INNOVATIVE: Journal Of Social Science Research
Volume 3 Nomor 4 Tahun 2023 Page 8667-8679
E-ISSN 2807-4238 and P-ISSN 2807-4246
Website: https://j-innovative.org/index.php/Innovative

Analisis Prediksi Pendapatan Penduduk dengan Metode K-Nearest Neighbor, Decision Tree, Naive Bayes, Ensemble Methods, dan Linear Regression

Eri Mardiani[1], Nur Rahmansyah[2], Endah Tri Esti Handayani[3] , Sari Ningsih[4], Deny Hidayatullah[5], Dhieka Avrilia Lantana[6], Yuni Latifah[7], Alica Dwi Fahira[8], Keysha Belynda Tyva Panggabean[9], Imelta Natalia Ginting[10]
(1), (3), (4), (5), (6) Universitas Nasional, Indonesia
(2) Program Studi Animasi Politeknik Negeri Media Kreatif, Indonesia
(7), (8), (9), (10) UPN Veteran Jakarta, Indonesia
Email: erimardiani1@gmail.com[1✉]

Abstrak

Data mining bermula dari peningkatan data yang cukup pesat dilihat dari segi volume serta variasi data yang dihasilkan oleh berbagai sumber, dan jumlahnya yang sangat besar, serta kompleksitasnya data hingga pembuatannya yang cepat. Dengan data bisa menghasilkan prediksi yang membantu pemerintah dalam mengambil keputusan dan kebijakan di masa mendatang. Selain itu prediksi dapat membantu pemerintah dalam perencanaan kegiatan yang akan dilakukan untuk mencapai tujuan, karena prediksi ini dapat memberikan output terbaik sehingga diharapkan resiko kesalahan yang disebabkan oleh kesalahan perencanaan dapat ditekan seminimal mungkin. Prediksi biasanya digunakan untuk menemukan informasi dari sejumlah data yang besar sehingga diperlukan data mining. Data mining dapat digunakan untuk menggali informasi dari data yang besar sehingga didapatkan informasi yang dapat digunakan dalam memprediksi sesuatu. Dalam data mining terdapat banyak teknik dalam pengerjaannya, untuk menemukan pola atau informasi yang tersembunyi diantaranya adalah Klasterisasi (clustering), Regresi (regression), Asosiasi (association), dan Klasifikasi (classification)

Kata Kunci : *Data mining, income*

Abstract

Data mining begins with a fairly rapid increase in data in terms of the volume and variety of data produced by various sources, and the numbers are very large, as well as the complexity of the data and its rapid creation. Data can produce predictions that help the government in making decisions and policies in the future. Apart from that, predictions can help the government in planning activities that will be carried out to achieve goals, because these predictions can provide the best output so it is hoped that the risk of errors caused by planning errors can be reduced to a minimum. Predictions are usually used to find information from large amounts of data, so data mining is needed.Data mining can be used to dig up information from large amounts of data to obtain information that can be used to predict something. In data mining, there are many techniques used to find hidden patterns or information, including clustering, regression, association and classification.
Keywords: *Keywords contain basic ideas or concepts that represent the field under study; The number of keywords is between 3-5 Phrases and are sorted alphabetically*

PENDAHULUAN

Di era digital saat ini, ekonomi merupakan salah satu industri terpenting dan berkembang sangat cepat. Hal ini didukung oleh perkembangan teknologi yang semakin pesat di seluruh dunia. Itulah sebabnya setiap orang yang berkecimpung dalam bisnis bergerak dan saling bersaing untuk menjadi yang terbaik di bidangnya.[10]

Penelitian ini mempunyai tujuan agar dapat membandingkan metode K-NN, Naive Bayes, Decision Tree, Ensemble Methods, dan Linear Regression yang dilakukan untuk klasifikasi apakah seorang penduduk memiliki pendapatan lebih besar atau sama dengan $50.000 USD per tahun atau tidak. Sedangkan aplikasi yang digunakan adalah aplikasi data mining orange yang merupakan aplikasi data mining open source yang terbukti dapat membantu dalam hal penganalisaan data. Untuk melakukan hal tersebut kita akan menunjukkan prosesnya mulai dari akuisisi data sampai prediksi.[1]

Data Mining menggunakan Teknik Klasifikasi dengan 5 Model Algoritma[2]

1. K-Nearest Neighbor (k-NN) [3]

Algoritma K-Nearest Neighbor (k-NN) merupakan sebuah model algoritma yang digunakan untuk melakukan klasifikasi terhadap objek yang berdasarkan dari data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Ataupun dapat dipahami juga bahwa k- nearest neighbor adalah salah satu algoritma yang paling sederhana dan banyak Titik data akan diklasifikasikan berdasarkan kesamaan kelompok tertentu dari titik data lain yang berdekatan. Sehingga, algoritma ini akan memberikan hasil yang kompetitif.

