Proposal Stage 0

Agi Rahmawandi as Data Scientist
Shan Ramadhan as Data Analyst
Muhammad Muqorrobin as Business Analyst
I Gusti Ngurah Agung Hari Vijaya Kusuma as PM/DE

August 29, 2025

Submission Links

- Repository: github.com/4Kings-Rakamin
- ManPro & Timeline: GoogleSheet_Timeline

1 Pendahuluan

Dalam perekrutan karyawan, pengambilan keputusan yang tepat, cepat, efektif serta efisien sangat penting untuk memastikan bahwa perusahaan mendapatkan kandidat yang paling sesuai dengan kebutuhan dan budaya organisasi. Dengan kemajuan teknologi dan analisis data, perusahaan kini dapat memanfaatkan data historis yang diolah menjadi model prediksi untuk meningkatkan proses perekrutan mereka lebih baik lagi.

Project ini disusun dengan pendekatan data-driven strategy menggunakan machine learning untuk membantu perusahaan memprediksi pelamar yang potensial. Dengan model prediksi ini, perusahaan dapat melakukan pengambilan keputusan secara cepat dan tepat tanpa adanya bias subjektif.

1.1 Problem Statement

Proses rekrutmen, perusahaan kerap menghadapi berbagai tantangan, mulai dari tingginya jumlah pelamar, keterbatasan waktu untuk melakukan penilaian, hingga adanya subjektivitas yang dapat memengaruhi konsistensi keputusan.

Menurut CareerBuilder, hampir tiga perempat perusahaan yang melakukan rekrutmen yang buruk melaporkan rata-rata kerugian biaya sebesar USD 14.900. Selain itu, 74% pengusaha menyatakan pernah merekrut orang yang tidak tepat untuk suatu posisi. Kondisi ini sering kali disebabkan oleh human error dalam proses seleksi, terutama karena banyaknya kandidat yang harus disaring serta adanya tenggat waktu yang ketat, sehingga memperbesar peluang terjadinya kesalahan. (Swita, 2023)

Dalam dunia kerja yang serba cepat, kecepatan menjadi salah satu faktor kunci bagi perusahaan untuk mempertahankan keunggulan kompetitif. Namun, masih banyak perusahaan yang menjalankan proses rekrutmen secara manual, sehingga proses perekrutan menjadi kurang efisien. Oleh karena itu, dibutuhkan solusi berbasis kecerdasan buatan (AI) yang mampu mempercepat proses seleksi, mengurangi beban operasional, sekaligus meningkatkan kualitas keputusan rekrutmen.

Kompleksitas parameter penilaian seperti kompetensi, keterampilan, serta faktor demografi semakin menambah risiko perusahaan gagal menemukan kandidat terbaik. Untuk itu, penelitian ini memanfaatkan dataset berisi 1.500 data historis kandidat (recruitment data.csv), yang mencakup informasi usia, gender, tingkat pendidikan, pengalaman kerja, jumlah perusahaan sebelumnya, jarak tempat tinggal dari kantor, skor wawancara, skor keterampilan, skor kepribadian, strategi rekrutmen, hingga keputusan akhir perekrutan (Hiring Decision) dan mengasilkan model prediksi yang akurat.

1.2 Goals, Objectives, and Business Metrics

Tujuan utama dari proyek ini adalah membangun model machine learning yang dapat membantu proses pengambilan keputusan perekrutan karyawan secara lebih cepat, objektif, dan konsisten. Model ini diharapkan mampu memberikan rekomendasi kandidat potensial serta mengurangi bias subjektif dalam proses seleksi.

Objectives:

1. Mengembangkan model prediktif berdasarkan data historis recruitment untuk mengklasifikasikan kandidat apakah diterima atau tidak
2. Mempercepat proses identifikasi kandidat dan mengurangi bias subjektif dalam mengambil keputusan.
3. Menghindari salah rekrut karyawan yang berpotensi merugikan peruhahaan.
4. Menurunkan biaya rekrutmen agar lebih efesien

Business Metrics:

1. **Akurasi Model:** Akurasi $\geq 86\%$ untuk memastikan keputusan perekrutan yang lebih tepat.
2. **Precision:** Precision $\geq 80\%$, agar recruiter dapat mengurangi jumlah kandidat "false positive" yang tidak sesuai.
3. **AUC:** Nilai AUC $\geq 85\%$ untuk menjamin kestabilan performa model pada berbagai threshold dalam proses shortlisting kandidat.
4. **Efisiensi Waktu:** Mengurangi waktu screening kandidat sebesar 60% dalam waktu 3 bulan setelah model prediksi mulai digunakan.
5. **Reduksi Biaya:** Menurunkan biaya operasional terkait dengan proses perekrutan sebesar 15% dalam waktu 3 bulan setelah model prediksi mulai digunakan.

1.3 Gap Analysis

Saat ini belum tersedia sistem prediksi keputusan perekrutan berbasis data yang cepat dan objektif. Proses pengambilan keputusan HR masih didominasi oleh penilaian subjektif, membutuhkan waktu yang relatif lama, serta tidak memberikan insight yang jelas terkait faktor-faktor yang memengaruhi keputusan, seperti kompetensi, keterampilan, dan aspek demografi kandidat. Kondisi ini berpotensi menurunkan konsistensi dan efektivitas dalam proses seleksi karyawan baru.

Dengan adanya model machine learning yang diusulkan, diharapkan dapat mengisi gap ini dengan menyediakan alat bantu yang mampu menganalisis data historis secara cepat dan akurat. Selain itu EDA juga dilakukan untuk menemukan aspek aspek apa saja yang berpengaruh besar dalam menghasilkan keputusan. Dimana outputnya yaitu Insight dan Model machine learning ini akan membantu HR dalam mengidentifikasi kandidat potensial berdasarkan fitur-fitur yang relevan, sehingga mengurangi ketergantungan pada penilaian subjektif dan mempercepat proses seleksi. Selain itu, model ini juga akan memberikan wawasan tentang faktor-faktor kunci yang memengaruhi keputusan perekrutan, sehingga perusahaan dapat mengoptimalkan strategi rekrutmen mereka ke depannya.

1.4 Ideal Condition & Expected Impact

Sistem mampu mengidentifikasi kandidat potensial secara cepat, objektif, dan tepat sasaran melalui pemetaan kandidat yang layak direkrut. Dampaknya, proses pengambilan keputusan menjadi lebih efisien sehingga HR dapat memfokuskan waktu pada kandidat yang benar-benar potensial dan sesuai kebutuhan perusahaan.

2 Tinjauan Pustaka

2.1 Proses Rekrutmen

Proses rekrutmen adalah serangkaian langkah yang diambil oleh perusahaan untuk menarik, menilai, dan memilih kandidat yang paling sesuai untuk mengisi posisi yang tersedia. Proses ini biasanya dimulai dengan identifikasi kebutuhan tenaga kerja, diikuti oleh pencarian kandidat melalui berbagai saluran seperti iklan lowongan kerja, agen perekrutan, dan media sosial. Setelah kandidat ditemukan, mereka akan melalui tahap seleksi yang melibatkan penilaian kualifikasi, wawancara, dan tes keterampilan. Akhirnya, keputusan perekrutan dibuat berdasarkan evaluasi menyeluruh dari semua kandidat yang telah melalui proses seleksi. Proses rekrutmen yang efektif tidak hanya memastikan bahwa perusahaan mendapatkan karyawan yang berkualitas, tetapi juga membantu dalam membangun budaya organisasi yang positif dan mendukung tujuan bisnis jangka panjang. (Mathis et al., 2017)

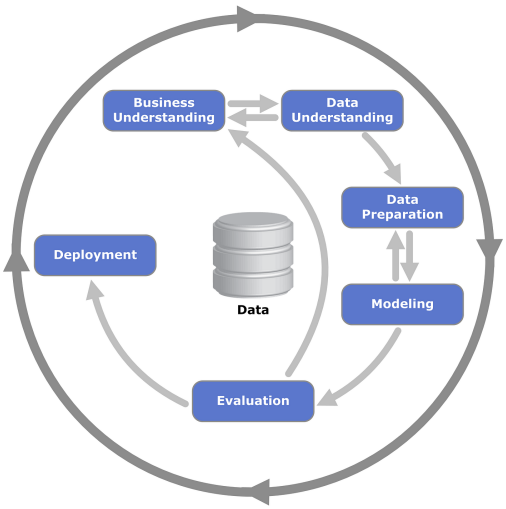
Adapun beberapa faktor yang mempengaruhi keputusan perekrutan meliputi:

1. Demografis: usia, gender, latar belakang pendidikan. (Ng & Burke, 2005)
2. Pengalaman kerja: jumlah tahun pengalaman dan variasi perusahaan sebelumnya. (Ployhart, 2006)
3. Kompetensi teknis dan soft skills: hasil tes keterampilan (skill score), wawancara, dan penilaian kepribadian. (Schmidt & Hunter, 1998)
4. Faktor eksternal: jarak tempat tinggal ke kantor sering menjadi pertimbangan dalam retensi. (Hausknecht et al., 2009)
5. Strategi rekrutmen: pendekatan organisasi (job fairs, rekrutmen online, campus hiring) dapat memengaruhi kualitas kandidat. (Breaugh, 2013).

Berdasarkan studi literatur, faktor-faktor di atas secara signifikan mempengaruhi keputusan perekrutan. Dimana hal tersebut dapat menjadi pertimbangan penting dalam pengembangan model prediktif untuk proses perekrutan.

2.2 CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) adalah metodologi standar yang digunakan dalam proyek data mining dan analisis data. Metodologi ini terdiri dari enam fase utama yang membantu dalam mengorganisir dan mengelola proyek data secara sistematis. (Chumbar, 2020)



Gambar 1: Alur Kerja CRISP-DM

Gambar 1 menggambarkan alur kerja CRISP-DM yang terdiri dari enam fase utama, yaitu:

1. Business Understanding: Memahami tujuan bisnis dan kebutuhan proyek.
2. Data Understanding: Mengumpulkan dan memahami data yang tersedia.
3. Data Preparation: Membersihkan dan mempersiapkan data untuk analisis.
4. Modeling: Membangun model prediktif menggunakan teknik machine learning.
5. Evaluation: Mengevaluasi model untuk memastikan bahwa tujuan bisnis tercapai.
6. Deployment: Menerapkan model dalam lingkungan produksi untuk digunakan dalam pengambilan keputusan bisnis.

2.3 EDA (Exploratory Data Analysis)

Exploratory Data Analysis (EDA) adalah proses awal dalam analisis data yang bertujuan untuk memahami karakteristik dan pola dalam dataset. EDA melibatkan berbagai teknik visualisasi dan statistik untuk mengeksplorasi data, mengidentifikasi outlier, dan menemukan hubungan antara variabel. Proses ini sangat penting karena membantu dalam mengarahkan langkah-langkah selanjutnya dalam analisis data, seperti pemilihan fitur dan pemodelan. (Tukey, 1977)

2.4 T-Test

T-Test adalah metode statistik yang digunakan untuk membandingkan rata-rata dari dua kelompok data. Uji ini membantu menentukan apakah perbedaan antara kedua kelompok tersebut signifikan secara statistik atau hanya terjadi secara kebetulan. T-Test dapat digunakan dalam berbagai konteks, seperti membandingkan hasil tes antara dua kelompok siswa

atau mengevaluasi efektivitas dua metode pengajaran yang berbeda. Hasil dari uji T-Test memberikan nilai p-value yang digunakan untuk menilai signifikansi perbedaan antara kedua kelompok. (De Veaux et al., 2011)

Rumus untuk menghitung nilai T-Test (t) adalah sebagai berikut:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1}$$

Dimana:

- t = nilai T-Test
- $\bar{X}_1$ = rata-rata kelompok pertama
- $\bar{X}_2$ = rata-rata kelompok kedua
- s_1^2 = varians kelompok pertama
- s_2^2 = varians kelompok kedua
- n_1 = ukuran sampel kelompok pertama
- n_2 = ukuran sampel kelompok kedua

Singkatnya uji t-test ini memudahkan kita dalam membandingkan dua kelompok data untuk menentukan apakah ada perbedaan yang signifikan antara keduanya. Apabila p-value lebih kecil dari tingkat signifikansi (misalnya 0,05), maka kita menolak hipotesis nol dan menyimpulkan bahwa ada perbedaan yang signifikan antara kedua kelompok tersebut. Ini berguna apabila fitur numerik ingin dibandingkan terhadap target kategorikal.

