Analysis Insurance Costs for Smokers Using Linear Regression

Bernadheta Dinar Koen Hardayani[1], Albertus Dwiyoga Widiantoro[2]
[1,2]Department of Information System, Faculty of Computer Science
Soegijapranata Catholic University, Semarang, Indonesia
[1]22g40006@student.unika.ac.id , [2]*yoga@unika.ac.id*

Abstract— This study explores the effect of smoking habits on insurance costs using the linear regression method. Data was collected from various sources that included information on smoking habits and insurance costs. Using RapidMiner, linear regression analysis was conducted to predict the pattern of insurance costs influenced by smoking status. The results show that smokers tend to pay higher insurance premiums than non-smokers, in line with the health risks they carry. This study highlights the importance of including smoking habits in insurance pricing models, which can help insurance companies set more appropriate premiums based on individual risk. Thus, the results of this study not only deepen the understanding of the relationship between smoking and insurance costs, but also provide practical guidance for insurance companies in developing pricing policies that are more responsive to customers' risk profiles.

Keywords— Insurance, Linear Regression, Rapidminer, Smokers, Tobacco

I. INTRODUCTION

Insurance is an important mechanism in managing financial risks that may arise due to unforeseen events. In this process, insurance companies consider two main factors: the likelihood of an event occurring and its probability. This is particularly evident in the context of cyberattacks, where the capabilities of the threat group determine the potential for future harmful consequences. To manage this risk, insurance companies categorize customers into risk groups with the aim of forming a sustainable group, where premium payments by group members cover the cost of claims that may occur in the future. By reallocating risks and responsibilities to insurance companies, companies can mitigate adverse scenarios by covering the consequences or claims of customer companies using the financial resources generated from these groups. This approach not only helps manage risk, but also ensures the financial sustainability of the insurance company[1].

Health insurance is an important instrument to ensure access to health services for individuals and families. There are several relevant types of health insurance, including nongroup insurance, small group insurance, and insurance from public providers. Nongroup insurance is purchased independently on the insurance market, where premiums for tobacco users can be up to 50% more expensive than non-users. The insurance marketplace allows individuals and families to purchase health insurance without going through a group insurance provider, in accordance with the Affordable Care Act (ACA) rules[2]. In addition, small group insurance is provided for employee groups with less than 50 people, with similar rules that allow premium increases for tobacco users unless they join a smoking cessation program. Understanding these different types of insurance is important for choosing the right insurance and understanding the impact of premium policies on costs, especially for tobacco users[3].

The dynamic and personalized pricing of health insurance presents a major challenge for stakeholders in the healthcare and insurance sectors. The increase in healthcare costs, both absolute and relative, demands a deeper understanding of the causes behind the rise in healthcare expenses. Both patients and healthcare industry players play crucial roles in

influencing supply, demand, and pricing. Several studies have shown that predictive modeling of costs using machine learning can provide accurate results[4]. The integration of flow data and reinforcement learning techniques can facilitate the development of dynamic pricing models that can adapt in real-time to changes in market conditions, regulatory policies, and individual risk profiles. Linear regression methods can be used to predict future healthcare costs based on historical data. Therefore, the formulation of dynamic pricing models and personalized premiums becomes an important area for further research to optimize health insurance pricing strategies based on machine learning analysis and other relevant methodologies[5].

RapidMiner Studio uses predictive and descriptive data analysis techniques to provide information to every user so that they can make informed decisions. Statistical methods are also used in RapidMiner. Descriptive or inferential statistics are used to process data. Descriptive statistics are statistics used to analyze data by collecting and describing the data itself, without intending to draw conclusions about the general public[6].

This study aims to analyze the application of linear regression techniques in predicting insurance costs for individuals with smoking habits by utilizing historical data that includes insurance costs and smoking patterns and develop a more accurate and fair prediction model in determining the cost of insurance premiums. This research focuses on the important role of data analysis in understanding the relationship between smoking habits and insurance costs, hoping to enrich the understanding of health insurance as a financial protection for high-risk individuals, especially smokers who are likely to face serious health problems such as heart disease and cancer[7].

