seminar nasional 2024 UNIVERSITAS MA CHUNG

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 73-82, 2024

ISSN: 2808-7771 online

Penerapan Model CRISP-DM pada Analisis Pendapatan Menggunakan Metode Klasifikasi

Tjok Istri Vicky Savitri[1], Wilbert Bryan Wibowo[2], dan Trianggoro Wiradinata[3]

[1,2,3] School of Information Technology, Universitas Ciputra Surabaya
CitraLand CBD Boulevard, Surabaya, Indonesia, 60219

Korespondensi: Wilbert Bryan Wibowo (wbryanwibowo@student.ciputra.ac.id)

Received: 24 Juli 2024 – *Revised*: 31 Agustus 2024 - *Accepted*: 05 Sept 2024 - *Published*: 10 Sept 2024

Abstrak. Di era globalisasi ini, prediksi klasifikasi pendapatan dibutuhkan untuk membantu pemerintah dalam mengalokasikan sumber daya untuk berbagai layanan publik, pembangunan infrastruktur, kesehatan, pendidikan, dan program sosial lainnya. Dengan memahami pola pendapatan dan kebutuhan masyarakat, pemerintah dapat merencanakan dan mendistribusikan anggaran secara lebih efektif dan efisien, serta memastikan bahwa layanan dan program yang disediakan tepat sasaran dan memberikan manfaat maksimal bagi masyarakat. Data *Census Income* mencakup berbagai atribut demografis dan ekonomi, termasuk usia, jenis kelamin, pendidikan, status pernikahan, pekerjaan, ras, jam kerja per minggu, dan asal negara. Penelitian ini menggunakan teknik *machine learning* untuk mengklasifikasikan individu berdasarkan tingkat pendapatan mereka, apakah di atas atau di bawah $50.000 per tahun. Metode klasifikasi yang digunakan meliputi *Logistic Regression*, *K-Nearest Neighbors* (KNN), dan *Naive Bayes*. Penelitian ini menggunakan sebanyak 30.162 data dengan pembagian 80% sebagai data latih dan 20% sebagai data tes. Hasil penelitian menunjukkan akurasi untuk *Logistic Regression* sebesar 81%, *KNN* sebesar 79%, dan *Naive Bayes* sebesar 77%. Hasil penelitian juga menunjukkan bahwa faktor-faktor seperti tingkat pendidikan, jam kerja per minggu, dan jenis pekerjaan memiliki pengaruh signifikan terhadap pendapatan individu. Temuan ini dapat membantu pemerintah dan pembuat kebijakan dalam merumuskan strategi untuk mengurangi kesenjangan pendapatan dan meningkatkan kesejahteraan ekonomi masyarakat. Dapat disimpulkan bahwa penggunaan *Logistic Regression* terbukti paling akurat dalam memprediksi pendapatan.

Kata kunci: Model CRISP-DM, Prediksi Pendapatan, Klasifikasi Pendapatan, Klasifikasi

Citation Format: Savitri, T.I.V., Wibowo, W.B., & Wiradinata, T. (2024). Penerapan Model CRISP-DM pada Analisis Pendapatan Menggunakan Metode Klasifikasi. *Prosiding SENAM 2024: Seminar Nasional Sistem Informasi & Informatika Universitas Ma Chung*. 4, 73-82. Malang: Ma Chung Press.

PENDAHULUAN

Peningkatan kualitas hidup dan kesetaraan ekonomi merupakan fokus utama dalam berbagai kebijakan pemerintah di seluruh dunia. Dalam upaya mencapai tujuan tersebut, analisis data demografi dan ekonomi menjadi sangat krusial (Chakrabart & Biswas, 2018). Salah satu alat yang digunakan untuk mendapatkan wawasan mendalam mengenai kondisi ekonomi masyarakat adalah melalui *Census Income*. *Census Income* merupakan sebuah dataset yang dikumpulkan oleh Biro Sensus Amerika Serikat dengan tujuan menganalisis

informasi mengenai distribusi pendapatan di berbagai kelompok masyarakat. Informasi yang diperoleh dari dataset ini mencakup berbagai aspek kehidupan individu, seperti usia, jenis kelamin, pendidikan, status pernikahan, pekerjaan, dan pendapatan (Becker *et al*., 1996).