2.5 Chi Square Test

Chi Square Test adalah metode statistik yang digunakan untuk menguji hubungan antara dua variabel kategorikal. Uji ini membandingkan frekuensi yang diamati dalam setiap kategori dengan frekuensi yang diharapkan jika tidak ada hubungan antara variabel. Hasil dari uji Chi Square memberikan nilai p-value yang digunakan untuk menentukan apakah hubungan antara variabel tersebut signifikan secara statistik. (Agresti, 2018)

Rumus untuk menghitung nilai Chi Square (χ^2) adalah sebagai berikut:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

Dimana:

- χ^2 = nilai Chi Square
- O_i = frekuensi yang diamati dalam kategori ke-i
- E_i = frekuensi yang diharapkan dalam kategori ke-i

- $\sum$ = penjumlahan untuk semua kategori

Apabila p-value lebih kecil dari tingkat signifikansi (misalnya 0,05), maka kita menolak hipotesis nol dan menyimpulkan bahwa ada hubungan yang signifikan antara kedua variabel tersebut. Ini berguna apabila fitur kategorikal ingin dibandingkan terhadap target kategorikal.

2.6 Standard Scaler

Standard Scaler adalah teknik normalisasi data yang digunakan untuk mengubah fitur numerik sehingga memiliki rata-rata (mean) nol dan standar deviasi (standard deviation) satu. Proses ini membantu dalam mengurangi skala variabilitas antar fitur, sehingga model machine learning dapat belajar lebih efektif. Standard Scaler sangat berguna ketika fitur-fitur dalam dataset memiliki rentang nilai yang berbeda-beda, karena dapat meningkatkan konvergensi dan kinerja model. (Jain & Bhandare, 2016) Rumus untuk menghitung nilai yang telah dinormalisasi (z) menggunakan Standard Scaler adalah sebagai berikut:

$$z = \frac{(X - \mu)}{\sigma} \tag{3}$$

Dimana:

- z = nilai yang telah dinormalisasi
- X = nilai asli dari fitur
- μ = rata-rata (mean) dari fitur
- σ = standar deviasi (standard deviation) dari fitur
- $X - \mu$ = selisih antara nilai asli dan rata-rata
- $\frac{(X-\mu)}{\sigma}$ = hasil pembagian selisih dengan standar deviasi

Dengan menggunakan Standard Scaler, setiap fitur numerik dalam dataset akan diubah sehingga memiliki distribusi yang seragam, yang pada gilirannya dapat meningkatkan performa model machine learning.

2.7 Logistic Regression

Logistic Regression adalah metode statistik yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen (fitur) dengan variabel dependen biner (target). Metode ini digunakan untuk memprediksi probabilitas kejadian suatu peristiwa, seperti apakah seorang kandidat akan diterima atau ditolak dalam proses perekrutan. Logistic Regression menggunakan fungsi logit untuk mengubah output linier menjadi probabilitas yang berada dalam rentang 0 hingga 1. (Hosmer et al., 2013) Rumus untuk menghitung probabilitas (p) menggunakan Logistic Regression adalah sebagai berikut:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}} \tag{4}$$

Dimana:

- p = probabilitas kejadian (misalnya, kandidat diterima)
- e = basis dari logaritma natural (sekitar 2,718)
- β_0 = intercept (konstanta)
- $\beta_1, \beta_2, ..., \beta_n$ = koefisien regresi untuk masing-masing fitur
- $X_1, X_2, ..., X_n$ = nilai dari fitur-fitur independen
- $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$ = kombinasi linier dari fitur-fitur

Logistic Regression sangat berguna dalam berbagai aplikasi, termasuk analisis risiko kredit, diagnosis medis, dan prediksi perilaku konsumen. Dengan kemampuannya untuk memberikan interpretasi yang jelas melalui koefisien regresi, metode ini memungkinkan pemahaman yang lebih baik tentang faktor-faktor yang mempengaruhi keputusan biner.

2.8 Decision Tree

Decision Tree adalah algoritma machine learning yang digunakan untuk klasifikasi dan regresi. Algoritma ini bekerja dengan membagi data menjadi subset berdasarkan fitur-fitur tertentu, sehingga membentuk struktur pohon yang terdiri dari simpul (nodes) dan cabang (branches). Setiap simpul mewakili keputusan berdasarkan nilai fitur, sementara cabang menghubungkan simpul-simpul tersebut. Proses pembagian data berlanjut hingga mencapai simpul daun (leaf nodes) yang memberikan prediksi akhir. Decision Tree sangat populer karena kemampuannya untuk menangani data kategorikal dan numerik, serta memberikan interpretasi yang mudah dipahami.

Rumus untuk menghitung impurity menggunakan Gini Index (G) adalah sebagai berikut:

$$G = 1 - \sum_{i=1}^{C} p_i^2 \tag{5}$$

Dimana:

- G = nilai Gini Index
- C = jumlah kelas dalam target
- p_i = proporsi dari kelas ke-i dalam subset data
- $\sum_{i=1}^{C} p_i^2$ = penjumlahan kuadrat proporsi untuk semua kelas
- $1 - \sum_{i=1}^{C} p_i^2$ = hasil pengurangan dari 1 dengan penjumlahan kuadrat proporsi

Decision Tree sangat berguna dalam berbagai aplikasi, termasuk diagnosis medis, analisis risiko kredit, dan prediksi perilaku konsumen. Dengan kemampuannya untuk memberikan interpretasi yang jelas melalui struktur pohon, metode ini memungkinkan pemahaman yang lebih baik tentang faktor-faktor yang mempengaruhi keputusan klasifikasi.

2.9 Random Forest

Random Forest adalah algoritma ensemble learning yang digunakan untuk klasifikasi dan regresi. Algoritma ini bekerja dengan membangun sejumlah besar pohon keputusan (decision trees) secara acak dan menggabungkan hasil prediksi dari masing-masing pohon untuk menghasilkan prediksi akhir yang lebih akurat dan stabil. Random Forest mengurangi risiko overfitting yang sering terjadi pada pohon keputusan tunggal dengan cara mengambil rata-rata (untuk regresi) atau mode (untuk klasifikasi) dari hasil prediksi semua pohon dalam hutan. (Breiman, 2001)

Rumus untuk menghitung prediksi akhir ($\hat{y}$) dalam Random Forest adalah sebagai berikut:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(X) \tag{6}$$

Dimana:

- $\hat{y}$ = prediksi akhir dari Random Forest
- T = jumlah pohon keputusan dalam hutan
- $h_t(X)$ = prediksi dari pohon keputusan ke-t untuk input X
- $\sum_{t=1}^{T} h_t(X)$ = penjumlahan prediksi dari semua pohon keputusan
- $\frac{1}{T} \sum_{t=1}^{T} h_t(X)$ = hasil pembagian penjumlahan dengan jumlah pohon untuk mendapatkan rata-rata prediksi

Random Forest sangat berguna dalam berbagai aplikasi, termasuk diagnosis medis, analisis risiko kredit, dan prediksi perilaku konsumen. Dengan kemampuannya untuk menangani data yang kompleks dan memberikan interpretasi yang jelas melalui fitur penting (feature importance), metode ini memungkinkan pemahaman yang lebih baik tentang faktor-faktor yang mempengaruhi keputusan klasifikasi atau regresi.

2.10 XGBoost

XGBoost (Extreme Gradient Boosting) adalah algoritma machine learning yang digunakan untuk klasifikasi dan regresi. Algoritma ini merupakan implementasi dari teknik boosting yang menggabungkan beberapa model lemah (weak learners), biasanya pohon keputusan, untuk membentuk model yang lebih kuat dan akurat. XGBoost dikenal karena kemampuannya dalam menangani data besar dan kompleks, serta memberikan performa yang tinggi melalui optimasi paralel dan regularisasi.

Rumus untuk menghitung prediksi akhir ($\hat{y}$) dalam XGBoost adalah sebagai berikut:

$$\hat{y} = \sum_{k=1}^{K} f_k(X) \tag{7}$$

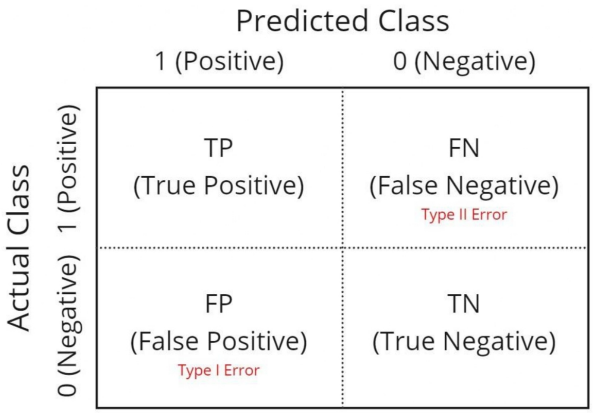
Dimana:

- $\hat{y}$ = prediksi akhir dari XGBoost
- K = jumlah model lemah (pohon keputusan) yang digabungkan
- $f_k(X)$ = prediksi dari model lemah ke-k untuk input X
- $\sum_{k=1}^{K} f_k(X)$ = penjumlahan prediksi dari semua model lemah

XGBoost sangat berguna dalam berbagai aplikasi, termasuk diagnosis medis, analisis risiko kredit, dan prediksi perilaku konsumen. Dengan kemampuannya untuk menangani data yang kompleks dan memberikan interpretasi yang jelas melalui fitur penting (feature importance), metode ini memungkinkan pemahaman yang lebih baik tentang faktor-faktor yang mempengaruhi keputusan klasifikasi atau regresi.

2.11 Confusion Matrix

Confusion Matrix adalah alat evaluasi yang digunakan untuk menilai kinerja model klasifikasi. Matriks ini menyajikan hasil prediksi model dalam bentuk tabel yang menunjukkan jumlah prediksi benar dan salah untuk setiap kelas. Confusion Matrix terdiri dari empat komponen utama: True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Dengan menggunakan Confusion Matrix, kita dapat menghitung berbagai metrik evaluasi seperti akurasi, presisi, recall, dan F1-score, yang memberikan gambaran lebih lengkap tentang kinerja model. Gambar 2 menunjukkan struktur dasar dari Confusion Matrix.(Han et al., 2012)



Gambar 2: Struktur Confusion Matrix

Confusion matrix memberikan informasi tentang jumlah prediksi yang benar dan salah untuk setiap kelas. Dari confusion matrix, kita dapat menghitung berbagai metrik evaluasi model seperti akurasi, presisi, recall, dan F1-score. Metrik-metrik ini membantu dalam menilai kinerja model klasifikasi secara keseluruhan.

2.12 Akurasi

Akurasi adalah metrik evaluasi yang digunakan untuk mengukur seberapa baik model klasifikasi dalam memprediksi kelas yang benar. Akurasi dihitung sebagai rasio antara jumlah prediksi yang benar (baik True Positive maupun True Negative) dengan total jumlah prediksi yang dibuat oleh model. Metrik ini memberikan gambaran umum tentang kinerja model, namun perlu diingat bahwa akurasi saja tidak selalu mencerminkan kinerja model secara menyeluruh, terutama pada dataset yang tidak seimbang. Oleh karena itu, akurasi sering digunakan bersama dengan metrik lain seperti presisi, recall, dan F1-score untuk mendapatkan evaluasi yang lebih komprehensif. (Jain & Bhandare, 2016)

Rumus untuk menghitung akurasi (A) adalah sebagai berikut:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Dimana:

- A = akurasi
- TP = True Positive (jumlah prediksi benar untuk kelas positif)
- TN = True Negative (jumlah prediksi benar untuk kelas negatif)
- FP = False Positive (jumlah prediksi salah untuk kelas positif)
- FN = False Negative (jumlah prediksi salah untuk kelas negatif)
- $TP + TN$ = total prediksi yang benar
- $TP + TN + FP + FN$ = total prediksi yang dibuat oleh model
- $\frac{TP+TN}{TP+TN+FP+FN}$ = hasil pembagian total prediksi benar dengan total prediksi

Akurasi memberikan informasi tentang seberapa sering model membuat prediksi yang benar. Nilai akurasi berkisar antara 0 hingga 1, dimana nilai 1 menunjukkan bahwa model membuat prediksi yang benar untuk semua kasus, sedangkan nilai 0 menunjukkan bahwa model tidak pernah membuat prediksi yang benar. Meskipun akurasi adalah metrik yang berguna, penting untuk mempertimbangkan konteks dan karakteristik dataset saat menilai kinerja model.

