JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 653 – 660 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.2933
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

COMPARISON FEATURE EXTRACTION USING ARTIFICIAL NEURAL NETWORK ALGORITHM ON SMOKER PREDICTION

Arie Satia Dharma[1*], Cynthia Veronika Pardede[1], Jonggi Vegas Sitorus[1]
[1]Informatics and Electrical Engineering, Institut Teknologi Del
email: *ariesatia@del.ac.id

Abstract: The habit of smoking is dangerous because of the addictive substances that make cigarettes addictive. Its addictive nature poses a significant risk, affecting personality with stress, depression and nervous disorders. Body factors that indicate smoking include blood sugar levels, dental caries, and hemoglobin. To address this, research has been conducted with focused efforts to understand and address the risks associated with smoking and its impact on overall health. This research aims to choose the best method for predicting smokers by using feature selection techniques. The feature selection algorithms uses for that are Analysis of Variance (ANOVA), Recursive Feature Elimination (RFE), and Genetic Algorithm (GA) to select optimal attributes and uses the k-fold cross validation technique as the validation of the Artificial Neural Network algorithm. The data includes various parameters such as age, height, weight, vision, blood pressure, cholesterol, triglycerides, hemoglobin, AST, ALT, GTP, gender, dental caries and tartar. Hearing ability, urine protein content, and tartar were selected. The results showed that using the Analysis of Variance method showed higher accuracy (77.101%) compared to the Genetic Algorithm method (74.64%) and the Recursive Feature Elimination method (76.08%). Selection of relevant attributes increases the predictions and insights of the Artificial Neural Network model about the effects of smoking on health.

Keywords: artificial neural network; analysis of variance; genetic algorithm; recursive feature elimination; smoker prediction

Abstrak: Kebiasaan merokok berbahaya karena adanya zat adiktif yang membuat rokok menjadi ketagihan. Sifatnya yang membuat ketagihan menimbulkan risiko yang signifikan, mempengaruhi kepribadian dengan stres, depresi, dan gangguan saraf. Faktor tubuh yang mengindikasikan kebiasaan merokok antara lain kadar gula darah, karies gigi, dan hemoglobin. Untuk mengatasi hal ini, penelitian telah dilakukan dengan upaya terfokus untuk memahami dan mengatasi risiko yang terkait dengan merokok dan dampaknya terhadap kesehatan secara keseluruhan. Penelitian ini bertujuan untuk memilih metode terbaik dalam memprediksi perokok dengan menggunakan teknik seleksi fitur. Metode seleksi fitur yang digunakan adalah Analysis of Variance (ANOVA), Recursive Feature Elimination (RFE), dan Genetic Algorithm (GA) untuk memilih atribut yang optimal dan menggunakan teknik k-fold cross validation sebagai validasi algoritma Artificial Neural Network. Data tersebut mencakup berbagai parameter seperti umur, tinggi badan, berat badan, penglihatan, tekanan darah, kolesterol, trigliserida, hemoglobin, AST, ALT, GTP, jenis kelamin, karies gigi dan karang gigi. Kemampuan pendengaran, kandungan protein urin, dan karang gigi dipilih. Hasil penelitian menunjukkan bahwa penggunaan metode Analysis of Variance menunjukkan akurasi yang lebih tinggi (77,101%) dibandingkan dengan metode Genetic Algorithm (74,64%) dan metode Recursive Feature Elimination (76,08%). Pemilihan atribut yang relevan meningkatkan prediksi dan wawasan model Jaringan Syaraf Tiruan tentang dampak merokok terhadap kesehatan.

Kata kunci: artificial neural network; analysis of variance; genetic algorithm; prediksi perokok; recursive feature elimination

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 653 – 660 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.2933
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

Smoking is the activity of smoking tobacco rolls wrapped in palm leaves or paper that are burned and then the smoke is taken into the body and exhaled again. In daily life we can find people smoking in public places and even in our own homes around the neighborhood [1]. Cigarettes contain many harmful substances, including nicotine, which gives smokers an addictive effect. Smokers will generally experience weight loss than nonsmokers, even though their caloric intake is the same or more than nonsmokers. This can occur because when burning cigarettes, nicotine will enter the blood circulation by 25% and enter the human brain in approximately 15 seconds which then nicotine will be accepted by acetylcholine-nicotinic receptors to spur the dopaminergic system so that it will affect appetite suppression which causes changes in nutritional status [2]. WHO states that more than 6 million people die from active smoking. Data from Indonesia also shows a high prevalence of smoking among adolescents and young adults, which includes university students [3].