2. Naïve Bayes[5]

Salah satu metode data mining ialah klasifikasi Naive Bayes. Naive Bayes Classifier adalah metode klasifikasi yang berakar pada teorema Bayes. Naive bayes merupakan metode

pengklasifikasian berdasarkan probabilitas sederhana dan dirancang agar dapat dipergunakan dengan asumsi antar variabel penjelas saling bebas (independen). Pada algoritma ini pembelajaran lebih ditekankan pada pengestimasian probabilitas. tujuan dari metode Naïve Bayes adalah untuk menemukan probabilitas ketika kita mengetahui probabilitas tertentu lainnya. Hasil dari perhitungan data mining menggunakan metode klasifikasi Naïve Bayes akan makin berguna jika penyajiannya menarik dan dapat dipahami dengan baik oleh penerima data.

3. Decision Tree[9]

Decision tree adalah algoritma machine learning yang menggunakan seperangkat aturan untuk membuat keputusan dengan struktur seperti pohon yang memodelkan kemungkinan hasil, biaya sumber daya, utilitas dan kemungkinan konsekuensi atau resiko. Konsepnya adalah dengan cara menyajikan algoritma dengan pernyataan bersyarat, yang meliputi cabang untuk mewakili langkah-langkah pengambilan keputusan yang dapat mengarah pada hasil yang menguntungkan. Klasifikasi ini menggunakan observasi pada node untuk menemukan target pada leaves. Decision Tree merupakan salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi oleh manusia dengan kemampuannya untuk mem-break down proses pengambilan keputusan yang kompleks menjadi lebih simple.[6]

4. Ensemble Method

Ensemble Method adalah algoritma dalam pembelajaran mesin (machine learning) dimana algoritma ini sebagai pencarian solusi prediksi terbaik dibandingkan dengan algoritma yang lain karena metode ensemble ini menggunakan beberapa algoritma pembelajaran untuk

pencapaian solusi prediksi yang lebih baik daripada algoritma yang bisa diperoleh dari salah satu pembelajaran algoritma konstituen saja. Tidak seperti ensemble statistika dalam mekanika

statistika biasanya selalu tak terbatas. Ensemble Pembelajaran hanya terdiri dari seperangkat model alternatif yang bersifat terbatas, namun biasanya memungkinkan untuk menjadi lebih banyak lagi struktur fleksibel yang ada diantara alternatif model itu sendiri. Evaluasi prediksi dari

ensemble biasanya memerlukan banyak komputasi daripada evaluasi prediksi model tunggal

(single model).

5. Linear Regression[4]

Linear regression merupakan salah satu algoritma yang memodelkan suatu persamaan untuk menghitung estimasi. Pada metode ini bertujuan untuk mencari pola pada nilai

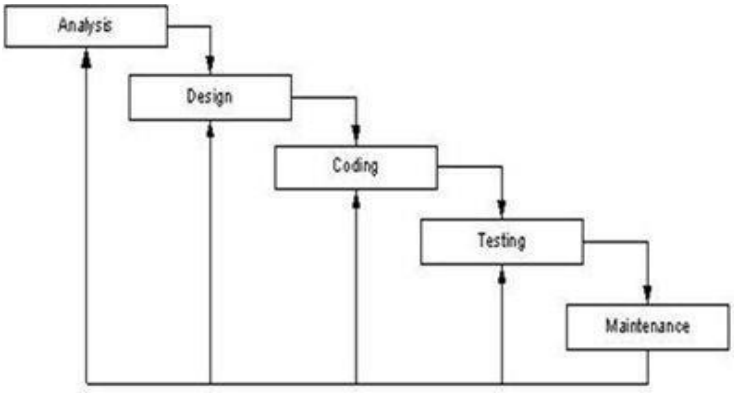
numerik,

sehingga data yang dibutuhkan berupa data numerik agar dapat diolah dengan model algoritma ini. Linear regression memprediksi nilai data yang tidak diketahui dengan menggunakan nilai data lain yang terkait dan diketahui. Secara matematis memodelkan variabel yang tidak diketahui atau tergantung dan variabel yang dikenal atau independen sebagai persamaan linier..

METODE PENELITIAN

Data yang dikumpulkan untuk penelitian ini yaitu primer dan sekunder. Teknik pengumpulan data primer berasal dari studi literatur seperti buku maupun jurnal. Adapun Teknik pengumpulan data sekunder pada penelitian ini diperoleh dari media online maupun sumber-sumber lainnya. Studi pustaka akan digunakan oleh peneliti guna menjabarkan dan menganalisis data yang sudah dikumpulkan yang memiliki kaitan dengan penelitian ini .[7]

Selain metode kualitatif , kami juga menggunakan metode waterfall karena dibutuhkan pendekatan yang sistematis dan berurutan dalam membangun sistem. Aliran cascading adalah bahwa sistem ditulis secara berurutan. Sistem yang dihasilkan menghasilkan kualitas yang sangat baik karena implementasinya bertahap dan tidak terfokus pada fase tertentu [12]