2.13 Presisi

Presisi adalah metrik evaluasi yang digunakan untuk mengukur seberapa akurat model klasifikasi dalam memprediksi kelas positif. Presisi dihitung sebagai rasio antara jumlah prediksi benar untuk kelas positif (True Positive) dengan total jumlah prediksi yang dibuat untuk kelas positif (True Positive + False Positive). Metrik ini sangat penting dalam situasi di mana biaya kesalahan positif (False Positive) tinggi, seperti dalam diagnosis medis atau deteksi penipuan. Dengan demikian, presisi memberikan gambaran tentang kualitas prediksi model dalam mengidentifikasi kasus positif. (Jain & Bhandare, 2016)

Rumus untuk menghitung presisi (P) adalah sebagai berikut:

$$P = \frac{TP}{TP + FP} \tag{9}$$

Dimana:

- P = presisi
- TP = True Positive (jumlah prediksi benar untuk kelas positif)
- FP = False Positive (jumlah prediksi salah untuk kelas positif)
- $TP + FP$ = total prediksi yang dibuat untuk kelas positif
- $\frac{TP}{TP+FP}$ = hasil pembagian jumlah prediksi benar untuk kelas positif dengan total prediksi untuk kelas positif

2.14 Recall

Recall adalah metrik evaluasi yang digunakan untuk mengukur seberapa baik model klasifikasi dalam mengidentifikasi semua kasus positif yang sebenarnya. Recall dihitung sebagai rasio antara jumlah prediksi benar untuk kelas positif (True Positive) dengan total jumlah kasus positif yang sebenarnya (True Positive + False Negative). Metrik ini sangat penting dalam situasi di mana biaya kesalahan negatif (False Negative) tinggi, seperti dalam diagnosis medis atau deteksi penipuan. Dengan demikian, recall memberikan gambaran tentang kemampuan model dalam menangkap semua kasus positif yang ada. (Jain & Bhandare, 2016)

Rumus untuk menghitung recall (R) adalah sebagai berikut:

$$R = \frac{TP}{TP + FN} \tag{10}$$

Dimana:

- R = recall
- TP = True Positive (jumlah prediksi benar untuk kelas positif)
- FN = False Negative (jumlah prediksi salah untuk kelas negatif)
- $TP + FN$ = total kasus positif yang sebenarnya
- $\frac{TP}{TP+FN}$ = hasil pembagian jumlah prediksi benar untuk kelas positif dengan total kasus positif yang sebenarnya
- $0 \leq R \leq 1$ = nilai recall berkisar antara 0 hingga 1

2.15 F1-Score

F1-Score adalah metrik evaluasi yang digunakan untuk mengukur keseimbangan antara presisi dan recall dalam model klasifikasi. F1-Score dihitung sebagai rata-rata harmonis dari presisi dan recall, yang memberikan gambaran lebih komprehensif tentang kinerja model, terutama dalam situasi di mana terdapat ketidakseimbangan kelas. Metrik ini sangat berguna ketika kita ingin memastikan bahwa model tidak hanya akurat dalam memprediksi kelas positif, tetapi juga mampu menangkap semua kasus positif yang sebenarnya. (Jain & Bhandare, 2016)

Rumus untuk menghitung F1-Score ($F1$) adalah sebagai berikut:

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{11}$$

Dimana:

- $F1$ = F1-Score
- P = presisi
- R = recall
- $P \times R$ = hasil perkalian antara presisi dan recall
- $P + R$ = hasil penjumlahan antara presisi dan recall
- $2 \times \frac{P \times R}{P + R}$ = hasil perkalian 2 dengan rasio antara hasil perkalian presisi dan recall dengan hasil penjumlahan presisi dan recall
- $0 \leq F1 \leq 1$ = nilai F1-Score berkisar antara 0 hingga 1

F1-Score memberikan informasi tentang keseimbangan antara presisi dan recall. Nilai F1-Score berkisar antara 0 hingga 1, dimana nilai 1 menunjukkan bahwa model memiliki presisi dan recall yang sempurna, sedangkan nilai 0 menunjukkan bahwa model tidak memiliki presisi atau recall sama sekali. Dengan menggunakan F1-Score, kita dapat menilai kinerja model klasifikasi secara lebih menyeluruh, terutama dalam konteks dataset yang tidak seimbang.

2.16 ROC-AUC

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) adalah metrik evaluasi yang digunakan untuk menilai kinerja model klasifikasi biner. ROC curve adalah grafik yang menunjukkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai threshold prediksi. AUC adalah luas di bawah kurva ROC, yang memberikan gambaran tentang kemampuan model dalam membedakan antara kelas positif dan negatif. Nilai AUC berkisar antara 0 hingga 1, dimana nilai 1 menunjukkan bahwa model memiliki kemampuan prediksi yang sempurna, sedangkan nilai 0,5 menunjukkan bahwa model tidak lebih baik dari tebakan acak. (Fawcett, 2006) Rumus untuk menghitung True Positive Rate (TPR) dan False Positive Rate (FPR) adalah sebagai berikut:

$$TPR = \frac{TP}{TP + FN} \tag{12}$$

$$FPR = \frac{FP}{FP + TN} \tag{13}$$

Dimana:

- TPR = True Positive Rate (rasio prediksi benar untuk kelas positif)
- FPR = False Positive Rate (rasio prediksi salah untuk kelas positif)
- TP = True Positive (jumlah prediksi benar untuk kelas positif)
- FN = False Negative (jumlah prediksi salah untuk kelas negatif)
- FP = False Positive (jumlah prediksi salah untuk kelas positif)
- TN = True Negative (jumlah prediksi benar untuk kelas negatif)
- $TP + FN$ = total kasus positif yang sebenarnya
- $FP + TN$ = total kasus negatif yang sebenarnya
- $\frac{TP}{TP+FN}$ = hasil pembagian jumlah prediksi benar untuk kelas positif dengan total kasus positif yang sebenarnya
- $\frac{FP}{FP+TN}$ = hasil pembagian jumlah prediksi salah untuk kelas positif dengan total kasus negatif yang sebenarnya

ROC-AUC memberikan informasi tentang kemampuan model dalam membedakan antara kelas positif dan negatif. Nilai AUC berkisar antara 0 hingga 1, dimana nilai 1 menunjukkan bahwa model memiliki kemampuan prediksi yang sempurna, sedangkan nilai 0,5 menunjukkan bahwa model tidak lebih baik dari tebakan acak. Dengan menggunakan ROC-AUC, kita dapat menilai kinerja model klasifikasi secara lebih menyeluruh, terutama dalam konteks dataset yang tidak seimbang.

2.17 Hyperparameter Tuning

Hyperparameter tuning adalah proses mengoptimalkan parameter-parameter yang tidak dipelajari oleh model selama pelatihan, tetapi memiliki dampak signifikan terhadap kinerja model. Hyperparameter ini dapat mencakup berbagai aspek seperti jumlah pohon dalam Random Forest, laju pembelajaran (learning rate) dalam XGBoost, atau jumlah iterasi dalam algoritma boosting. Proses tuning melibatkan pencarian kombinasi hyperparameter yang menghasilkan kinerja terbaik pada data validasi, sering kali menggunakan teknik seperti grid search atau random search. Dengan melakukan hyperparameter tuning, kita dapat meningkatkan akurasi dan generalisasi model, sehingga menghasilkan prediksi yang lebih andal pada data baru. (Bergstra & Bengio, 2012)

2.18 Cross Validation

Cross Validation adalah teknik evaluasi model yang digunakan untuk menilai kinerja model machine learning dengan membagi dataset menjadi beberapa subset (folds). Proses ini melibatkan pelatihan model pada beberapa subset data dan menguji kinerjanya pada subset yang tersisa. Salah satu metode Cross Validation yang paling umum adalah K-Fold Cross Validation, dimana dataset dibagi menjadi K bagian yang sama, dan model dilatih dan diuji K kali, masing-masing kali menggunakan satu bagian sebagai data uji dan sisanya sebagai data latih. Teknik ini membantu dalam mengurangi overfitting dan memberikan estimasi kinerja model yang lebih akurat dan stabil. (Arlot & Celisse, 2010)

2.19 Feature Importance

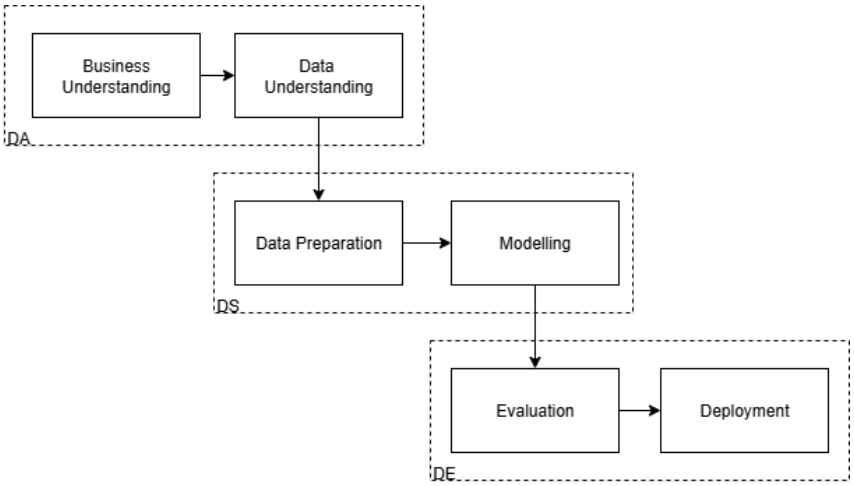
Feature Importance adalah teknik yang digunakan untuk mengukur kontribusi masing-masing fitur dalam mempengaruhi prediksi model machine learning. Dengan mengetahui fitur mana yang paling penting, kita dapat memahami faktor-faktor yang paling berpengaruh terhadap hasil prediksi, serta melakukan seleksi fitur untuk meningkatkan kinerja model. Beberapa algoritma, seperti Random Forest dan XGBoost, secara otomatis menghitung feature importance berdasarkan seberapa sering dan seberapa efektif fitur tersebut digunakan dalam pembentukan pohon keputusan.(Molnar, 2020)

2.20 SHAP Values (Shapley Additive Explanations)

SHAP Values (Shapley Additive Explanations) adalah metode interpretasi model machine learning yang didasarkan pada teori permainan. SHAP Values memberikan penjelasan tentang kontribusi masing-masing fitur terhadap prediksi individu dengan menghitung nilai Shapley, yang merupakan rata-rata kontribusi marginal dari fitur tersebut di semua kemungkinan kombinasi fitur. Metode ini memungkinkan kita untuk memahami bagaimana setiap fitur mempengaruhi hasil prediksi, baik secara positif maupun negatif, sehingga memberikan wawasan yang lebih mendalam tentang perilaku model. SHAP Values sangat berguna dalam konteks model yang kompleks dan sulit diinterpretasikan, seperti ensemble models atau deep learning. (Lundberg & Lee, 2017)

3 Metodologi Penelitian

Metodologi penelitian yang akan digunakan dalam proyek ini adalah Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM adalah sebuah model proses yang terstruktur dan berulang yang terdiri dari enam fase utama, yaitu Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Setiap fase memiliki tujuan dan aktivitas spesifik yang membantu dalam mengelola project ini secara efektif. Gambar 3 menunjukkan diagram alur metodologi yang berlandaskan prinsip CRISP-DM yang akan diikuti dalam penelitian ini.



Gambar 3: Diagram Alur Metodologi

Setiap fase dalam metodologi CRISP-DM memiliki peran penting dalam memastikan keberhasilan proyek ini. Fase Business Understanding bertujuan untuk memahami konteks bisnis dan mengidentifikasi tujuan yang ingin dicapai melalui analisis data. Fase Data Understanding melibatkan eksplorasi awal terhadap dataset untuk menilai kualitas data, mengidentifikasi pola, dan mendeteksi potensi masalah seperti data yang hilang atau outlier.

Fase Data Preparation fokus pada pembersihan dan transformasi data agar siap digunakan dalam proses pemodelan. Ini termasuk penanganan data yang hilang, normalisasi, dan encoding variabel kategorikal. Fase Modeling adalah tahap di mana berbagai algoritma machine learning diterapkan untuk membangun model prediktif berdasarkan data yang telah dipersiapkan.

Fase Evaluation melibatkan penilaian kinerja model menggunakan metrik yang relevan untuk memastikan bahwa model memenuhi kebutuhan bisnis yang telah ditetapkan. Terakhir, fase Deployment adalah tahap di mana model yang telah dievaluasi dan disetujui diintegrasikan ke dalam sistem bisnis yang ada, serta dipantau secara berkelanjutan untuk memastikan performa yang optimal di lingkungan nyata.

3.1 Peran Tim dalam Metodologi

Data analyst berperan penting dalam setiap fase pertama, mulai dari memahami kebutuhan bisnis, melakukan eksplorasi data, menyiapkan data untuk analisis, membangun dan mengevaluasi model, hingga memastikan bahwa model yang dihasilkan dapat diimplementasikan secara efektif dalam konteks bisnis.

Data scientist akan lebih fokus pada fase Modeling dan PreProcessing, di mana mereka akan menerapkan teknik machine learning yang lebih kompleks, melakukan tuning hyperparameter, serta mengevaluasi model dengan metrik yang lebih mendalam untuk memastikan bahwa model tidak hanya akurat tetapi juga dapat diinterpretasikan dan diandalkan.