Penelitian ini menggunakan data *Census Income* untuk memprediksi pendapatan individu berdasarkan aspek-aspek tertentu seperti pendidikan dan jenis pekerjaan mempengaruhi tingkat pendapatan seseorang. Guna mendukung penelitian ini, dibutuhkan metode klasifikasi *data mining* yakni *Logistic Regression*, *K-Nearest Neighbour*, dan *Naive Bayes*.

Dalam kajian terkini, banyak peneliti yang berfokus pada peningkatan metode analisis data untuk mengidentifikasi faktor-faktor yang paling mempengaruhi distribusi pendapatan. Berdasarkan penelitian yang dilakukan oleh Boyko *et al*. (2021), menyoroti pentingnya penggunaan teknik clustering dan klasifikasi dalam memahami pola distribusi pendapatan dan faktor-faktor yang berkontribusi pada ketimpangan ekonomi. Melalui pendekatan ini, diharapkan bisa memberikan wawasan yang lebih mendalam dan mendukung pengambilan keputusan yang berbasis data.

Melalui analisis ini, diharapkan dapat diketahui bagaimana variabel-variabel tertentu seperti pendidikan dan jenis pekerjaan mempengaruhi tingkat pendapatan seseorang, yang akan membantu merancang kebijakan yang lebih efektif untuk meningkatkan kualitas hidup dan kesetaraan ekonomi.

MASALAH

Dalam penyediaan layanan publik, seringkali ditemukan berbagai permasalahan yang berdampak pada kualitas dan efisiensi layanan yang diberikan kepada masyarakat. Salah satu masalah utama adalah distribusi anggaran yang tidak merata (Chatterjee & Turnovsky, 2012). Hal ini dapat menyebabkan ketimpangan dalam pelayanan publik di berbagai wilayah, terutama di daerah terpencil atau kurang berkembang.

Distribusi anggaran yang tidak adil dapat mengakibatkan kurangnya investasi pada infrastruktur dasar seperti jalan, jembatan, fasilitas kesehatan, dan pendidikan di daerah-daerah tersebut. Akibatnya, masyarakat di wilayah-wilayah tersebut mengalami kesulitan untuk mendapatkan layanan yang setara dengan yang ada di wilayah perkotaan atau daerah yang lebih maju. Ketimpangan ini berdampak negatif pada kualitas hidup dan kesejahteraan masyarakat secara keseluruhan.

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 73-82, 2024
ISSN: 2808-7771 online

Pemerintah memperoleh pemasukan dari berbagai sumber, namun salah satu sumber utama adalah pajak. Pajak adalah kontribusi wajib dari warga negara dan badan usaha kepada negara yang digunakan untuk membiayai berbagai kebutuhan publik, seperti infrastruktur, pendidikan, kesehatan, dan keamanan. Dengan pemerintah yang dapat mengontrol pajak penghasilan dari setiap jenis pekerjaan, diharapkan tercipta kesejahteraan bagi semua golongan pekerja. Pengaturan yang tepat atas pajak penghasilan memungkinkan mereka yang berpenghasilan lebih tinggi untuk secara tidak langsung membantu mereka yang berpenghasilan lebih rendah. Hal ini dilakukan melalui redistribusi pendapatan yang digunakan untuk membangun infrastruktur dasar seperti jalan, jembatan, fasilitas kesehatan, dan pendidikan di daerah-daerah yang kurang berkembang (Boesen, 2023). Dengan demikian, tidak hanya memperbaiki kesejahteraan individu, tetapi juga meningkatkan kualitas hidup dan kesejahteraan masyarakat secara keseluruhan.