II. METHOD

2.1 Data Collection

The data collection process in this research is an important step to ensure the accuracy of the analysis. The data used comes from kaggle.com which is the main source of data collection and the title of the dataset is 'Healthcare Insurance" created by Arun Jangir. This dataset was selected for its ability to provide comprehensive and up-to-date information on insurance components. The data selection focused on important aspects in the analysis of health insurance, such as age, gender, Body Mass Index, number of dependent children, smoking status, region, and the cost of insurance to be covered. These selection criteria were based on their relevance to the purpose of the study, which was to identify patterns of health insurance costs using the Linear Regression method. In this context, the researcher decided to use detailed data as it allows for a more in-depth observation of the linear relationship of the independent variable (volume) and the dependent variable (Adj. Close) affecting insurance costs under various conditions. The data contains several attributes that will be shown in Table 1.

Tabel 1. Attribute on the data

Column name	Type	Description
Age	Integer	Age of the insured
Gender	Integer	Gender of the insured
Body Mass Index	Integer	Ideal weight based on height and weight
Number of child dependents	Integer	Number of child dependents covered

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v11i2.12998

Column name	Type	Description
Smokers	Integer	The insured is an active smoker or not
Region	Integer	Geographic area
Insurance cost	Polynominal	Insurance costs covered by the insured

Linear regression has an advantage that lies in its ability to provide estimates based on statistical analysis[8]. By using statistical methods that are established and accepted in the scientific community, the results of the analysis are not only objectively interpretable, but also repeatable and verifiable. In addition, linear regression also makes it possible to identify previously unseen patterns, which can assist insurance companies in optimizing marketing and retention strategies. For example, linear regression can show that smokers with a higher BMI may have higher insurance costs due to greater health risks. By knowing these patterns, insurance companies can target more effective advertising and retention campaigns to groups that have higher health risks.

2.2 Pre-processing

1. Gender question : for female 1 and male 2
2. Smokers : for passive smoker replace with number 1 and for active smokers use number 2
3. Region : for some existing regions southwest uses the number 1 southeast, uses the number 2, northwest uses the number 3 and northeast uses the number 4.
4. Ideal Weight Index : For numbers that have a comma above 5, round up and for numbers that have a comma below 5, round down.
5. Charges : The number of charges is converted from general using accounting format.

2.3 Research flowchart

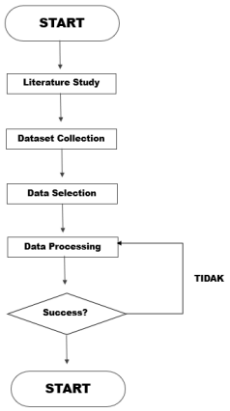

Figure 1. Flowchart

In this study, the research process begins with literature study, which is essential for reviewing sources that provide relevant information related to the problem being studied. This method enables the researcher to explore other research references from reliable sources, and it is considered a scientific research activity because data collection is conducted through note-taking, literature reviews, or reading. After gathering references and conducting the literature study, the next step is dataset collection, where the researcher selects the dataset to be used in the study. The dataset for this research was sourced from the Kaggle.com website, a trusted platform for various datasets. Following dataset collection, the researcher proceeds with dataset selection, where decisions are made regarding the appropriate method to apply in the research. This decision takes into account the objectives of the study and the availability of relevant datasets. Once the dataset is selected, the researcher conducts data filtering to identify and resolve issues such as missing variables or other errors that may affect the quality of the data. After filtering, the implementation stage becomes crucial, where the main goal is to extract new insights from the data. In this research, the linear regression method was chosen to predict and analyze the data using RapidMiner software this approach was selected to analyze the dataset effectively. Finally, data analysis is performed by processing the dataset within

RapidMiner to obtain meaningful results from the analysis. This series of steps forms the backbone of the research process, ensuring that the study is carried out systematically and accurately[9].

2.4 Rapidminer

RapidMiner is a software designed to make data processing easier. Available in a free version for academic and research purposes, as well as a higher-priced commercial version for business use, RapidMiner offers a wide range of essential functions. The software allows processing data from various sources, including text, images, and graphs, using various data mining algorithms to obtain useful information[10]. RapidMiner provides visualization features that allow users to view analysis results in the form of graphs or tables, facilitating data understanding and decision making. RapidMiner also supports integration with various operating systems and programming languages[11].