Data engineer akan berperan utama dalam fase Deployment, di mana mereka akan memastikan bahwa model yang telah dikembangkan dapat diintegrasikan dengan lancar ke dalam infrastruktur teknologi yang ada. Mereka juga akan bertanggung jawab untuk membangun pipeline data yang efisien, mengelola penyimpanan data, serta memastikan bahwa sistem dapat menangani beban kerja yang diperlukan untuk menjalankan model secara real-time atau batch processing sesuai kebutuhan bisnis.

3.2 TimeLine Project

Berikut adalah rincian timeline proyek yang direncanakan untuk setiap fase dalam metodologi CRISP-DM, beserta estimasi waktu yang dibutuhkan untuk menyelesaikan masing-masing fase. Tabel 1 merangkum jadwal proyek secara keseluruhan.

Table 1: Timeline Project

Milestone	Aug	September				Oct
	W4	W1	W2	W3	W4	W1
Project Initiation & Problem Framing	PM & DA					
Data Acquisition & Preparation		DA & DS				
Model Development & Experimentation			DA & DS			
Model Evaluation & Interpretability				DS & BA		
Deployment & Business Integration					DS & DE	
Final Presentation						All Role

Agar lebih jelas lagi, berikut adalah penjabaran dari setiap fase beserta estimasi waktu yang dibutuhkan untuk per stage sesuai jadwal yang diberikan Rakamin Academy. Gambar 4 menunjukkan Timeline Stage0 Project secara keseluruhan.

No	Nama Aktivitas	Nama Task	Role	start date	due date	PIC	Progress Task
		Riset terkait bisnis	Bussiness Analyst	08/23/2025	08/30/2025		Done
		Riset terkait industri	Data Analyst	08/23/2025	08/30/2025		Done
		Identifikasi Kebutuhan data	Project Manager	08/23/2025	08/30/2025		Done
	Melakukan riset terkait industri dari dataset yang dipilih	Riset terkait data	Project Manager	08/23/2025	08/30/2025		Done
		Problem Statement	Bussiness Analyst	08/23/2025	08/30/2025		Done
	Penyusunan Problem Statement & Business Understanding	Tujuan Bisnis	Data Analyst	08/23/2025	08/30/2025		Done
		Draft TimeLine Project	Bussiness Analyst	08/23/2025	08/30/2025		Done
	Penyusunan Project Timeline	Validasi Workflow (Crisp DM)	Data Analyst	08/23/2025	08/30/2025		Done
		Identifikasi Resiko	Bussiness Analyst	08/23/2025	08/30/2025		Done
		Review Dampak Bisnis	Bussiness Analyst	08/23/2025	08/30/2025		Done
	Risk & Feasibility Analysis	Dokumentasi	Project Manager	08/23/2025	08/30/2025		Done
		Memilih Dataset, Pembahasan EDA, Workflow & Timeline	Data Engineer	08/23/2025	08/30/2025		In Pro...
		Review Mentoring	All	08/23/2025	08/30/2025		In Pro...
	Sesi Mentoring	Pembahasan Hasil Stage 0	All	08/23/2025	08/30/2025		In Pro...
		Page Organisasi (4Kings)	Project Manager	08/23/2025	08/30/2025		Done
	Dokumentasi (Github)	Repository untuk Stage 0	Project Manager	08/23/2025	08/30/2025		Done
Stage 0	Evaluasi & Laporan	Penyusunan Proposal	All	08/23/2025	08/30/2025		In Pro...

Gambar 4: Timeline Stage0 Project

Timeline Stage 0 Project dimulai pada minggu ke-4 bulan Agustus dengan dengan detail seperti pada Gambar 4. Dibuat juga kolom progress task untuk menandai progress mana yang belum dikerjakan, sedang dikerjakan dan belum mulai dikerjakan. Gambar 5 menunjukkan Timeline Stage1 Project dengan progress task.

		Pengecekan Dataset	Data Analyst	08/30/2025	09/06/2025		Not Yet
		Relevansi Dataset dengan Bisnis	Bussiness Analyst	08/30/2025	09/06/2025		Not Yet
	Finalisasi Pemilihan Dataset	Dokumentasi Dataset	Project Manager	08/30/2025	09/06/2025		Not Yet
		Analasis Univariate&Multivariate pada Fitur	Data Analyst	08/30/2025	09/06/2025		Not Yet
	Exploratory Data Analysis (EDA)	Visualisasi & Insight	Data Scientist	08/30/2025	09/06/2025		Not Yet
		Handling missing values & duplicates	Data Scientist	08/30/2025	09/06/2025		Not Yet
		Outlier detection	Data Scientist	08/30/2025	09/06/2025		Not Yet
	Preprocessing Data	Memilih outlier handling (IQR/ZScore)	Project Manager	08/30/2025	09/06/2025		Not Yet
		Uji Statistik (Chi Square/Anova)	Bussiness Analyst	08/30/2025	09/06/2025		Not Yet
		Pemilihan Fitur yang relevan	Data Analyst	08/30/2025	09/06/2025		Not Yet
	Feature Selection & Engineering	Standarisasi, Encoding &Transform Log (Opsional)	Project Manager	08/30/2025	09/06/2025		Not Yet
		Hasil Riset, Saran Model & Evaluasi Metric	All	08/30/2025	09/06/2025		Not Yet
		Review Mentoring	All	08/30/2025	09/06/2025		Not Yet
	Sesi Mentoring	Pembahasan Hasil Stage 1	All	08/30/2025	09/06/2025		Not Yet
		Transisi Proposal ke Report (LaTeX)	Project Manager	08/30/2025	09/06/2025		Not Yet
	Dokumentasi (Github)	Repository untuk Stage 1	Project Manager	08/30/2025	09/06/2025		Not Yet
Stage1	Evaluasi & Laporan	Mengisi Final Report	All	08/30/2025	09/06/2025		Not Yet

Gambar 5: Timeline Stage1 Project dengan Progress Task

Gambar 6 menunjukkan Timeline Stage2 Project dengan progress task.

		Menentukan Baseline Model (ex : Logistic Regression, DecisionTree dl	Data Scientist	09/13/2025	09/20/2025		Not Yet
		Model Evaluasi & Interpretasi Bisnis	Data Scientist	09/13/2025	09/20/2025		Not Yet
		Deployment Kasar Coba" Techstack	Data Engineer	09/13/2025	09/20/2025		Not Yet
	Baseline Model Development	Review Baseline model	Project Manager	09/13/2025	09/20/2025		Not Yet
		Mencoba algoritma alternatif (ex : Random Forest, XGBoost, dll.)	Data Scientist	09/13/2025	09/20/2025		Not Yet
	Experiment & Komparasi Model	Analisis & Dokumentasi Hasil Experiment	Project Manager	09/13/2025	09/20/2025		Not Yet
		Menentukan Tuning terbaik (ex : bisa RandomSearch/Grid dll)	Data Scientist	09/13/2025	09/20/2025		Not Yet
		Cross Validation Setup	Data Scientist	09/13/2025	09/20/2025		Not Yet
	Hyperparameter Tuning	Review Hasil Tuning	Project Manager	09/13/2025	09/20/2025		Not Yet
		Uji Statistik (Chi Square/Anova)	Data Analyst	09/13/2025	09/20/2025		Not Yet
		Pemilihan Fitur yang relevan	Project Manager	09/13/2025	09/20/2025		Not Yet
	Feature Selection & Engineering	Standarisasi, Encoding &Transform Log (Opsional)	Data Scientist	09/13/2025	09/20/2025		Not Yet
		Exporting Model	Data Engineer	09/13/2025	09/20/2025		Not Yet
		Menggunakan Teknik Processing yang sama untuk data baru	Data Engineer	09/13/2025	09/20/2025		Not Yet
	Pipeline Model	Review Pipeline	Project Manager	09/13/2025	09/20/2025		Not Yet
	Dokumentasi (Github)	Repository untuk Stage 2	Project Manager	09/13/2025	09/20/2025		Not Yet
Stage 2	Evaluasi & Laporan	Mengisi Final Report	All	09/13/2025	09/20/2025		Not Yet

Gambar 6: Timeline Stage2 Project dengan Progress Task

Gambar 7 menunjukkan Timeline Stage3 Project dengan progress task.

		Evaluasi dengan metrik (Accuracy, Precision, Recall, F1, AUC)	Data Scientist	09/20/2025	09/27/2025		Not Yet
	Model Performance Evaluation	Bandingkan baseline dan model model lain hasil Experiment & Tuning	Project Manager	09/20/2025	09/27/2025		Not Yet
		Shap Values fitur importance	Data Scientist	09/20/2025	09/27/2025		Not Yet
	Feature Importance & Model Tradeoff	Bias & Variance Tradeoff untuk model. Memilih Sesuai dengan Solusi Bisnis	Bussiness Analyst	09/20/2025	09/27/2025		Not Yet
		Review dampak bisnis dari model	Bussiness Analyst	09/20/2025	09/27/2025		Not Yet
		Penyelarasan hasil dengan tujuan Project	Project Manager	09/20/2025	09/27/2025		Not Yet
	Business Impact	Review Hasil Tuning	Project Manager	09/20/2025	09/27/2025		Not Yet
	Dokumentasi (Github)	Repository untuk Stage 3	Project Manager	09/20/2025	09/27/2025		Not Yet
Stage 3	Evaluasi & Laporan	Mengisi Final Report	All	09/20/2025	09/27/2025		Not Yet

Gambar 7: Timeline Stage3 Project dengan Progress Task

Gambar 8 menunjukkan Timeline Stage4 Project dengan progress task.

		Implementasi model ke API	Data Engineer	09/27/2025	10/04/2025		Not Yet
		Pembuatan dashboard interaktif	Data Scientist	09/27/2025	10/04/2025		Not Yet
	Deployment & API Development	Koordinasi Deployment	Project Manager	09/27/2025	10/04/2025		Not Yet
		Menentukan strategi monitoring model	Data Scientist	09/27/2025	10/04/2025		Not Yet
	Monitoring & Maintenance Strategy	Dokumentasi strategi monitoring	Project Manager	09/27/2025	10/04/2025		Not Yet
		Menyusun Pitch Deck Presentasi	All	09/27/2025	10/04/2025		Not Yet
		Visualisasi hasil & insight	All	09/27/2025	10/04/2025		Not Yet
	Final Presentation Preparation	Dokumentasi source code & pipeline	Project Manager	09/27/2025	10/04/2025		Not Yet
	Dokumentasi (Github)	Repository untuk Stage 4 & Readme Interaktif untuk masing" Repository	Project Manager	09/27/2025	10/04/2025		Not Yet
Stage 4	Final Presentation Delivery	Presentasi hasil proyek & Penyerahan Laporan	All	09/27/2025	10/04/2025		Not Yet

Gambar 8: Timeline Stage4 Project dengan Progress Task

3.3 Risk & Feasibility Analysis

Dalam menjalankan proyek ini, terdapat beberapa risiko yang perlu diidentifikasi dan dianalisis untuk memastikan kelancaran proses project. Diharapkan dengan memahami potensi risiko yang ada, tim dapat merancang strategi mitigasi yang efektif guna mengurangi dampak negatif yang mungkin timbul. Tabel 2 merangkum berbagai aspek risiko, potensi risiko yang mungkin dihadapi, strategi mitigasi yang dapat diterapkan, serta penilaian kelayakan dari masing-masing risiko tersebut.

Table 2: Risk-Feasibility Analysis

Aspek	Potensi Risiko	Strategi Mitigasi	Kelayakan
Data	Data kandidat bisa tidak lengkap, tidak seimbang (imbalanced), atau mengandung bias (gender, usia).	Lakukan preprocessing, balancing data, feature engineering, serta audit fairness.	Layak jika dilakukan data cleaning & monitoring.
Teknis	Model bisa overfitting atau performa rendah di data baru.	Gunakan cross-validation, regularisasi, dan retraining berkala.	Layak dengan pipeline validasi yang baik.
Operasional	HR sulit mengadopsi sistem baru, lebih percaya screening manual.	Berikan pelatihan, buat antarmuka user-friendly, dan jelaskan transparansi model.	Layak jika ada kolaborasi dengan HR.
Etika & Regulasi	Risiko diskriminasi dalam keputusan perekrutan (misalnya gender bias).	Terapkan fairness metrics, hindari variabel sensitif sebagai faktor utama.	Layak dengan pengawasan etis & regulasi.
Ekonomi	Biaya implementasi dan maintenance model cukup tinggi.	Bandingkan cost vs benefit (efisiensi waktu, cost per hire, kualitas kandidat).	Layak jika ROI positif dalam 1–2 tahun.
Keberlanjutan	Model bisa usang (model drift) seiring perubahan tren pasar tenaga kerja.	Monitoring performa model, retraining dengan data terbaru setiap periode tertentu.	Layak dengan komitmen maintenance rutin.

