BAB IV
HASIL DAN PEMBAHASAN

4.1 Data Understanding

Dataset yang digunakan dalam penelitian ini diperoleh langsung dari bagian pemasaran Universitas XYZ. Data tersebut mencakup periode dari tahun 2020 hingga 2023, memberikan peneliti gambaran menyeluruh tentang peminatan program studi selama beberapa tahun terakhir. Dengan jumlah total record sebanyak 13.404, setiap entri data mewakili informasi lengkap tentang calon mahasiswa termasuk berbagai fitur. Rentang waktu yang dipilih tidak hanya memberikan penelitian akses terhadap data yang relevan dalam kurun waktu yang cukup luas, tetapi juga memungkinkan analisis yang lebih mendalam terhadap perubahan yang terjadi dari tahun ke tahun. Dengan pemahaman yang mendalam tentang karakteristik dataset, peneliti dapat mengarahkan penelitian dengan lebih tepat sasaran, serta merencanakan strategi analisis yang efektif untuk menemukan wawasan dan pengetahuan yang bernilai.

Penelitian ini akan menghasilkan beberapa output utama. Nilai akurasi dari *K-Nearest Neighbor* adalah salah satunya (KNN) yang dihitung untuk mengevaluasi kinerja model dalam memprediksi peminatan program studi. Akurasi ini akan memberikan indikasi seberapa baik model dapat mengklasifikasikan data baru berdasarkan data historis. Selain itu, grafik dan persentase pertumbuhan juga akan dihasilkan untuk menggambarkan distribusi peminatan program studi dari tahun ke tahun, serta untuk menunjukkan tren dan perubahan yang terjadi. Persentase pertumbuhan ini akan memberikan pemahaman lebih lanjut tentang dinamika peminatan program studi di Universitas XYZ selama periode yang diteliti. Diharapkan bahwa temuan ini akan memberikan pemahaman yang komprehensif tentang tren spesialisasi program studi dan membantu Universitas XYZ dalam merumuskan strategi pemasaran dan perencanaan akademik yang lebih efektif.

Tabel 4. 1 ***Data Understanding***

nomor	tahun	jenis_kelamin	jenis_sekolah	asal_sekolah	jalur_seleksi	jurusan_sekolah	prodi
1	2020	PRIA	SMAS	SUMATERA SELATAN	BEASISWA	IPA	ILMU KOMUNIKASI
2	2020	PRIA	SMAS	SUMATERA SELATAN	REGULER	IPA	MANAJEMEN
3	2020	PRIA	SMAN	JAKARTA SELATAN		IPS	PSIKOLOGI
4	2020	PRIA	SMAN	JAKARTA SELATAN	REGULER	IPS	ARSITEKTUR
5	2020	PRIA	SMKN	BANTEN	REGULER		TEKNIK SIPIL
6	2020	PRIA	SMAN	JAWA TIMUR	BEASISWA	IPS	ILMU KOMUNIKASI
7	2020	PRIA	MAS	SUMATERA SELATAN	BEASISWA	IPA	DESAIN KOMUNIKASI VISUAL
8	2020	PRIA	SMAN	TANGERANG SELATAN	REGULER	IPA	ILMU KOMUNIKASI
9	2020	PRIA	SMAS	KALIMANTAN TENGAH		IPA	TEKNIK SIPIL
10	2020	WANITA	SMAN	KALIMANTAN UTARA		IPA	PSIKOLOGI
11	2020	WANITA	SMKN	TANGERANG SELATAN	REGULER	TATA BOGA	PSIKOLOGI
12	2020	WANITA	SMAN	JAKARTA SELATAN	REGULER	IPS	PSIKOLOGI
13	2020	PRIA	SMAN	BOGOR	BEASISWA	IPA	AKUNTANSI
14	2020	PRIA	SMAN	TANGERANG		IPS	PSIKOLOGI
15	2020	WANITA	SMAN	TANGERANG SELATAN	REGULER	IPA	MANAJEMEN
16	2020	WANITA	SMAS	TANGERANG	REGULER	IPA	AKUNTANSI
…	…	…	…	…	…	…	…
13404	2023	WANITA	SMAS	TANGERANG SELATAN	REGULER	IPA	PSIKOLOGI

4.2 *Data Selection*

Tahap ini melibatkan pemilihan variabel-variabel yang relevan untuk membangun model prediksi peminatan program studi dengan menggunakan metode *K-Nearest Neighbour* (KNN). Proses pemilihan variabel ini bertujuan untuk menangkap karakteristik yang signifikan dari calon mahasiswa baru yang dapat mempengaruhi pemilihan program studi mereka. Variabel-variabel yang dipilih dalam proses ini meliputi berbagai

aspek yang mencakup latar belakang pendidikan, jalur penerimaan, dan preferensi individu calon mahasiswa. Variabel-variabel yang dipilih adalah sebagai berikut:

1. Tahun Penerimaan

   Tahun akademik saat calon mahasiswa mengisi data kuisioner peminatan program studi. Variabel ini penting untuk melihat peminatan program studi dari tahun ke tahun dan memahami bagaimana perubahan dalam kebijakan penerimaan atau kondisi eksternal mempengaruhi pilihan calon mahasiswa.

2. Jenis Kelamin

   Jenis kelamin calon mahasiswa. Variabel ini digunakan untuk mengidentifikasi pola peminatan berdasarkan gender, yang dapat memberikan wawasan tentang kecenderungan pemilihan program studi antara calon mahasiswa laki-laki dan perempuan.

3. Jenis Sekolah

   Jenis sekolah asal calon mahasiswa. Variabel ini membantu dalam memahami pengaruh latar belakang pendidikan terhadap pemilihan program studi.

4. Asal Sekolah

   Lokasi sekolah asal calon mahasiswa. Variabel ini berguna untuk mengidentifikasi pengaruh geografis dan institusional terhadap peminatan program studi. Misalnya, sekolah-sekolah dari daerah tertentu mungkin memiliki kecenderungan untuk memilih program studi tertentu di universitas.