Gambar 1. Metode Waterfall

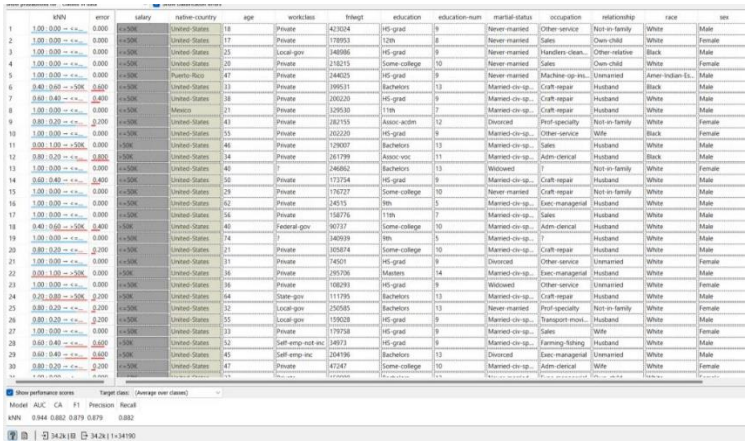
Pada tahap pengembangan sebuah sistem, metode ini bersifat sistematis, serta rangkaian dalam sistem perancangan perangkat lunak dilakukan secara urut. Metode waterfall atau disebut juga metode air terjun merupakan metode yang sering dipakai untuk melakukan pembaharuan pada sistem aplikasi yang sedang berjalan. Terdapat beberapa tahapan pada metode ini, yaitu analisis aplikasi, desain aplikasi, implementasi aplikasi, pengujian aplikasi, dan pemeliharaan [8][11]

HASIL DAN PEMBAHASAN

Hasil

A. K-Nearest Neighbor

Implementasi menggunakan K-Nearest Neighbor (KNN)



Gambar 2 Tabel Prediction

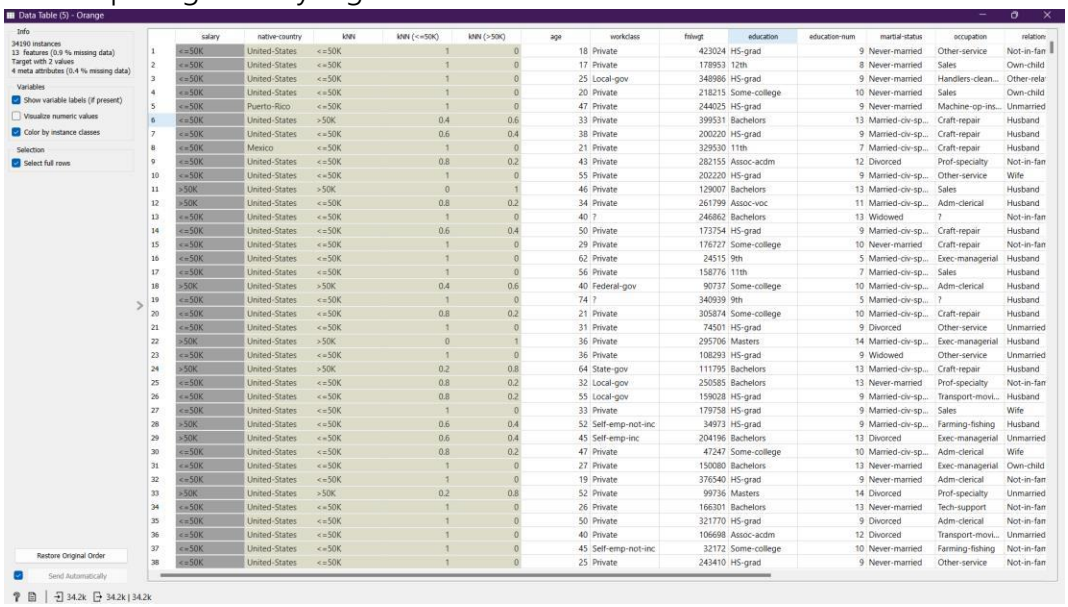
Berdasarkan hasil penelitian menggunakan Predictions dapat dilihat tabel nilai prediksi dan nilai aktualnya. Jika dilihat secara manual oleh tabel pada Predictions, nilai prediksi menggunaka metode K-NN merupakan hampir keseluruhan nilainya mendeskripsikan dengan benar kepada nilai aktualnya. Hal tersebut didukung dengan nilai K-NN yang berada di bawah, nilai K-NN memperoleh nilai AUC sebesar 94,4% yang berarti sangat baik karena hampir mendekati 100%, nilai CA sebesar 88,2%, nilai F1 sebesar 87,9%, nilai Precisions sebesar 87,9%, dan nilai Recall sebesar 88,2%.

