendation.

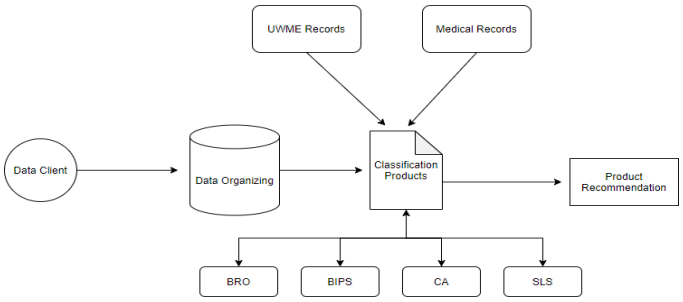


Fig. 6: Data Modeling Classification of Insurance Products

In the classification sample,

P(BRO|X) > P(CA|X), P(BRO|X) > P(SLS|X), P(BRO|X) > P(BIPS|X).

Client data that is entered into the product recommendation system refers to Fig. 6, the system will perform data processing or mining into 4 types of products. The first stage of the system will perform data matching and requirements tests on BRO products with CA where the captured requirements are similar. When the client data is compared and there is one part where the CA cannot cover, then the BRO product will make a comparison with other products, namely SLS and BIPS. When the data classification process with all types of products is completed, and it is found that from the existing submission categories, only BRO product types meet the requirements, the system will display a message recommending BRO products to continue applying for an insurance policy.

Fig. 7: Classification/Data Mining for Getting Product Recommendation

In Fig. 7. it is explained that after the client data is entered into the system, the system will carry out data organizing to ensure the data needed for the classification process is appropriate. The data entered will be carried out by a matching process for all types of products. Later, as a result of matching and classification, the type of insurance product will appear which indeed has the closest level of coverage to be able to proceed to the policy application stage.

3.2.5 Evaluation

Researchers evaluate the model that has been built by comparing the level of accuracy of the model with the pre-trained classification that was done previously. The process of evaluating this classification model is carried out several stages:

A. Load the classification model you want to evaluate

The classification model generated in the previous process is done manually so that it affects the level of accuracy and precision of insurance product recommendations, while the design of this system is integrated in a system where there is data mining (AI) related to insurance product classification. This can speed up the submission process and increase accuracy in recommending insurance products according to client needs.

B. Reading the dataset verification and testing file the evaluation of the data recognition

model utilizes the training dataset to test the ability of the pattern recognition model and customer data. The introduction of the previous data was done manually which reduces the effectiveness of data processing. This can affect the application of an insurance policy. In this new design, pattern recognition and customer data are processed at the "Product Classification" stage whose data will collaborate with UWME, Medical Records and all types of insurance products available. The verification process will be more transparent and faster with a good level of accuracy because AI can adapt data that allows for an update to behave from the previous data recognition.

3.2.6 Deployment

As part of the scope, this research analyzes the things that need to be considered in connection with the application of the classification model that has been built. In addition, the researcher also proposes the steps needed to apply the research results in the form of:

a) Adjustment of business processes required

Explaining the business processes related to the use case of the application of classification technology for product type recommendations at the xyz insurance company.

b) Preparation of data, applications, and supporting devices

Describes supporting needs in the form of additional client datasets, supporting applications, and other related devices such as smartphones.

c) Development of a product type classification system

Describe the proposed product type classification system based on client data based on mobile apps (OS & IOS) along with the required technology. The classification carried out by the system is also sourced from job data, age, medical history and other supporting data in accordance with the data entered by the client.

d) Additional training on classification models

Explains adjustments/developments that need to be made to the classification model in order to increase the accuracy of the model, especially in the Customer Data environment, other/specific Health History.

e) Strategy for implementing product recommendation classification

Conducting analysis to formulate the required product classification strategy based on mobile apps which includes the algorithm of the classification model that has been prepared.

4. Result and Discussion

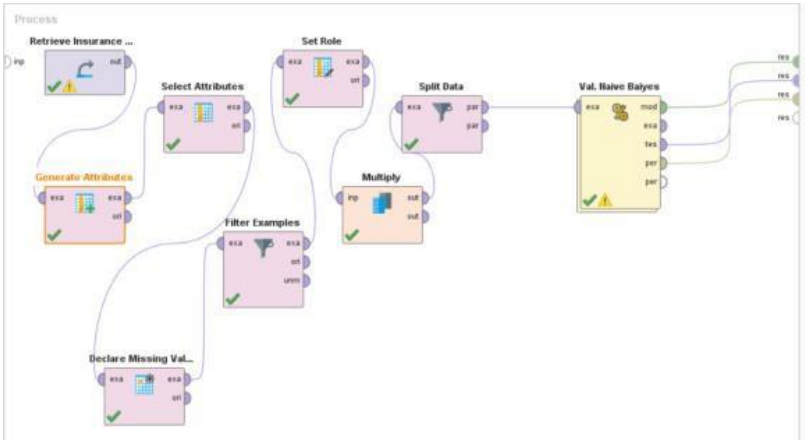


Fig. 8: Implementation Data on Rapidminer Tools

Add 1 column named 'product_insurance' using the Generate Attributes operator as a customer recommendation to find out the type of product to be used in accordance with age coverage, medical record coverage, and policy holder job class. the contents of function expressions are:

If (age <=55 && medical_hstry != "low", "SLS", if(medical_hstry != "high" && occupation_cls <=3, "BIPS", if(occupation_cls >=0 && occupation_cls

<=2, "CA", "BRO")))

Correlation Matrix (Correlation Matrix) | ExampleSet (Retrieve Insurance Data_Denni Dwi)

Attributes	usia	jenis_kelamin	status_pernikahan	status_perokok	provinsi	kelas_pekerjaan	riwayat_kesehatan	gaji
usia	1	-0.019	?	0.034	?	-0.008	?	?
jenis_kelamin	-0.019	1	?	-0.082	?	0.067	?	?
status_pernikahan	?	?	1	?	?	?	?	?
status_perokok	0.034	-0.082	?	1	?	0.006	?	?
provinsi	?	?	?	?	1	?	?	?
kelas_pekerjaan	-0.008	0.067	?	0.006	?	1	?	?
riwayat_kesehatan	?	?	?	?	?	?	1	?
gaji	?	?	?	?	?	?	?	1

Fig. 9: Data Correlation

It can be seen from Fig. 9. on the correlation that the datasets that have been processed have positive correlation values for the attributes age, gender, marital status, smoking status, province, job class, health history, and salary.

BRO insurance products can cover those aged 0 to more than 56 years, medical record coverage of low, medium, and high medical history, covers occupations of class 1 to 4. SLS insurance products can cover only those aged 0 to 55 years, medical record coverage from medical history is only medium and high, covers occupations of class 1 to 4. BIPS insurance products can cover those aged 0 to more than 56 years, medical record coverage from medical history is only low and medium, covers occupations class 1 to 3. CA insurance products can cover those aged 0 to more than 56 years, medical record coverage of low, medium, and high medical history, covers occupations of class 1 to 2.

You can see Fig. 10. shows that below the datasets have the attributes age, gender, marital status, smoking status, province, job class, health history and salary. In the health history attribute, there are 4 missing values because before modeling, researchers will clean the data by using replace missing values.

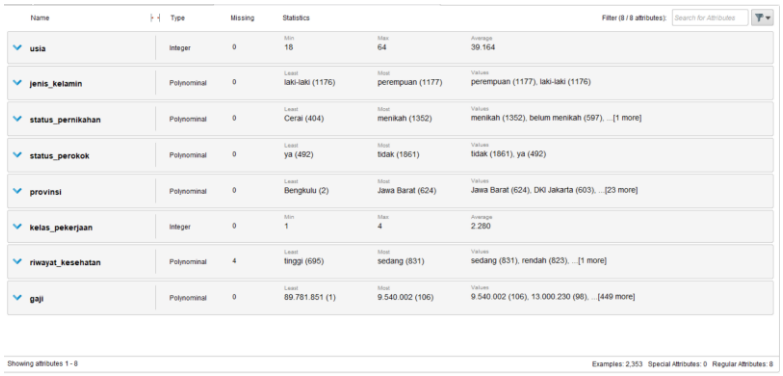


Fig. 10: Datasets Preparation

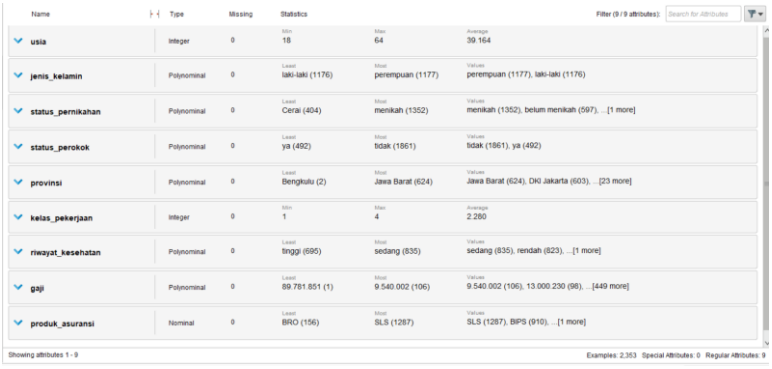


Fig. 11: Data Cleansing

Confusion matrix is an N x N matrix used to evaluate the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. Reason Confusion matrix is used to provide information about the errors made by the classifier and the types of errors that are being made. So that the classification model is disorganized and confused when making predictions. helps overcome the limitations of applying classification accuracy alone. In addition, it can overcome the problem of very unbalanced classification and one class dominating the other classes. Confusion matrix is very suitable for calculating Recall, Precision, Specificity, Accuracy and AUC-ROC Curve.

