Final Project Presentation

Kelompok Suka Kedistrak

Nama : Putrija BR Malau (211402063)
Gabryelle Ninna Siahaan (211402087)
Erli Gurning (211402123)
Vincent Enrique Shie (211402111)

Daftar Isi

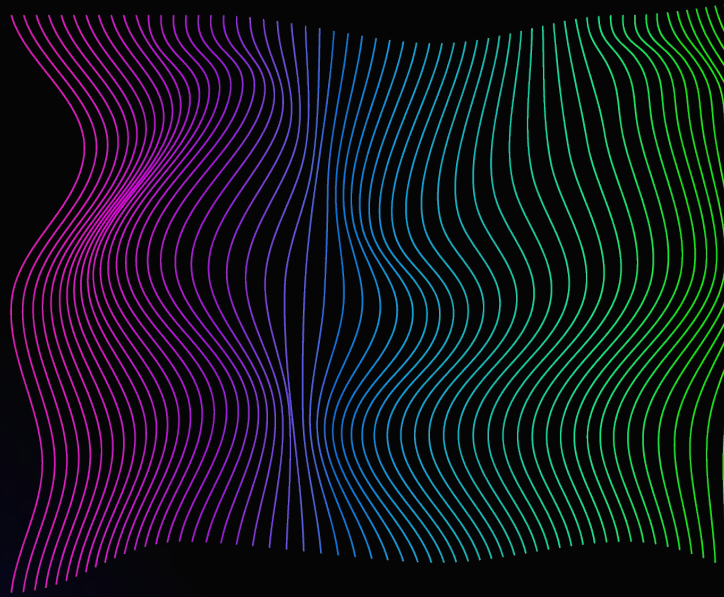
01 Latar Belakang

02 Eksplorasi Data dan visualisasi

03 Modeling

04 Kesimpulan

Latar Belakang



Latar Belakang

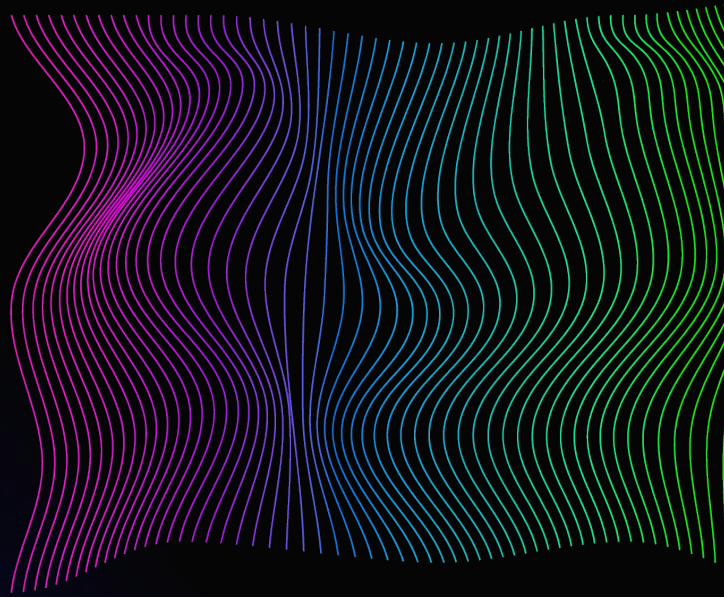
Sumber data : https://drive.google.com/file/d/14ySas0DRD6eJ7ysw22JSN2G_OZ5tLHIR/view?usp=sharing

Problem : classification

Tujuan :

- Diabetes adalah penyakit kronis yang ditandai dengan tingginya kadar gula darah. Glukosa merupakan sumber energi utama bagi sel tubuh manusia. Akan tetapi, pada penderita diabetes, glukosa tersebut tidak dapat digunakan oleh tubuh. Penyakit diabetes sangat salah satu penyakit tidak menular yang banyak diderita oleh penduduk di seluruh dunia.
- Proyek ini akan membahas mengenai memprediksi seseorang sedang mengidap penyakit diabetes berdasarkan data pengecekan kesehatan rutin, usia, suara mendengkur, tekanan darah, indeks massa tubuh, waktu tidur, aktivitas fisik, tekanan darah, keturunan diabetes, stress, kuantitas memakan junk food, usia, frekuensi kencing, kehamilan, alkohol, merokok dan pdiabetes yang saling berkorelasi.

