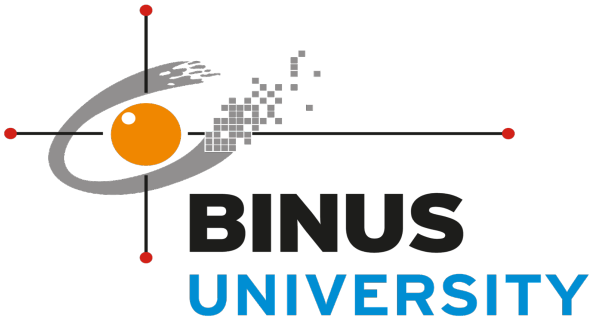
Assurance of Learning

DATA MINING



Dosen Pembimbing:

D6198 – FEPRI PUTRA PANGHURIAN, S.Kom, M.T.I.

Disusun Oleh:

Alfonsus Adrian – 2540118290

Devy Ivana Tati – 2540096371

Eric Sunjaya – 2540122451

Kurniadi – 2540128410

Valentio Stanley Gunadi – 2540096011

Vivian Agatha Andrea – 2540125560

PROGRAM STUDI COMPUTER SCIENCE

FAKULTAS COMPUTER SCIENCE

UNIVERSITAS BINA NUSANTARA

2023

DAFTAR ISI

BAB I

PENDAHULUAN

1.1. Latar Belakang

Perkembangan suatu kota mendorong masyarakat untuk bermigrasi dan mencari peluang di kawasan urban, menciptakan fenomena urbanisasi. Dampak dari fenomena ini termasuk peningkatan jumlah penduduk di kota tersebut.

Salah satu kota yang menjadi tujuan migrasi adalah Jakarta. Sebagaimana dilaporkan dalam berita detikfinance, "Jakarta merupakan salah satu kota terpadat penduduk di Indonesia. Selain berfungsi sebagai pusat pemerintahan, Jakarta juga menjadi pusat bisnis yang menarik para pencari kerja untuk datang ke kota ini [1]."

Menurut data Badan Pusat Statistik (BPS), pada tahun 2022, tingkat kepadatan penduduk di Ibu Kota telah mencapai 16.158 orang per km² [2][3]. Kepadatan penduduk ini juga menyebabkan semakin menyempitnya lahan dari waktu ke waktu. Kendala terbatasnya lahan ini berperan signifikan terhadap kenaikan nilai tanah di wilayah tersebut, yang menjadi hambatan bagi banyak orang dalam membeli tanah.

Salah satu solusi untuk mengatasi permasalahan ini adalah dengan menyewa kamar di hunian vertikal atau rumah, karena harganya lebih terjangkau dan lebih mudah dibandingkan mencari lahan untuk dibeli secara pribadi [1]. Selain itu, menyewa rumah juga memberikan fleksibilitas yang lebih besar bagi masyarakat yang sering bermigrasi, terutama jika mereka belum siap untuk tinggal dalam jangka panjang atau harus sering berpindah tempat tinggal.

Dalam proyek ini, kelompok kami mengembangkan sebuah program menggunakan bahasa Python untuk mengotomatisasi penilaian kelayakan peminjaman uang guna menyewa hunian. Program ini mempertimbangkan informasi pemohon, seperti tingkat pendidikan, pendapatan, jumlah pinjaman, jangka waktu pinjaman, riwayat kredit, dan tanggungan pemohon. Harapannya, proyek ini dapat membantu pihak pemberi pinjaman dalam seleksi pemohon pinjaman untuk menyewa hunian, sehingga dana pinjaman diberikan hanya kepada mereka yang memenuhi kriteria yang tepat.

1.2. Tujuan

Berikut merupakan tujuan dari program kita:

1. Memastikan penilaian kelayakan pemohon dilakukan secara konsisten, objektif, dan mengeliminasi pengaruh kesalahan atau kelalaian manusia.
2. Menyediakan dasar yang kuat untuk mengambil keputusan pemberian pinjaman kepada pemohon berdasarkan faktor yang ditetapkan dan informasi yang diberikan.
3. Meningkatkan efisiensi bisnis karena pekerja dimudahkan dalam menilai pemohon yang sesuai dan dapat berfokus untuk melakukan pekerjaan lainnya.