5. Jalur Seleksi

   Jalur penerimaan calon mahasiswa. Variabel ini penting untuk melihat perbedaan peminatan berdasarkan jalur seleksi yang diikuti oleh calon mahasiswa. Setiap jalur seleksi mungkin memiliki karakteristik atau standar yang berbeda yang dapat mempengaruhi pilihan program studi.

6. Jurusan Sekolah

   Jurusan yang diambil calon mahasiswa saat di sekolah jenjang menengah atas. Variabel ini digunakan untuk memahami bagaimana latar belakang jurusan di sekolah menengah mempengaruhi pemilihan program studi di universitas. Misalnya, calon mahasiswa dari jurusan IPA mungkin lebih cenderung memilih program studi sains atau teknik.

7. Program Studi yang Dipilih

   Program studi pilihan calon mahasiswa pada saat mendaftar. Variabel ini adalah variabel utama yang menjadi target prediksi dalam model KNN. Informasi ini mencerminkan preferensi awal calon mahasiswa dalam memilih program studi di universitas.

Selain variabel-variabel tersebut, proses ini juga menambahkan variabel label sebagai acuan untuk mengelompokkan program studi berdasarkan tingkat peminatan. Variabel label ini dikategorikan menjadi dua kelompok utama Minat dan Kurang Diminati. Label Minat mencakup program studi yang dianggap lebih diminati oleh calon mahasiswa baru. Program studi yang termasuk dalam kategori ini adalah Ilmu Komunikasi, Manajemen, dan Psikologi. Ketiga program studi ini sering kali menjadi pilihan utama karena salah satu faktor utamanya adalah popularitas di kalangan calon mahasiswa. Di sisi lain, Label Kurang Diminati mencakup program studi yang dianggap kurang diminati oleh calon mahasiswa baru. Program studi dalam kategori ini adalah Sistem Informasi, Informatika, Teknik Sipil, Arsitektur, Akuntansi, Desain Komunikasi Visual, dan Desain Produk. Meskipun program studi ini memiliki prospek yang baik dan relevansi industri yang kuat, mereka mungkin kurang diminati oleh sebagian calon mahasiswa baru karena berbagai alasan seperti persepsi tentang tingkat kesulitan, kurangnya informasi, atau minat pribadi yang lebih rendah dalam bidang-bidang tersebut. Hasil dari proses seleksi data ditampilkan pada Gambar 4.1 di bawah ini.

	tahun	jenis_kelamin	jenis_sekolah	asal_sekolah	jalur_seleksi	jurusan_sekolah	prodi	minat
0	2020	PRIA	SMAS	SUMATERA	BEASISWA	IPA	ILMU KOMUNIKASI	DIMINATI
1	2020	PRIA	SMAS	SUMATERA	REGULER	IPA	MANAJEMEN	DIMINATI
2	2020	PRIA	SMAN	JAKARTA	NaN	IPS	PSIKOLOGI	DIMINATI
3	2020	PRIA	SMAN	JAKARTA	REGULER	IPS	ARSITEKTUR	KURANG DIMINATI
4	2020	PRIA	SMKN	BANTEN	REGULER	NaN	TEKNIK SIPIL	KURANG DIMINATI
...	...	...	...	...	...	...	...	...
13399	2023	WANITA	MAS	JAWA BARAT	REGULER	IPA	PSIKOLOGI	DIMINATI
13400	2023	PRIA	PAKET C	JAWA BARAT	REGULER	NaN	MANAJEMEN	DIMINATI
13401	2023	PRIA	SMAS	JAWA BALI	REGULER	IPA	MANAJEMEN	DIMINATI
13402	2023	WANITA	SMAN	TANGERANG	REGULER	IPA	ILMU KOMUNIKASI	DIMINATI
13403	2023	WANITA	SMAS	TANGERANG SELATAN	REGULER	IPA	PSIKOLOGI	DIMINATI

13404 rows × 8 columns

Gambar 4. 1 *Data Selection*

4.3 *Data Cleaning*

Prosedur ini bertujuan untuk menjamin data yang bebas dari kesalahan untuk penyelidikan. atau inkonsistensi yang dapat mempengaruhi hasil analisis. Proses pembersihan data ini merupakan langkah penting untuk meningkatkan kualitas data dan memastikan akurasi prediksi peminatan program studi calon mahasiswa baru. Langkah-langkah utama dalam pembersihan data meliputi menghilangkan *missing values* dan menghapus duplikat data.

1. Menghilangkan *Missing Values*

   Menghilangkan *missing values* adalah langkah kritis dalam proses *data cleaning*. *Missing values* dapat muncul dalam berbagai bentuk, seperti sel kosong atau entri dengan nilai "null" atau "N/A". Identifikasi *missing values* dilakukan dengan menggunakan fungsi-fungsi statistik yang mendeteksi sel-sel kosong atau tidak lengkap dalam dataset. Setelah mengidentifikasi *missing values*, penyebabnya dianalisis untuk menentukan pendekatan yang tepat dalam menanganinya. Dalam konteks penelitian ini, *missing values* dapat disebabkan oleh berbagai faktor, seperti kesalahan saat pengumpulan data atau entri yang tidak diisi oleh calon mahasiswa.

Beberapa strategi dapat digunakan untuk menangani missing values, termasuk isi nilai yang hilang atau gunakan alternatif yang sesuai seperti rata-rata atau median untuk catatan yang tidak memilikinya. Namun, dalam penelitian ini, diputuskan untuk menghapus record yang memiliki *missing values* guna memastikan data yang digunakan sepenuhnya lengkap dan akurat. Langkah ini diambil karena jumlah record yang memiliki *missing values* relatif kecil dan tidak signifikan terhadap ukuran keseluruhan dataset.

Proses penghapusan *missing values* dimulai dengan mengidentifikasi dan mencatat semua record yang tidak lengkap. Setelah itu, record tersebut dihapus dari dataset. Dari total 13.404 record data calon mahasiswa baru, proses pembersihan ini mengakibatkan pengurangan jumlah data menjadi 12.008 record. Dengan menghilangkan 1.396 record yang memiliki *missing values*, dataset yang tersisa lebih lengkap dan konsisten, yang akan meningkatkan akurasi dan keandalan model prediksi yang dikembangkan.

