COMPARISON OF KNN, NAIVE BAYES, DECISION TREE, ENSEMBLE, REGRESSION METHODS FOR INCOME PREDICTION

Eri Mardiani[1*]; Nur Rahmansyah[2]; Andy Setiawan[3]; Zakila Cahya Ronika[4]; Dini Fatihatul Hidayah[5]; Atira Syakira[6]

Informatika[1]
Universitas Nasional[1]
www.unas.ac.id[1]

Animasi[2]
Politeknik Negeri Media Kreatif[2]
www.polimedia.ac.id[2]

Akuntansi[3,4,5,6]
UPN Veteran Jakarta[3,4,5,6]
www.upnvj.ac.id[3,4,5,6]

erimardiani1@gmail.com[1*], nur_rahmansyah@polimedia.ac.id[2], andysetiawan2285@upnvj.ac.id[3], zakilacahya@gmail.com[4], dinifatihatulhidayah@gmail.com[5], atirasyakiraa@gmail.com[6]

Abstract—*Using the income classification dataset, we performed data analysis with the help of data mining to gather interesting information from the available data. Currently, data processing can be done using many tools. One of the tools that we use for data processing is the orange application. By using the dataset we looked at the welfare level ranging from marital status, school, gender, and from all fields related to income ranging from sales, to daily life to find out the income earned by employees or workers from several countries such as the United States, Cambodia, United Kingdom, Puerto-Rico, Canada, Germany, Outer US (Guam-USVI-etc). The purpose of this analysis is to determine the hourly income in one week that can affect the income classification. The classification technique uses various classification models, namely the K-Nearest Neighbor (KNN) algorithm model, Naïve Bayes, Decision Tree, Essemble Method and Linear Regression algorithm. The results of the analysis based on the test results of various algorithm models can be concluded that the best algorithm model for measuring workers' income is to use the Naive Bayes Decision. Analysis of variables based on Hours-per-Week and Capital-Gain affects Income Classification which determines whether the income earned is more than 50 thousand/50 K and the analysis results in a prediction of a person's income level.*

Keywords: *algorithm method comparison, data mining, income classification, orange.*

Intisari—*Menggunakan dataset income klasifikasi, kami melakukan analisis data dengan bantuan data mining untuk mengumpulkan informasi menarik dari data yang tersedia. Saat ini pengolahan data dapat dilakukan dengan menggunakan banyak tools. Salah satu tools yang kami gunakan untuk pengolahan data adalah aplikasi orange. Dengan menggunakan dataset kami melihat berdasarkan tingkat kesejahteraan mulai dari status pernikahan, sekolah, jenis kelamin, dan dari segala bidang yang berhubungan dengan pendapatan mulai dari penjualan, hingga kehidupan sehari-hari untuk mengetahui pendapatan yang didapat karyawan atau pekerja dari beberapa negara seperti Amerika Serikat, Kamboja, Inggris, Puerto-Rico, Kanada, Jerman, AS Terluar (Guam-USVI-dll). Tujuan dari analisis ini untuk mengetahui pendapatan per jam dalam satu minggu yang dapat mempengaruhi klasifikasi pendapatan. Teknik klasifikasi menggunakan berbagai model klasifikasi, yaitu model algoritma K-Nearest Neighbor (KNN), Naïve Bayes, Decision Tree, Essemble Method dan algoritma Linear Regression. Hasil analisis berdasarkan hasil pengujian dari berbagai macam model algoritma dapat disimpulkan bahwa model algoritma yang paling bagus untuk mengukur pendapatan pekerja adalah dengan menggunakan Naive Bayes Decision.*

Vol. 20, No. 2 September 2023 | DOI https://doi.org/10.33480/techno.v20i2.4613

Analisis variabel berdasarkan Hours-per-Week dan Capital-Gain mempengaruhi Income Classification yang menentukan apakah penghasilan yang didapat lebih dari 50 ribu/50 K dan analisa tersebut menghasilkan prediksi tingkat pendapatan seseorang.

***Kata Kunci**: data mining, klasifikasi pendapatan, orange.*

INTRODUCTION

There are many technological advances available with computers. Computers not only help us get work done, but they can also be fun and useful tools to use. To further maximize work, internet-based technology really supports online work (Mardiani, et al., 2023).

