PRAKTIKUM MACHINE LEARNING

UNIT 2 : DATA EXAMINING (BAGIAN 2)

adult_klasifikasi.ipynb

Eka Praja Wiyata Mandala, S.Kom, M,Kom, CADS

Unit 2 : Data Examining

3. Analisis Korelasi

```
[ ]  # Korelasi antara variabel numerik
     correlation_matrix = df[numeric_features].corr()
     plt.figure(figsize=(12, 10))
     sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
     plt.title('Correlation Matrix of Numeric Features')
     plt.show()
```

Unit 2 : Data Examining

catatan:

Observasi Utama:

- Sebagian besar korelasi antar variabel sangat lemah (warna biru muda)
- Tidak ada korelasi yang sangat kuat antar variabel (tidak ada warna merah tua selain diagonal)

Korelasi Spesifik:

- Korelasi tertinggi adalah antara 'education_num' dan 'hours_per_week' (0.15)
- 'age' memiliki korelasi lemah positif dengan 'capital_gain' (0.078)
- 'fnlwgt' memiliki korelasi sangat lemah dengan semua variabel lain

Unit 2 : Data Examining

Implikasi:

- Multikolinearitas tidak menjadi masalah besar dalam dataset ini Setiap variabel cenderung memberikan informasi yang unik Decision tree mungkin perlu mempertimbangkan interaksi antar variabel

Insight untuk Modeling:

- Feature selection (mungkin) tidak terlalu efektif karena rendahnya korelasi

Unit 2 : Data Examining

4. Analisis Hubungan dengan Variabel Target

```
[ ]  # Boxplot numerik vs target
     for feature in numeric_features:
         plt.figure(figsize=(10, 6))
         sns.boxplot(x='income', y=feature, data=df)
         plt.title(f'{feature} vs Income')
         plt.show()
```

Unit 2 : Data Examining

cara baca:

Variabel yang Ditampilkan:

- Sumbu Y: 'fnlwgt' (final weight) dan fitur lainnya. disini fokus pada fnlweight dulu
- Sumbu X: 'income' dengan dua kategori (<=50K dan >50K)

Struktur Box Plot:

- Kotak menunjukkan interquartile range (IQR) - dari kuartil pertama (Q1) ke kuartil ketiga (Q3)
- Garis horizontal di dalam kotak adalah median
- Whiskers (garis vertikal) menunjukkan range data di luar IQR
- Titik-titik di atas whiskers adalah outlier

Unit 2 : Data Examining

Perbandingan antara Dua Kategori Income:

- Kedua kategori income menunjukkan distribusi 'fnlwgt' yang sangat mirip
- Median (garis tengah kotak) untuk kedua kategori hampir sama
- IQR (ukuran kotak) juga sangat mirip untuk kedua kategori

Outlier:

- Kedua kategori memiliki banyak outlier yang ditunjukkan oleh titik-titik di atas whiskers
- Outlier tersebar dari sekitar 0.6 x 10^6 hingga 1.4 x 10^6 untuk kedua kategori

Distribusi:

- Distribusi 'fnlwgt' cenderung miring ke atas (right-skewed) untuk kedua kategori income
- Mayoritas nilai 'fnlwgt' terkonsentrasi di bawah 0.4 x 10^6

Unit 2 : Data Examining

Interpretasi:

- Tidak ada perbedaan signifikan dalam distribusi 'fnlwgt' antara orang dengan pendapatan <=50K dan >50K
- Nilai 'fnlwgt' tidak tampak menjadi prediktor kuat untuk kategori pendapatan

Implikasi:

- 'fnlwgt' mungkin tidak terlalu informatif dalam membedakan antara dua kategori pendapatan
- Faktor-faktor lain mungkin lebih berpengaruh dalam menentukan kategori pendapatan

Unit 2 : Data Examining

Catatan:

- 'fnlwgt' adalah bobot statistik yang menunjukkan berapa banyak orang di populasi yang direpresentasikan oleh satu record
- Nilai yang tinggi menunjukkan bahwa record tersebut mewakili lebih banyak orang dalam populasi

Kesimpulannya, berdasarkan plot ini, 'fnlwgt' tidak menunjukkan perbedaan distribusi yang jelas antara dua kategori pendapatan, yang mengindikasikan bahwa variabel ini mungkin tidak terlalu berguna sebagai prediktor tunggal untuk income dalam model prediktif.