Dengan melakukan analisis risiko ini, tim proyek dapat lebih siap dalam menghadapi tantangan yang mungkin muncul selama pelaksanaan proyek. Setiap risiko yang diidentifikasi telah diberikan strategi mitigasi yang spesifik, sehingga dapat diatasi dengan cara yang paling efektif. Selain itu, penilaian kelayakan dari setiap risiko membantu dalam menentukan prioritas tindakan yang perlu diambil untuk memastikan keberhasilan proyek secara keseluruhan.

3.4 Penjelasan Dataset

Dataset yang digunakan dalam proyek ini adalah `recruitment_data.csv`, berisi informasi kandidat dan faktor yang dipertimbangkan dalam proses perekrutan. Tujuan pemodelan adalah memprediksi keputusan perekrutan (*HiringDecision*) berdasarkan atribut kandidat.

3.4.1 Ringkasan Dataset

- **Jumlah rekaman (baris):** 1,500
- **Jumlah fitur (prediktor):** 10
- **Target:** `HiringDecision` (biner: 0 = tidak diterima, 1 = diterima)
- **Sifat data:** Sintetis (dibuat untuk tujuan pendidikan/proyek data sains)

3.4.2 Definisi Variabel

Berikut fitur dan target yang tersedia, beserta tipe data, rentang/kategori, dan keterangan singkat. Tabel 3 merangkum definisi variabel dalam dataset.

Table 3: Definisi Variabel Dataset

Nama Fitur	Tipe	Rentang	Keterangan
Age	Numerik	20-50	Umur kandidat
Gender	Kategorikal	0/1	0 = Laki-laki, 1 = Perempuan
EducationLevel	Kategorikal	1/2/3/4	1 = S1 (Tipe 1), 2 = S1 (Tipe 2), 3 = S2, 4 = S3/PhD
ExperienceYears	Numerik	0-15	Lama pengalaman kerja (tahun)
PreviousCompanies	Numerik	1-5	Jumlah perusahaan tempat bekerja sebelumnya
DistanceFromCompany	Numerik	1-50	Jarak dari rumah ke perusahaan
InterviewScore	Numerik	0-100	Skor hasil wawancara
SkillScore	Numerik	0-100	Skor keterampilan teknis
PersonalityScore	Numerik	0-100	Skor aspek kepribadian
RecruitmentStrategy	Kategorikal	1/2/3	1 = Agresif, 2 = Moderat, 3 = Konservatif
HiringDecision	Target	0/1	Target: 0 = tidak diterima, 1 = diterima

3.4.3 Catatan Kodefikasi dan Pra-pemrosesan

- **Gender**: dikodekan sebagai 0 (Laki-laki) dan 1 (Perempuan).
- **EducationLevel**: ordinal 1–4 dengan pemetaan spesifik (S1 Tipe 1, S1 Tipe 2, S2, S3/PhD). Jika korelasi kuat, dapat di *one-hot*.

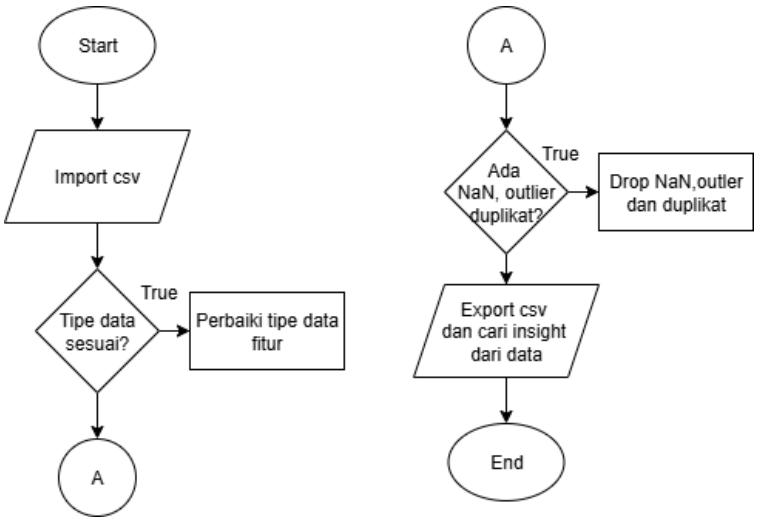
- **RecruitmentStrategy**: kategorikal 1–3. Umumnya di-*one-hot* untuk model linear; model pohon dapat menggunakan kode numeriknya langsung.
- **Skor (Interview/Skill/Personality)**: berada pada skala 0–100; pertimbangkan penskalaan (*standardization/min-max*) untuk model sensitif skala.
- **Fitur numerik lain** (`Age`, `ExperienceYears`, `PreviousCompanies`, `DistanceFromCompany`): periksa outlier, distribusi, dan lakukan transformasi/penanganan jika diperlukan.

3.4.4 Sumber dan Lisensi

Dataset ini dibagikan oleh **Rabie El Kharoua** dengan lisensi **CC BY 4.0**. Dataset bersifat *exclusive synthetic* dan ditujukan untuk keperluan edukasi/proyek data sains. Penggunaan diperbolehkan dengan mencantumkan atribusi yang tepat kepada pemilik dataset. DOI dan rincian penyedia data tercantum pada kartu data sumbernya. (Kharoua, 2024)

3.5 EDA (Exploratory Data Analysis)

EDA adalah langkah awal yang penting dalam analisis data untuk memahami struktur, pola, dan karakteristik dataset. Gambar 9 menunjukkan flowchart EDA yang akan dilakukan dalam proyek ini.



Gambar 9: Flowchart EDA

Dengan mengikut langkah EDA yang terstruktur, tim dapat memperoleh wawasan yang mendalam tentang dataset, mengidentifikasi potensi masalah, dan menyiapkan data dengan baik untuk tahap pemodelan selanjutnya. EDA membantu memastikan bahwa model yang dibangun didasarkan pada pemahaman yang kuat tentang data, sehingga meningkatkan peluang keberhasilan proyek secara keseluruhan.

3.5.1 Handle Tipe Data, NaN, & Duplikasi

Data wrangling adalah proses penting dalam persiapan data untuk analisis dan pemodelan. Proses ini melibatkan beberapa langkah kunci yang bertujuan untuk membersihkan, mengubah, dan mengorganisir data agar siap digunakan. Dengan menggunakan df.info(), kita dapat memperoleh gambaran umum tentang struktur dataset, termasuk jumlah entri, tipe data setiap kolom, dan informasi tentang nilai yang hilang. Berikut adalah hasil dari df.info() pada dataset yang digunakan dalam proyek ini.

Listing 1: Info Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Age                   1500 non-null   int64
 1   Gender                1500 non-null   int64
 2   EducationLevel        1500 non-null   int64
 3   ExperienceYears       1500 non-null   int64
 4   PreviousCompanies     1500 non-null   int64
 5   DistanceFromCompany   1500 non-null   float64
 6   InterviewScore        1500 non-null   int64
 7   SkillScore            1500 non-null   int64
 8   PersonalityScore      1500 non-null   int64
 9   RecruitmentStrategy   1500 non-null   int64
 10  HiringDecision        1500 non-null   int64
dtypes: float64(1), int64(10)
memory usage: 129.0 KB
```

Dari hasil df.info(), kita dapat melihat bahwa dataset terdiri dari 1500 entri dengan 11 kolom. Semua kolom memiliki tipe data numerik (int64 dan float64), dan tidak ada nilai yang hilang (non-null count sama dengan total entries untuk setiap kolom). Ini menunjukkan bahwa dataset sudah cukup bersih dari segi kelengkapan data, namun masih perlu dilakukan pemeriksaan lebih lanjut terhadap distribusi nilai, outlier, dan potensi inkonsistensi lainnya. Walaupun beberapa fitur numerik memiliki makna kategorikal seperti gender, education level, dan recruitment strategy, hal tersebut tidaklah menjadi masalah karena jika dia bertipe object pada akhirnya akan diubah menjadi numerik juga.

Selanjutnya mengecek apakah ada nilai duplikasi pada dataset. Dengan menggunakan df.duplicated().sum(), kita dapat menghitung jumlah baris yang duplikat dalam dataset. Berikut adalah hasil dari pengecekan duplikasi pada dataset yang digunakan dalam proyek ini.

Listing 2: Cek Duplikasi Dataset

```
df.duplicated().sum()