2.5 Operator

Operators in RapidMiner serve a variety of purposes that aid in the data analysis process and each operator has a complete set of elements that define its function, including a description of the expected input, a description of the output provided, the actions performed by the operator on the input, as well as a number of parameters that can control the actions performed by the operator. These operators also have various functions, such as data preprocessing, modeling, validation, and visualization. Examples of operators in RapidMiner include Retrive, Outliner, Replace Missing Value, Linear Regression, Apply Model, and others[12].

In this research, the focus is on utilizing machine learning techniques for predictive modeling of healthcare costs, specifically using linear regression. The Read CSV operator in RapidMiner is employed to read and process the data from CSV files, converting it into an ExampleSet that can be used for analysis[13]. The Split Data operator divides the dataset into training and test sets, ensuring that the model is trained on one subset of data and its performance is evaluated on a separate subset, which helps prevent overfitting. Linear Regression is then applied to model the relationship between independent variables (such as lifestyle, medical history, and health data) and dependent variables (such as healthcare costs), with the aim of predicting future costs based on historical data. The Apply Model operator is used to apply the trained model on new data, allowing for predictions and further analysis[14]. Lastly, the Performance of the model is assessed using metrics like Mean Squared Error (MSE), R-squared, and Root Mean Squared Error (RMSE) to evaluate its accuracy and effectiveness in predicting healthcare costs[12]. This methodology is key to understanding and optimizing insurance pricing strategies, as it allows for personalized and dynamic pricing models based on individual risk profiles.

2.6 Linear Regression

Linear regression analysis is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. In simple linear regression, which involves only one independent variable, insurance costs can be predicted based on daily cigarette consumption. The goal of this method is to find the best straight line that represents the relationship between these variables. Linear regression is a statistical technique used to model the relationship between a dependent variable (Y) and one or more independent variables (X). The basic equation in linear regression is:

$$Y = \beta 0 + \beta 1X + \varepsilon$$

Where:

- Y is the dependent variable that we want to predict or explain.
- X is the independent variable or explanatory variable used to predict Y.

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v11i2.12998

- β0 is the intercept that shows the value of Y when X is equal to zero.
- β1 is the regression coefficient that shows the average change in Y for every one unit change in X.
- ε is an error term that represents other variables that affect Y but are not included in the model.

The interpretation of the regression coefficient (β1) provides important information about the relationship between the independent and dependent variables. If β1 is positive it indicates a positive relationship between the independent and dependent variables which means, as the value of X increases, the value of Y will also increase. If β1 is negative it indicates a negative relationship between the independent and dependent variables which means, as the value of X increases, the value of Y will decrease. Linear regression not only allows us to understand the relationship between variables, but also allows us to predict the value of Y for a given value of X.

III. RESULTS AND DISCUSSION

The data processing process begins by selecting the read csv operator to enter the data that has been selected, then importing the dataset.

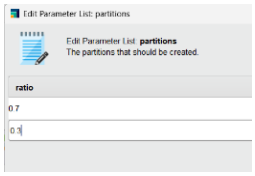

Figure 2. Testing and Trainning Data

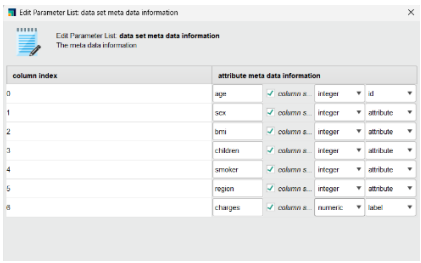

Figure 3. Attribute data information

In Figure 3, please change the attributes in the age column with id and charges column with numeric and label because these variables are independent variables that will be predicted. Next, add the Split Data operator which will be used to divide the dataset into 2 parts, namely training data which is used to learn the model and produce the best parameters to predict data and testing data which functions to measure the performance of the model and detect the dataset[15]. Connect out on the Read CSV operator with exa on the split data.

LinearRegression

```
  1334293.558 * bmi
+ 156085610.379 * children
+ 527449381.559 * smoker
- 227437016.819 * region
+ 605432206.218
```

Picture 4. Linear Regression Result

At the next stage, please connect the existing ports.