Eksplorasi Data dan Visualisasi



Business Understanding

- Diabetes adalah penyakit kronis yang terjadi ketika pankreas tidak memproduksi cukup insulin atau ketika tubuh tidak dapat menggunakan insulin yang dihasilkan secara efektif. Insulin adalah hormon yang mengatur glukosa darah.
- Faktor-faktor yang menyebabkan seseorang terkena penyakit diabetes yaitu genetik, obesitas, kurangnya aktivitas fisik, pola makan, tekanan darah tinggi, kolesterol, stress dan lain-lain.
- Pertumbuhan penduduk penderita diabetes di seluruh dunia terus meningkat tiap tahunnya. Penderita diabetes tidak memandang gender, usia dan latar belakang, Sehingga kita perlu mengantisipasi resiko penyakit diabetes.

Data Cleansing

- Pada dataset yang dimiliki, perlu dilakukan data cleansing karena terdapat beberapa missing value pada dataset yang dimiliki.

- Terdapat missing value pada kolom BMI sebanyak 4 buah data, kolom pregnacies sebanyak 42 data, kolom pdiabetes sebanyak 1, dan diabetic sebanyak 1, sehingga kami melakukan penghapusan data yang memiliki nilai null.

Terdapat penulisan data yang memiliki makna yang sama namun ditulis dengan format berbeda pada kolom diabetic, regularmedicine, BPevel, dan Pdiabetes. Sehingga kami menyeragamkan value yang dimiliki.

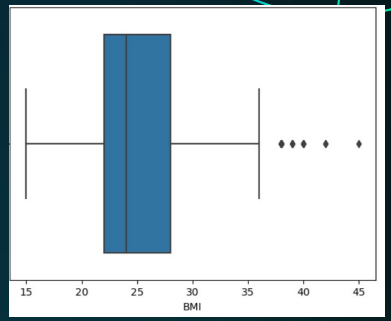
```
['no' 'yes' ' no']
no     641
yes    263
 no      1
Name: Diabetic, dtype: int64
['no' 'yes' 'o']
no     581
yes    323
o        1
Name: RegularMedicine, dtype: int64
['high' 'normal' 'low' 'Low' 'High' 'normal ']
normal     667
high       205
low         25
High         5
Low          2
normal       1
Name: BPLevel, dtype: int64
['0' 'yes']
0      891
yes     14
Name: Pdiabetes, dtype: int64
```

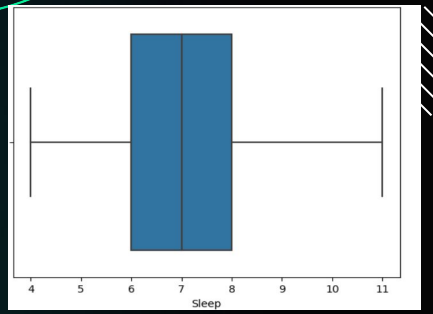
Data Cleansing

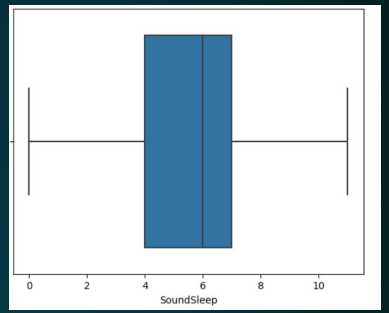
- Kami memiliki outlier namun hanya sedikit, sehingga kami tidak menghapus outlier karena masih memiliki hubungan dengan data kami.

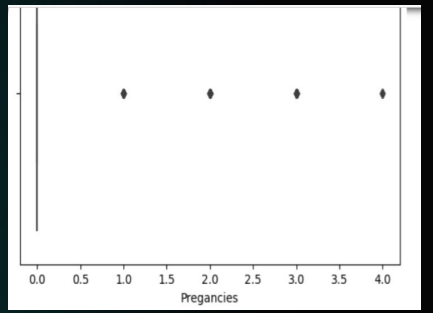
- Kami memiliki data duplikat yang berjumlah 632 data dari 905 data. Namun kami tidak menghapus data duplikat dengan alasan data duplikat hampir 70% dari seluruh data, yang apabila dihapus akan mengubah akurasi data.

