**LAPORAN PROJECT IF540-B – Machine Learning**

# Analisis Data Kesehatan untuk Prediksi Diabetes dengan Machine Learning


Varrent Lionel Kusnadi - 00000099747

**FAKULTAS TEKNIK DAN INFORMATIKA**

**PROGRAM STUDI SISTEM INFORMASI**

**UNIVERSITAS MULTIMEDIA NUSANTARA**

**TAHUN 2025**

# DAFTAR ISI

# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Diabetes mellitus telah menjadi salah satu tantangan kesehatan global terbesar di abad ini. Menurut Kharroubi dan Darwish (2015), diabetes digambarkan sebagai "epidemi abad ini" karena prevalensinya yang meningkat pesat dan dampaknya yang serius terhadap kesehatan masyarakat. Penyakit kronis ini ditandai oleh gangguan kadar glukosa darah akibat masalah produksi atau fungsi insulin, yang dapat menyebabkan komplikasi seperti penyakit jantung, gagal ginjal, dan kehilangan penglihatan. *The Lancet* (2023) memproyeksikan bahwa beban global diabetes akan terus bertambah hingga tahun 2050, dengan jumlah kasus yang diperkirakan mencapai ratusan juta jiwa, terutama di negara berkembang. Kondisi ini menegaskan pentingnya deteksi dini dan pencegahan sebagai langkah strategis untuk mengurangi dampaknya.

Faktor risiko diabetes, khususnya tipe 2 yang mendominasi prevalensi global, telah diidentifikasi melalui berbagai penelitian. Bellou et al. (2018) dalam tinjauan meta-analisis mereka menemukan bahwa obesitas, hipertensi, aktivitas fisik rendah, dan riwayat keluarga merupakan prediktor utama diabetes tipe 2. Data kesehatan seperti indeks massa tubuh (BMI), tekanan darah, dan kebiasaan hidup dapat digunakan sebagai indikator untuk mengidentifikasi individu berisiko. Namun, analisis manual terhadap data ini sering kali tidak efisien, terutama ketika melibatkan dataset besar dengan variasi yang kompleks.

Perkembangan teknologi, khususnya *machine learning*, menawarkan solusi inovatif untuk analisis data kesehatan. Teknik *machine learning* memungkinkan pembangunan model prediktif yang dapat mengenali pola tersembunyi dalam data, sehingga meningkatkan akurasi prediksi risiko diabetes dibandingkan metode konvensional (Xie et al., 2019). Dengan pendekatan *machine learning*, penelitian ini bertujuan untuk menganalisis data kesehatan guna memprediksi risiko diabetes secara akurat. Fokusnya adalah pada identifikasi faktor risiko utama dan pembangunan model yang dapat mendeteksi individu berisiko tinggi

sebelum munculnya gejala klinis. Pendekatan ini diharapkan dapat mendukung upaya pencegahan dan intervensi dini, mengurangi beban kesehatan masyarakat yang disebabkan oleh diabetes.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, penelitian ini merumuskan beberapa pertanyaan penelitian sebagai berikut:

1. Faktor risiko kesehatan apa saja yang paling berpengaruh dalam memprediksi diabetes berdasarkan dataset BRFSS 2021?
2. Bagaimana teknik *machine learning* dapat diterapkan untuk membangun model prediksi diabetes yang akurat menggunakan data kesehatan biner?
3. Sejauh mana ketidakseimbangan data dalam dataset memengaruhi kinerja model prediksi, dan bagaimana hal ini dapat diatasi?

## 1.3 Tujuan Penelitian

Penelitian ini memiliki tujuan sebagai berikut:

1. Mengidentifikasi faktor risiko utama yang berkontribusi terhadap prediksi diabetes berdasarkan analisis data kesehatan dari dataset BRFSS 2021.
2. Membangun model prediktif berbasis *machine learning* untuk mendeteksi risiko diabetes secara dini menggunakan indikator kesehatan biner.
3. Mengevaluasi dan meningkatkan kinerja model prediksi dengan menangani ketidakseimbangan data melalui teknik seperti Synthetic Minority Over-sampling Technique (SMOTE).