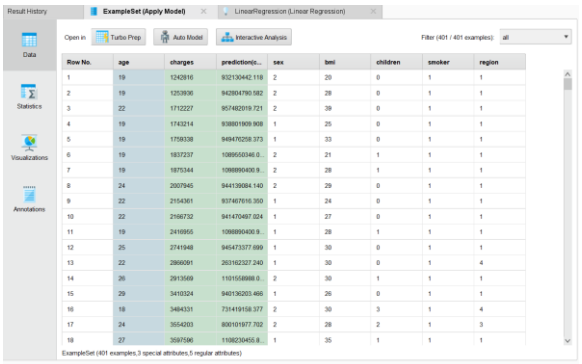

Figure 5. Data testing prediction result

In Figure 5, the prediction results obtained against insurance testing data on prediction charges. This research also uses another assessment of the linear regression model, namely by using the performance operator.

root_mean_squared_error

```
root_mean_squared_error: 3891556800.697 +/- 0.000
```

Figure. 6 RMSE Result

The results of the linear regression model analysis conducted with various proportions of training data and testing data. In the proportion of 90% training data and 10% testing data, the model produces a

regression coefficient with an intercept value of -7234882.824, bmi by 127003705.547, children by 409464634.376, smoker by 308779085.834, and region by 130651261.060. Furthermore, for the proportion of 80% training data and 20% testing data, the model produces a regression coefficient with an intercept value of - 2174491.400, bmi of 433593128.909, smoker of 257896891.342, and region of 1100146236.607. In the proportion of 70% training data and 30% testing data, the model produces a regression coefficient with an intercept value of 1334293.558, bmi of 156085610.379, children of 527449381.559, smoker of -227437016.819, and region of 605432206.218. The RMSE value for this scenario is 3891556800.697.

Tabel 2. **RMSE Value Calculation**

Data *Training*	Data *Testing*	Value	RMSE
90%	**10%**	-7234882.824 * bmi + 127003705.547 * children + 409464634.376 * smoker - 308779805.834 * region + 1306561261.060	44530460 19.440 +/- 0.000
80%	**20%**	- 2174491.400 * bmi + 433593128.909 * smoker - 257896891.342 * region + 1100146236.607	38249497 99.743 +/- 0.000
70%	**30%**	1334293.558 * bmi + 156085610.379 * children + 527449381.559 * smoker - 227437016.819 * region + 605432206.218	38915568 00.697 +/- 0.000

RMSE analysis shows that the model with 80% training data and 30% testing data has the best performance in predicting data. This is evidenced by the lowest RMSE value of 9138322.118 ± 0.000, which shows the average prediction error of this model is around 9,138,322.118. Compared to other models, the RMSE value of this model is lower. The model with 90% training data and 30% testing data has an RMSE value of 9170163.991 ± 0.000, while the model with 70% training data and 30% testing data has an RMSE value of 11089148.439 ± 0.000. Based on the RMSE value, it can be concluded that the model with 80% training data and 30% testing data has the best prediction accuracy.

IV. CONCLUSION

This study reveals that smoking habits significantly impact insurance premiums, with smokers paying higher rates due to increased health risks like heart disease and cancer, as confirmed by linear regression analysis. The predictive model developed, based on an 80/20 training/testing data split, achieved an RMSE of 9,138,322.118, indicating high accuracy in estimating premiums based on factors like BMI, dependent children, smoking status, and region. This model supports fairer, data-driven premium policies and encourages healthier behavior. Although promising, its real-world application requires further testing across diverse datasets, considering regional, socioeconomic, and regulatory factors.

ACKNOWLEDGMENT

With deep gratitude, I would like to express my sincere thanks to Mr. Albertus Dwiyoga Widiantoro, S.Kom., M.Kom., for his attentive guidance, profound expertise, and constructive feedback, which have been invaluable at every stage of this research. Without his support, I would not have been able to achieve such satisfying results. I also extend my heartfelt thanks to Soegijapranata Catholic University, particularly the Faculty of Computer

Science, Program of Information Systems, for providing the facilities and resources that greatly facilitated the progress of this research. I am also deeply grateful to Arun Jangir, who generously made the "Healthcare Insurance" dataset available on Kaggle, which served as a critical foundation for the analysis in this study.

Most importantly, I would like to express my deepest appreciation to my parents, who have given me unwavering support, prayers, and priceless love throughout this journey. Without them, I would not be where I am today. Thank you for providing everything I needed to grow and develop.