Kolom pregnancies memiliki tipe data float namun semua value yang dimiliki adalah integer, sehingga Kami mengubah tipe data pregnancies yang semula dari float menjadi integer .



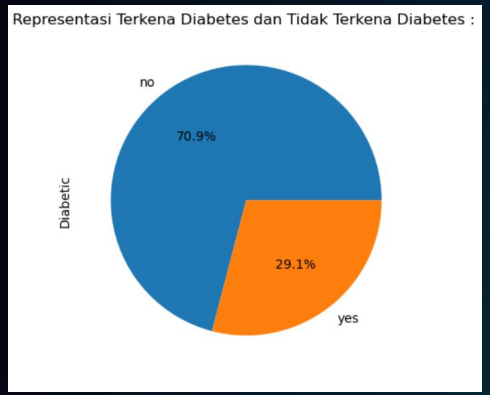


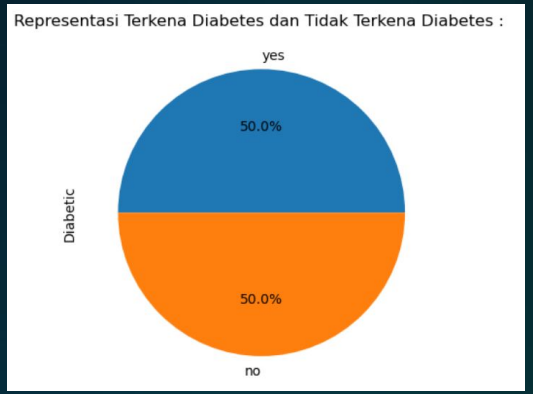




Exploratory Data Analysis

- Data yang kita miliki tidak imbalance antara yang terkena diabetes dan tidak diabetes, sehingga kami melakukan oversampling data.

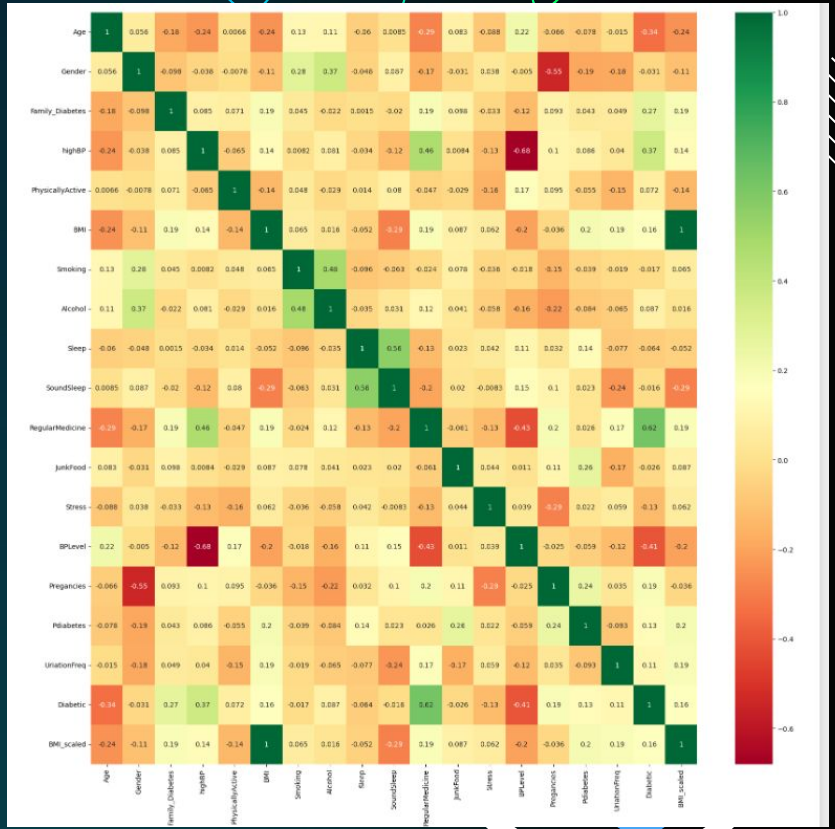




Exploratory Data Analysis

Dari hasil eksplorasi data, ditemukan beberapa insight yang ditemukan yaitu hubungan antar kolom yang dapat menyebabkan penyakit diabetes. Berikut merupakan