		Predicted		
		<=50K	**>50K**	**Σ**
Actual	**<=50K**	24465	1512	**25977**
	>50K	2523	5690	**8213**
	Σ	**26988**	**7202**	**34190**

Gambar 3 Confusion KNN

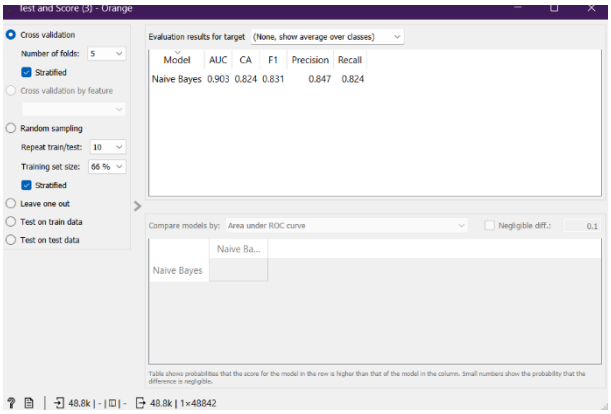
Selanjutnya, penilaian metode K-NN dapat dilihat dengan evaluasi menggunakan Confusion Matrix. Pada gambar diatas dijelaskan bahwa untuk variabel <=50K menunjukkan hasil prediksi dengan metoda K-NN memiliki nilai benar sebanyak 24,465 sedangkan nilai salah sebanyak 2.523. Kemudian, untuk variabel >50K menunjukkan hasil prediksi dengan metode K-NN memiliki nilai benar sebanyak 5.690, sedangkan nilai salah sebanyak 1.512.

Hasil penelitian dengan Confusion Matrix juga dapat dilihat menggunakan ikon Data Table, untuk melihat secara manual dan lebih jelas, manakah nilai prediksi dari metode K-NN yang benar dan manakah nilai prediksi dari metode K-NN yang salah. Untuk hasil Data Table ditampilkan seperti gambar yang tertera di bawah.



Gambar 4 Hasil Data Table

B. Naïve Bayes



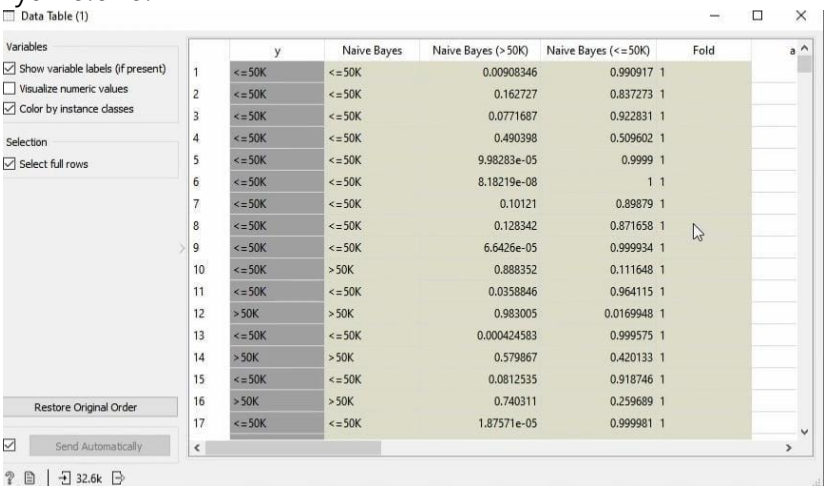
Gambar 5 Data Tabel Naïve Bayes

Berdasarkan hasil penelitian menggunakan Test and Score dapat dilihat pada gambar diatas, nilai yang diperoleh menggunakan metode Naive Bayes. Nilai Naive Bayes memperoleh nilai AUC sebesar 90,3% yang berarti sangat baik karena hampir mendekati 100%, nilai CA sebesar 82,4%, nilai F1 sebesar 83,1%, nilai Precisions sebesar 84,7%, dan nilai Recall sebesar 82,4%.

		Predicted		
		<=50K	>50K	Σ
Actual	<=50K	31129	6026	37155
	>50K	2555	9132	11687
	Σ	33684	15158	48842

Gambar 6 Confusion Naïve Bayes

Selanjutnya, penilaian metode Naive Bayes dapat dilihat dengan evaluasi menggunakan Confusion Matrix. Pada gambar diatas dijelaskan bahwa untuk variabel <=50K menunjukkan hasil prediksi dengan metode Naive Bayes memiliki nilai benar sebanyak 31.129, sedangkan nilai salah sebanyak 2555. Kemudian, untuk variabel >50K menunjukkan hasil prediksi dengan metode Naive Bayes memiliki nilai benar sebanyak 9.132, sedangkan nilai salah sebanyak 6.026.