1.3. Manfaat

Berikut merupakan manfaat dari program kita:

1. Memperpendek waktu penilaian pemohon secara manual guna meningkatkan efisiensi waktu dan biaya.
2. Mengurangi risiko pemberian pinjaman kepada pemohon yang tidak memenuhi kriteria tertentu, sehingga meningkatkan keamanan penyaluran dana.
3. Menjamin keputusan pemberian pinjaman yang akurat, konsisten, dan objektif.

BAB III
METODE PENELITIAN

3.1. Logistic Regression

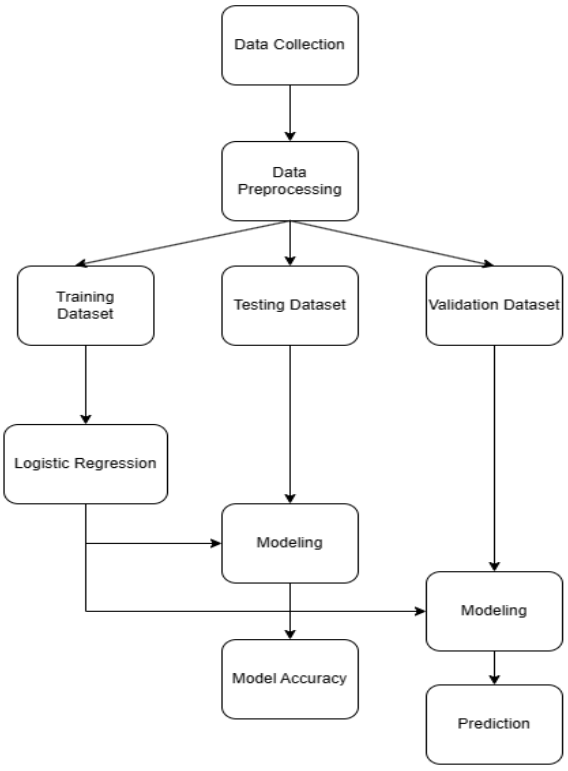
Logistic Regression adalah model statistikal yang sering digunakan untuk melakukan klasifikasi dan analisis prediksi. Logistic regression mengestimasi probabilitas dari suatu kejadian berdasarkan kumpulan data variabel independen tertentu. Logistic regression memiliki komponen integral lainnya seperti decision boundary, sigmoid/logistic function, cost function, dan algoritma gradient descent. Decision boundary adalah sebuah garis atau batas yang digunakan untuk membagi kelas-kelas yang diprediksi dalam ruang fitur. Sigmoid/logistic function digunakan untuk mengklasifikasikan data point berdasarkan dengan input dari masing-masing feature/atribut.

Rumus dari sigmoid function adalah $F(z) = \frac{1}{1+e^{-z}}$ dimana z adalah kombinasi linear dari masing-masing input feature dan bobotnya. Apabila dijabarkan secara matematis kombinasi linear z akan berbentuk seperti ini $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$, dimana β adalah nilai bobot dari masing-masing feature sedangkan x adalah nilai input dari masing-masing instance/baris. Cost function adalah sebuah fungsi yang digunakan untuk menghitung perbedaan antara probabilitas yang diprediksi dengan true class labels dari training data. Rumus cost function adalah $-\frac{1}{m}\Sigma_{i=1}^{m}\left[y_i log(p_i) + (1 - y_i) log(1 - p_i) b\right]$, dimana $y_i$merupakan true class label (0 atau 1) untuk instance ke-i dan p_i merupakan probabilitas yang diprediksi.

Untuk mencari nilai bobot dari suatu feature dapat menggunakan algoritma gradient descent. Gradient descent adalah algoritma yang digunakan untuk meminimalisir cost function secara iteratif. Rumus dari gradient descent adalah $\beta_i = \beta_i - \alpha\frac{\partial Cost}{\partial \beta_i}$, dimana β_i adalah nilai bobot suatu feature saat ini. α adalah learning rate, sebuah hyperparameter yang menentukan besaran langkah yang akan diambil pada tiap iterasi. $\frac{\partial Cost}{\partial \beta_i}$ merupakan turunan parsial dari cost function terhadap bobot β_i. $\frac{\partial Cost}{\partial \beta_i}$ dapat

dijabarkan kembali menjadi $\Sigma_{j=1}^{m}(h_{\beta}(x^{(j)}) - y^{(j)})x_i^{(j)}$, dimana m adalah jumlah dari training examples, $h_{\beta}(x^{(j)})$ adalah probabilitas yang diprediksi pada training example ke j, $y^{(j)}$ adalah label true class untuk training example ke j, dan $x_i^{(j)}$ adalah fitur/atribut ke i dari training example ke j.