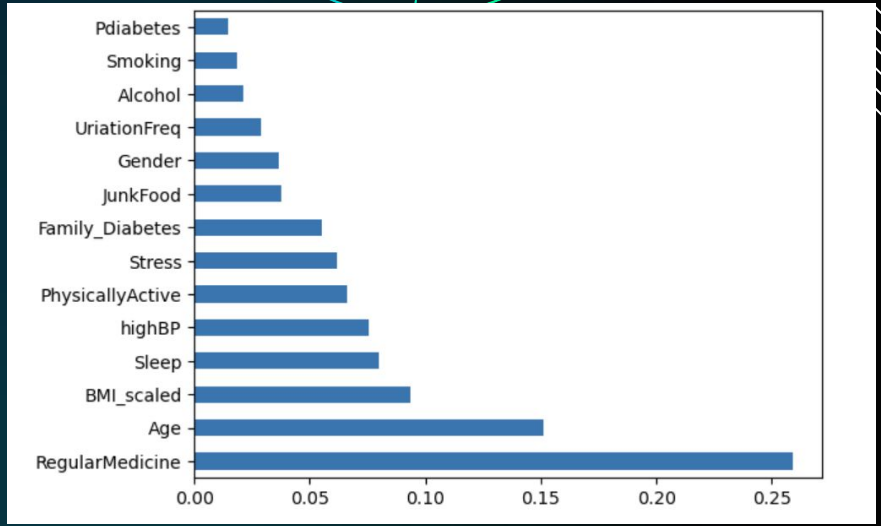
1. Konsumsi obat harian memiliki korelasi yang cukup dengan tekanan darah tinggi yaitu : 0.46
2. Alkohol dan gender memiliki korelasi yang cukup yaitu : 0.37
3. Diabetes dan tekanan darah tinggi memiliki korelasi cukup yaitu : 0.37
4. Konsumsi alkohol dan penggunaan rokok memiliki korelasi cukup yaitu : 0.48
5. Suara tidur dan tidur memiliki korelasi yang kuat yaitu : 0.56
6. Konsumsi obat harian dan diabetes memiliki korelasi yang kuat yaitu : 0.62



Exploratory Data Analysis

Dari hasil eksplorasi data, ditemukan beberapa insight yang ditemukan yaitu :

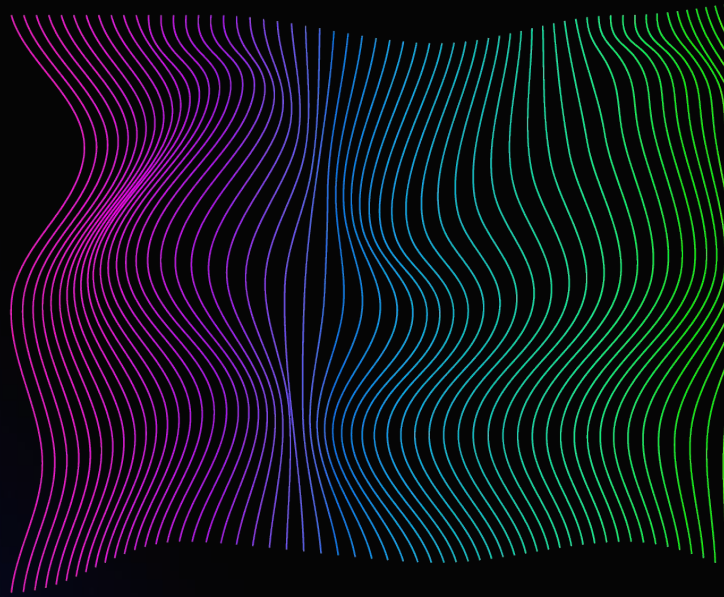
1. Jika Konsumsi obat sehari-hari, usia semakin tua, indeks massa tubuh, durasi tidur dan tekanan darah tinggi semakin tinggi maka seseorang akan cenderung terkena penyakit diabetes.
2. Seseorang harus menjaga pola hidup sehat dan memperhatikan 14 data disamping agar tetap stabil sehingga jauh dari resiko terkena diabetes