2. Menghapus Data Duplikat

Menghapus data duplikat adalah langkah krusial untuk memastikan bahwa setiap entri dalam dataset adalah unik dan valid. Data duplikat adalah entri yang identik yang muncul lebih dari sekali dalam dataset dengan informasi yang sama. Identifikasi duplikat dilakukan dengan menggunakan fungsi-fungsi khusus dalam perangkat lunak pengolahan data, seperti *drop_duplicates* di *Python* atau fitur serupa di perangkat lunak lain.

Setelah mengidentifikasi data duplikat, penyebab kemunculan duplikat dianalisis untuk memahami mengapa duplikasi terjadi. Penyebab umum termasuk kesalahan saat pengumpulan data, kesalahan input manual, atau duplikasi data saat integrasi dari berbagai sumber. Memahami penyebab ini membantu dalam mengembangkan prosedur untuk mencegah terjadinya duplikasi di masa mendatang.

Setelah duplikat diidentifikasi, langkah selanjutnya adalah menghapus record yang duplikat sehingga hanya satu entri unik yang tersisa. Dalam penelitian ini, penghapusan duplikat dilakukan secara

otomatis menggunakan fungsi perangkat lunak yang sesuai. Proses ini mengakibatkan pengurangan jumlah data lebih lanjut. Dari total 12.008 record yang telah dibersihkan dari *missing values*, identifikasi dan penghapusan duplikat mengurangi jumlah record menjadi 6.109.

Berikut merupakan jumlah data sebelum dan sesudah proses cleaning dilakukan.

```
0s  df.isnull().sum()

    tahun              0
    jenis_kelamin      0
    jenis_sekolah     12
    asal_sekolah       1
    jalur_seleksi    400
    jurusan_sekolah 1044
    prodi              1
    minat              1
    dtype: int64
```

Gambar 4. 2 *Data Before Cleaning*

```
0s  [50] df.dropna(inplace=True)
         df.drop_duplicates(inplace=True)

0s  df.isnull().sum()

    tahun             0
    jenis_kelamin     0
    jenis_sekolah     0
    asal_sekolah      0
    jalur_seleksi     0
    jurusan_sekolah   0
    prodi             0
    minat             0
    dtype: int64

1s  [53] print(df.shape)

    (6109, 8)
```

Gambar 4. 3 *Data After Cleaning*

4.4 *Data Transformation*

Pada tahap berikutnya, proses transformasi data dilakukan untuk mempersiapkan data agar dapat digunakan dalam model prediksi peminatan program studi dengan metoda *K-Nearest Neighbor* (KNN). Proses transformasi data ini sangat penting karena metode KNN bekerja dengan data numerik, sehingga semua variabel dalam dataset harus diubah ke dalam

format numerik. Pada penelitian ini, semua variabel yang digunakan merupakan tipe data objek kategorikal yang perlu diubah menjadi numerik.

Untuk mengubah variabel kategorikal menjadi numerik, digunakan teknik encoding. Teknik yang digunakan adalah *Label Encoding.* Teknik ini memberikan label numerik unik untuk setiap kategori dalam variabel. Misalnya, untuk variabel Jenis Kelamin, kategori "Laki-laki" dapat diubah menjadi 0 dan kategori "Perempuan" dapat diubah menjadi 1. Label *Encoding* sederhana dan cocok untuk variabel dengan urutan kategori yang inheren.

```
[54] df.head()
```

	tahun	jenis_kelamin	jenis_sekolah	asal_sekolah	jalur_seleksi	jurusan_sekolah	prodi	minat
0	2020	0	4	44	0	9	4	0
1	2020	0	4	44	1	9	6	0
3	2020	0	3	11	1	10	1	1
5	2020	0	3	17	0	10	4	0
6	2020	0	1	44	0	9	2	1

Gambar 4. 4 ***Data After Transformation***

Dengan melakukan transformasi data ini, dataset yang awalnya terdiri dari variabel-variabel kategorikal kini telah diubah menjadi bentuk numerik yang dapat diproses oleh model KNN. Transformasi data memastikan bahwa informasi yang terkandung dalam variabel kategorikal tetap terwakili dengan benar dalam bentuk numerik. Proses ini penting untuk meningkatkan akurasi dan efisiensi model prediksi, karena data numerik diperlukan oleh metode KNN untuk menghitung jarak antara titik data.

4.5 Penerapan Metode KNN

Metode *K-Nearest Neighbour* (KNN) diimplementasikan pada tahap ini untuk mengantisipasi peminatan calon mahasiswa baru pada program studi di Universitas XYZ. Metode KNN adalah algoritma klasifikasi yang digunakan untuk mengelompokkan data berdasarkan kedekatan atau kemiripan antara titik data. Proses penerapan KNN melibatkan beberapa

langkah penting yang memastikan model prediksi bekerja dengan optimal dan menghasilkan hasil yang akurat.

4.5.1 Menentukan Variabel X dan Y

Langkah pertama dalam penggunaan metode *K-Nearest Neighbor* (KNN) adalah menentukan variabel input (X) dan variabel output (Y) setelah data penerimaan mahasiswa baru ditransformasi. Variabel-variabel yang digunakan untuk memprediksi peminatan program studi calon mahasiswa baru disebut variabel X. Variabel-variabel tersebut mencakup Tahun Penerimaan, Jenis Kelamin, Jenis Sekolah, Asal Sekolah, Jalur Seleksi, Jurusan Sekolah, dan Program Studi yang Dipilih. Setiap variabel memiliki peran penting dalam membentuk profil calon mahasiswa, yang pada gilirannya akan digunakan oleh model KNN untuk melakukan prediksi.