In addition, with internet technology work is much faster and easier to complete than using traditional computers. The development of information technology has become so rapid. Internet technology connects thousands of computer networks of individuals and organizations throughout the world (Laksono, et al., 2023).

As time progresses, more and more data is needed by every human being. With so much data available, it becomes increasingly difficult to analyze the data (Indriyawati & Khoirudin, 2019). Therefore, humans need the help of data mining to collect interesting information from available data. By using data mining, you can discover interesting knowledge from large amounts of data stored in databases, data warehouses, or other information storage places (Karo, et al., 2020).

One of the important problems in data mining is classification which involves finding rules that limit given data into predetermined classes. In the data mining domain where trillions of data are used, the execution time of existing algorithms can take a long time. Therefore, we need automatic tools that can help us convert this huge amount of data into information (Djamaludin, et al., 2022).

Several previous studies were used as references. Research with the Orange Application is known to be beginner-friendly and the data analysis process is simple (Hozairi, et al., 2021). This is because Orange does not require coding skills to operate it. You just have to choose the existing features according to your needs. Let's say you want to create a classification or regression model. You just need to add a widget like KNN or Naive Bayes and provide data to the model by connecting the data source to the model by drawing connecting lines (Ratra & Gulia, 2020).

The next reference research is regarding research with Orange as a comprehensive component-based framework for machine learning and data mining. Oranges have been used in science, industry, and learning. Scientifically, it is used as a testing platform for new machine learning algorithms, as well as to apply new techniques in genetics and other fields of bioinformatics. Orange provides an overview using data visualization, classification, evaluation, unsupervised learning, association, visualization using Qt, and prototype implementation are some of the well-known features of Orange (Wiguna & Rifai, 2021).

The next reference research regarding data prediction can be carried out using several algorithms, including the K-nearest neighbor algorithm and the Neural Network algorithm. In this research, we will compare how the K-Nearest Neighbor and Neural Network algorithms predict household income in the census conducted in Bereau (Priyanti, 2019).

Processing results using orange data mining using income classification dataset with target variable income <=50k and income >50k based on 14 features consisting of Age, Workclass, Fnlwgt, Education, Education-num, Marital-status, Occupation, Relationship, Race , Sex, Capital-gain, Capital-loss, Hours-per week, and Native country.

This research uses quantitative research because we analyze data based on the numbers that we process and then we can conclude the algorithm (Marutho, 2019).

MATERIALS AND METHODS

This research uses quantitative methods related to numbers or nominal values which are often used in survey research or opinion polls. Qualitative methods focus on natural, real, subjective and interactive events with participants. Mixed methods are a combination of quantitative and qualitative techniques so that the results are complete, useful, balanced and informative (Marinu, 2023).

The data collected for this research is primary and secondary. Primary data is taken from the Kaggle website, namely the Income Classification dataset https://www.kaggle.com/datasets/lodetomasi1995/income-classification. The secondary data collection techniques in this research were obtained from online media and other sources. Literature studies will be used by researchers to describe and analyze the data that has been collected which is related to this research (Dachi, 2023).

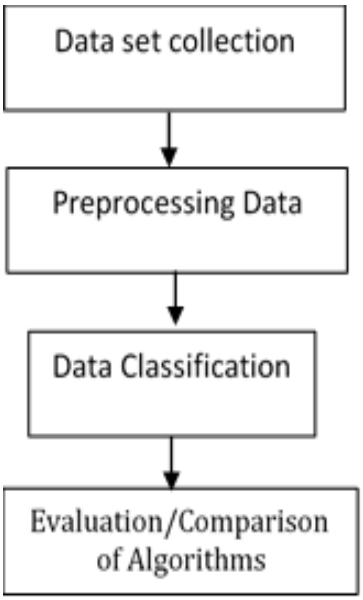

Figure 1. Research Stages

The stages in this research consist of 4 stages, namely dataset collection, data testing, prediction process, performance evaluation and method comparison results (Susetyoko, et al., 2022).