#output
np.int64(0)
```

Dari hasil pengecekan duplikasi, kita dapat melihat bahwa tidak ada baris yang duplikat dalam dataset (jumlah duplikasi adalah 0). Ini menunjukkan bahwa setiap entri dalam dataset adalah unik, yang merupakan kondisi ideal untuk analisis data dan pemodelan. Dengan tidak adanya duplikasi, kita dapat melanjutkan ke tahap berikutnya dalam proses data wrangling dengan keyakinan bahwa data yang kita miliki sudah bersih dari masalah duplikasi.

3.5.2 Analisis Fitur Numerik

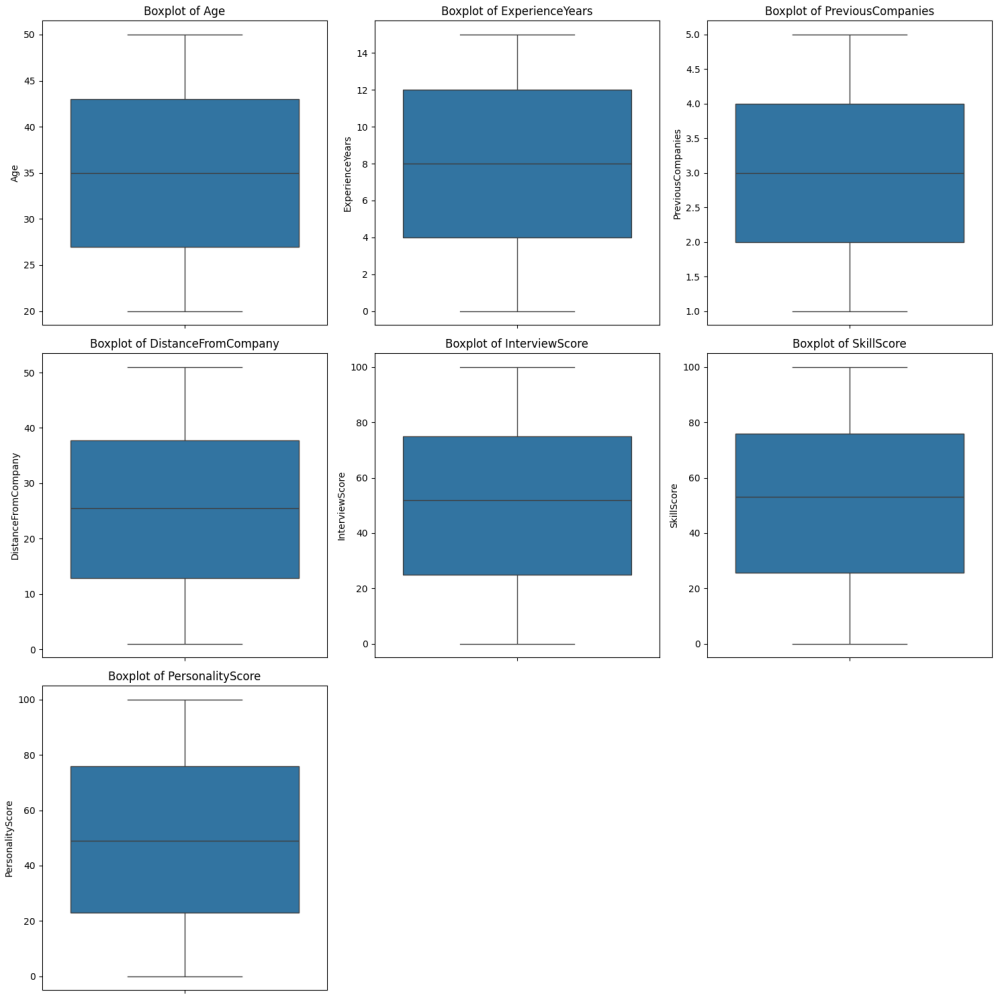
Fitur numerik dalam dataset ini meliputi `Age`, `ExperienceYears`, `PreviousCompanies`, `DistanceFromCompany`, `InterviewScore`, `SkillScore`, dan `PersonalityScore`. Untuk memahami karakteristik dari fitur-fitur ini, kita dapat melakukan analisis statistik deskriptif dan visualisasi distribusi data. Tabel 4 merangkum statistik deskriptif dari fitur numerik dalam dataset.

Table 4: Statistik Deskriptif Fitur Numerik

Fitur	Count	Mean	Std	Min	25%	50%	75%	Max
Age	1500	35.15	9.25	20.00	27.00	35.00	43.00	50.00
ExperienceYears	1500	7.69	4.64	0.00	4.00	8.00	12.00	15.00
PreviousCompanies	1500	3.00	1.41	1.00	2.00	3.00	4.00	5.00
DistanceFromCompany	1500	25.51	14.57	1.00	12.84	25.50	37.74	50.99
InterviewScore	1500	50.56	28.63	0.00	25.00	52.00	75.00	100.00
SkillScore	1500	51.12	29.35	0.00	25.75	53.00	76.00	100.00
PersonalityScore	1500	49.39	29.35	0.00	23.00	49.00	76.00	100.00

Dari tabel statistik deskriptif, kita dapat melihat bahwa fitur-fitur numerik memiliki variasi yang cukup besar dalam nilai rata-rata, standar deviasi, dan rentang nilai. Misalnya, `Age` memiliki rata-rata sekitar 35 tahun dengan rentang dari 20 hingga 50 tahun, sementara `ExperienceYears` memiliki rata-rata sekitar 7.69 tahun dengan rentang dari 0 hingga 15 tahun. Fitur-fitur seperti `InterviewScore`, `SkillScore`, dan `PersonalityScore` menunjukkan distribusi yang luas dengan nilai minimum 0 dan maksimum 100, yang menunjukkan adanya variasi signifikan dalam penilaian kandidat.

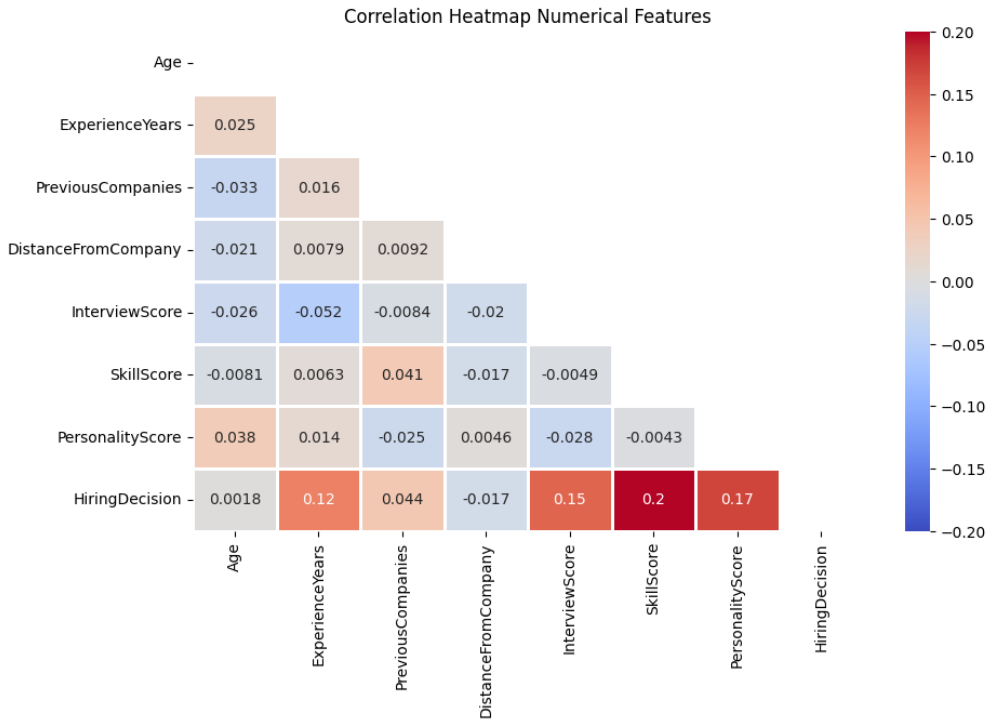
Untuk visualisasi distribusi fitur numerik, kita dapat menggunakan boxplot.boxplot memberikan informasi tentang median, kuartil, dan potensi outlier. Gambar 10 menunjukkan boxplot dari beberapa fitur numerik utama.



Gambar 10: Boxplot Fitur Numerik

Dari Gambar 10, dapat dilihat bahwa semua fitur numerik memiliki tipe distribusi normal serta bersih dari outlier. Hal ini menunjukkan bahwa data sudah cukup baik untuk digunakan dalam pemodelan tanpa perlu penanganan khusus terhadap outlier.

Agar kita lebih memahami hubungan antar fitur numerik, kita dapat menggunakan heatmap untuk memvisualisasikan korelasi antar fitur. Gambar 11 menunjukkan heatmap dari korelasi fitur numerik dalam dataset.



Gambar 11: Heatmap Korelasi Fitur Numerik

Dari Gambar 11, kita dapat melihat bahwa sebagian besar fitur numerik memiliki korelasi yang rendah hingga sedang satu sama lain, dengan nilai korelasi berkisar antara -0.03 hingga 0.2. Tidak ada fitur yang menunjukkan korelasi sangat tinggi (di atas 0.8), yang mengindikasikan bahwa tidak ada multikolinearitas yang signifikan di antara fitur-fitur tersebut. Fitur-fitur seperti `InterviewScore`, `SkillScore`, dan `PersonalityScore` menunjukkan korelasi positif yang signifikan dengan label target (Hiring Decision), yang masuk akal karena ketiga fitur ini berkaitan dengan penilaian kandidat.

Agar memiliki insight yang kuat terhadap data, digunakan juga visualisasi barplot untuk melihat hubungan antara fitur numerik dengan target. Adapun pertanyaan yang dapat dijawab melalui analisis ini antara lain:

1. Apakah kandidat yang diterima memiliki usia yang lebih tinggi dibandingkan yang tidak diterima?

2. Apakah kandidat yang diterima memiliki pengalaman kerja yang lebih banyak dibandingkan yang tidak diterima?

3. Apakah kandidat yang diterima memiliki riwayat bekerja di lebih banyak perusahaan dibandingkan yang tidak diterima?

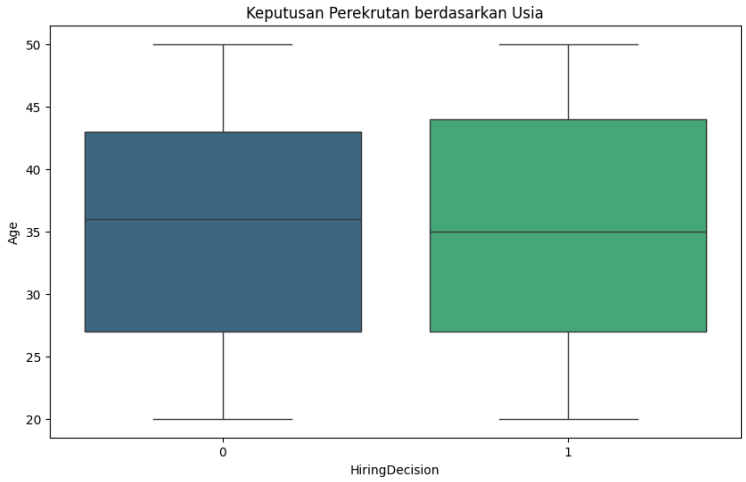
4. Apakah kandidat yang diterima tinggal lebih dekat ke perusahaan dibandingkan yang tidak diterima?

5. Apakah kandidat yang diterima memiliki Skor Wawancara yang lebih tinggi dibandingkan yang tidak diterima?

6. Apakah kandidat yang diterima memiliki skor Keterampilan yang lebih tinggi dibandingkan yang tidak diterima?

7. Apakah kandidat yang diterima memiliki skor Kepribadian yang lebih tinggi dibandingkan yang tidak diterima?

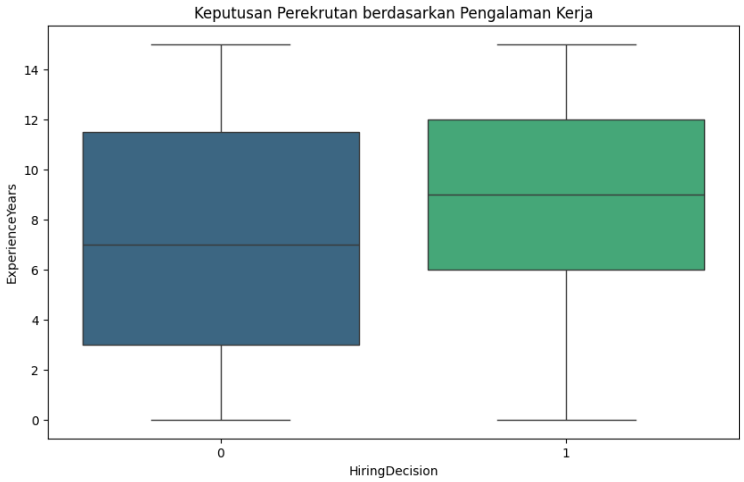
Untuk menjawab pertanyaan-pertanyaan tersebut, kita dapat membuat bar plot dan box plot yang menunjukkan perbandingan rata-rata dari fitur numerik berdasarkan keputusan perekrutan (HiringDecision). Untuk menjawab pertanyaan pertama, kita dapat membuat box plot yang menunjukkan perbandingan rata-rata usia kandidat yang diterima dan tidak diterima. Gambar 12 menunjukkan box plot dari perbandingan rata-rata usia berdasarkan HiringDecision dalam dataset.



Gambar 12: Box Plot Perbandingan Rata-Rata Usia Berdasarkan HiringDecision

Dapat dilihat pada Gambar 12 bahwa tidak ada perbedaan yang signifikan dalam usia antara kandidat yang diterima dan tidak diterima. Rata-rata usia untuk kedua kelompok tersebut tampak cukup mirip, menunjukkan bahwa usia mungkin bukan faktor penentu utama dalam keputusan perekrutan.

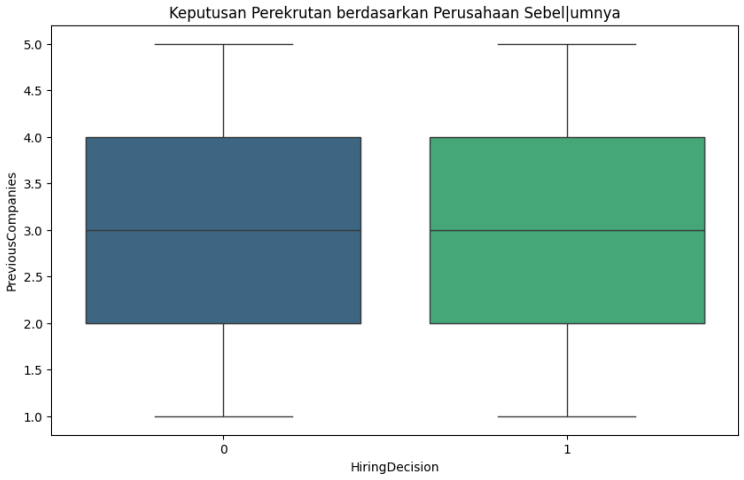
Untuk menjawab pertanyaan kedua, kita dapat membuat box plot yang menunjukkan perbandingan rata-rata pengalaman kerja kandidat yang diterima dan tidak diterima. Gambar 13 menunjukkan box plot dari perbandingan rata-rata pengalaman kerja berdasarkan HiringDecision dalam dataset.



Gambar 13: Box Plot Perbandingan Rata-Rata Pengalaman Kerja Berdasarkan HiringDecision

Dapat dilihat pada Gambar 13 bahwa kandidat yang diterima cenderung memiliki pengalaman kerja yang sama dengan kandidat yang tidak diterima. Namun, dapat dilihat bahwa 50% dari kandidat yang diterima memiliki pengalaman kerja lebih dari 6 tahun yang menandakan bahwa perusahaan cendrung menerima kandidat yang memiliki pengalaman kerja lebih dari besar walaupun hal tersebut bukan menjadi faktor penentu utama dalam keputusan perekrutan karena 50% dari kandidat yang tidak diterima juga berasal dari rentang pengalaman kerja yang tinggi.

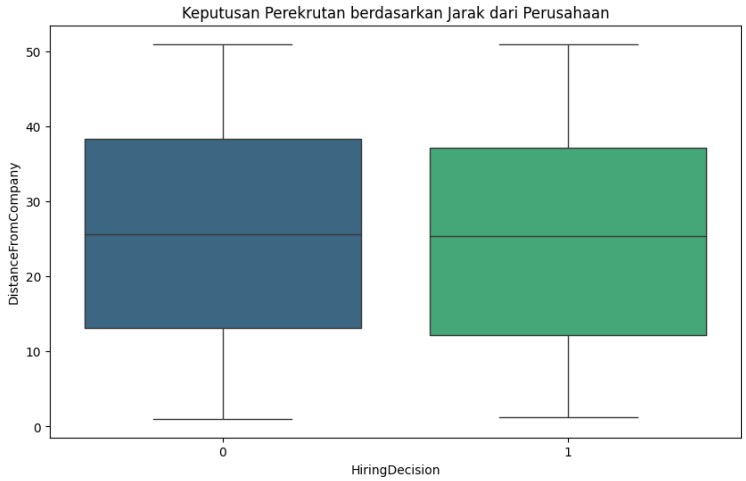
Untuk menjawab pertanyaan ketiga, kita dapat membuat box plot yang menunjukkan perbandingan rata-rata jumlah perusahaan sebelumnya tempat kandidat bekerja berdasarkan HiringDecision. Gambar 14 menunjukkan box plot dari perbandingan rata-rata jumlah perusahaan sebelumnya berdasarkan HiringDecision dalam dataset.



Gambar 14: Box Plot Perbandingan Rata-Rata Jumlah Perusahaan Sebelumnya Berdasarkan HiringDecision

Dapat dilihat pada Gambar 14 bahwa tidak ada perbedaan yang signifikan dalam jumlah perusahaan sebelumnya antara kandidat yang diterima dan tidak diterima. Rata-rata jumlah perusahaan sebelumnya untuk kedua kelompok tersebut tampak cukup mirip, menunjukkan bahwa jumlah perusahaan sebelumnya mungkin bukan faktor penentu utama dalam keputusan perekrutan.

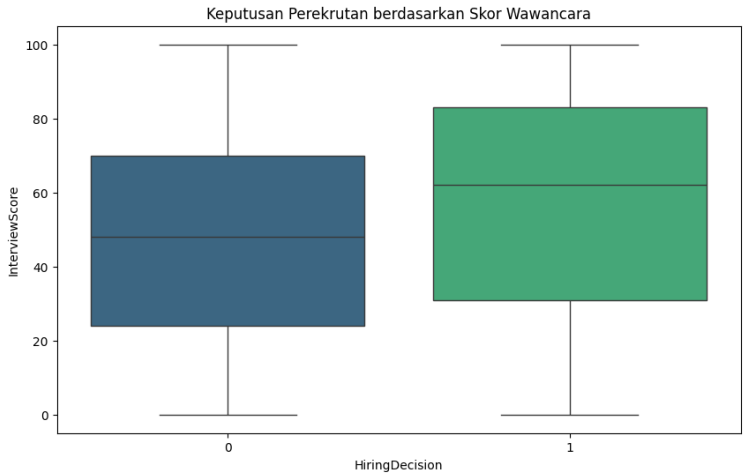