3.2. Alur Diagram



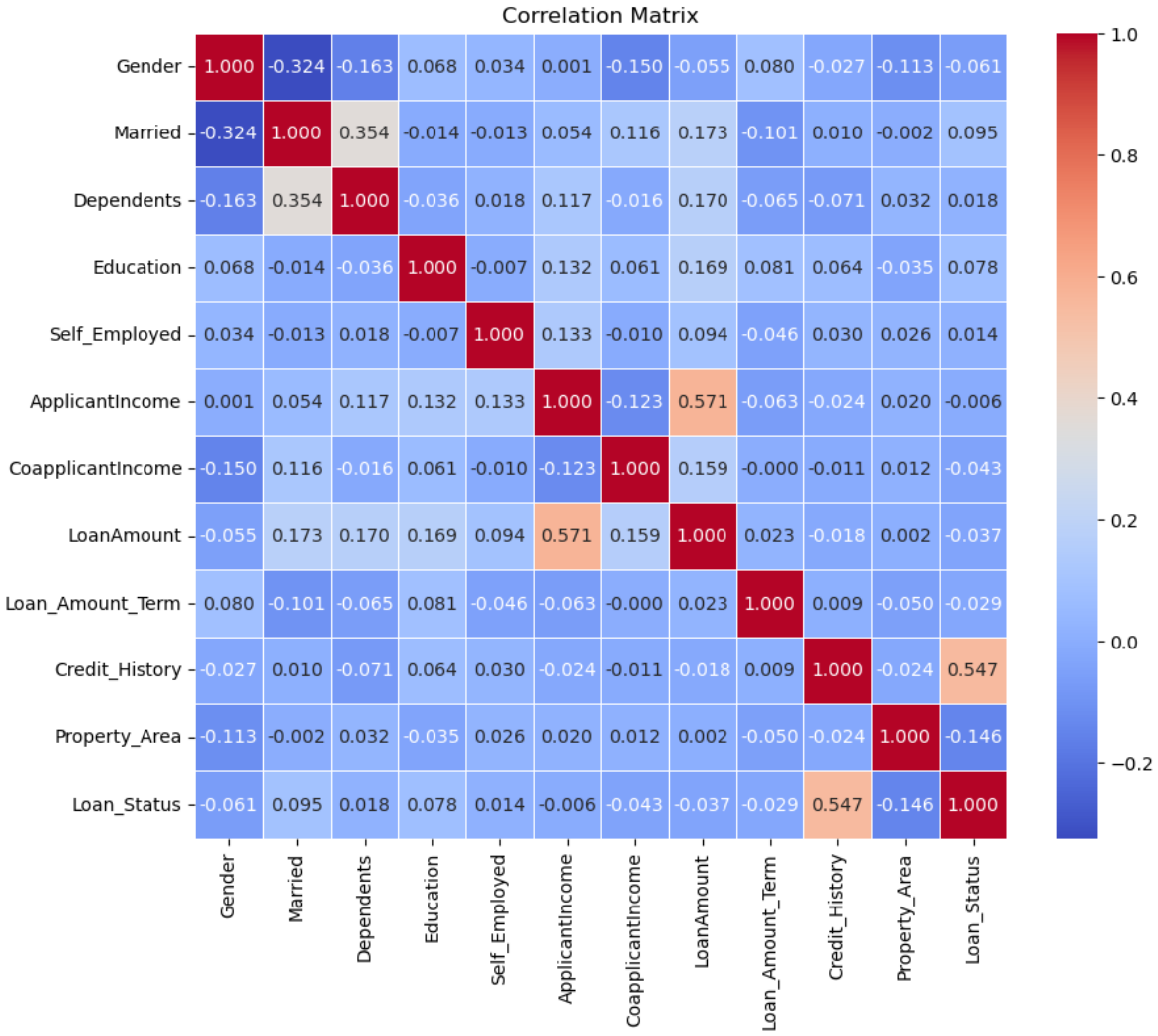
Gambar 3.1 Alur diagram

Gambar alur diagram 1.1 menyatakan bahwa memiliki beberapa proses yang harus dilalui untuk mencapai prediksi yang diinginkan. Bermulai dari *data collection, data preprocessing,* dan dilanjutkan dengan *training, testing,* dan *validation*. *Training* akan menggunakan algoritma Logistic Regression yang kemudian akan mencari akurasi dengan data *test* dan mendapatkan prediksi dengan menggunakan data validasi atau *real-life*.