Langkah selanjutnya adalah menetapkan variabel Y. Variabel Y adalah label atau kategori yang ingin diprediksi oleh model, yaitu minat calon mahasiswa terhadap program studi tertentu. Penentuan variabel Y sangat penting karena model KNN akan belajar dari data historis untuk membuat prediksi yang akurat tentang minat program studi berdasarkan variabel X.

```
[55] X = df.drop('minat', axis=1)  # Fitur
     y = df['minat']  # Label

     print("X : ", X.shape)
     print("y : ", y.shape)
```

```
X :  (6109, 7)
y :  (6109,)
```

Gambar 4. 5 Menentukan Variabel X dan Y

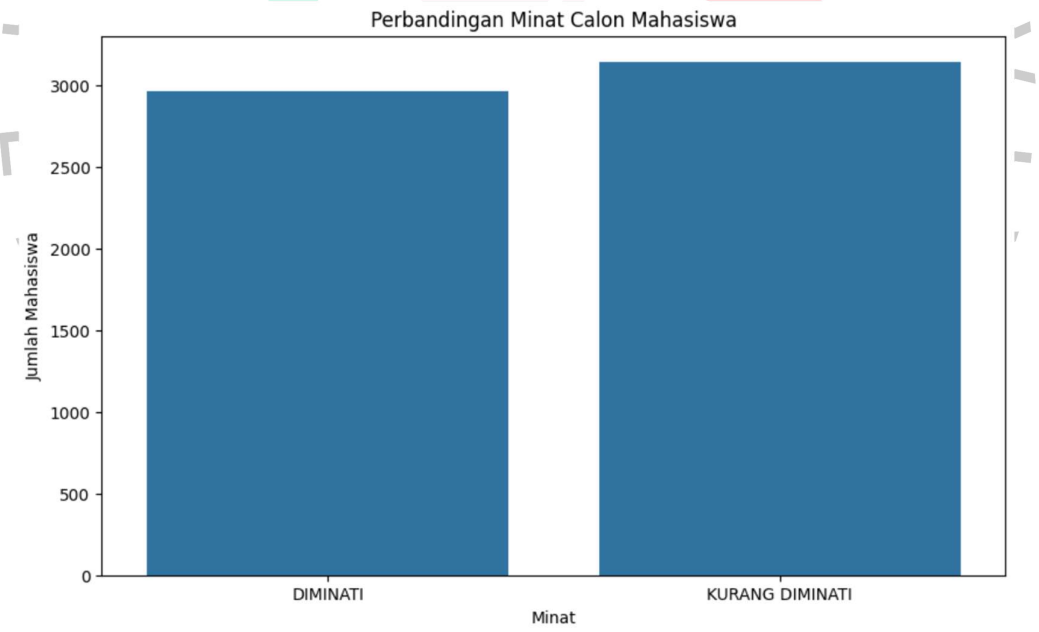
4.5.2 Mengatasi *Imbalance Data* Dengan *Smote*

Ketidakseimbangan data adalah masalah umum dalam machine learning yang terjadi ketika jumlah sampel dalam kategori yang berbeda sangat tidak merata. Dalam penelitian ini, terdapat ketidakseimbangan antara jumlah calon mahasiswa yang memilih

program studi dalam kategori "Minat" dan "Kurang Diminati". Ketidakseimbangan ini dapat menyebabkan model *K-Nearest Neighbor* (KNN) menjadi bias terhadap kategori mayoritas, sehingga mengurangi akurasi prediksi untuk kategori minoritas. Teknik *Synthetic Minority Over Sampling Technique* (SMOTE) digunakan untuk mengatasi masalah ini. SMOTE, sebuah metode oversampling, menciptakan sampel sintetis baru dari data minoritas yang ada. Dengan demikian, SMOTE membantu dalam menyeimbangkan distribusi data tanpa menambahkan sampel yang sama persis, yang dapat menyebabkan *overfitting*.

```
[45] # Menggunakan SMOTE untuk oversampling kategori minoritas
     smote = SMOTE(random_state=42)
     X_resample, y_resample = smote.fit_resample(X, y)
```

Gambar 4. 6 *Oversampling with* SMOTE



Gambar 4. 7 Variabel Minat setelah di SMOTE

4.5.3 *Scaling Dataset*

Menskalakan dataset adalah tahap selanjutnya dalam menggunakan pendekatan *K-Nearest Neighbor* (KNN) setelah ketidakseimbangan data dikoreksi menggunakan SMOTE. *Scaling*

sangat penting dalam algoritma KNN karena metode ini menggunakan jarak *Euclidean* untuk menghitung kedekatan antar titik data. Perbedaan skala antara fitur-fitur dapat mempengaruhi perhitungan jarak ini, sehingga mempengaruhi kinerja model.

Untuk melakukan scaling, digunakan *Standard Scaler*. Dengan scaling, fitur-fitur yang memiliki rentang nilai yang sangat berbeda dapat disesuaikan sehingga semuanya berada dalam skala yang sama, memungkinkan model untuk memproses data secara lebih efektif dan akurat.

Proses *scaling* dengan *Standard Scaler* melibatkan dua langkah utama, yaitu *fitting* dan *transforming*. Pertama, *scaler di-fit* pada data training untuk menghitung *mean* dan *standard deviation* dari setiap fitur. Kedua, *scaler* tersebut digunakan untuk mentransformasi data training dan data testing berdasarkan mean dan standard deviation yang telah dihitung. Data yang dilakukan scaling pada data training dan data testing dipastikan konsisten oleh langkah ini, yang penting untuk menjaga integritas model prediksi.

```
[46] # Feature Scaling menggunakan StandardScaler
     scaler = StandardScaler()
     X_resample = scaler.fit_transform(X_resample)
```