Selain itu, sistem perpajakan yang adil dapat mendorong pembangunan yang inklusif dan berkelanjutan, memastikan bahwa semua warga negara memiliki akses yang sama terhadap pelayanan publik dan peluang ekonomi. Hal ini akan menciptakan masyarakat yang lebih sejahtera dan merata, di mana setiap orang memiliki kesempatan yang sama untuk berkembang dan mencapai kesejahteraan yang lebih baik.

METODE PELAKSANAAN

Analisis ini menggunakan klasifikasi dengan model *Logistic Regression, K-Nearest Neighbour,* and *Naive Bayes* untuk menentukan model paling bagus berdasarkan data *Census Income.* Metode klasifikasi dengan akurasi tertinggi akan memberikan prediksi pendapatan seseorang yang lebih akurat.

Cross-industry standard process for data mining

Tahapan penelitian ini dilakukan dengan cara mengadopsi sebuah standar proses *data mining* yang disebut *Cross-Industry Standard Process* for *Data Mining* atau CRISP-D/M. Standar proses tersebut terdiri atas lima fase yaitu *Business Understanding, Data Understanding, Data Preparation, Modeling,* dan *Evaluation.*

Gambar 1. Standar Proses Model CRISP-DM (Sumber: https://www.mygreatlearning.com/blog/why-using-crisp-dm-will-make-you-a-better-data-scientist)

Business Understanding

Census income merupakan data pendapatan yang dikumpulkan dan disusun oleh lembaga pemerintah selama sensus atau survei nasional. Data ini memberikan informasi mengenai distribusi pendapatan dan karakteristik sosio-ekonomi individu dan rumah tangga lainnya di suatu negara. Pemerintah dan pembuat kebijakan dapat bergantung pada data *census income* untuk merumuskan kebijakan yang berkaitan dengan redistribusi pendapatan, program kesejahteraan sosial, pendidikan, ketenagakerjaan, dan perpajakan. Selain itu data ini membantu dalam pengalokasian sumber daya untuk berbagai layanan publik, pembangunan infrastruktur, kesehatan, pendidikan, dan program sosial lainnya. Secara keseluruhan, dengan menentukan metode klasifikasi yang tepat dan mendapatkan akurasi yang tinggi dapat memprediksi pendapatan individu.

Data Understanding

Dataset penelitian merupakan data publik yang diunduh dari *kaggle.com* atau UCI Repository (Asuncion, 2007) tentang *census income*. Data pada *census income* sebanyak 30,162 entri dan memuat 10 atribut variabel prediktor dan 1 atribut variabel target yaitu:

Variabel prediktor:

1. *Age* atau umur individu dan tipe data berupa *integer*.
2. *Work class* atau kelas kerja merupakan data yang berisi *private* (swasta), *self-emp not-inc* (wirausaha tidak terinkorporasi), *local-gov* (pemerintah lokal), *state-gov* (pemerintah negara bagian), *self-emp-inc* (wirausaha dari korporasi), *federal-gov*

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 73-82, 2024
ISSN: 2808-7771 online

(pemerintah federal), *never-worded* (belum pernah bekerja) dan tipe data berupa *object*.