REFERENCES

[1] V. Matejka and J. Angel Huacan Soto Zurich, "A Framework for the Definition and Analysis of Cyber Insurance Requirements," *Univ. Zurich Master Proj.*, 2021, [Online]. Available: http://www.csg.uzh.ch/

[2] C. M. Kaplan and E. K. Kaplan, "State policies limiting premium surcharges for tobacco and their impact on health insurance enrollment," *Health Serv. Res.*, vol. 55, no. 6, pp. 983–992, 2020, doi: 10.1111/1475-6773.13577.

[3] V. U. Ekpu and A. K. Brown, "The Economic Impact of Smoking and of Reducing Smoking Prevalence: Review of Evidence," *Tob. Use Insights*, vol. 8, p. TUI.S15628, 2015, doi: 10.4137/tui.s15628.

[4] S. Albawi, L. Alshahrani, N. Albawi, R. Alharbi, and A. Alhakamy, "Prediction of healthcare insurance costs," *Comput. Informatics*, vol. 3, no. 1, p. 1250124, 2023, [Online]. Available: https://dergipark.org.tr/tr/pub/ci

[5] Md Mohtaseem Billa, "Medical Insurance Price Prediction Using Machine Learning," *J. Electr. Syst.*, vol. 20, no. 7s, pp. 2270–2279, 2024, doi: 10.52783/jes.3962.

[6] E. D. Madyatmadja, S. I. Jordan, and J. F. Andry, "Big data analysis using rapidminer studio to predict suicide rate in several countries," *ICIC Express Lett. Part B Appl.*, vol. 12, no. 8, pp. 757–764, 2021, doi: 10.24507/icicelb.12.08.757.

[7] A. Anggreini, D. Rochmah, and R. Wasir, "Adaptasi Sistem Asuransi Kesehatan Terhadap Perubahan Pola Penyakit: Indonesia," *An-Najat*, vol. 2, no. 2, 2024, [Online]. Available: https://jurnal.stikes-ibnusina.ac.id/index.php/an-Najat/article/view/1295

[8] L. Regression, *Linear Regression*.

[9] T. Maulana, R. Astuti, and F. Muhammad Basysyar, "Implementasi Algoritma Regresi Linear Untuk Memprediksi Pendapatan Pt Pln Berdasarkan Penggunaan Per Kelompok Pelanggan," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3196–3202, 2024, doi: 10.36040/jati.v7i6.8083.

[10] Ainurrohma, "Akurasi Algoritma Klasifikasi pada Software Rapidminer dan Weka," *Prism. Pros. Semin. Nas. Mat.*, vol. 4, pp. 493–499, 2021, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[11] D. Anugrah Pratama, I. Rizal Mutaqin, and K. Rafael Manuela, "Analisis Terjadinya Kanker Paru-Paru Pada Pasien Menggunakan Decision Tree: Penerapan Algoritma C4.5 Dan RapidMiner Untuk Menentukan Risiko Kanker Pada Gejala Pasien," *Jtmei*, vol. 2, no. 4, pp. 156–170, 2023, [Online]. Available: https://doi.org/10.55606/jtmei.v2i4.3004

[12] U. Lathifah and R. Danar Dana, "Implementasi Metode Linear

Regression Untuk Prediksi Harga Properti Real Estate Menggunakan Rapidminer," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 1129–1137, 2024, doi: 10.36040/jati.v8i1.8919.

[13] A. P. Ayudhitama and Utomo Pujianto, "Analisa 4 Algoritma Dalam Klasifikasi Liver Menggunakan Rapidminer," *J. Inform. Polinema*, vol. 6, no. 2, pp. 1–9, 2020, doi: 10.33795/jip.v6i2.274.

[14] R. C. Prihandari, "Data Mining: Konsep Dan Apikasi Menggunakan Rapidminer (Series: Supervised Learning Dan Unsupervised Learning)," p. 8, 2022, [Online]. Available: http://repository.uin-suska.ac.id/63073/1/REGITA CAHYANI PRIHANDARI.pdf

[15] Baiq Nurul Azmi, Arief Hermawan, and Donny Avianto, "Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 4, no. 4, pp. 281–290, 2023, doi: 10.35746/jtim.v4i4.298.