Unit 2 : Data Examining


```
[ ]  # Stacked bar plot kategorikal vs target
     for feature in categorical_features:
         if feature != 'income':
             plt.figure(figsize=(12, 6))
             df_temp = df.groupby([feature, 'income']).size().unstack()
             df_temp_perc = df_temp.div(df_temp.sum(axis=1), axis=0)
             df_temp_perc.plot(kind='bar', stacked=True)
             plt.title(f'{feature} vs Income')
             plt.xlabel(feature)
             plt.ylabel('Percentage')
             plt.legend(title='Income', loc='upper right')
             plt.xticks(rotation=45)
             plt.show()
```

Unit 2 : Data Examining

catatan:

Observasi Penting:

- 'Self-emp-inc' (self-employed incorporated) memiliki proporsi tertinggi untuk pendapatan >50K
- 'Never-worked' dan 'Without-pay' hampir seluruhnya memiliki pendapatan <=50K
- Pekerjaan pemerintah (Federal-gov, Local-gov, State-gov) memiliki proporsi pendapatan >50K yang cukup signifikan

Insight Spesifik:

- Sekitar 40% pekerja 'Federal-gov' memiliki pendapatan >50K
- 'Private' sector, yang mungkin merupakan kategori terbesar, memiliki proporsi pendapatan >50K yang lebih rendah dibanding pekerjaan pemerintah
- 'Self-emp-not-inc' (self-employed not incorporated) memiliki proporsi pendapatan >50K yang lebih rendah dibanding 'Self-emp-inc'

Unit 2 : Data Examining

Implikasi untuk Analisis:

- 'Workclass' tampaknya menjadi prediktor yang baik untuk level pendapatan
- Kategori seperti 'Never-worked' dan 'Without-pay' mungkin bisa digabungkan karena pola yang sangat mirip
- Perbedaan antara jenis pekerjaan pemerintah mungkin penting untuk dipertahankan dalam analisis

Pertimbangan untuk Modeling:

- 'Workclass' bisa menjadi fitur yang berguna dalam model prediksi pendapatan Mungkin perlu teknik encoding khusus untuk menangkap nuansa antar kategori

Unit 2 : Data Examining

5. Identifikasi Outlier

```
[ ]  # Box plot untuk mendeteksi outlier pada variabel numerik
     plt.figure(figsize=(15, 10))
     df[numeric_features].boxplot()
     plt.title('Box Plots of Numeric Features')
     plt.xticks(rotation=90)
     plt.show()
```

Unit 2 : Data Examining

penjelasan:

Analisis per Fitur:

a. fnlwgt:

- Memiliki range nilai terbesar
- Banyak outlier di atas whisker atas
- Distribusi miring ke kanan (right-skewed)
- IQR (kotak) relatif kecil dibanding range keseluruhan

Unit 2 : Data Examining

b. age, education_num, hours_per_week:

- Range nilai relatif kecil
- Distribusi terlihat lebih seimbang
- Sedikit atau tidak ada outlier yang terlihat

c. capital_gain dan capital_loss:

- Mayoritas nilai terkonsentrasi di sekitar 0
- Beberapa outlier signifikan di atas
- Menunjukkan bahwa sebagian besar orang tidak memiliki capital gain/loss, tapi beberapa memiliki nilai yang sangat tinggi

Unit 2 : Data Examining

6.Analisis dan Interpretasi

Implikasi untuk Analisis:

- Mungkin perlu normalisasi atau standardisasi fitur sebelum modeling, terutama untuk fnlwgt
- Perlu pertimbangan khusus untuk menangani outlier di fnlwgt, capital_gain, dan capital_loss
- Fitur seperti age, education_num, dan hours_per_week mungkin lebih mudah diinterpretasi dalam model

Potensi Preprocessing:

- Transformasi log mungkin berguna untuk fnlwgt, capital_gain, dan capital_loss
- Binning atau kategorisasi mungkin bermanfaat untuk capital_gain dan capital_loss

Insight Bisnis:

- Distribusi fnlwgt menunjukkan variasi besar dalam representasi populasi
- Pola capital_gain dan capital_loss menunjukkan ketimpangan ekonomi dalam sampel