Gambar 4. 8 *Scalling Dataset*

4.5.4 Menentukan *Data Traning* dan *Testing*

Tahap berikutnya adalah membagi dataset menjadi dua, membuat set data yang terpisah untuk pelatihan dan pengujian. Pembagian ini dilakukan untuk mengevaluasi kinerja model secara obyektif. Parameter *test_size* = 0.3 dan *random_state* = 42 digunakan dalam penelitian ini untuk memastikan bahwa pembagian tersebut konsisten dan dapat direproduksi. Parameter *test_size*=0.3 menunjukkan bahwa 30% dari total dataset dialokasikan sebagai data testing, sementara 70% sisanya digunakan sebagai data training. Ini memastikan bahwa *data testing* cukup besar untuk

memberikan evaluasi yang representatif terhadap kinerja model, sementara *data training* tetap cukup besar untuk memungkinkan model belajar dengan baik dari pola yang ada. Penggunaan *random_state*=42 memastikan bahwa pembagian dataset dilakukan secara acak namun konsisten setiap kali proses ini dijalankan, sehingga hasil evaluasi dapat dibandingkan secara adil dan konsisten.

```
[63] # Membagi dataset menjadi set pelatihan dan set pengujian
     X_train, X_test, y_train, y_test = train_test_split(X_resample,
     y_resample,test_size=0.3, random_state=42)
```

Gambar 4. 9 *Data Training and Test*

4.5.5 Menentukan Nilai K

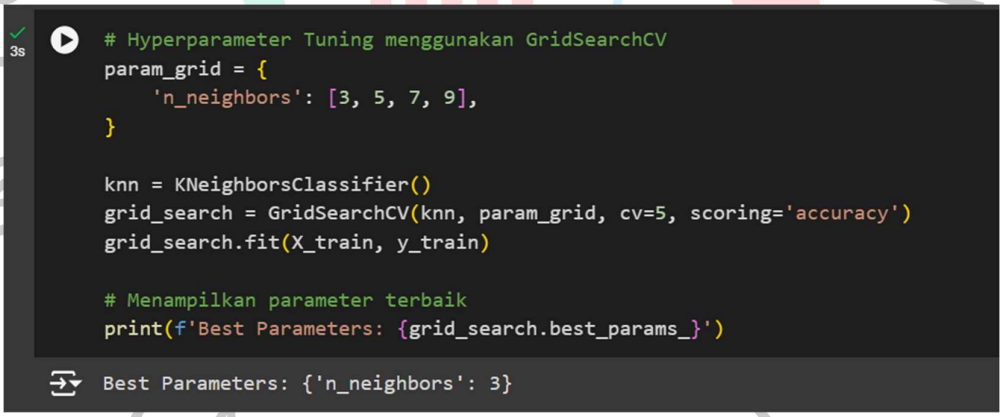
Kinerja model sangat dipengaruhi oleh pemilihan nilai k yang tepat. Jika k ditetapkan terlalu kecil, model akan menjadi terlalu sensitif terhadap *noise* sebaliknya, jika k ditetapkan terlalu besar, model akan mengabaikan informasi yang signifikan dalam data.

Untuk menentukan nilai k terbaik, digunakan *grid search* untuk menguji beberapa nilai k yang berbeda. Dalam penelitian ini, diuji nilai k yaitu 3, 5, 7, dan 9. Memilih nilai ganjil membantu menghindari masalah *tie* (seri) dalam pemungutan suara mayoritas di KNN. Dalam KNN, mayoritas label dari k tetangga terdekat menentukan hasil prediksi. Jika k bernilai genap, ada kemungkinan terjadi seri, di mana jumlah tetangga dari dua atau lebih kelas yang berbeda sama, sehingga sulit untuk menentukan kelas mayoritas.

Nilai k yang lebih kecil seperti 3 membuat model lebih fokus pada data yang sangat dekat dengan titik yang diprediksi, sehingga model bisa menangkap detail yang lebih spesifik. Namun, nilai k yang kecil juga membuat model lebih mudah terpengaruh oleh data yang tidak biasa atau salah. Model menjadi lebih stabil dan kurang terpengaruh oleh data yang tidak biasa karena penggunaan nilai k

yang lebih tinggi, seperti 9, meskipun ini bisa menyebabkan hilangnya beberapa rincian penting karena mempertimbangkan terlalu banyak data yang mungkin kurang relevan.

Setelah melakukan *grid search*, didapatkan bahwa nilai k terbaik adalah 3. Pemilihan k=3 menunjukkan bahwa model memberikan performa terbaik dalam hal akurasi prediksi peminatan program studi calon mahasiswa baru ketika mempertimbangkan tiga tetangga terdekat. Nilai k ini memberikan keseimbangan optimal antara sensitivitas terhadap data dan kemampuan generalisasi, sehingga memberikan hasil prediksi yang paling akurat. Langkah penting dalam mengembangkan model KNN adalah menentukan nilai k yang sesuai, memastikan bahwa hasil yang akurat dan berguna dapat diberikan oleh model bagi Universitas XYZ untuk memahami dan memprediksi minat program studi calon mahasiswa baru.



Gambar 4. 10 Menentukan Nilai K

4.5.6 Pelatihan Model KNN dan Classification Report

Dengan menetapkan nilai k, data pelatihan digunakan untuk mengajari model KNN. Proses pelatihan ini membantu model dalam memahami pola dan hubungan antara fitur input dan label minat. Sesudah pelatihan, model diuji dengan data uji untuk menilai kinerjanya.

Mengevaluasi model dilakukan dengan membuat laporan klasifikasi yang mencakup berbagai metrik evaluasi seperti precision, recall, f1-score, dan akurasi. Laporan klasifikasi ini menyediakan gambaran lengkap tentang performa model dalam memprediksi setiap kategori minat.

Berdasarkan hasil evaluasi, model KNN yang dilatih dengan nilai k terbaik memberikan akurasi sebesar 88%. Ini berarti model mampu memprediksi peminatan program studi calon mahasiswa baru dengan tingkat ketepatan yang tinggi. Akurasi yang dicapai menunjukkan bahwa model KNN yang dibangun efektif dalam mengenali pola-pola penting dalam data dan memberikan prediksi yang andal.

```
[69] # Pelatihan Model KNN dengan parameter terbaik
     best_knn = grid_search.best_estimator_
     best_knn.fit(X_train, y_train)

     # Evaluasi Model
     y_pred = best_knn.predict(X_test)
     accuracy = accuracy_score(y_test, y_pred)

     print(classification_report(y_test, y_pred))
     # buatkan %
     accuracy = accuracy * 100
     print(f'Accuracy: {accuracy:.2f}%')
```