3. *Education* atau pendidikan merupakan data yang berisi *HS-grad* (lulusan SMA), *some-college* (beberapa perguruan tinggi), *bachelors* (sarjana), *masters* (magister), *assoc-voc* (diploma vaksional) dan tipe data berupa *object.*
4. *Marital Status* atau status pernikahan merupakan data yang berisi *married-civ-spouse* (menikah dengan pasangan sipil), *never-married* (belum pernah menikah), *divorced* (bercerai), *seperated* (berpisah), *widowed*, *married-spouse-absent* (menikah tetapi tidak tinggal bersama pasangan), *married-af-spouse* (menikah dengan pasangan militer) dan tipe data berupa *object.*
5. *Occupation* atau pekerjaan *prof-speciality* (spesialisasi profesional), *craft-repair* (perbaikan kerajinan), *exec-managerial* (eksekutif-manajerial), *adm-clerial* (administratif-birokratis), *sales* (pramuniaga/promotor), *other-service* (layanan lainnya), *machine-op-inspct* (operasi mesin-pemeriksaan), *transport-moving* (transportasi-penggerakan), *handlers-cleaners* (penanganan-pembersihan), *farming-fishing* (pertanian-perikanan, *tech-support* (dukungan teknis), *protective-serv* (layanan perlindungan), *priv-house-serv* (pelayanan rumah pribadi), *armed-forces* (angkatan bersenjata) dan tipe data berupa *object*.
6. *Relationship* atau hubungan merupakan data yang berisi *husband* (suami), *not-in-family* (tidak dalam berkeluarga), *own-child* (memiliki anak), *unmarried* (tidak menikah), *wife* (istri), *other-relative* (kerabat lainnya) dan tipe data berupa *object.*
7. *Race* atau ras merupakan data yang berisi white (putih), black (hitam), asian-pac-islander (asia/pasifik), amer-indian-eskimo (indian amerika/eskimo), other (lainnya) dan tipe data berupa *object.*
8. *Gender* atau jenis kelamin merupakan data yang berisi *male* (laki-laki), *female* (perempuan) dan tipe data berupa *object.*
9. *Hours per Week* atau jam kerja per-minggu dan tipe data berupa *integer*.
10. *Native Country* atau negara asal merupakan data yang berisi Amerika Serikat, Meksiko, Filipina, Jerman, Puerto Riko, Kanada, India, El Salvador, Kuba, Inggris, Jamaika, Afrika Selatan, Cina, Italia, Republik Dominika, Vietnam, Guatemala, Jepang, Polandia, Kolombia, Iran, Taiwan, Haiti, Portugal, Nikaragua, Peru, Yunani, Perancis, Ekuador, Irlandia, Hong Kong, Kamboja, Trinidad &

Tobago, Thailand, Laos, Yugoslavia, Wilayah Luar AS (Guam, USVI, dll), Hongaria, Honduras, Skotlandia, Belanda dan tipe data berupa *object.*

Variabel target: *Income* atau pendapatan individu dan tipe data berupa *integer.*

Data Preparation

Pada tahap ini dilakukan pengecekan terhadap data - data agar valid dan akurat serta pengecekan *check missing values* agar dapat diidentifikasi apakah terdapat data yang hilang atau bersifat '?' atau *null* lalu pembersihan. Fase ini dapat jalankan dengan menggunakan *dropna* atau menghapus data dengan missing value sebanyak 7% dari total data. Data yang masih berupa *object* dapat ditransformasi menjadi *integer*.

Langkah selanjutnya adalah membagi data menjadi data *training* dan data *testing*. Kode di bawah digunakan untuk membagi dataset menjadi data latih (*training data*) dan data tes (*test data*) dengan perbandingan 80:20. Random state 1 menentukan seed atau bilangan acak awal untuk memastikan bahwa pembagian dataset bersifat deterministik. Ini memungkinkan hasil pembagian dataset akan sama jika kode dieksekusi ulang.

Pemilihan ambang $50.000 untuk pelabelan klasifikasi didasarkan pada data Census Income, yang secara tradisional menggunakan angka ini untuk memisahkan kelompok pendapatan rendah dan tinggi. Angka $50.000 sering kali dianggap sebagai batas yang signifikan dalam banyak studi ekonomi karena merepresentasikan median pendapatan rumah tangga di Amerika Serikat selama beberapa periode sensus terakhir. Selain itu, ambang ini membantu dalam mengidentifikasi kelompok populasi yang mungkin memerlukan perhatian lebih dalam kebijakan publik dan distribusi sumber daya.