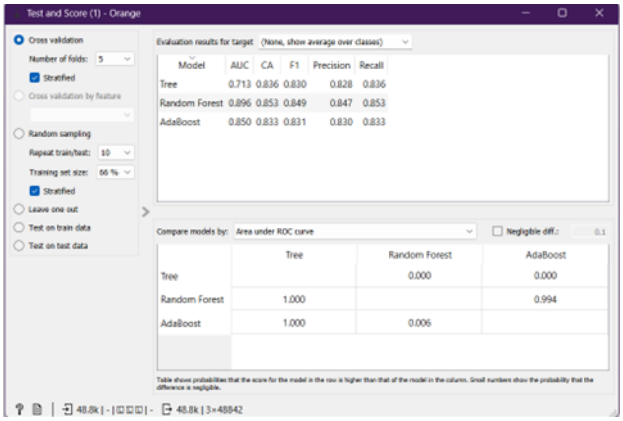


	y	Naive Bayes	Naive Bayes (>50K)	Naive Bayes (<=50K)	Fold
1	<=50K	<=50K	0.00908346	0.990917	1
2	<=50K	<=50K	0.162727	0.837273	1
3	<=50K	<=50K	0.0771687	0.922831	1
4	<=50K	<=50K	0.490398	0.509602	1
5	<=50K	<=50K	9.98283e-05	0.9999	1
6	<=50K	<=50K	8.18219e-08	1	1
7	<=50K	<=50K	0.10121	0.89879	1
8	<=50K	<=50K	0.128342	0.871658	1
9	<=50K	<=50K	6.6426e-05	0.999934	1
10	<=50K	>50K	0.888352	0.111648	1
11	<=50K	<=50K	0.0358846	0.964115	1
12	>50K	>50K	0.983005	0.0169948	1
13	<=50K	<=50K	0.000424583	0.999575	1
14	>50K	>50K	0.579867	0.420133	1
15	<=50K	<=50K	0.0812535	0.918746	1
16	>50K	>50K	0.740311	0.259689	1
17	<=50K	<=50K	1.87571e-05	0.999981	1

Gambar 6 Hasil Data Table Naïve Bayes

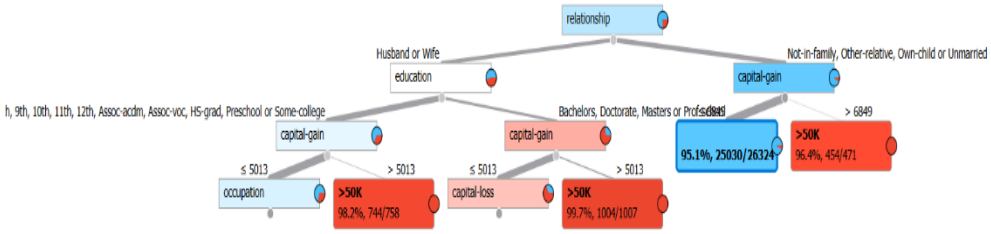
Hasil penelitian dengan Test and Score juga dapat dilihat menggunakan ikon Data Table yang tertera pada gambar diatas, untuk melihat secara manual dan lebih jelas, manakah nilai prediksi dari metode Naive Bayes yang benar dan manakah nilai prediksi dari metode Naive Bayes yang salah.

c. Decision Tree dan Ensemble Tree



Gambar 7 Test And Score

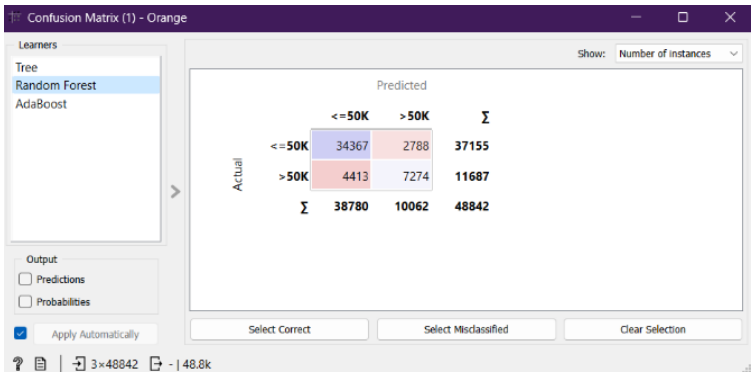
Berdasarkan hasil penelitian menggunakan Test and Score dapat dilihat pada gambar diatas, niilai yang diperoleh menggunakan metode Tree, Random Forest, dan AdaBoost. Untuk nilai Tree memperoleh nilai AUC sebesar 71,3% yang berarti sangat baik karena hampir mendekati 100%, nilai CA sebesar 83,6%, nilai F1 sebesar 83%, nilai Precisions sebesar 82,8%, dan nilai Recall sebesar 83,6%. Untuk nilai Random Forest memperoleh nilai AUC sebesar 89,6% yang berarti sangat baik karena hampir mendekati 100%, nilai CA sebesar 85,3%, nilai F1 sebesar 84,9%, nilai Precisions sebesar 84,7%, dan nilai Recall sebesar 85,3%. Untuk nilai AdaBoost memperoleh nilai AUC sebesar 85% yang berarti sangat baik karena hampir mendekati 100%, nilai CA sebesar 83,3%, nilai F1 sebesar 83,1%, nilai Precisions sebesar 83%, dan nilai Recall sebesar 83,3%. Dari ketiga metode, disimpulkan bahwa nilai metode terbaik untuk prediksi adalah metode Random Forest.