The first step is collecting a dataset. This is done first to develop research objectives and research contributions (Yuwono, et al., 2021). Second is data testing, namely testing existing data with applications that are useful for compiling data as a source of data classification. Third is the prediction process for the results of data testing. Fourth is the process of evaluating the performance of each method tested to produce predictions using KNN, Decision Tree, Essembled Method, Naive Bayes and linear regression models (Giri, 2018). Fifth is the process of method comparison results and analyzing the method comparison results (Indrapras, et al., 2022).

RESULTS AND DISCUSSION

The following is the workflow for the income classification data set technique. Income classification, this data set is obtained from the Kaggle site and is used to present information regarding the predicted value of a group of attributes https://www.kaggle.com/datasets/lodetomasi1995/income-classification.

In the data mining domain where trillions of data are used, the execution time of existing algorithms can take a long time. Therefore, we need automatic tools that can help us convert this huge amount of data into information.

From Figure 2 we can see the data table dataset. We can use the KNN method to get predictions by connecting Preprocess with KNN, Data Sample and Predictions.

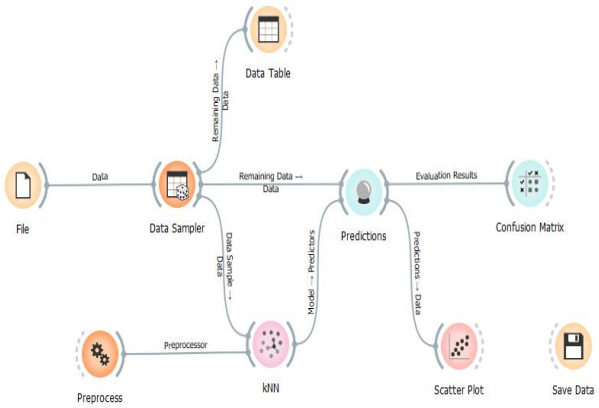

Figure 2. Workflow for the KNN Algorithm

From Figure 3 the Data Sampler is connected to the Data Table and Naive Bayes, and the test score to calculate the success rate and results of the Test and Score.

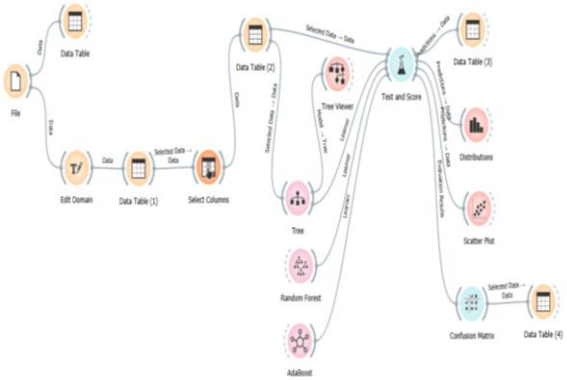

Figure 3. Workflow for the KNN Algorithm

From Figure 4, the Data Sampler is connected to Column to arrange the data domain manually, add a Data Table and connect to Tree Viewer, Data Table, Random Forest, AdaBoost, Distributions, Scatter Plot, Confusion Matrix to see the results

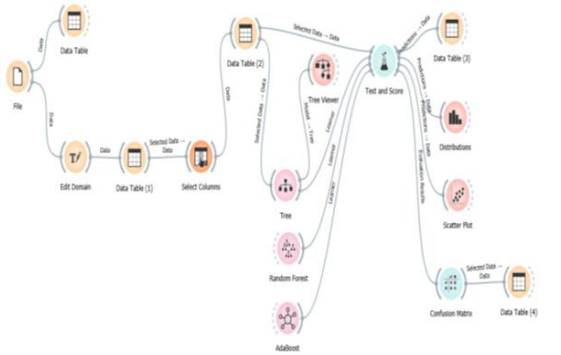

Figure 4. Workflow for Decision Tree and Esemble Algorithms

From Figure 5, the Data Sample is connected to Column and select numeric variables as features

Vol. 20, No. 2 September 2023 | DOI https://doi.org/10.33480/techno.v20i2.4613

and capital gain as target, add the rank and correlation widgets, finally add the Test and Score & Scatter Plot widgets then connect the two to get the results from linear regression