Modeling

Penelitian ini menggunakan metode klasifikasi *Logistic Regression, K-Nearest Neighbors*, dan *Naive Bayes* yang akan dibandingkan tingkat akurasi atau kinerjanya. Pada *KNN* untuk mendapatkan optimal nilai k dapat dilakukan dengan menggunakan akar dari jumlah data. Kemudian metode klasifikasi dengan akurasi paling tinggi akan digunakan sebagai prediksi.

Predict digunakan untuk melakukan prediksi target variabel untuk *train* data yang diberikan menggunakan model yang telah dilatih sebelumnya. Dalam konteks kode yang diberikan, ini akan menghasilkan prediksi untuk *train* data yang diberikan oleh tiga model

yang berbeda (*classifier K-Nearest Neighbors* (*KNN*), model *Logistic Regression*, dan model *Naive Bayes*).

Data Preparation

Untuk mengevaluasi kinerja model dapat dilakukan dengan membuat *Confusion matrix.* Ini merupakan *machine learning* yang dapat memberikan gambaran tentang hasil prediksi model yaitu akurasi, precision f1 score, dan recall. Akurasi adalah proporsi prediksi yang benar dari keseluruhan prediksi. Presisi merupakan proporsi prediksi positif yang benar dari semua prediksi positif yang dibuat oleh model. Recall adalah proporsi prediksi positif yang benar dari semua kasus aktual positif. F1 Score merupakan rata-rata harmonis dari presisi dan recall (Yacouby & Axman, 2020). Ini memberikan ukuran keseimbangan antara presisi dan recall.

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

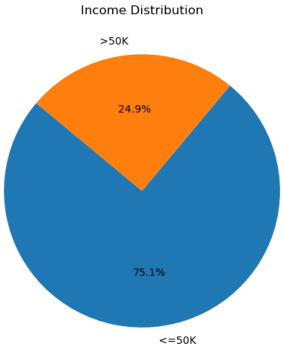
$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ score = 2\ x\ \frac{Precision\ X\ Recall}{Precision + Recall} \quad (4)$$

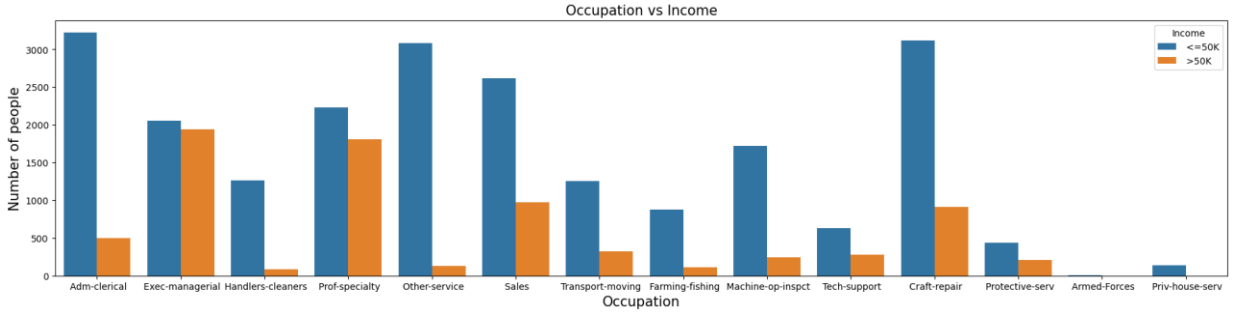
HASIL DAN PEMBAHASAN

Pie chart seperti diilustrasikan pada Gambar 2 merupakan visualisasi dari proporsi pendapatan individu dari data *census income*. Dengan membaginya menjadi 2 kategori yaitu pendapatan lebih dari *$50k* dan pendapatan kurang dari atau sama dengan *$50k*. Dari visualisasi ini, dapat disimpulkan bahwa sebagian besar, yaitu 75.1%, dari data pada dataset *census income* menunjukkan pendapatan kurang dari atau sama dengan *$50k*.