Gambar 8 Hasil Tree Viewer

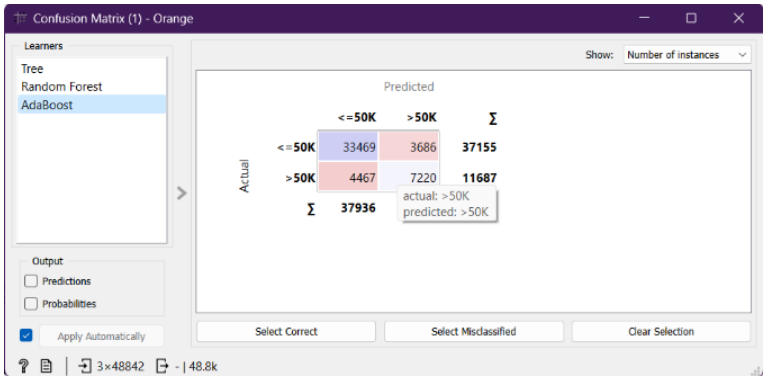
Untuk melihat hasil evaluasi Tree bisa digunakan Tree Viewer, pada gambar dijelaskan bahwa variabel-variabel terbagi ke dalam beberapa klasifikasi. Penjelasan dalam gambar diatas terdiri dari beberapa cabang. Cabang Utama yang menjelaskan mengenai hubungan

atau relationship, memiliki dua cabang lainnya yang pertama mengenai status pernikahan yang berpendidikan dan yang masih single dengan memiliki pendapatan sendiri. Dalam status pernikahan terdapat dua cabang mengenai beberapa lama waktu menempuh pendidikan dan status pernikahan yang sudah mempunyai gelar pendidikan (bachelor, master, etc). Sedangkan untuk pribadi yang masih single dan mendapatkan pendapatan sendiri memiliki dua cabang dimana nilai yang kurang dari 6849 sebesar 95,1% dan yang lebih dari 6849 sebesar 96,4%. Lalu, untuk status pernikahan yang sedang mengambil gelar memiliki dua cabang yaitu kurang dari 5013 dengan hasil occupation (pekerjaan) dan yang lebih dari 5013 dengan hasil pendapatan lebih dari $50.000 memiliki hasil sebesar 98,2%. Lalu untuk status pernikahan yang sudah mendapatkan gelar memiliki dua cabang yang pertama kurang dari 5013 menghasilkan capital-loss dan yang lebih dari 5013 dengan hasil pendapatan lebih dari $50.000 mendapatkan nilai sebesar 99.7%.



Gambar 9 Confusion Tree Viewer

Selanjutnya, penilaian metode Tree dapat dilihat dengan evaluasi menggunakan Confusion Matrix. Pada gambar diatas dijelaskan bahwa untuk variabel <=50K menunjukkan hasil prediksi dengan metode Tree memiliki nilai benar sebanyak 34.084, sedangkan nilai salah sebanyak 4.959. Kemudian, untuk variabel >50K menunjukkan hasil prediksi dengan metode Tree memiliki nilai benar sebanyak 6.728, sedangkan nilai salah sebanyak 3.071.

