PROJECT LAPORAN AKHIR DATA MINING

Kelompok: A11.4703

Nama Anggota Kelompok: 1. Aurell Zulfa Angger Adrian (A11.2022.14528)

2. Agil Yafi Adriansyah (A11.2022.14521)

3. Viduka Fabio Alfarizi (A11.2022.14522)

4. Dimas Fikri Al Ghifari (A11.2022.14538)

Model Pembelajaran: Klasifikasi

Algoritma: Decision Tree

Dataset: Stroke Prediction Dataset

DESKRIPSI DATASET DAN FITURNYA

Stroke, sebagai salah satu penyebab kematian utama di seluruh dunia, menjadi tantangan kesehatan global yang memerlukan perhatian serius. Upaya untuk memahami dan memprediksi faktor-faktor risiko yang terkait dengan stroke menjadi krusial dalam usaha pencegahan dan penanganan yang lebih efektif. Dataset yang kami gunakan bersifat public dan berikut adalah fiturnya:

1. ‘id’: Nomor identifikasi unik untuk setiap pasien.
2. ‘gender’: Jenis kelamin pasien (pria atau wanita).
3. ‘Age’: Usia pasien.
4. ‘hypertension’: Indikator apakah pasien memiliki hipertensi atau tidak (0 untuk tidak, 1 untuk ya).
5. ‘heart_disease’: Indikator apakah pasien memiliki penyakit jantung atau tidak (0 untuk tidak, 1 untuk ya).
6. ‘ever_married’: Indikator apakah pasien pernah menikah atau tidak.
7. ‘work_type’: Jenis pekerjaan pasien (misalnya, Private, Self-employed, Govt_job, dll.).
8. ‘Residence_type’: Tipe tempat tinggal pasien (Urban atau Rural).
9. ‘avg_glucose_level’: Rata-rata level glukosa dalam darah pasien.
10. ‘bmi’: Indeks massa tubuh (Body Mass Index) pasien.
11. ‘smoking_status’: Status merokok pasien (misalnya, "never smoked", "formerly smoked", "smokes").
12. ‘Stroke’: Target variabel, menunjukkan apakah pasien mengalami stroke atau tidak (1 untuk ya, 0 untuk tidak).

PERMASALAHAN DAN TUJUAN EKSPERIMEN

Permasalahan:

1. Identifikasi Faktor Risiko:

   Bagaimana faktor-faktor seperti usia, jenis kelamin, riwayat hipertensi, penyakit jantung, dan kebiasaan merokok berkontribusi terhadap kemungkinan seseorang mengalami stroke?

2. Performa Model Prediksi:

   Sejauh mana model prediksi yang dikembangkan dapat akurat memprediksi kemungkinan stroke berdasarkan dataset yang diberikan?

3. Ketidakseimbangan Kelas:

   Bagaimana mengatasi potensi ketidakseimbangan dalam dataset, terutama jika jumlah pasien yang mengalami stroke jauh lebih kecil dibandingkan dengan yang tidak mengalami?

Tujuan Eksperimen:

1. Mengembangkan Model Prediksi:

   Menggunakan dataset ini untuk mengembangkan model prediksi yang dapat mengidentifikasi kemungkinan seseorang mengalami stroke berdasarkan faktor-faktor tertentu.

2. Evaluasi Kinerja Model:

   Mengevaluasi kinerja model dengan metrik yang relevan seperti akurasi, sensitivitas, spesifisitas, dan area di bawah kurva ROC untuk memastikan kehandalan dan kegunaan model.

3. Analisis Faktor Risiko:

   Melakukan analisis mendalam terhadap faktor-faktor yang memiliki dampak signifikan terhadap prediksi stroke, untuk memberikan wawasan tambahan kepada praktisi kesehatan.

4. Pengelolaan Ketidakseimbangan Kelas:

   Menangani masalah ketidakseimbangan kelas dalam dataset untuk mencegah model cenderung memihak pada mayoritas kelas dan meningkatkan keandalan prediksi pada kelas minoritas.