Table 3. Confusion Matrix

		Observed	
		True	False
Predicted Class	True	True Positive (TP)	False Positive (FP)
	False	False Negative (FN)	True Negative (TN)

Table 4. The Calculation Value of Precision, Recall, f-1 Score Positive

No.	Value	Formula	Calculation	Score
1.	Precision	TP / (TP+FP)	872/ (872+183)	0,8265
2.	Recall	TP / (TP+FN)	872/ (872+78)	0,9179
3.	Flscore	2*precision*recall / (precision+recall)	2 * 0,8265 * 0,9179 / (0,8265+0,9179)	0,87

Table 5. The Calculation Value of Precision, Recall, f-1 Score Negative

No.	Value	Formula	Calculation	Score
1.	Precision	TN / (TN+FN)	514/ (514+78)	0,8682
2.	Recall	TN / (TN+FP)	514/ (514+183)	0,7374
3.	Flscore	2*precision*recall / (precision+recall)	2*0,8682*0,7374 (0,8682+0,7374)	2,05

From the results of tables 4 and 5, it shows that the table that is related to precision and recall calculations in calculating accuracy can be seen in table 4.5, which is the result of calculating precision and recall for positive and negative data in showing the percentage results.

The Following on Fig. 12. that ROC (Receiver Operating Characteristics) is a kind of performance measuring tool for classification problems in determining the threshold of a model. Naïve Bayes algorithm, the default threshold is 0.5 and the ROC graph from the Naïve Baiyes model has AUC = 0.946 + / - 0.021.

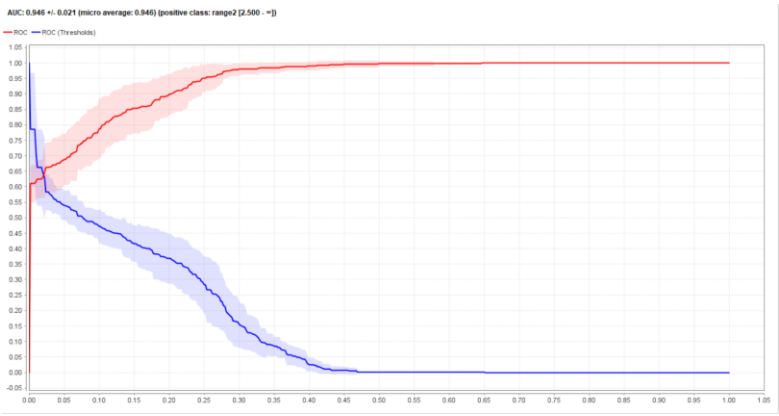


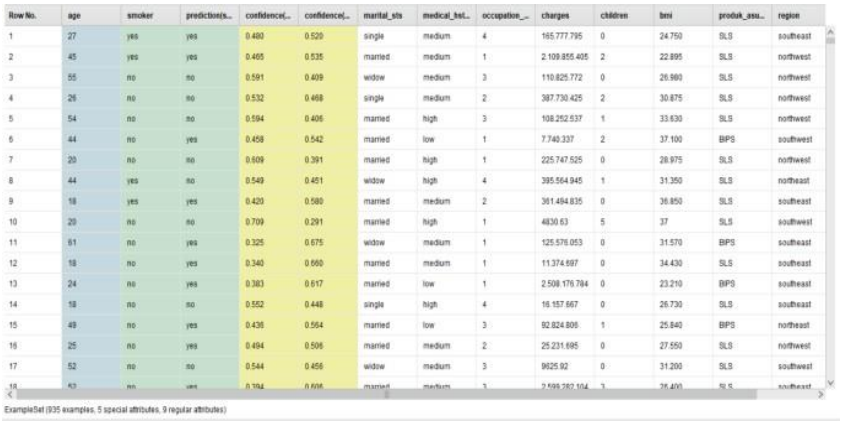
Fig. 12: Graphic Result of ROC AUC

Accuracy results:

Table 6. Accuracy Naïve Baiyes

No	Class	Prediction Result	Precentage
1	Precision Positive	0,8265 * 100	82,65%
2	Recall Positive	0,9179 * 100	91,79%
3	Precision Negative	0,8682 * 100	86,82%
4	Recall Negative	0,7374 * 100	73,74%

By using the Naive Bayes development model, it can be seen that the level of accuracy obtained is 92,75%, where the level of data accuracy good enough and could be more accurate with variety data that entered the system.