Gambar 2. Jumlah distribusi pendapatan

Grafik batang seperti diilustrasikan pada Gambar 3 menunjukkan perbandingan jumlah orang berdasarkan jenis pekerjaan dan pendapatan mereka, dengan kategori pendapatan dibagi menjadi dua yaitu *≤$50k* dan *>$50k*. Jumlah orang dengan pekerjaan *exec-managerial* tergolong cukup banyak dengan pendapatan *≤$50k* maupun *>$50k*, dengan distribusi yang cukup merata. Pekerjaan *craft-repair* dan *other-service* juga dimiliki oleh banyak orang dengan pendapatan *≤$50k*, tetapi relatif sedikit dengan pendapatan *$50k*. Pekerjaan seperti *handlers-cleaners*, *farming-fishing*, dan *priv-house-serv* memiliki jumlah tenaga kerja yang relatif kecil di kedua kategori pendapatan.



Gambar 3. Visualisasi Occupation vs Income (Pekerjaan vs Pendapatan)

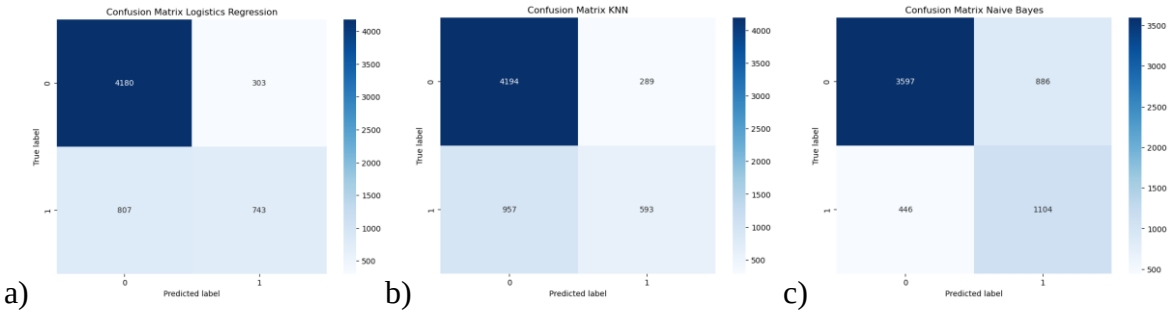
Pemodelan data yang ditunjukkan pada Gambar 4 menggunakan model *Logistic Regression, K Neighbors Classifier, and GaussianNB*. Pada KNN terdapat Parameter metric menentukan cara menghitung jarak antara titik data. Dalam hal ini, jarak Euclidean digunakan. Jarak *Euclidean* adalah jarak linier (garis lurus) antara dua titik dalam ruang multidimensi. Parameter *n_neighbors* menentukan jumlah tetangga terdekat yang akan dipertimbangkan dalam proses klasifikasi. Dalam hal ini, nilai 174 menunjukkan bahwa algoritma akan mempertimbangkan 174 titik data terdekat dari titik yang sedang dievaluasi untuk menentukan kelasnya.

a) `LogisticRegression()` b) `KNeighborsClassifier(metric='euclidean', n_neighbors=174)` c) `GaussianNB()`

Gambar. 4. (a) Model Logistic Regression; (b) Model KNN; (c) Model Naive Bayes

Predict digunakan untuk melakukan prediksi target variabel untuk *train* data yang diberikan menggunakan model yang telah dilatih sebelumnya. Dalam konteks kode yang diberikan, ini akan menghasilkan prediksi untuk *train* data yang diberikan oleh tiga model yang berbeda (*classifier K-Nearest Neighbors* (*KNN*), model *Logistic Regression*, dan model *Naive Bayes*). Kemudian untuk mengevaluasi hasil klasifikasi dapat menggunakan confusion matrix seperti yang diilustrasikan pada Gambar 5.