The World Health Organization (WHO) also records that around 225,700 people in Indonesia die every year due to smoking or diseases related to substances contained in cigarettes [4]. According to the World Health Organization (WHO) in 2008, around 5.4 million people lose their lives every year due to causes related to smoking. More recent data from the WHO in 2018 revealed an even higher toll, with tobacco responsible for more than 7 million deaths annually. Of these deaths, more than 6 million can be attributed to direct tobacco use, while about 890,000 are caused by exposure to secondhand smoke (secondhand smoke). It is noteworthy that the majority of smokers, around 80%, live in low- and middle-income countries, with a total of around 1.1 billion people worldwide. The 2018 Basic Health Research Data (RISKESDAS) shows that the prevalence of smoking among 10-18 year olds is 9.1%, and this figure has continued to increase since 2013. In addition, the percentage of tobacco consumption (both smoking and chewing) among Indonesians who are aged 15 years and over is 62.9% for boys and 4.8% for girls. Given this concerning data, the development of websites to identify smokers and nonsmokers can play an important role in raising awareness about smoking and potentially helping individuals make informed decisions about their health [1]. Therefore, this study aims to develop a website that can effectively determine whether someone is a smoker or a nonsmoker.

This study uses machine learning to identify a smoker. Machine learning is the scientific study of the algorithms and statistical models that computer systems use to perform specific tasks without being explicitly programmed. Machine learning algorithms are used in applications such as data mining, image processing, predictive analytics and more. The goal of machine learning is to allow computers to learn from data and make predictions or decisions based on that learning and the selection of algorithms depends on the specific problem and characteristics of the data [5][6][7]. Based on the available data, decisions can be made on what to do through data processing using an algorithm. The results of the accuracy of the predictions depend on the amount of data, variables, and the value of each specified attribute [8]. In this study, the machine learning algorithm used is an artificial neural net-

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 653 – 660 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.2933
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

work algorithm and using several feature selections such as analysis of variance (ANOVA) which includes the filter method, recursive feature elimination (RFE) which includes the wrapper method, and genetic algorithm (GA) which includes the embedded method in predicting smokers.

In a previous study entitled "Comparing different feature selection algorithms for cardiovascular disease prediction", an experiment was conducted to develop an effective feature selection method for predicting cardiovascular disease using data mining techniques. In this study, experiments were conducted using filter, wrapper, and embedded methods. Wrapper method-based feature selection provides the best accuracy (0.7320) with the artificial neural network algorithm [9]. In another study, entitled "Developing a model to predict the start of combustion in HCCI engine using ANN-GA approach", which conducted research to develop a predictive model for the start of combustion on the HCCI machine. By using genetic algorithm (GA) on the artificial neural network algorithm, it can increase the accuracy from 0.89 to 0.96 [10]. In another study, entitled "Machine learning techniques with ANOVA for the prediction of breast cancer", which conducted a study to predict breast cancer prediction. The results showed that artificial neural network achieved the highest accuracy using ANOVA, which was 87.4% [11] (Thakur et al., 2022).

This research create a classification system that can help someone identify smokers or not by using a machine learning approach. We expected to find out the best method of the Artificial Neural Network algorithm with feature selection in case studied of predicting smokers.

METHOD

Image 1. is a diagram of the proposed system in this study.