BAB IV

PEMBAHASAN

Gambar alur diagram 3.1 memberikan alur singkat yang menjelaskan tentang langkah-langkah yang dilakukan untuk membuat klasifikasi dan prediksi dengan menggunakan logistic regression. Langkah pertama yang harus dilakukan adalah menyiapkan datasets yang akan digunakan. Kami menggunakan dataset "Home Loan Approval" dari situs Kaggle. Dataset ini mengandung 12 features yang merupakan detail informasi yang diisi oleh pemohon kredit saat mengisi formulir peminjaman kredit rumah. Rincian atributnya adalah *Loan_ID, Gender*, *Married*, *Education*, *Dependents*, *Self_Employed, ApplicantIncome*, *CoapplicantIncome, LoanAmount*, *Loan_Amount_Term, Credit_History, Property_Area, Loan_Status.*



Gambar 4.1 Correlation Matrix

Berdasarkan correlation matrix diatas, atribut yang akan digunakan adalah *Education, Credit_History,* dan *Married* (total tiga atribut). Selain itu, kita akan melakukan prediksi dengan menggunakan lima atribut yang memiliki nilai korelasi tertinggi (positif), yaitu *Education, Credit_History, Married, Dependents,* dan *Self_Employed.* Untuk mengetahui lebih rinci mengenai atribut yang dipakai dapat membaca tabel 4.1.

Nama Atribut	Deskripsi	Tipe
Education	Level edukasi	*Categorical* (*Graduate*/*Not Graduate*)
Married	Status pernikahan	*Categorical* (Y/N)
Credit_History	Riwayat peminjaman kredit	*Categorical* (0/1)
Dependents	Jumlah tanggungan dalam keluarga tidak termasuk dirinya	*Categorical* (0/1/2/3+)
Self_Employed	Wiraswasta	*Categorical* (Y/N)

Tabel 4.1 Kamus data yang akan dipakai

Selanjutnya data yang akan kita gunakan perlu melalui proses preprocessing yang mencakup import dan read dataset yang sudah kita kumpulkan, konversi variabel kategorik menjadi numerik, menghilangkan null/missing value, membuat correlation matrix, dan melakukan feature selection. Akan dilakukan pembagian data dari dataframe df_train, di mana beberapa kolom dijadikan fitur (x) dan kolom "Loan_Status" dijadikan target (y). Kemudian, dilakukan standarisasi fitur menggunakan StandardScaler, yang mengurangi nilai data dengan mean dan membaginya dengan standar deviasi untuk memperkecil rentang nilainya. Setelah itu, dilakukan pemodelan dengan membuat model regresi logistik pada data pelatihan yang telah diskalakan. Selanjutnya, dilakukan prediksi pada data pengujian, dan hasilnya dimasukkan ke dalam dataframe x_test sebagai kolom baru. Setelah itu, dibuat confusion matrix untuk menggambarkan kinerja model klasifikasi, yang mencakup informasi tentang true positives (TP), false positives (FP), true negatives (TN), dan false negatives (FN) pada dataset. Selanjutnya, dilakukan pencetakan akurasi sebagai ukuran seberapa baik model yang dibuat dalam melakukan prediksi.