Row No.	age	smoker	prediction(s...	confidence(...	confidence(...	marital_sts	medical_hst...	occupation_...	charges	children	bmi	produk_asu...	region
1	27	yes	yes	0.480	0.520	single	medium	4	165.777.795	0	24.750	SLS	southeast
2	45	yes	yes	0.465	0.535	married	medium	1	2.109.855.405	2	22.895	SLS	northwest
3	55	no	no	0.591	0.409	widow	medium	3	110.825.772	0	26.980	SLS	northwest
4	26	no	no	0.532	0.468	single	medium	2	387.730.425	2	30.875	SLS	northwest
5	54	no	no	0.594	0.406	married	high	3	108.252.537	1	33.630	SLS	northwest
6	44	no	yes	0.458	0.542	married	low	1	7.740.337	2	37.100	BPS	southwest
7	20	no	no	0.609	0.391	married	high	1	225.747.525	0	28.975	SLS	northwest
8	44	yes	no	0.549	0.451	widow	high	4	395.564.945	1	31.350	SLS	northeast
9	18	yes	yes	0.420	0.580	married	medium	2	361.494.835	0	36.850	SLS	southeast
10	20	no	no	0.709	0.291	married	high	1	4830.63	5	37	SLS	southwest
11	61	no	yes	0.325	0.675	widow	medium	1	125.576.053	0	31.570	BPS	southeast
12	18	no	yes	0.340	0.660	married	medium	1	11.374.597	0	34.430	SLS	southeast
13	24	no	yes	0.383	0.617	married	low	1	2.508.176.784	0	23.210	BPS	southeast
14	18	no	no	0.552	0.448	single	high	4	16.157.667	0	26.730	SLS	southeast
15	49	no	yes	0.436	0.564	married	low	3	92.824.806	1	25.840	BPS	northeast
16	25	no	yes	0.494	0.506	married	medium	2	25.231.695	0	27.550	SLS	northwest
17	52	no	no	0.544	0.456	widow	medium	3	9625.92	0	31.200	SLS	southwest
18	52	no	yes	0.394	0.606	married	medium	3	2.599.282.104	3	26.400	SLS	southeast

ExampleSet (935 examples, 5 special attributes, 9 regular attributes)

Fig. 13: Product Recommendation for Customer

5. Conclusion

This research is entitled classification of insurance policy submissions for product types at insurance company. The types of insurance products that use as sample in this classification are BRO, SLS, BIPS, and CA. Based on age data that we have are from 0 to more than 56 years. Then, the type of work according to class we set from 1 to 4. This research was conducted using Naïve Baiyes method and the variables that used were Gender, Occupation, Marital Status, Smoker Status, Pre- Existing Conditions, Age, Account Holder Salary.

The product classification listed in Fig. 7 are the result of classification based on customer data such as age, gender, occupation class and others data support. The data collected and transfer into mining process which will generates recommendations for each customer according to their insurance needs as shown on Fig. 5 and Table 2.

Acknowledgement

This research was completed by conducting research and development on the M.M.S.I thesis of Denni Dwi Kurnianto as the author.

Reference

Betrisandi. (2017). *Klasifikasi Nasabah Asuransi Jiwa menggunakan Algoritma Naive Bayes berbasis backward elimination. Vol. 9 No. 1*.

Chapman, P. & C., Julian & Kerber, Randy & Khabaza, Tom & Reinartz, Thomas & Sheare, & Colin & Wirth, R. (2014). *CRISP-DM 1.0 step-by-step data mining guide.*

E. Sukmana. (2021). *DIGITALISASI PUSTAKA*.

Han, J., & Kamber, Micheline. (2006). *Data Mining Concepts and Techniques (2nd Edition)*. Morgan Kaufmann. https://www.researchgate.net/publication/262562891_Data_Mining_Concepts_and_Techniques_2nd_Edition

Kumar, N., Srivastava, J. D., Singh, G. G., & Bisht, H. (2019). *Artificial Intelligence in Insurance Sector*. https://www.researchgate.net/publication/337305024

Lavendelis, E. (2013). *Multi-Agent Architecture for Intelligent Insurance Systems Multi-Agent Architecture for Intelligent Insurance Systems*.

Ngai, E. & X. L., & Chau, D. (2009). *Application of data mining techniques in customer relationship management: A literature review and classification*.

Pupkov, K. A. (2019). Intelligent Systems and Human Being. *Procedia Computer Science*, *150*, 540–543. https://doi.org/10.1016/j.procs.2019.02.090

Tina R. Patil, & S. S. Sherekar. (2013). *Performance analysis of naive bayes and J48 Classication Algorithm for data classication. Vol. 6, No. 2*.

Turban, E., Aronson, J. A., & Ting Peng Liang. (2005). *Decision Support Systems and Intelligent Systems*.

Wang, H., & Wang, J. (2014). An Effective Image Representation Method Using Kernel Classification. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2014-December*, 853–858. https://doi.org/10.1109/ICTAI.2014.131