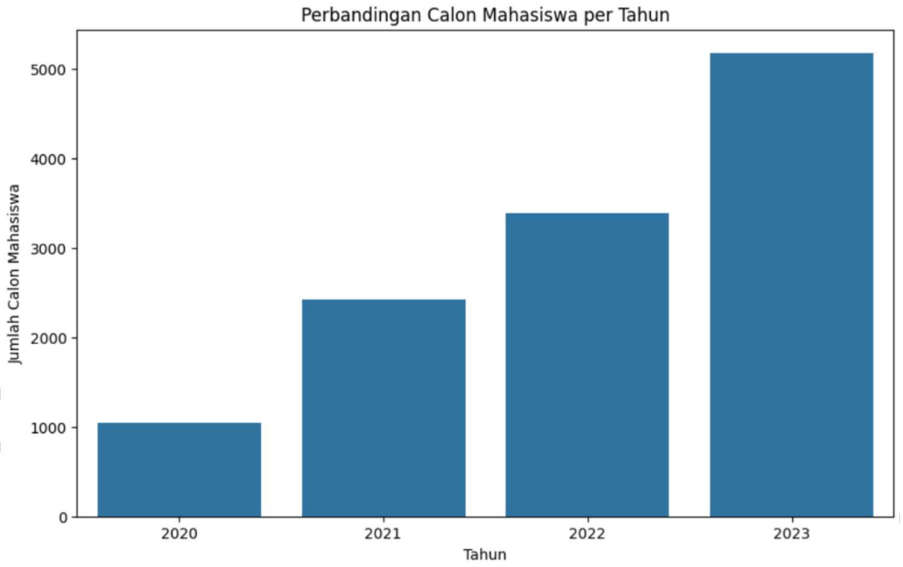
```
              precision    recall  f1-score   support

           0       0.85      0.93      0.89       946
           1       0.92      0.84      0.88       941

    accuracy                           0.88      1887
   macro avg       0.88      0.88      0.88      1887
weighted avg       0.88      0.88      0.88      1887

Accuracy: 88.18%
```

Gambar 4. 11 Akurasi Skor

4.6 *Data Visualization*

Visualisasi data adalah alat yang berguna untuk analisis dan pemahaman data. Visualisasi data digunakan untuk memberikan gambaran yang jelas tentang jumlah peminat program studi dan pertumbuhan peminat program studi di Universitas XYZ dari tahun 2020 hingga 2023. Dengan visualisasi data, informasi yang terdapat dalam data dapat diungkap dengan

lebih baik, memungkinkan pembuatan keputusan yang lebih tepat dan pemahaman yang mendalam tentang pola dan tren yang muncul difasilitasi oleh visualisasi data.



Gambar 4. 12 Perbandingan Data Calon Mahasiswa

Grafik yang ditampilkan menunjukkan perbandingan Banyak calon mahasiswa baru yang mengajukan pendaftaran ke Universitas XYZ dari tahun 2020 hingga 2023. Berikut adalah penjelasan mengenai data dalam grafik tersebut:

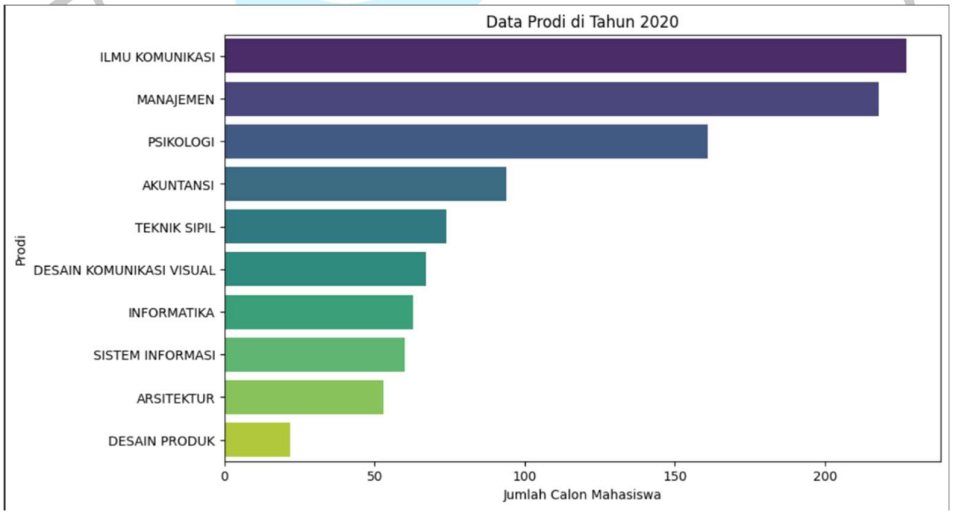
1. Pada tahun 2020, jumlah calon mahasiswa yang mendaftar adalah yang paling sedikit dibandingkan tahun-tahun lainnya, dengan total sekitar 1.000 calon mahasiswa. Ini mungkin disebabkan oleh berbagai faktor, termasuk kemungkinan dampak awal pandemi COVID-19 yang menyebabkan ketidakpastian dalam rencana pendidikan banyak calon mahasiswa.
2. Jumlah calon mahasiswa meningkat signifikan pada tahun 2021, dengan total mencapai sekitar 2.500 calon mahasiswa. Peningkatan ini menunjukkan adanya pemulihan minat dalam melanjutkan pendidikan tinggi setelah periode ketidakpastian pada tahun sebelumnya.
3. Tren peningkatan jumlah calon mahasiswa berlanjut pada tahun 2022, dengan total mencapai sekitar 3.500 calon mahasiswa. Pertumbuhan ini

dapat mencerminkan stabilisasi situasi pandemi dan adaptasi masyarakat terhadap kondisi normal baru, termasuk dalam sektor pendidikan.

4. Tahun 2023 menunjukkan puncak jumlah pendaftaran, dengan total sekitar 5.000 calon mahasiswa. Angka ini mencerminkan pertumbuhan yang paling tinggi dalam periode yang diteliti, menunjukkan kemungkinan keberhasilan strategi pemasaran dan promosi Universitas XYZ, serta meningkatnya calon mahasiswa tertarik untuk meneruskan studi mereka di universitas tersebut.