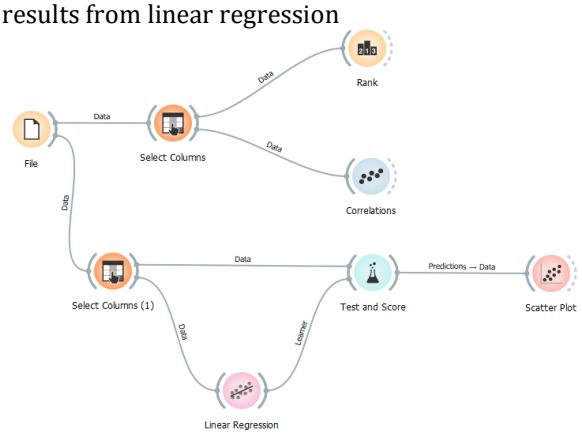

Figure 5. Workflow for the Linear Regression Algorithm

K-Nearest Neighbor (KNN)

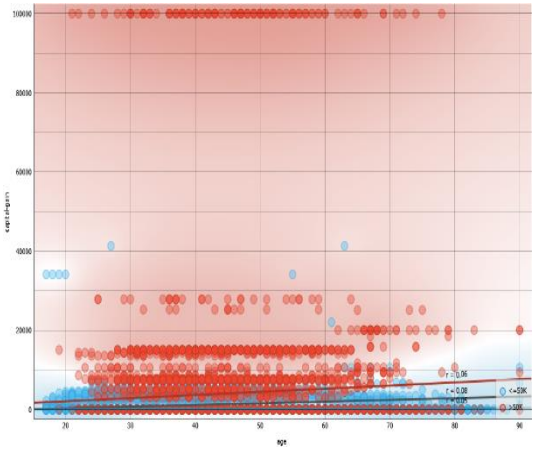

Figure 6. K-Nearest Neighbor (KNN)

The KNN model shows AUC of 84.7%, CA of 82.7%, F1 of 82.5% and Recall of 82.7%. This shows that the KNN prediction model shows good predictions seen from its AUC of 84.7%, which means that anything above 50% shows good prediction results. This means that the income analysis in the kNN model has good predictions. Then, from the results of the Confusion Matrix using the existing Train Data, it shows that KNN's prediction for Income <=50k is actually 19,967 for income <=50k. Then there is a prediction error of 2,280 because KNN predicts income >50k when the actual income is <=50k. The prediction error was 2,782 because KNN predicted income <=50k, so actual income should be >50k. Then, there are KKN predictions for actual income >50k against income >50k totaling 4,275 Data Trains.

From the Scatter Plot above, the Age and Capital-Gain variables influence Income with the highest point being Age 78 and Capital-Gain being 99,999, entering the Income >50k class. The Age and Capital-Gain variables influence Income with the lowest point being Age 17 and Capital Gain being 0, entering the Income <=50k class. If you look at the Scatter Plot above, it is a distribution of Income with the variables Age and Capital-Gain, the relationship between these two variables to Income is, if you look at the Age, the taller or more mature a person is and the higher the Capital Gain, the Income obtained will be greater. namely above 50k. On the other hand, if you look at the lower or immature age and the lower the capital-gain, it will affect the income below 50k.

Naive Bayes

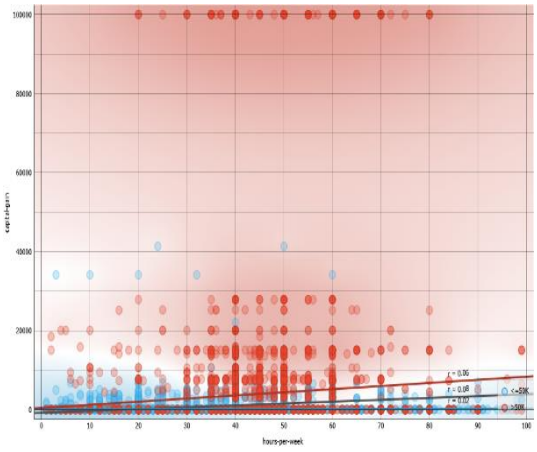

Figure 7 Naïve Bayes

The Naïve Bayes model shows AUC of 0.903, CA of 0.824, F1 of 0.831, Precision of 0.846, and Recall of 0.824. This means that the level of precision in predicting or classifying income below 50k and above 50k is above 84%. Then, the Data Table results connected to the Confusion Matrix show that the Naive Bayes prediction for Income <=50k Actual against Income <=50k is 20,751. However, there is a Naive Bayes prediction error of 3,969 because Naive Bayes predicts Income >50k, the Actual Income should be <=50k and there is a prediction error of 1,750 because kNN predicts Income <=50k, the Actual Income should be >50k.