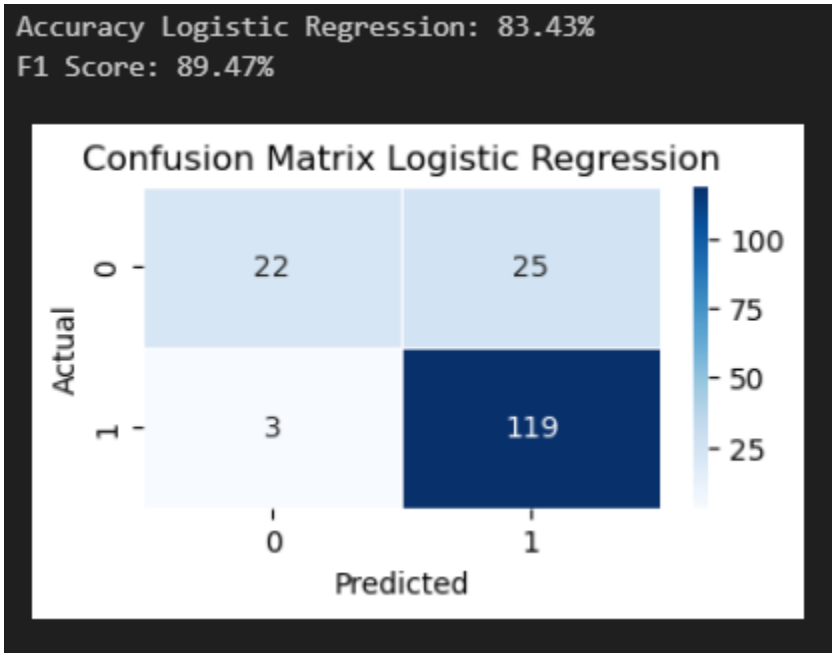
Dilakukan juga prediksi probabilitas kelas positif pada pengujian, diikuti dengan perhitungan kurva presisi-recall beserta area di bawah kurva tersebut.

Setelah proses pelatihan model selesai, model diaplikasikan ke kasus real life dari loan_sanction_test.csv yang disimpan sebagai df_test. Yang dimana akan menghasilkan kolom baru bernama “prediction” yang berisikan prediksi apakah permohonan peminjaman kredit dapat dilakukan apa tidak.

BAB V

PENUTUP

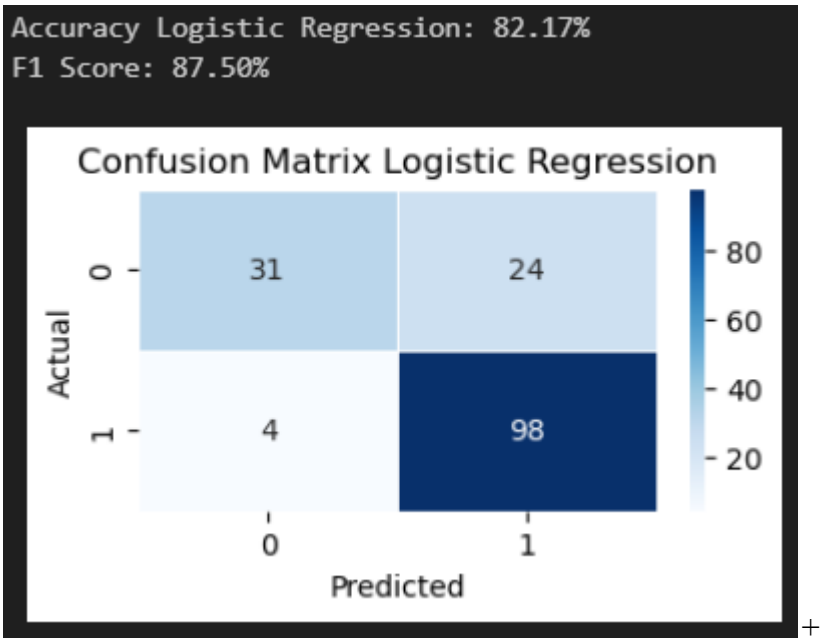
Dataset “Home Loan Approval” dapat digunakan sebagai dasar untuk mengambil keputusan kelayakan pemberian pinjaman setelah melalui algoritma Logistic Regression berdasarkan tiga atribut dan lima atribut. Berdasarkan gambar 5.1, akurasi dengan menggunakan tiga atribut dengan korelasi tertinggi dan algoritma Logistic Regression adalah 83.43% yang dimana F1 skornya adalah 89.47%. F1 score adalah cara untuk mengevaluasi klasifikasi yang dapat didefinisikan sebagai rata-rata harmonis dari presisi dan recall. F1 score merupakan ukuran statistik untuk menguji keakuratan model. Dengan begitu, semakin tinggi nilai F1 maka keakuratan model juga semakin terjamin. Confusion matrixnya menyatakan bahwa hasil prediksi nilai “1” yang nilai sebenarnya “1” (True Positive) sejumlah 119 data dan prediksi “0” yang nilai sebenarnya “1” (False Positive) sejumlah 3 data. Sedangkan, untuk prediksi “0” yang nilai sebenarnya “0” (True Negative) sejumlah 22 data dan prediksi “1” yang nilai sebenarnya “0” (False Positive) sejumlah 25 data. Dari confusion matrix tersebut dapat ditentukan akurasinya dari penjumlahan True Positive dan True Negative yang kemudian dibagi dengan total keseluruhan. Hasil akurasi prediksi pada model ini adalah 83.43%