Untuk menjawab pertanyaan keempat, kita dapat membuat box plot yang menunjukkan perbandingan rata-rata jarak dari rumah ke perusahaan berdasarkan HiringDecision. Gambar 15 menunjukkan box plot dari perbandingan rata-rata jarak berdasarkan HiringDecision dalam dataset.



Gambar 15: Box Plot Perbandingan Rata-Rata Jarak dari Rumah ke Perusahaan Berdasarkan HiringDecision

Dapat dilihat pada Gambar 15 bahwa tidak ada perbedaan yang signifikan dalam jarak dari rumah ke perusahaan antara kandidat yang diterima dan tidak diterima. Rata-rata jarak untuk kedua kelompok tersebut tampak cukup mirip, menunjukkan bahwa jarak dari rumah ke perusahaan mungkin bukan faktor penentu utama dalam keputusan perekrutan.

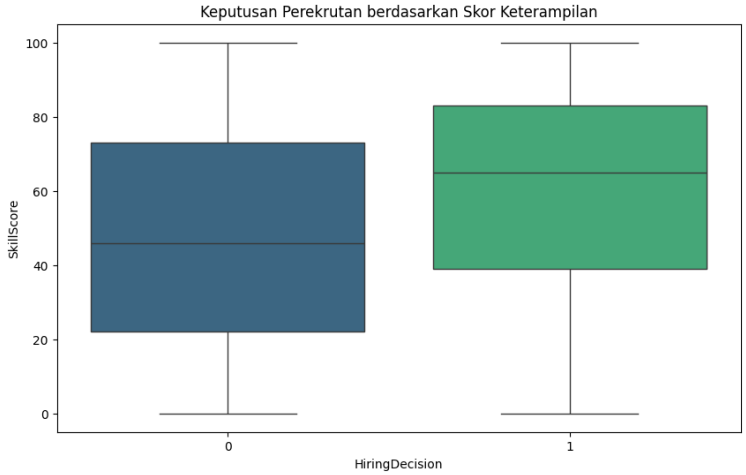
Untuk menjawab pertanyaan kelima, kita dapat membuat box plot yang menunjukkan perbandingan rata-rata skor wawancara berdasarkan HiringDecision. Gambar 16 menunjukkan box plot dari perbandingan rata-rata skor wawancara berdasarkan HiringDecision dalam dataset.



Gambar 16: Box Plot Perbandingan Rata-Rata Skor Wawancara Berdasarkan HiringDecision

Dapat dilihat pada Gambar 16 bahwa kandidat yang diterima cenderung memiliki skor wawancara yang lebih tinggi dibandingkan dengan kandidat yang tidak diterima. Rata-rata skor wawancara untuk kandidat yang diterima tampak lebih tinggi, menunjukkan bahwa skor wawancara mungkin menjadi faktor penting dalam keputusan perekrutan.

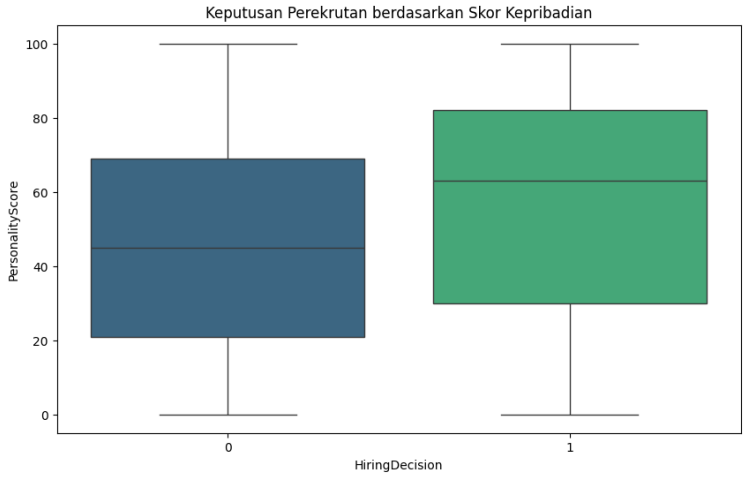
Untuk menjawab pertanyaan keenam, kita dapat membuat box plot yang menunjukkan perbandingan rata-rata skor keterampilan berdasarkan HiringDecision. Gambar 17 menunjukkan box plot dari perbandingan rata-rata skor keterampilan berdasarkan HiringDecision dalam dataset.



Gambar 17: Box Plot Perbandingan Rata-Rata Skor Keterampilan Berdasarkan HiringDecision

Dapat dilihat pada Gambar 17 bahwa kandidat yang diterima cenderung memiliki skor keterampilan yang lebih tinggi dibandingkan dengan kandidat yang tidak diterima. Rata-rata skor keterampilan untuk kandidat yang diterima tampak lebih tinggi, menunjukkan bahwa skor keterampilan mungkin menjadi faktor penting dalam keputusan perekrutan.

Untuk menjawab pertanyaan ketujuh, kita dapat membuat box plot yang menunjukkan perbandingan rata-rata skor kepribadian berdasarkan HiringDecision. Gambar 18 menunjukkan box plot dari perbandingan rata-rata skor kepribadian berdasarkan HiringDecision dalam dataset.



Gambar 18: Box Plot Perbandingan Rata-Rata Skor Kepribadian Berdasarkan HiringDecision

Dapat dilihat pada Gambar 18 bahwa kandidat yang diterima cenderung memiliki skor kepribadian yang lebih tinggi dibandingkan dengan kandidat yang tidak diterima. Rata-rata skor kepribadian untuk kandidat yang diterima tampak lebih tinggi, menunjukkan bahwa skor kepribadian mungkin menjadi faktor penting dalam keputusan perekrutan.

Dengan analisa yang dilakukan sebelumnya, kita sudah mengetahui bahwa data sudah bersih dan siap untuk dilakukan pemodelan, kita juga sudah mengetahui fitur-fitur mana saja yang memiliki korelasi yang cukup tinggi dengan target dan mana fitur yang tidak memiliki korelasi sama sekali. Namun uji statistik tetap perlu dilakukan untuk memastikan fitur-fitur yang akan digunakan dalam pemodelan adalah fitur yang benar-benar memiliki korelasi yang signifikan dengan target. Karena label target adalah kategorikal (1/0) dan fitur yang akan diuji adalah numerik, maka uji statistik yang tepat adalah uji t-test. Berikut adalah hasil uji t-test yang dilakukan pada fitur-fitur numerik dalam dataset. Tabel 5 merangkum hasil uji t-test untuk setiap fitur numerik.

Table 5: Hasil Uji T-Test Fitur Numerik

Fitur	T-Stat	P-Value	Kesimpulan
Age	0.0716	9.43×10^{-1}	Tidak ada perbedaan signifikan ($p >= 0.05$)
ExperienceYears	4.7770	1.95×10^{-6}	Ada perbedaan signifikan ($p < 0.05$)
PreviousCompanies	1.7056	8.83×10^{-2}	Tidak ada perbedaan signifikan ($p >= 0.05$)
DistanceFromCompany	-0.6500	5.16×10^{-1}	Tidak ada perbedaan signifikan ($p >= 0.05$)
InterviewScore	5.7146	1.32×10^{-8}	Ada perbedaan signifikan ($p < 0.05$)
SkillScore	8.0515	1.7×10^{-15}	Ada perbedaan signifikan ($p < 0.05$)
PersonalityScore	6.6436	4.27×10^{-11}	Ada perbedaan signifikan ($p < 0.05$)

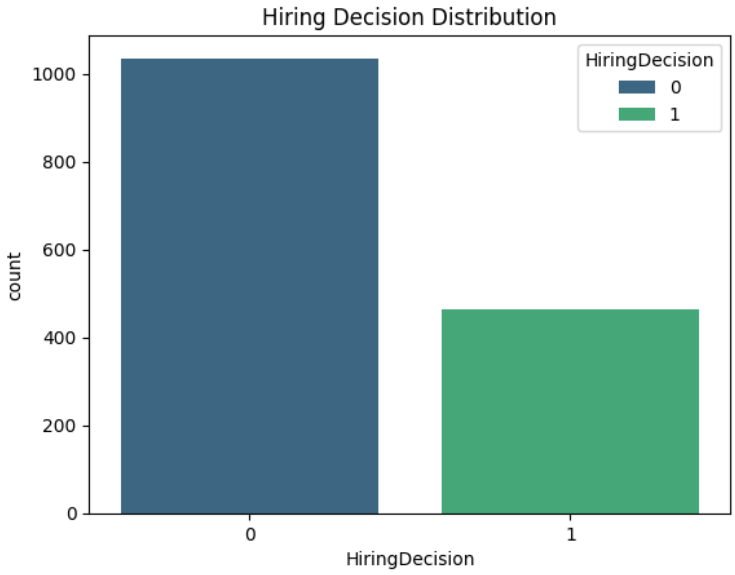
Dari tabel hasil uji t-test, kita dapat melihat bahwa fitur-fitur seperti `ExperienceYears`, `InterviewScore`, `SkillScore`, dan `PersonalityScore` menunjukkan perbedaan yang signifikan antara kelompok kandidat yang diterima dan tidak diterima (HiringDecision) ($p < 0.05$). Ini mengindikasikan bahwa fitur-fitur ini memiliki pengaruh yang kuat terhadap keputusan perekrutan dan sebaiknya dipertimbangkan dalam pemodelan. Sebaliknya, fitur seperti `Age`, `PreviousCompanies`, dan `DistanceFromCompany` tidak menunjukkan perbedaan yang signifikan ($p >= 0.05$), sehingga mungkin kurang relevan untuk dimasukkan dalam model prediksi.

3.6 Analisis Fitur Kategorikal

Fitur kategorikal dalam dataset ini meliputi Gender, EducationLevel, dan RecruitmentStrategy. Untuk memahami karakteristik dari fitur-fitur ini, kita dapat melakukan analisis frekuensi dan visualisasi distribusi data. Adapun beberapa pertanyaan yang dapat dijawab melalui analisis ini antara lain:

1. Berapa perbandingan jumlah kandidat yang diterima dan tidak ?
2. Apakah EducationLevel berperan besar dalam keputusan hiring ?
3. Berapa perbandingan pria dan wanita yang diterima dan tidak ?
4. Apakah RecruitmentStrategy berperan besar dalam keputusan hiring ?

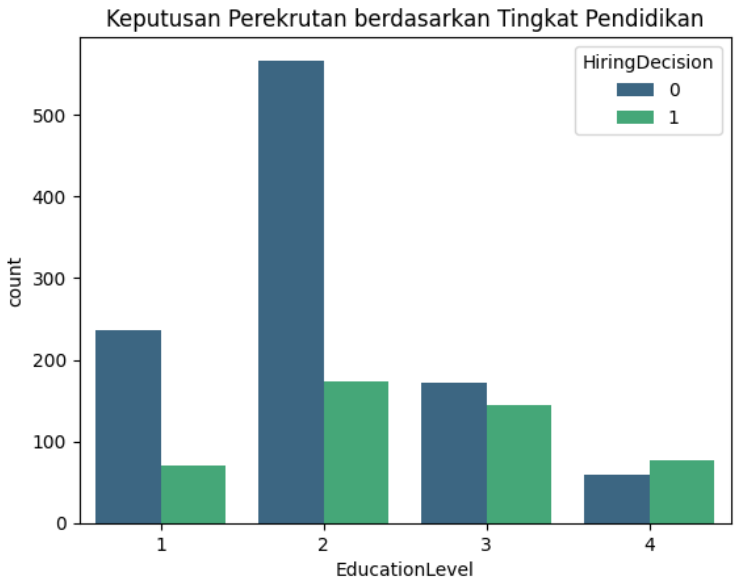
Untuk menjawab pertanyaan-pertanyaan tersebut, kita dapat menggunakan visualisasi seperti bar plot. Untuk menjawab pertanyaan pertama, kita dapat membuat bar plot yang menunjukkan jumlah kandidat yang diterima dan tidak diterima. Gambar 19 menunjukkan bar plot dari distribusi HiringDecision dalam dataset.



Gambar 19: Bar Plot Distribusi HiringDecision

Didapat perbandingnan 31:69, artinya 31% kandidat diterima dan 69% tidak diterima. Hal ini menunjukkan bahwa proses perekrutan cukup selektif, dengan hanya sekitar sepertiga dari total kandidat yang berhasil diterima. Informasi ini penting untuk memahami tingkat persaingan di antara kandidat dan dapat membantu dalam merancang strategi perekrutan yang lebih efektif.

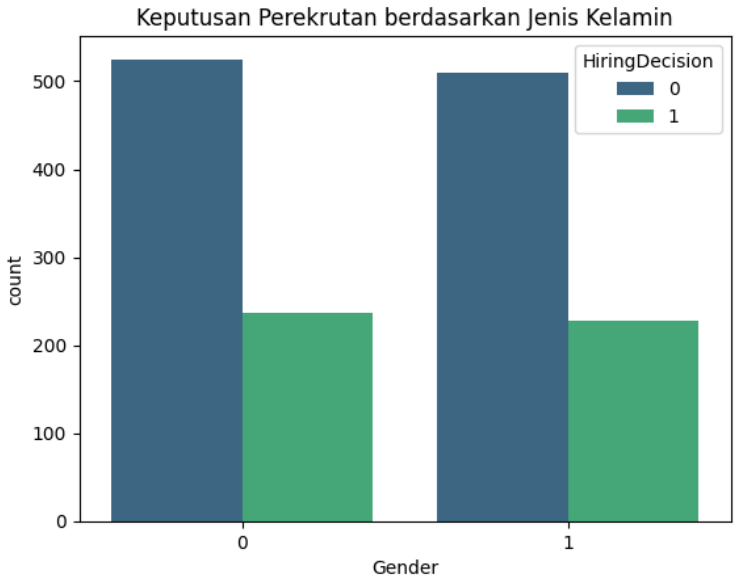
Untuk menjawab pertanyaan kedua, kita dapat membuat bar plot yang menunjukkan hubungan antara EducationLevel dan HiringDecision. Gambar 20 menunjukkan bar plot dari distribusi EducationLevel berdasarkan HiringDecision.



Gambar 20: Bar Plot EducationLevel vs HiringDecision

Dari Gambar 20, kita dapat melihat bahwa kandidat dengan EducationLevel yang lebih tinggi cenderung memiliki peluang lebih besar untuk diterima (HiringDecision = 1). Namun ada beberapa temuan unik, seperti kandidat dengan EducationLevel 4 (S3/PhD) memiliki peluang diterima yang lebih rendah dibandingkan dengan kandidat dengan EducationLevel 2 (S1 Tipe 2) dan 3 (S2). Hal ini menunjukkan bahwa meskipun tingkat pendidikan yang lebih tinggi umumnya dianggap sebagai keunggulan, faktor lain seperti pengalaman kerja, keterampilan, dan penilaian wawancara juga memainkan peran penting dalam keputusan perekrutan.

Untuk menjawab pertanyaan ketiga, kita dapat membuat bar plot yang menunjukan hubungan antara Gender dan HiringDecision. Gambar 21 menunjukan bar plot dari distribusi Gender berdasarkan HiringDecision.