The Scatter Plot above depicts the distribution of Income with the Hours-per-Week and Capital-Gain variables. The relationship between these two variables and Income is, if you look at Hours-per Week, the higher a person's working hours, the higher the income they get, too. but because there is a Capital-Gain factor, even though you have high Hours-per Week, if the Capital-Gain is low then the Income you get will be low too. On the other hand, if the Capital-Gain is high even though the Hours-per-Week is low, the Income will be high too. If you look at the predictions for

Income Under 50k and Above 50k, you can see the distribution, on average a high Capital-Gain will have an Income Above 50k (in red) and the distribution for Income Under 50k is only spread across Capital-Gain 50,000 and below.

Decision Tree

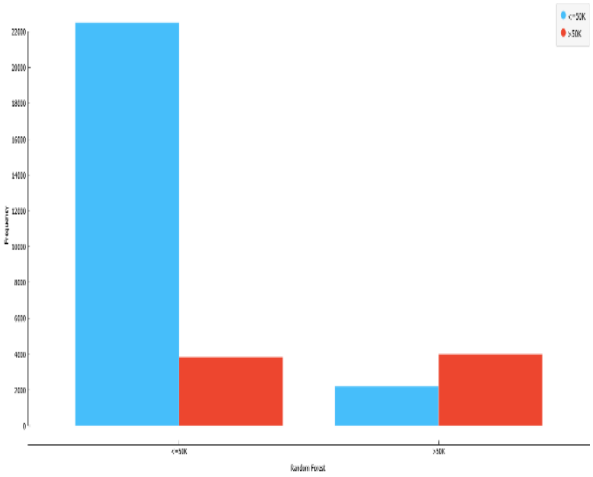

Figure 8 Decision Tree

Based on the 3 algorithm models used, namely Tree, Random Forest, and AdaBoost, the best algorithm model is Random Forest because it has the highest AUC, namely 0.821 or 82%. Meanwhile, the AdaBoost algorithm model has a result of 0.755 or 75% and the smallest is the Tree algorithm model, namely 0.673 or 67%.

The Confusion Matrix results show that the Decision Tree prediction for Actual Income Under 50k against Income Above 50k is 22,481 data. Then, there was a Decision Tree prediction error of 2,239 data because the Decision Tree predicted Income Above 50k, when the Actual Income should have been under 50k. Apart from that, there was a prediction error of 3,902 data because the Decision Tree predicted Income Under 50k, when the Actual Income should be Above 50k. The Confusion Matrix results also show that the Decision Tree prediction for Actual Income Above 50k against Income Under 50k is 3,939 data.

In Tree Viewer, the first benchmark taken is Capital-Gain. If the Capital-Gain is below 6,849 then it is in the Under 50k category and if it is above 6,849 then it is in the Above 50k category. Then the Under 50k category is divided into two based on age. If younger than 29 years, then enter the Under 50k category and if older than 29 years then enter the Under 50k category. The second Under 50k is further divided based on Education-num, if the education period is less than 12 years then it is in the Under 50k category and if it is above 12 years then it is in the Above 50k category.

In the Scatter Plot, the Age and Capital-Gain variables influence Income with the highest point Age = 78 and Capital-Gain = 99,999 Income Above 50k. The Age and Capital-Gain variables influence Income with the lowest point being Age = 17 and Capital-Gain = 0 entering the Income Under 50k class. The Scatter Plot is a distribution of Income with Age and Capital-Gain variables. If you look at the predictions for Income Under 50k and Above 50k, you can see the distribution, on average a high Capital-Gain will have an Income Above 50k (in red) and for the Income Under distribution 50k is only spread across Capital-Gains of 50,000 and under, it doesn't really look at the age, but those in the age range of 20 - 80 have high Capital-Gains, and 80-90 are in Capital-Gains under 30,000.