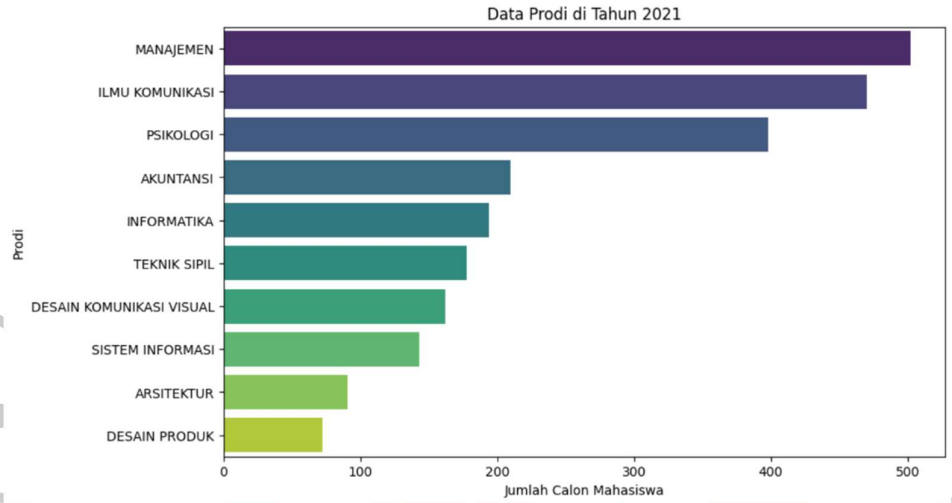
Secara keseluruhan, grafik tersebut menggambarkan tren peningkatan jumlah calon mahasiswa baru dari tahun ke tahun. Pertumbuhan yang konsisten setiap tahunnya menunjukkan adanya peningkatan kepercayaan dan minat calon mahasiswa terhadap Universitas XYZ. Faktor-faktor seperti adaptasi terhadap situasi pandemi, peningkatan kualitas promosi universitas, serta berbagai program dan fasilitas yang ditawarkan oleh universitas dapat berkontribusi terhadap tren positif ini.

Selanjutnya, dipaparkan grafik jumlah peminat program studi di Universitas XYZ dari tahun 2020 hingga 2023. Visualisasi data akan membantu kita memahami tren minat mahasiswa baru terhadap berbagai program studi.



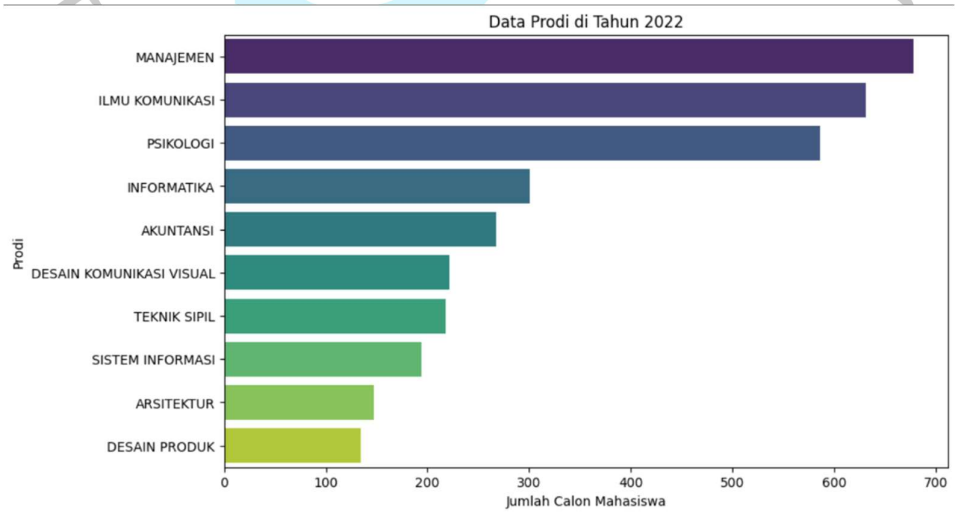
Gambar 4. 13 Jumlah Peminat Tahun 2020

Pada tahun 2020, data menunjukkan bahwa program studi dengan jumlah peminat tertinggi adalah Ilmu Komunikasi, yang memiliki jumlah peminat tertinggi. Posisi kedua ditempati oleh program studi Manajemen, sementara program studi Psikologi berada di posisi ketiga dalam hal jumlah peminat.



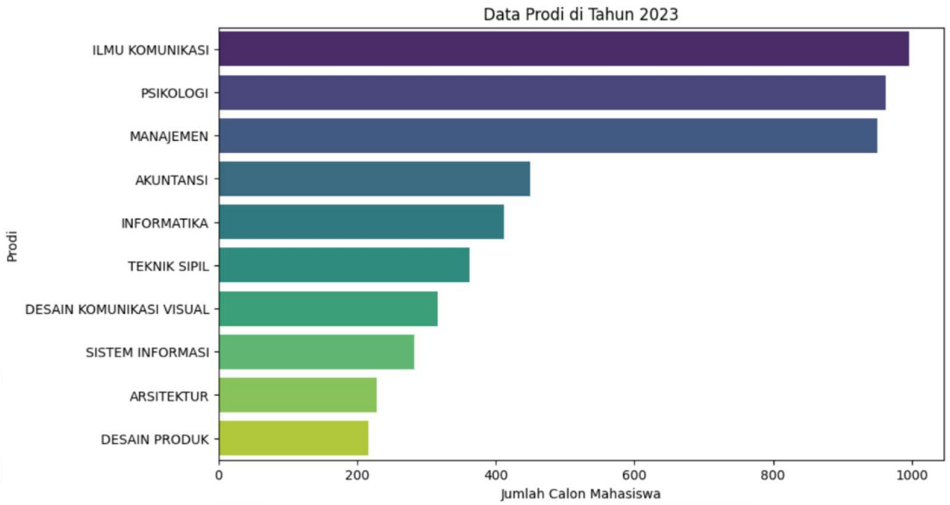
Gambar 4. 14 Jumlah Peminat Tahun 2021

Pada tahun 2021, data menunjukkan perubahan dalam jumlah peminat program studi. Program studi dengan jumlah peminat tertinggi adalah Manajemen. Posisi kedua ditempati oleh Ilmu Komunikasi, sementara Psikologi tetap berada di posisi ketiga dalam hal jumlah peminat.