Gambar 5.1 Akurasi, F1 Score, dan Confusion Matrix dengan menggunakan algoritma Logistic Regression dan tiga atribut yang memiliki nilai korelasi tertinggi

Berdasarkan gambar 5.2, dengan menggunakan lima atribut dengan korelasi tertinggi dan algoritma Logistic Regression menghasilkan nilai akurasi sebesar 82.17% yang dimana F1 skornya adalah

87.50%. Confusion matrixnya menyatakan bahwa hasil prediksi nilai “1” yang nilai sebenarnya “1” (True Positive) sejumlah 98 data dan prediksi “0” yang nilai sebenarnya “1” (False Positive) sejumlah 4 data. Sedangkan, untuk prediksi “0” yang nilai sebenarnya “0” (True Negative) sejumlah 31 data dan prediksi “1” yang nilai sebenarnya “0” (False Positive) sejumlah 24 data. Dari confusion matrix tersebut dapat ditentukan akurasinya dari penjumlahan True Positive dan True Negative yang kemudian dibagi dengan total keseluruhan. Hasil akurasi pada model dengan 5 atribut ini adalah 82.17%

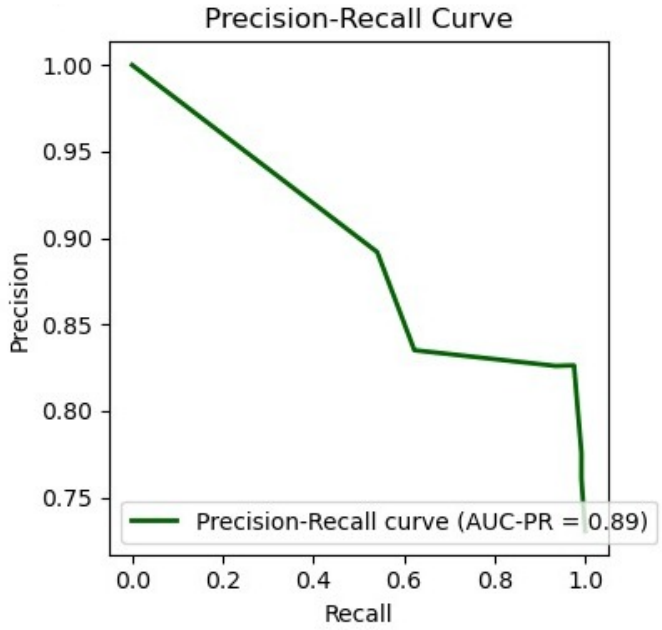


Gambar 5.2 Akurasi, F1 Score, dan Confusion Matrix dengan menggunakan algoritma Logistic Regression dan lima atribut yang memiliki nilai korelasi positif

Selain menggunakan f1 score dan confusion matrix dalam mengevaluasi model yang kami buat, kami juga menggunakan metode lainnya seperti Precision-Recall Curve. Kedua metode ini cocok untuk mengevaluasi model binary classification. Precision-Recall Curve adalah suatu metrik yang digunakan untuk mengukur tradeoff antara precision dan juga recall untuk threshold berbeda pada suatu model binary classification.