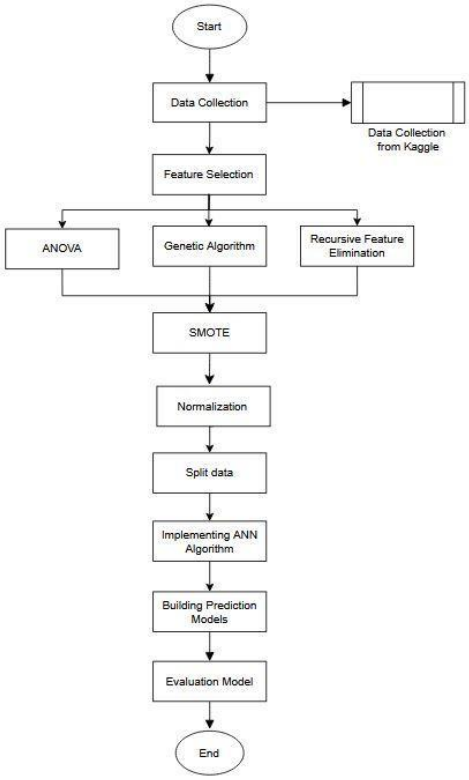

Image 1. Research Stages

Data Collection

The dataset used in this study is the Smoking Dataset from Basic Health Biological Signals obtained from Kaggle. The data will be used to predict cigarettes with a total of 55,692 rows and 26 attributes. The attributes namely *gender, age, height, weight, waist, eyesight (left), eyesight (right), hearing (left), hearing (right), systolic, relaxation, fasting blood sugar, cholesterol, triglyceride,* HDL, LDL, *hemoglobin, urine protein, serum creatinine,* AST, ALT, Gtp, *oral, dental caries,* dand *tartar*. The dataset obtained will be preprocessed data which aims to ensure that the data to be used is simple and complete, so that errors such as missing values, inconsistent data, and data

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 653 – 660 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.2933
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

noise do not occur when implemented in the machine learning analysis process. Therefore, the data preprocessing step was applied to remove outliers and fill in the missing values for more reliable data analysis [12].

Analysis of Variance

This filter-based approach to feature selection does not consider the possibility of features influencing each other. So, although some features may have high scores independently, if they are highly correlated with each other, we should consider discarding one of them so as not to interfere with the statistical analysis results [13]. One of this approach is ANOVA.

ANOVA (Analysis of Variance) is a statistical analysis technique that aims to compare the means of several groups or treatments. This method involves comparison of variances and is an extension of the t-test. The use of ANOVA is intended to identify whether there is a significant difference between the means of different groups or treatments. ANOVA is a subset of comparative analysis that involves more than two means. A one-way ANOVA is used to compare more than two means, while a t-test is used to compare two means.

Recursive Feature Elimination

This method uses a wrapper algorithm which involves a model training process to evaluate each feature subset. The features that provide the best model performance will be selected.

The RFE (Recursive Feature Elimination) technique is a feature selection method that uses cross-validation to repeatedly select a subset of features based on their rating. The main goal of RFE is to reduce dependencies and collinearity between features in the data. The process of removing features is done iteratively by building the model, first with all the features, then removing the one with the lowest ranking feature, and building the model again with the remaining features. This can help improve the efficiency and performance of classification models and minimize problems caused by feature dependability and collinearity [14].

Genetic Algorithm

Genetic Algorithm (GA) is an adaptive heuristic search algorithm that is commonly used to find optimal approximate solutions to optimization problems with large search spaces, and can be applied effectively to feature selection in optimization problems [15]. The process of GA can be divided into seven stages:

1. Solution Encoding: Each potential solution is encoded into a chromosome, where each chromosome represents a combination of features.
2. Initialization: The population size is set, and an initial population of chromosomes is randomly generated.
3. Fitness Evaluation: The fitness of each chromosome is evaluated based on a fitness function that measures the quality of the solution.
4. Termination Condition Checking: The algorithm checks if a termination condition is met, such as reaching a maximum number of iterations or finding an optimal solution.
5. Selection: Chromosomes with high fitness values are selected to create a new population for the next generation.
6. Crossover: The selected chromosomes undergo crossover, where genetic information is exchanged between pairs of chromosomes to create new offspring.
7. Mutation: Random changes are intro-

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 653 – 660 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.2933
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

du ced to the offspring chromosomes to maintain genetic diversity in the population.