In the Distributions menu, the Random Forest variable influences Income where the Random Forest prediction for Under 50k predicts Under 50k with 26,383 data and Above 50 with 3,902 data. Apart from that, Random Forest Above 50k predictions show Under 50k predictions of 3,939 data and Above 50k predictions of 6,178 data.

Linear Regression

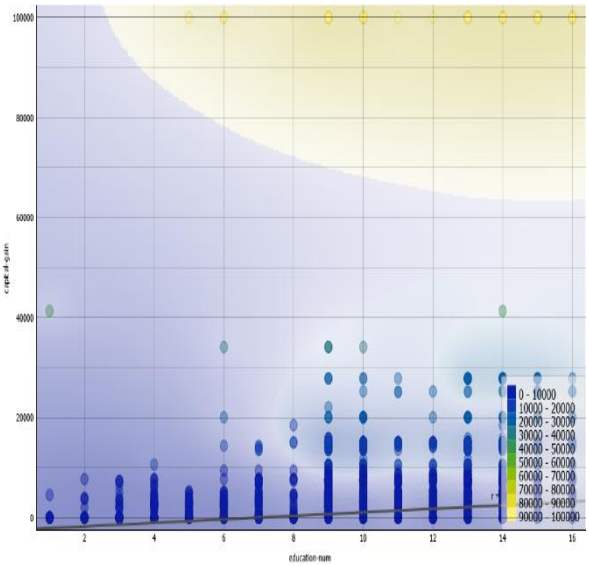

Figure 9 Linear Regression

Based on the creation of a linear regression algorithm model, in Test & Score there are MSE, RMSE, and MAE models which are values for calculating the level of Regression strength. If we use R2, the correlation between education-num data and Capital-Gain, the effect is only 1.5%. Meanwhile, if we add the house-per-week variable the effect becomes 1.9%. This means that it has increased by 0.4%, which means the level of accuracy has increased (better). Then, in the Scatter Plot, the regression points that accumulate between the x axis = Education-num and the y axis = Capital-

Gain, show that the education-num variable influences Capital-Gain. This means that the longer you study, the higher your Capital Gain. From the Scatter Plot, it can be seen that the highest point is education-num = 16 and Capital-Gain = 99,999. Meanwhile, the lowest point is education-num = 1 and Capital-Gain = 0. This means that the results of the regression analysis are, education-num influences the high and low Capital-Gain figures.

Result of Model Comparison

Based on testing various algorithm models, the comparison of test results can be seen in Table 1.

Tabel 1. Result of Model Comparison

Model	AUC	CA	F1	Precision	Recall
KNN	0.847	0.827	0.825	0.847	0.827
NB	0,903	0,824	0,831	0,846	0,824
DT	0.673	0,794	0.788	0.785	0,794
RF	0.821	0.811	0.803	0.800	0.811

Based on table 1, it can be seen that the Naïve Bayes model shows an AUC of 0.90, CA of 0.824, F1 of 0.831, Precision of 0.846, and Recall of 0.824. This means that the level of precision in predicting or classifying income below 50 thousand and above 50 thousand is above 84%

CONCLUSION

From the results of the analysis based on the results of various models ranging from kNN, Naive Bayes, Decision Tree & Ensemble Method and Linear Regression, it can be concluded that the best algorithm model for measuring income is using the Naive Bayes Decision model. This can be seen from the results of AUC, CA, F1, Precision and Recall which have values above 0.5 or above 50%, which shows that the prediction results from the Naive Bayes model are good. It can also be seen in the AUC results of the Naive Bayes model with the number 0.903 and the precision results of 0.846, which means that the prediction from the Naive Bayes model is almost close to 1 or almost accurate.

This means that it can be concluded that variable analysis based on Hours-per-Week and Capital-Gain influences Income Classification which will determine whether the Income is above 50k or below 50k. Because in this analysis our group determined the target to be Income by choosing 2 variables as a means of measuring it, namely Hours-per-Week and Capital-Gain with Numeric data.

These results can be used to predict a person's income level, and are also useful in various ways, such as determining a person's eligibility to receive financial aid programs. However, it is important to ensure that the model is accurate and unbiased in its predictions, and that the data used is representative of the population from which it is studied.

REFERENCE

Dachi, J. (2023). Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit.