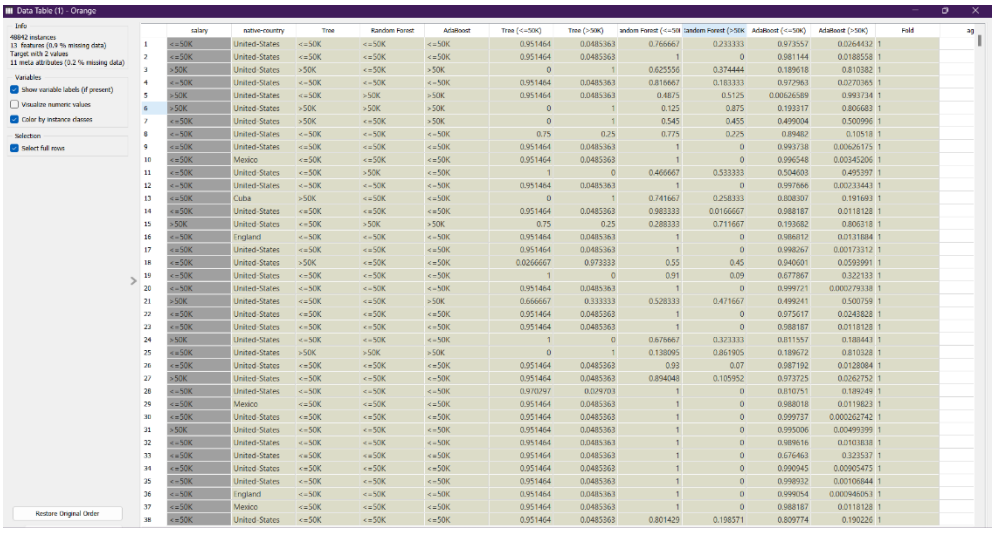


Gambar 10 Confusion Random Forest

Untuk penilaian metode Random Forest dapat dilihat dengan evaluasi menggunakan Confusion Matrix. Pada gambar diatas dijelaskan bahwa untuk variabel <=50K menunjukkan hasil prediksi dengan metode Random Forest memiliki nilai benar sebanyak 34.367, sedangkan nilai salah sebanyak 4.413.Kemudian, untuk variabel >50K menunjukkan hasil prediksi dengan metode Random Forest memiliki nilai benar sebanyak 7.274, sedangkan nilai salah sebanyak 2.728.

Untuk penilaian metode AdaBoost dapat dilihat dengan evaluasi menggunakan Confusion Matrix. Pada gambar diatas dijelaskan bahwa untuk variabel <=50K menunjukkan hasil prediksi dengan metode AdaBoost memiliki nilai benar sebanyak 33.469, sedangkan nilai salah sebanyak 4.467. Kemudian, untuk variabel >50K menunjukkan hasil prediksi dengan metode AdaBoost memiliki nilai benar sebanyak 7.220, sedangkan nilai salah sebanyak 3.686.

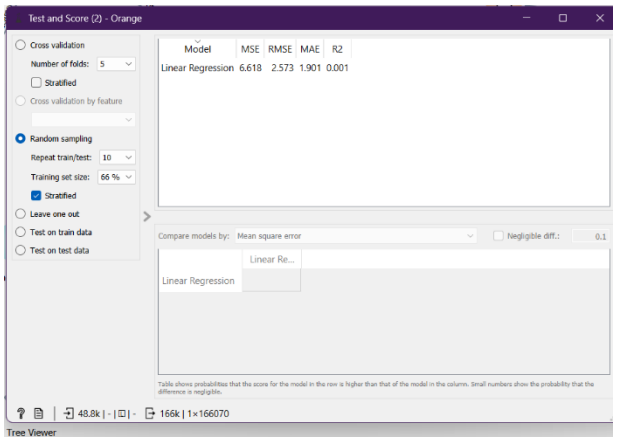
Hasil penelitian dengan Test and Score juga dapat dilihat menggunakan



Gambar 11 Data Table

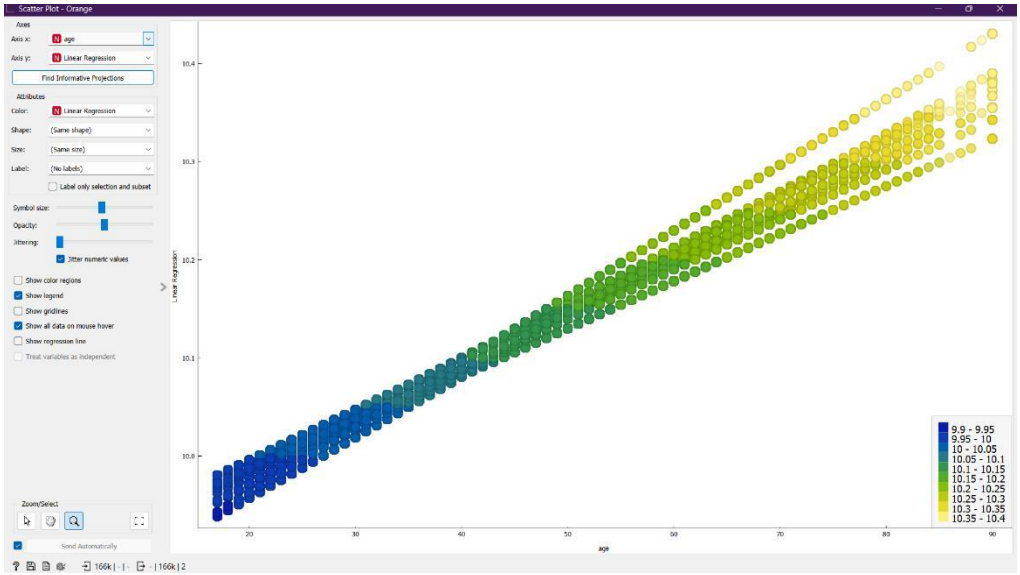
ikon Data Table yang tertera pada gambar diatas, untuk melihat secara manual dan lebih jelas, manakah nilai prediksi dari ketiga metode yang benar dan manakah nilai prediksi dari ketiga metode yang salah.

D. Linear Regression



Gambar 12 Test and Score

Berdasarkan hasil penelitian menggunkanakan Test and Score dapat dilihat pada gambar diatas, nilai yang diperoleh menggunakan metode Linear Regression. Nilai Linear Regression memperoleh nilai MSE sebesar 6.618 yang berarti bahwa pada saat nanti kita melakukan prediksi akan terdapat error sebesar 6.618, nilai MRSE sebesar 2.573, nilai MAE sebesar 1.901, dan nilai R2 sebesar 0.001.