Gambar 4. 15 Jumlah Peminat Tahun 2022

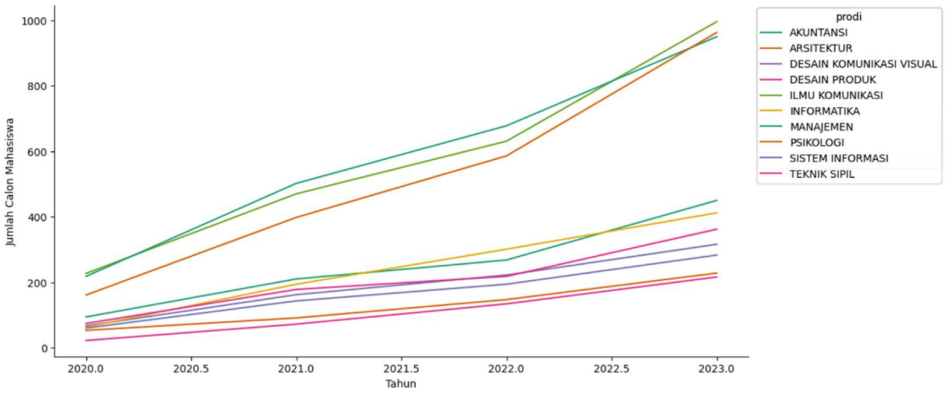
Pada tahun 2022, program studi Manajemen memiliki jumlah peminat tertinggi, diikuti oleh Ilmu Komunikasi. Kedua program studi ini hampir menyentuh 700 peminat. Psikologi tetap berada di posisi ketiga dalam hal jumlah peminat.



Gambar 4. 16 Jumlah Peminat Tahun 2023

Pada tahun 2023, terjadi perubahan signifikan dalam jumlah peminat program studi. Ilmu Komunikasi kembali menduduki peringkat pertama dengan hampir menyentuh 1000 peminat. Psikologi menempati posisi kedua, diikuti oleh Manajemen, keduanya juga hampir mencapai jumlah peminat yang sama.

Sehingga berikut merupakan grafik pertumbuhan selama 2020 hingga 2023 secara keseluruhan yang ditampilkan pada Gambar 4.17 Data Peminat Keseluruhan.



Gambar 4. 17 Jumlah Peminat Keseluruhan

Grafik tersebut menggambarkan peningkatan jumlah calon mahasiswa baru Universitas XYZ pada berbagai program studidari tahun 2020 hingga 2023. Sumbu horizontal (x) digunakan untuk menunjukkan tahun, sedangkan sumbu vertikal (y) menunjukkan jumlah mahasiswa baru (count). Berdasarkan grafik, beberapa program studi menunjukkan tren peningkatan yang stabil dan signifikan dari tahun ke tahun. Misalnya, Akuntansi dan Informatika menunjukkan peningkatan yang konsisten, sementara Desain Produk mengalami peningkatan yang sangat signifikan, terutama pada tahun 2021 dan 2023. Di sisi lain, program studi seperti Arsitektur dan Manajemen menunjukkan peningkatan yang lebih moderat. Grafik ini memberikan pemahaman yang jelas tentang preferensi mahasiswa terhadap berbagai program studi serta perubahan tren dari masa ke masa yang bisa digunakan untuk analisis lebih lanjut terkait peminatan dan strategi penerimaan mahasiswa baru di Universitas XYZ.

Setelah membahas jumlah peminat tiap tahunnya, penting untuk menganalisis pertumbuhan peminat di setiap program studi. Pemahaman terhadap pertumbuhan ini akan memberikan wawasan yang lebih mendalam mengenai dinamika minat calon mahasiswa dan membantu Universitas XYZ dalam mengidentifikasi program studi yang mengalami peningkatan atau penurunan signifikan. Berikut ini adalah analisis pertumbuhan peminat program studi dari tahun 2020 hingga 2023.

prodi tahun	AKUNTANSI	ARSITEKTUR	DESAIN KOMUNIKASI VISUAL	DESAIN PRODUK	ILMU KOMUNIKASI	INFORMATIKA	MANAJEMEN	PSIKOLOGI	SISTEM INFORMASI	TEKNIK SIPIL
2020	-	-	-	-	-	-	-	-	-	-
2021	123.4%	71.7%	141.79%	227.27%	107.05%	207.94%	130.28%	147.2%	138.33%	140.54%
2022	27.62%	61.54%	37.04%	86.11%	34.26%	55.15%	35.06%	47.24%	35.66%	22.47%
2023	67.91%	55.1%	42.34%	61.19%	57.84%	36.88%	40.12%	64.33%	45.88%	66.06%

Gambar 4. 18 Data Persentase Pertumbuhan

Tabel ini menunjukkan persentase peningkatan jumlah peminat mahasiswa baru pada berbagai program studi di Universitas XYZ. Pada tahun 2020, tidak ada data yang tersedia untuk semua program studi. Namun, pada tahun 2021, semua program studi mengalami peningkatan signifikan, dengan Desain Produk mencatat peningkatan tertinggi sebesar 227.27%, diikuti oleh Informatika (207.94%) dan Desain Komunikasi Visual (141.79%).

Selanjutnya, pada tahun 2022, terjadi penurunan persentase peningkatan di beberapa program studi dibandingkan tahun sebelumnya. Meskipun demikian, Desain Produk tetap menunjukkan peningkatan tinggi sebesar 86.11%, sementara Teknik Sipil mencatat peningkatan terendah sebesar 22.47%.

Pada tahun 2023, beberapa program studi mengalami peningkatan kembali, dengan Desain Produk mencatat peningkatan sebesar 61.19% dan Teknik Sipil meningkat menjadi 66.06%.