Djamaludin, M. A., Triayudi, A., & Mardiani, E. (2022). Analisis Sentimen Tweet KRI Nanggala 402 di Twitter menggunakan Metode Naïve Bayes Classifier. *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*.

Giri, G. A. (2018). Klasifikasi Musik Berdasarkan Genre dengan Metode K-Nearest Neighbor. *Jurnal Ilmu Komputer VOL. XI No. 2*.

Hozairi, H., Anwari, A., & Alim, S. (2021). Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes. *Nero (Networking Engineering Research Operation)*.

Indrapras, K., Alfaris, M., Falentina, & Triana, A. (2022). Analisis Big Data dan Official Statistics dalam Melakukan Nowcasting Pertumbuhan Ekonomi Indonesia Sebelum dan Selama Pandemi COVID-19. *Seminar Nasional Official Statistics 2022, https://prosiding.stis.ac.id/index.php/semnasoffstat*.

Indriyawati, H., & Khoirudin. (2019). Penerapan Metode Regresi Linier Dalam Koherensi Pengolahan Data Bahan Baku Tiandra Store Guna Meningkatkan Mutu Produksi. *Sintak Prosiding,https://www.unisbank.ac.id/ojs/index.php/sintak/article/view/7603*.

Karo, G. E., Erwansyah, K., & Suharsil. (2020). Implementasi Data Mining Dalam Mengestimasi Pendapatan Pada Pt Citosarana Jasa. *Jurnal Cyber Tech*.

Laksono, R. A., Achmadi, S., & Sasmito, A. P. (2023). Implementasi Data Mining Menggunakan Metode Least Square Untuk Memprediksi Jumlah Pendapatan. *JATI (Jurnal Mahasiswa Teknik Informatika). Vol. 7 No. 5, Oktober 2023*.

Mardiani, E., Rahmansyah, N., Ningsih, S., Lantana, D. A., Wirawan, A. S., Wijaya, S. A., & Putri, D. N. (2023). Komparasi Metode KNN, Naive Bayes. Decision Tree, Ensemble, Linear Regression Terhadap Analisis Performa Pelajar Sma. *Jurnal INNOVATIVE: Journal Of Social Science Research*, 13880-13892.

Marinu, W. (2023). Pendekatan Penelitian Pendidikan: Metode Penelitian Kualitatif, Metode Penelitian Kuantitatif dan Metode Penelitian Kombinasi (Mixed Method). *Jurnal Pendidikan Tambusai, Volume 7 Nomor 1 Tahun 2023. Halaman 2896-2910*, 2896-2910.

Marutho, D. (2019). Perbandingan Metode Naïve Bayes, Knn, Decision Tree Pada Laporan Water Level Jakarta. *Jurnal Ilmiah Infokam*, Vol 15, No 2.

Priyanti, E. (2019). Komparasi Klasifikasi Pada Prediksi Pendapatan Rumah Tangga. *JURNAL SWABUMI Vol.7 No.2 September 2019, pp.114~121*, 114-121.

Ratra, R., & Gulia, P. (2020). Experimental evaluation of open source data mining tools (WEKA and Orange). International Journal of Engineering Trends and Technology,. *International Journal of Engineering Trends and Technology, 68(8), 30-35.*, 30-35.

Susetyoko, R., Yuwono, W., Purwantini, E., & Ramadijanti, N. (2022). Perbandingan Metode Random Forest, Regresi Logistik,Naïve Bayes, dan Multilayer Perceptron Pada Klasifikasi Uang Kuliah Tunggal (UKT). *Jurnal Infomedia: Teknik Informatika, Multimedia & Jaringan*.

Wiguna, R. A., & Rifai, A. I. (2021). Analisis Text Clustering Masyarakat Di Twitter Mengenai Omnibus Law Menggunakan Orange Data Mining. *Journal of Information Systems and Informatics*, 2656-5935.

Yuwono, L., Fadillah, M. E., Indrayani, M., Maesarah, W., Ramadhan, A., & Panjaitan, F. S. (2021). Klasifikasi Pendapatan Pedagang Kaki Lima Dan Pelaku Usaha Online Akibat DampakCovid-19 Menggunakan Metode Naive Bayes. *Bulletin of Applied Industrial Engineering Theory*.