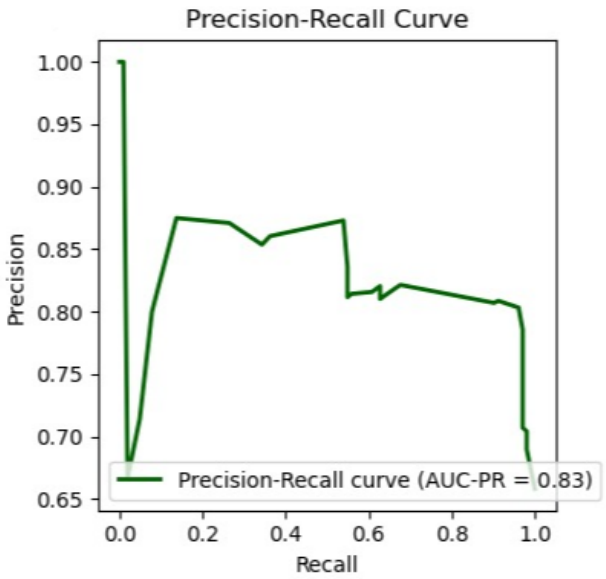
Kami menggunakan metode ini karena dataset yang kami gunakan imbalance. Total kelas positif yang terdapat pada dataset training kami 69% sedangkan kelas negatif hanya 31% sehingga metode seperti AUC-ROC kurang cocok untuk digunakan. Precision-Recall Curve dibuat dengan menghitung nilai Precision dan Recall pada masing-masing threshold. Precision adalah proporsi dari klasifikasi positif yang benar (True Positive) dibagi dengan total prediksi klasifikasi positif yang dibuat (True Positive + False Positive). Recall adalah proporsi klasifikasi positif yang benar (True Positive) dibagi dengan total klasifikasi yang benar-benar positif (True Positive + False Negative).

Area Under the Curve adalah suatu angka yang merepresentasikan seberapa baik performa suatu model dengan mengintegrasikan tradeoff antara precision dan recall di berbagai thresholds. Angka ini didapat dengan menjumlahkan seluruh luas wilayah dibawah kurva. Performa suatu model dapat dikatakan baik apabila nilai AUC-PR nya lebih dari 0.5 dan mendekati 1. Apabila suatu model memiliki nilai AUC-PR sama dengan 0.5, maka performa dari model tersebut sama saja dengan random classifier/random chance. Berdasarkan gambar 5.3 kita dapat melihat bahwa nilai dari AUC-PR nya adalah 0.89 yang berarti model yang menggunakan 3 feature/atribut memiliki performa yang baik.



Gambar 5.3 Precision-Recall Curve tiga atribut

Berdasarkan gambar 5.4, kita dapat melihat bahwa nilai AUC-PR dari model yang menggunakan 5 feature/atribut adalah 0.83 yang berarti walaupun model ini memiliki performa yang cukup baik, tetapi tidak sebaik model yang menggunakan 3 feature/atribut saja.



Gambar 5.4 Precision-Recall Curve lima atribut

Berdasarkan hasil yang ditampilkan, algoritma yang menggunakan tiga atribut akan menghasilkan akurasi yang lebih besar dibandingkan lima atribut, namun tidak signifikan. Hal ini dapat disebabkan karena meminimalisirkan adanya *overfitting* yang dimana *noise* atau data yang tidak relevan akan cenderung lebih sedikit. Semakin *simple* data yang digunakan maka akan mengurangi waktu dan parameter atau jumlah iterasi yang digunakan sehingga untuk *training* data akan menjadi lebih akurat dibandingkan data yang kompleks. Setelah melalui tahap *pre-processing* sebelumnya akan lebih meyakinkan bahwa minimnya data *na* yang lolos, namun jika datanya besar dan kompleks, akan tetap memiliki peluang data *na* yang lolos. Dengan begitu, data yang kurang relevan akan muncul dan mengganggu nilai akurasi data tersebut.

DAFTAR PUSTAKA

[1] Fadilah, Ilyas. “Lahan Kosong Makin Sempit, Wagub DKI: Hunian Harus Vertikal, Tidak Ada Pilihan.” Detikfinance, finance.detik.com/properti/d-6279901/lahan-kosong-makin-sempit-wagub-dki-hunian-harus-vertikal-tidak-ada-pilihan. Accessed 15 Dec. 2023.

[2] Ahdiat, Adi. “Penduduk DKI Bertambah Seribu Orang per Km Persegi Dalam 10 Tahun | Databoks.” Databoks.katadata.co.id, databoks.katadata.co.id/datapublish/2023/03/03/penduduk-dki-bertambah-seribu-orang-per-km-persegi-dalam-10-tahun#:~:text=Menurut%20data%20Badan%20Pusat%20Statistik. Accessed 15 Dec. 2023.

[3] “BPS Provinsi DKI Jakarta.” Bps.go.id, 2019, jakarta.bps.go.id/indicator/12/124/1/3-1-1-penduduk-laju-pertumbuhan-penduduk-distribusi-persentase-penduduk-kepadatan-penduduk-rasio-jenis-kelamin-penduduk-menurut-provinsi-kabupaten-kota-kecamatan.html.