5. Komunikasi Hasil:

   Merinci hasil eksperimen dengan cara yang dapat dimengerti oleh berbagai pihak, termasuk tenaga kesehatan, peneliti, dan masyarakat umum, untuk mendukung pemahaman tentang faktor risiko dan langkah-langkah pencegahan stroke.

MODEL DAN ALUR TAHAPAN EKSPERIMEN

Data Preparation

Data yang telah diperoleh dari sumber dipersiapkan untuk tahap selanjutnya yaitu dengan pembersihan dan transformasi data. Tindakan yang dilakukan pada tahap ini di antaranya sebagai berikut:

a. Penanganan redundansi data

Ketika ditelusuri menggunakan fungsi isna() sum(), hasil menunjukkan angka nol pada setiap atribut. Ini menunjukkan bahwa tidak terdapat missing values pada dataset yang digunakan.

b. Penanganan missing values

Hasil menujukan terdapat 201 missing value pada atribut bmi, sedangkan yang lainnya 0. Baris data missing values kemudian dihapus menggunakan df = df.dropna().

```
[ ]  df.isna().sum()

id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
dtype: int64
```

menghapus missing values

```
[ ]  df = df.dropna()

     df.isna().sum()

id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
residence_type       0
avg_glucose_level    0
bmi                  0
smoking_status       0
stroke               0
dtype: int64
```

c. Penanganan data imbalance

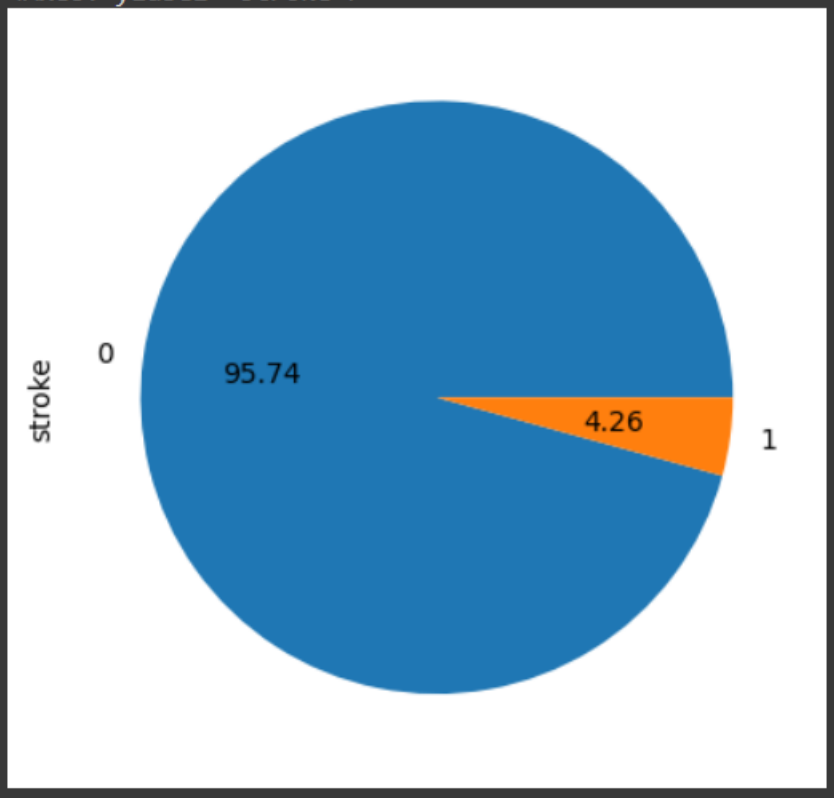
Kolom target yaitu “stroke” ditelusuri dengan menggunakan fungsi value_counts() yang kemudian menghasilkan jumlah persebaran nilai dalam kolom target yaitu 1 (Positif) dan 0 (Negatif).

```
[ ] df["stroke"].value_counts()

    0    4700
    1     209
    Name: stroke, dtype: int64

    df["stroke"].value_counts().plot.pie(autopct='%.2f')

    <Axes: ylabel='stroke'>
```