Gambar 21: Bar Plot Gender vs HiringDecision

Dari Gambar 21, kita dapat melihat bahwa kandidat Pria maupun Wanita memiliki proporsi yang seimbang baik yang diterima maupun yang tidak diterima. Hal ini menunjukkan bahwa proses perekrutan dalam dataset ini tidak menunjukkan bias yang signifikan terhadap jenis kelamin, dan keputusan perekrutan lebih dipengaruhi oleh faktor-faktor lain seperti keterampilan, pengalaman, dan penilaian wawancara.

Untuk menjawab pertanyaan keempat, kita dapat membuat bar plot yang menunjukkan hubungan antara RecruitmentStrategy dan HiringDecision. Gambar 22 menunjukkan bar plot dari distribusi RecruitmentStrategy berdasarkan HiringDecision.

Gambar 22: Bar Plot RecruitmentStrategy vs HiringDecision

Dari Gambar 22, kita dapat melihat bahwa kandidat yang direkrut melalui strategi agresif (RecruitmentStrategy = 1) memiliki peluang lebih besar untuk diterima (HiringDecision = 1) dibandingkan dengan strategi moderat (2) dan konservatif (3). Hal ini menunjukkan bahwa strategi perekrutan yang lebih proaktif dan intensif cenderung menghasilkan kandidat yang lebih sesuai dengan kebutuhan perusahaan, sehingga meningkatkan peluang mereka untuk diterima dalam proses seleksi.

Dengan analisa yang dilakukan sebelumnya, kita sudah mengetahui bahwa beberapa fitur kategorikal memiliki korelasi yang signifikan dengan target, seperti EducationLevel dan RecruitmentStrategy. Namun uji statistik tetap perlu dilakukan untuk memastikan fitur-fitur yang akan digunakan dalam pemodelan adalah fitur yang benar-benar memiliki korelasi yang signifikan dengan target. Karena label target adalah kategorikal (1/0) dan fitur yang akan diuji juga kategorikal, maka uji statistik yang tepat adalah uji Chi-Square. Berikut adalah hasil uji Chi-Square yang dilakukan pada fitur-fitur kategorikal dalam dataset. Tabel 6 merangkum hasil uji Chi-Square untuk setiap fitur kategorikal.

Table 6: Hasil Uji Chi-Square Fitur Kategorikal

Fitur	Chi^2	P-Values	Kesimpulan
Gender	0.9750	9.75×10^{-1}	Tidak ada perbedaan signifikan ($p >= 0.05$)
EducationLevel	103.67	2.52×10^{-22}	Ada perbedaan signifikan ($p < 0.05$)
RecruitmentStrategy	489.68	4.65×10^{-62}	Ada perbedaan signifikan ($p < 0.05$)

Dari tabel 6 dapat kita lihat bahwa analisa kita sebelumnya terbukti benar terkait perbedaan signifikan, terutama pada fitur Gender yang tidak memiliki perbedaan signifikan. Sedangkan pada fitur EducationLevel dan RecruitmentStrategy terbukti memiliki perbedaan yang signifikan. Dengan demikian, fitur-fitur ini dapat dipertimbangkan untuk dimasukkan dalam model prediksi karena memiliki pengaruh yang kuat terhadap keputusan perekrutan.

3.7 Kesimpulan Hasil EDA

Berdasarkan analisis data yang telah dilakukan, dapat disimpulkan beberapa hal penting terkait fitur-fitur dalam dataset dan hubungannya dengan keputusan perekrutan (HiringDecision):

1. Fitur numerik seperti `InterviewScore`, `SkillScore`, dan `PersonalityScore` menunjukkan perbedaan yang signifikan antara kandidat yang diterima dan tidak diterima. Fitur-fitur ini memiliki pengaruh yang kuat terhadap keputusan perekrutan dan sebaiknya dipertimbangkan dalam pemodelan.

2. Fitur kategorikal seperti `EducationLevel` dan `RecruitmentStrategy` juga menunjukkan perbedaan yang signifikan dengan keputusan perekrutan. Kandidat dengan tingkat pendidikan yang lebih tinggi dan yang direkrut melalui strategi agresif cenderung memiliki peluang lebih besar untuk diterima.

3. Dari total 1500 kandidat yang ada, hanya sekitar 31% yang diterima, menunjukkan bahwa proses perekrutan cukup selektif.

4. Tidak ada bias signifikan terhadap Usia dan jenis kelamin dalam proses perekrutan, dengan proporsi pria dan wanita yang diterima dan tidak diterima cukup seimbang serta rentang usia yang seimbang menandakan usia dan gender bukan faktor penentu utama dalam keputusan perekrutan. Sehingga tim perekrutan dapat dikatakan profesional dan objektif dalam menilai kandidat.

5. Fitur-fitur seperti `Age`, `PreviousCompanies`, dan `DistanceFromCompany` tidak menunjukkan perbedaan yang signifikan dengan keputusan perekrutan, sehingga mungkin kurang relevan untuk dimasukkan dalam model prediksi.

6. Berdasarkan hasil uji statistik, fitur-fitur yang memiliki perbedaan signifikan dengan keputusan perekrutan dapat diprioritaskan dalam pemodelan untuk meningkatkan akurasi prediksi.

Dengan demikian, fitur-fitur yang telah diidentifikasi sebagai signifikan dapat digunakan untuk membangun model prediksi yang lebih akurat dalam menentukan keputusan perekrutan di masa depan. Langkah selanjutnya adalah melakukan pemodelan menggunakan fitur-fitur tersebut dan mengevaluasi kinerja model yang dihasilkan.

3.8 Feature Selection

3.9 Preprocessing

Daftar Pustaka

Agresti, A. (2018). *Statistical methods for the social sciences* (5th ed.). Pearson.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Breaugh, J. A. (2013). Employee recruitment. *Annual Review of Psychology*, *64*, 389–416. https://doi.org/10.1146/annurev-psych-113011-143757

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chumbar, S. (2020). *The crisp-dm process: A comprehensive guide* [Accessed: 2025-08-26]. https://medium.com/@shawn.chumbar/the-crisp-dm-process-a-comprehensive-guide-4d893aecb151

De Veaux, R. D., Velleman, P. F., & Bock, D. E. (2011). *Intro stats* (3rd ed.). Pearson Education.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Hausknecht, J. P., Rodda, J., & Howard, M. J. (2009). Targeted employee retention: Performancebased and jobrelated differences in reported reasons for staying. *Human Resource Management*, *48*(2), 269–288. https://doi.org/10.1002/hrm.20279

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387

Jain, A. K., & Bhandare, S. (2016). Min max normalization based data perturbation method for privacy protection. *International Journal of Computer Applications*, *136*(5), 1–4. https://doi.org/10.5120/ijca2016908569

Kharoua, R. E. (2024). Predicting hiring decisions in recruitment data. https://doi.org/10.34740/KAGGLE/DSV/8715385

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 4765–4774.

Mathis, R., Jackson, J., Valentine, S., & Meglich, P. (2017). *Human resource management*. Cengage Learning.

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*. Lulu.com. https://christophm.github.io/interpretable-ml-book/

Ng, E. S. W., & Burke, R. J. (2005). Person–organization fit and the war for talent: Does diversity management make a difference? *The International Journal of Human Resource Management*, *16*(7), 1195–1210. https://doi.org/10.1080/09585190500144038

Ployhart, R. E. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management*, *32*(6), 868–897. https://doi.org/10.1177/0149206306293625

Schmidt, F., & Hunter, J. (1998). The validity and utility of selection methods in personnel psychology. *Psychological Bulletin*, *124*(2), 262–274.

Swita, A. (2023). *Dampak proses rekrutmen yang buruk bagi perusahaan*. https://hrpods.co.id/recruitment-and-selection/dampak-proses-rekrutmen-yang-buruk-bagi-perusahaan-220928

Tukey, J. W. (1977). *Exploratory data analysis*. Addison–Wesley.