Preprocessing

- Untuk melakukan tahap modeling kami memerlukan data dengan tipe numerik. Namun, data kami sebagian besar dan target data merupakan tipe data objek. Kami melakukan encoding dengan LabelEncoder.
- Berikutnya kami melakukan normalisasi, dikarenakan pada kolom BMI memiliki rentang yang berbeda dari kolom yang lainnya. Kami menggunakan normalisasi MinMaxScaler.

Modeling

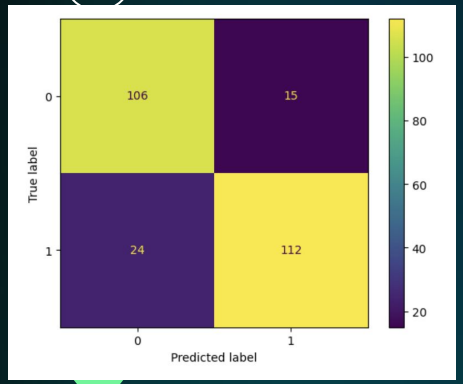


Train test split

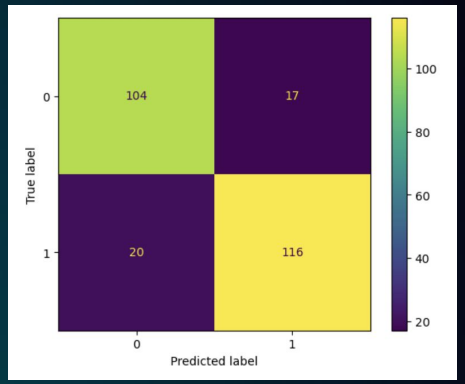
Menggunakan metode train_test_split yang merupakan bagian dari library Scikit-Learn di Python. Fungsinya adalah untuk memisahkan dataset menjadi data training dan data testing sesuai dengan proporsi yang ditentukan, dalam hal ini dengan proporsi 80% data untuk training dan 20% untuk testing, yang dapat diatur dengan argumen test_size=0.2.

Metrik evaluasi

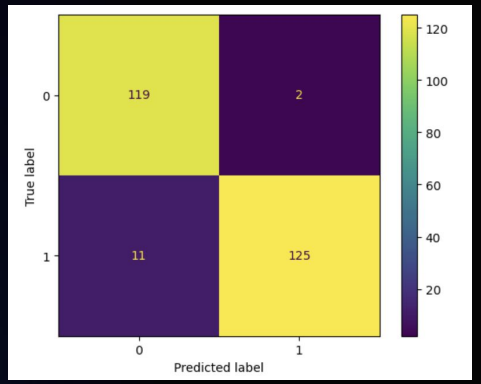
Menggunakan evaluasi **confusion matrix** memberikan gambaran yang lebih mendetail tentang bagaimana model klasifikasi Anda melakukan prediksi terhadap data uji yang memiliki label yang diketahui.



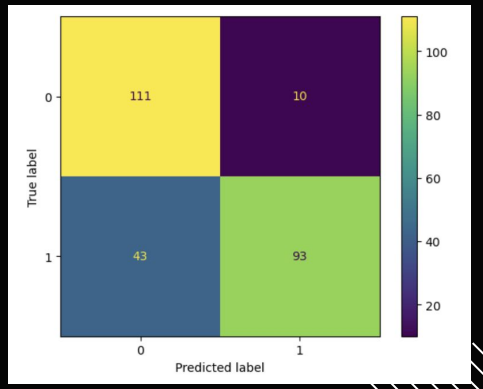
Confusion Matrix SVM



Confusion Matrix Logistic Regession



Confusion Matrix Decision Tree



Confusion Matrix Naive Bayes

Model yang dicoba

Nama Model	Akurasi	Presisi	Recall	AUC-ROC
Decision Tree	0.9494163424124513	0.9842519685039370	0.9191176470588235	0.9512943607194944
SVM	0.8482490272373541	0.8818897637795275	0.8235294117647058	0.8497812348079727
Logistic Regression	0.8560311284046692	0.8721804511278195	0.8529411764705882	0.8562226543509966
Naive Bayes	0.7937743190661478	0.9029126213592233	0.6838235294117647	0.8005894506562956