SMOTE

Synthetic Minority Over-sampl ing Technique (SMOTE) is a technique that creates more minority class instances near the existing ones. With the SMOTE, for each minority class it will increase by generating a new synthetic sample. The following are the process steps for carrying out the Synthetic Minority Oversampling Technique (SMOTE) as follows [16].

1. Identify the k-nearest neighbors in the minority class from the sample.
2. Then calculate the difference between the sample and k-neighbors. The difference is multiplied by a random value between 0 and 1 to generate a new value.
3. After that, create a synthetic sample using this new value and add it to the training data.

Artificial Neural Network

Artificial neural networks (ANNs) are mathematical tools, mimicking hum an biological neural networks, learning from experience and generalizing previous behavior as characteristics. An Artificial Neural Network (ANN) architecture consists of an input layer, one or more hidden layers and an output layer. ANNs are data-driven, adaptive methods that learn from examples while picking up subtle, hidden functional relationships that are unknown or difficult to explain.

The following are the process stages of the artificial neural network algorithm [17].

1. Initialize parameter and hyperparameter values, by initializing weights and bias parameters, hyperparameter threshold, learning rate, number of hidden layers, beta1, beta2, number of epochs and epsilons in the initial modeling.
2. Performs feedforward propagation, which moves forward to the output screen. This stage will receive input from the input layer and forwarded to the hidden layer. Furthermore, in the hidden layer, the input signal will be multiplied by the weight and added by the bias value. Then the results will be activated through the activation function whose results will be sent to the next layer until it reaches the output layer and produces a predictor value.
3. Carry out the backpropagation process, which moves from the output layer to the input layer. This stage calculates the partial derivative of the weight and bias of the layer. The partial derivatives obtained will be used during the parameter updating process to minimize the error value.
4. Calculation of the error value, from the results of the predictions made, the error value will be calculated through the loss equation.
5. Changing the weights and biases is done by changing the weights and biases according to the gradient of the error function in indicating the optimal direction of change in reducing the error value.

Evaluation Model

We evaluate our model with confusion matrix. The confusion matrix is usually represented in tabular form showing the frequency of true positives (observations that are correctly classified as class positive), false positives (obser vations that are incorrectly classified as class positive), true negatives (observations that are correctly classified as class negative), and false negatives (observations that are incorrectly classified as

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 653 – 660 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.2933
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

class negative). This frequency is used to calculate performance indicators such as precision, sensitivity, specificity and accuracy. The following formula is the accuracy formula using the confusion matrix.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

TP = True Positive; TN = True Negative
FP = False Positive; FN = False Negative

RESULT AND DISCUSSION

SMOTE Result

The SMOTE method which balances the data based on the majority class of the dataset. This stage is carried out because there is unbalanced data on the dataset where non-smoker data is more than smoker data. The results of the SMOTE method can be seen in Table 1.

Table 1. Result of Balancing Data

	Smoker Data	
	Non-Smoking	Smoking
Original	34732	19938
SMOTE	34732	34732

Based on the results shown in Table 1, the SMOTE method will balance the data in the majority class by increasing the number of samples by creating synthesis samples in the minority class, so that the minority and majority classes become more balanced. It can be seen that the amount of data from the dataset before data imbalance is obtained that the data in the minority class is 19,938 and in the majority class is 34,732. After the data imbalance is carried out, it is found that the amount of data in the minority data is the same as the majority data of 34,732.

Evaluation Model Result

In this research, three feature selection methods are compared, namely ANOVA, GA, and RFE on the Artificial Neural Network algorithm. In the ANOVA method, it is selected based on the k-best of the F-value. After the experiment, it is found that 10 features produce more optimal performance. The GA method produces the most optimal performance on 11 features and RFE produces the most optimal performance on 10 features.

The results of the three feature selections can be seen in Table 2.

Table 2. Result of Accuracy Model ANN

Classification	Feature Selection		
	ANOVA	GA	RFE
Artificial Neural Network	77,10%	74,64%	76,08%

Based on the results of Table 2, it can be seen that the ANOVA method produces higher accuracy compared to the two feature selection methods compared. The accuracy value obtained using ANOVA was 77.10% while using the GA method resulted accuracy was 74.64% and in RFE was 76.08%.