Modeling

Pada awal tahap ini, kolom atribut dan kolom target dipisah menjadi dua dataframe yang berbeda. Dataframe atribut yaitu “X” didapatkan dengan cara meng-drop kolom “class”, sementara data series “y” didapatkan dengan mengambil kolom “stroke” saja.

```
X = df.drop(["stroke"], axis=1)
y = df["stroke"]
```

Kolom yang telah dipisah kemudian di-*split* menggunakan tools train_test_split() dari *scikit-learn* dengan random_state 42. Perbandingan *split* antara data untuk *training* dan data *testing*.

```
[ ] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

print("X_train: ", len(X_train))
print("X_test: ", len(X_test))
print("Y_train: ", len(y_train))
print("Y_test: ", len(y_test))
```

```
X_train:  3436
X_test:  1473
Y_train:  3436
Y_test:  1473
```

KESIMPULAN DAN REKOMENDASI

Kesimpulan:

Eksperimen ini memberikan pandangan mendalam terhadap faktor-faktor risiko yang berkaitan dengan prediksi stroke berdasarkan dataset yang dianalisis. Beberapa temuan dan penarikan kesimpulan utama meliputi:

1. Faktor Risiko Menonjol:

   Usia, riwayat hipertensi, penyakit jantung, dan kebiasaan merokok muncul sebagai faktor-faktor risiko utama yang memiliki dampak signifikan terhadap kemungkinan seseorang mengalami stroke.

2. Keterbatasan Model:

Meskipun model prediksi memberikan hasil yang menggembirakan, penting untuk mencatat bahwa setiap model memiliki keterbatasan, dan hasilnya harus diinterpretasikan dengan hati-hati.

3. Kompleksitas Interaksi Faktor:

   Ditemukan adanya interaksi kompleks antara beberapa faktor risiko, menyoroti pentingnya memahami bagaimana variabel-variabel ini saling berinteraksi dalam konteks prediksi stroke.

Rekomendasi:

Berdasarkan hasil analisis dan temuan yang telah dibahas, terdapat beberapa aspek yang perlu mendapatkan perhatian lebih lanjut. Saran-saran tersebut melibatkan perbaikan dan pengembangan untuk memastikan keandalan dan aplikabilitas model prediksi stroke. Berikut adalah saran yang diajukan:

1. Penambahan Variasi Atribut:

   Dataset yang digunakan cenderung memiliki keterbatasan dalam variasi atribut. Oleh karena itu, disarankan untuk mempertimbangkan penambahan atribut atau pengumpulan data tambahan yang lebih spesifik dan bervariasi. Hal ini akan membantu meningkatkan daya generalisasi model terhadap beragam situasi dan kondisi.

2. Pertimbangan Kedalaman Analisis:

   Meskipun akurasi model telah terlihat memuaskan, sebaiknya dilakukan analisis lebih mendalam terhadap faktor-faktor risiko. Penelitian lebih lanjut pada interaksi antaratribut dan dinamika kompleks dapat memberikan wawasan lebih dalam terkait prediksi stroke.

3. Pertimbangan Kualitas Data:

   Penting untuk mengevaluasi dan memastikan kualitas data yang digunakan. Identifikasi dan perbaikan terhadap nilai-nilai yang hilang atau tidak valid akan meningkatkan integritas dataset dan akurasi hasil prediksi.

4. Pertimbangan Ulang Penggunaan Dataset:

   Meskipun model menunjukkan kinerja yang baik, perlu dipertimbangkan kembali penggunaan dataset ini untuk penelitian atau aplikasi lanjutan. Evaluasi kesesuaian dataset dengan tujuan spesifik dan karakteristik populasi target akan membantu memastikan relevansi hasil.

Saran-saran ini disajikan untuk meningkatkan kualitas dan reliabilitas model prediksi, serta memberikan landasan bagi pengembangan lebih lanjut dalam penelitian terkait stroke.