Hyperparameter Tuning

	Model	Grid	Random	Successive
Decision Tree	0.9494163	0.945525	0.945525	0.9416342
SVM	0.848249	0.937743	0.933852	0.9455253
Logistic Regression	0.8560311	0.821012	0.821012	0.8210117
Naïve Bayes	0.7937743	0.797665	0.797665	0.7976654

Model Final

Model Final : Decision Tree

Akurasi	: 0.9494163424124513
Presisi	: 0.984251968503937
Recall	: 0.9191176470588235
AUC-ROC	: 0.9512943607194944

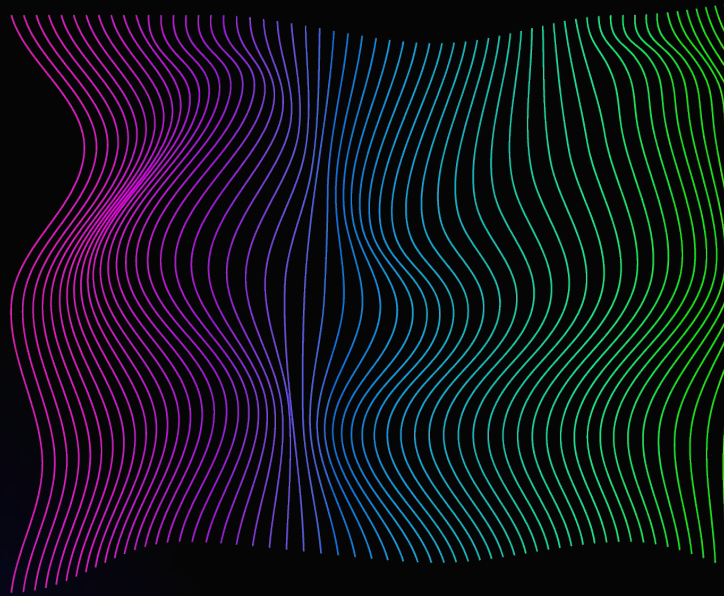
Dari Important Feature ditemukan ada 3 kolom yang tidak perlu digunakan pada model yaitu BPLevel, Pregancies, SoundSleep karena ada korelasinya terhadap kolom lain.

Maka kolom yang menjadi prediktor adalah:

1. RegularMedicine
2. Age
3. BMI
4. Sleep
5. highBP
6. PhysicallyActive
7. Stress
8. Family_Diabetes
9. JunkFood
10. Gender
11. UriationFreq
12. Alcohol
13. Smoking
14. Pdiabetes

Target tabel adalah Diabetic

Conclusion



Insight/trend menarik

Pola hidup, pola konsumsi serta usia sangat mempengaruhi kecenderungan seseorang terkena penyakit diabetes. Konsumsi obat sehari-hari perlu dikontrol, usia dapat menjadi faktor pendukung seseorang terkena diabetes, indeks massa tubuh perlu dikendalikan agar tetap ideal, durasi tidur harus tetap stabil dan terkontrol dan tekanan darah tinggi harus selalu stabil.

Saran kepada stakeholder

Dari data yang dipaparkan, ditemukan klasifikasi antara orang-orang yang menderita penyakit diabetes dan yang tidak menderita penyakit diabetes.

Seseorang perlu menjaga kadar Konsumsi obat sehari-hari agar tidak berlebih,

- Memperhatikan pola hidup dan dan pola makan di usia yang semakin tua, karena diabetes cenderung menyerang mereka yang mulai lanjut usia
- Kita harus menjaga pola makan agar indeks massa tubuh tetap stabil dan tidak melebihi indeks massa tubuh normalnya
- Sebaiknya kita menjaga durasi tidur yang normal untuk mencegah penyakit diabetes
- dan tekanan darah tinggi semakin tinggi maka beresiko terkena diabetes, maka hindari makanan dan pola hidup yang menyebabkan tekanan darah tinggi meningkat.

THANKYOU!