CONCLUSION

Based on the results obtained in this research, it was found that feature selection using the ANOVA method produced better results than two other algorithms in selection features for predicting smoking habits. The original dataset have been balancing process with SMOTE and selecting with ANOVA that results 69.464 datasets with 10 features. Then the classification with ANN algorithm result model with ANOVA feature selection that produces an accuracy of 77.101%.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 653 – 660 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.2933
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

BIBLIOGRAPHY

[1] D. P. Sekeronej, A. F. Saija, and N. E. Kailola, "TINGKAT PENGETAHUAN DAN SIKAP TENTANG PERILAKU MEROKOK PADA REMAJA DI SMK NEGERI 3 AMBON TAHUN 2019," *PAMERI Pattimura Med. Rev.*, vol. 2, no. 1, 2020.

[2] A. Bagaskoro and V. L. Amelia, "Hubungan Antara Konsumsi Rokok Dengan Status Nutrisi Pada Remaja," *J. Keperawatan Muhammadiyah*, vol. 5, no. 2, 2020.

[3] D. A. Supriyanto and T. Damayanti, "Correlation of Smoking Habit and Level of Nicotine Dependence in University Students," *Respir. Sci.*, vol. 3, no. 2, 2023.

[4] D. K. J. Cameng and Arfin, "Analisis Penerapan Kebijakan Earmarking Tax Dari Dana Bagi Hasil Cukai Hasil Tembakau Terhadap Kesehatan Masyarakat," *Simposium Nasional Keuangan Negara*. 2020.

[5] M. Batta, "Machine Learning Algorithms - A Review," *Int. J. Sci. Res.*, 2020.

[6] M. Zaffar, M. A. Hashmani, K. S. Savita, and S. A. Khan, "A review on feature selection methods for improving the performance of classification in educational data mining," *International Journal of Information Technology and Management*, vol. 20, no. 1–2. 2021.

[7] M. A. Arif, A. Jahan, M. I. Mau, and R. Tummarzia, "An Improved Prediction System of Students' Performance Using Classification model and Feature Selection Algorithm," *Int. J. Adv. Soft Comput. its Appl.*, vol. 13, no. 1, 2021.

[8] A. Triayudi and I. Fitri, "Comparison Of The Feature Selection Algorithm In Educational Data Mining," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 19, no. 6, 2021.

[9] N. Hasan and Y. Bao, "Comparing different feature selection algorithms for cardiovascular disease prediction," *Health Technol. (Berl).*, vol. 11, no. 1, 2021.

[10] M. Taghavi, A. Gharehghani, F. B. Nejad, and M. Mirsalim, "Developing a model to predict the start of combustion in HCCI engine using ANN-GA approach," *Energy Convers. Manag.*, vol. 195, 2019.

[11] B. Thakur, N. Kumar, and G. Gupta, "Machine learning techniques with ANOVA for the prediction of breast cancer," *Int. J. Adv. Technol. Eng. Explor.*, vol. 9, no. 87, 2022.

[12] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Frontiers in Energy Research*, vol. 9. 2021.

[13] D. P. M. Abellana and D. M. Lao, "A new univariate feature selection algorithm based on the best–worst multi-attribute decision-making method," *Decis. Anal. J.*, vol. 7, 2023.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 4, September 2024, hlm. 653 – 660 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i4.2933
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

[14] A. Thakkar and R. Lohiya, "Attack classification using feature selection techniques: a comparative study," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 1, 2021.

[15] S. Chen and C. Zhou, "Stock Prediction Based on Genetic Algorithm Feature Selection and Long Short-Term Memory Neural Network," *IEEE Access*, vol. 9, 2021.

[16] J. H. Seo and Y. H. Kim, "Machine-learning approach to optimize smote ratio in class imbalance dataset for intrusion detection," *Comput. Intell. Neurosci.*, vol. 2018, 2018.

[17] M. Hussain, M. Dhimish, S. Titarenko, and P. Mather, "Artificial neural network based photovoltaic fault detection algorithm integrating two bi-directional input parameters," *Renew. Energy*, vol. 155, 2020.