Gambar 3 Scatter Plotter

Untuk penilaian metode Linear Regression dapat dilihat dengan evaluasi menggunakan Scatter Ploter. Pada gambar diatas dijelaskan bahwa jika yang menjauhi garis, maka nilai prediksi terdapat kesalahan.

SIMPULAN

Prediksi atau peramalan (forecasting) adalah suatu perhitungan untuk meramalkan keadaan di masa mendatang melalui pengujian keadaan di masa lalu. Salah satu dari kegunaan prediksi adalah untuk membantu pemerintah dalam mengambil keputusan dan kebijakan di masa mendatang. Selain itu prediksi dapat membantu pemerintah dalam perencanaan kegiatan yang akan dilakukan untuk mencapai tujuan, karena prediksi ini dapat memberikan output terbaik sehingga diharapkan resiko kesalahan yang disebabkan oleh kesalahan perencanaan dapat ditekan seminimal mungkin. Prediksi biasanya digunakan untuk menemukan informasi dari sejumlah data yang besar sehingga diperlukan data mining. Data mining merupakan bidang dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistic, database dan visualisasi untuk penanganan permasalahan pengambilan informasi dari penyimpanan database yang besar.Data mining dapat digunakan untuk menggali informasi dari data yang besar sehingga didapatkan informasi yang dapat digunakan dalam memprediksi sesuatu. Dalam data mining terdapat banyak teknik dalam pengerjaannya, untuk menemukan pola atau informasi yang tersembunyi diantaranya adalah Klasterisasi (clustering), Regresi (regression), Asosiasi (association), dan Klasifikasi (classification).

DAFTAR PUSTAKA

[1] *https://www.kaggle.com*

[2] FH Pratama, A Triayudi, E Mardiani.(2022). Data Mining K-Medoids Dan K-Means Untuk Pengelompokan Potensi Produksi Kelapa Sawit di Indonesia. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika) 7 (4), 1294-1310

[3] Indriyanti, Ichsan, Nurul., Fatah, Haerul ,. Wahyuni,. Tri., Ermawati, Erni. Implementasi Orange Data Mining Untuk Prediksi Harga Bitcoin. JURNAL RESPONSIF, Vol. 4 No.2 Agustus 2022, pp. 118~125

[4] Indriyawati.Henny , Khoirudin.2019. PENERAPAN METODE REGRESI LINIER DALAM KOHERENSI PENGOLAHAN DATA BAHAN BAKU TIANDRA STORE GUNA MENINGKATKAN MUTU PRODUKSI. Proceeding SINTAK 2019 Universitas Semarang https://www.unisbank.ac.id/ojs/index.php/sintak/article/view/7603

[5] MA Djamaludin, A Triayudi, E Mardiani(.2022) Analisis Sentimen Tweet KRI Nanggala 402 di Twitter menggunakan Metode Naïve Bayes Classifier. Jurnal JTIK(Jurnal Teknologi Informasi dan Komunikasi)) 6 (2) 2022 pp 2580-1643

[6] Mardiani, E., Hayati, N., Purnama, M. I. W., Ningsih, S., Darusalam, U., Handayani, T. E.,

& Rahmansyah, N. (2023). *Kumpulan Latihan VB. Net*. Elex Media Komputindo.

[7] Mardiani, E., Rahmansyah, N., & Kurniati, I. (2023). Website Design at SDN Cipete Utara 07. *SITEKIN: Jurnal Sains, Teknologi Dan Industri*, *20*(2), 891–898.

[8] Mardiani, E., Rahmansyah, N., Kurniawan, H., & Sensuse, D. I. (2016). *Kumpulan Latihan SQL*. Elex Media Komputindo.

[9] Mardiani, E., Rahmansyah, N., Ningsih, S., Lantana, D. A., Wirawan, A. S. P., Wijaya, S. A., & Putri, D. N. (2023). Komparasi Metode Knn, Naive Bayes, Decision Tree, Ensemble, Linear Regression Terhadap Analisis Performa Pelajar Sma. *Innovative: Journal Of Social Science Research*, *3*(2), 13880–13892.

[10] Mardiani, E., & Ramadhan, F. A. (2023). Design Information System Sales of Nuts and Bolts at PT. Catur Naga Steelindo. *SITEKIN: Jurnal Sains, Teknologi Dan Industri*, *20*(2), 729–735.

[11] Matondang, Nurhafifah, Mardiani, E., Wahyudi, Praptiningsih, & Saebani, A. (2019). *Aplikasi Komputer*.

[12] Rahmansyah, N., Muliyani, D., Mardiani, E., & Rahman, A. (2022). Perancangan Sistem Transaksi Berbasis Web pada UKM Pangkas Rambut Tasik. *Jurnal Sistem Informasi Bisnis (JUNSIBI)*, *3*(1), 22–31.