ISSN: 2808-7771 online



a) b) c)

Gambar. 5. (a) *Confusion Matrix Logistic Regression*; (b) *Confusion Matrix* KNN; (c) *Confusion Matrix Naive Bayes*

Setelah menggunakan rumus akurasi didapatkan hasil untuk *Logistic Regression* memiliki akurasi sebesar 81%, *KNN* memiliki akurasi sebesar 79%, dan *Naive Bayes* memiliki akurasi sebesar 77%. Dari hasil ini, dapat disimpulkan bahwa metode klasifikasi *Logistic Regression* memiliki akurasi tertinggi dibandingkan dengan kedua metode lainnya, yaitu *KNN* dan *Naive Bayes*. *Logistic Regression* merupakan variabel dependen dan bersifat biner atau memiliki dua kemungkinan saja (Sperandei, 2014). Hal ini menunjukkan bahwa *Logistic Regression* adalah model yang paling baik dalam memprediksi target variable dalam dataset ini.

KESIMPULAN

Berdasarkan penelitian yang telah kami lakukan, dari 3 metode klasifikasi yang telah kami lakukan yaitu *Logistic Regression, K-Nearest Neighbors (KNN), dan Naive Bayes* dapat disimpulkan bahwa model *Logistic Regression* mencapai akurasi tertinggi di antara ketiga model, menunjukkan kinerjanya yang bagus dalam memprediksi tingkat pendapatan dalam dataset in dengan akurasi sebesar 81%. Presisi untuk kelas 0 adalah 0.936, yang menunjukkan bahwa 93.6% dari semua prediksi negatif adalah benar-benar negatif. Sementara itu, presisi untuk kelas 1 adalah 0.710, yang berarti 71% dari semua prediksi positif adalah benar-benar positif. Presisi yang lebih tinggi pada kelas 0 dibandingkan kelas 1 menunjukkan bahwa model lebih akurat dalam memprediksi kelas negatif. Recall untuk kelas 0 adalah 0.836, yang berarti model ini mampu mengidentifikasi 83.6% dari semua kasus negatif yang sebenarnya. Di sisi lain, recall untuk kelas 1 adalah 0.471, yang menunjukkan bahwa model ini hanya mampu mengidentifikasi 47.1% dari semua kasus positif yang sebenarnya. F1 Score adalah metrik yang menggabungkan presisi dan recall menjadi satu nilai tunggal. Untuk kelas 0, f1 score adalah 0.882, menunjukkan

keseimbangan yang baik antara presisi dan recall. Namun, untuk kelas 1, f1 score adalah 0.567. Temuan dari penelitian ini diharapkan dapat meningkatkan keandalan prediksi, mengidentifikasi faktor penentu pendapatan, dan membuat keputusan bisnis yang lebih baik berdasarkan data pendapatan yang akurat.

DAFTAR PUSTAKA

Asuncion, A. (2007). *UCI Machine Learning Repository*. UC Irvine Machine Learning Repository. http://ci.nii.ac.jp/naid/20001247967

Boesen, U. (2023, July 24). State infrastructure spending & state infrastructure revenue. *Tax Foundation*. https://taxfoundation.org/data/all/state/state-infrastructure-spending/

Boyko, N., Hetman, S., & Kots, I. (2021). Comparison of clustering algorithms for revenue and cost analysis. *COLINS*, 1866–1877. http://ceur-ws.org/Vol-2870/paper136.pdf

Chakrabart, N., & Biswas, S. (2018). A statistical approach to adult Census income level prediction. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. https://doi.org/10.1109/icacccn.2018.8748528

Chatterjee, S., & Turnovsky, S. J. (2012). Infrastructure and inequality. *European Economic Review*, 56(8), 1730–1745. https://doi.org/10.1016/j.euroecorev.2012.08.003

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 24(1), 12–18. https://doi.org/10.11613/bm.2014.003

Yacouby, R., & Axman, D. (2020). Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. *ACL Anthology*. https://doi.org/10.18653/v1/2020.eval4nlp-1.9

© 2024 by authors. Content on this article is licensed under a Creative Commons Attribution 4.0 International license. (http://creativecommons.org/licenses/by/4.0/).