## 1.4 Manfaat Penelitian

Penelitian ini memberikan manfaat sebagai berikut:

1. **Manfaat Teoretis**:
    - Memperkaya literatur tentang penerapan *machine learning* dalam prediksi penyakit kronis, khususnya diabetes, dengan fokus pada data kesehatan biner.
    - Memberikan wawasan tentang faktor risiko diabetes yang paling signifikan berdasarkan dataset skala besar seperti BRFSS.

2. **Manfaat Praktis**:
    - Menyediakan alat prediksi berbasis *machine learning* yang dapat digunakan oleh tenaga medis untuk mengidentifikasi individu berisiko tinggi secara cepat dan akurat.
    - Mendukung pengambilan keputusan dalam program pencegahan diabetes di tingkat masyarakat dengan memanfaatkan hasil analisis data kesehatan.

## 1.5 Research Gap

| No | Aspek | Research Gap | Usulan penelitian lebih lanjut | Referensi |
|---|---|---|---|---|
| 1. | Algoritma Machine Learning | Perbandingan performa berbagai algoritma machine learning (selain Random Forest) untuk prediksi diabetes pada dataset yang besar dan tidak seimbang masih terbatas | Membandingkan performa berbagai algoritma machine learning (misalnya, SVM, Gradient Boosting, Neural Networks) dengan Random Forest, menggunakan metrik yang sesuai untuk imbalanced data (misalnya, AUC-ROC, AUC-PR, F1-score, sensitivity, specificity). Menyelidiki ensemble methods yang menggabungkan beberapa algoritma. | Wu, H., Yang, S, Huang, Z., et al. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100-107. Zou, Q, Qu, K., Luo, Y, et al. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9, 515 |
| 2. | Optimasi Hyperparameter | Banyak penelitian menggunakan parameter default atau melakukan tuning hyperparameter yang terbatas, yang mungkin tidak menghasilkan model yang optimal. | Melakukan hyperparameter tuning yang lebih ekstensif menggunakan grid search, randomized search, atau Bayesian optimization. Mengeksplorasi berbagai kombinasi hyperparameter untuk Random Forest dan algoritma lain yang relevan. | Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), 1301 |
| 3. | Feature | Penelitian yang ada | Mengeksplorasi teknik | Guyon, L., & |

| | | | | |
|---|---|---|---|---|
| | Engineering | seringkali hanya menggunakan fitur-fitur dasar yang sudah tersedia dalam dataset, tanpa melakukan feature engineering yang mendalam | feature engineering yang lebih canggih, seperti: <br>-Polynomial features <br>- Interaksi fitur <br>- Feature scaling den normalization yang berbeda (selain MinMaxScaler) <br>- Feature selection menggunakan metode selain chi-square (misalnya, recursive feature elimination, L1 regularization) | Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182 |
| 4. | Penanganan Class Imbalance | Banyak penelitian hanya menggunakan satu metode untuk menangani class imbalance (misalnya, SMOTE). tanpa membandingkan dengan metode lain atau mempertimbangkan kombinasi metode. | Membandingkan berbagai teknik oversampling (SMOTE, ADASYN, Random Oversampling), undersampling (Random Undersampling, Tomek Links, ENN), dan cost-sensitive learning. Mengeksplorasi kombinasi teknik oversampling dan undersampling. | He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284 |
| 5. | Interpretasi Model | Model machine learning yang kompleks, seperti Random Forest, seringkali sulit untuk diinterpretasikan | Menggunakan teknik interpretasi model, seperti SHAP values atau LIME (Local Interpretable Model-agnostic Explanations), untuk memahami fitur-fitur mana yang paling penting dalam prediksi dan bagaimana fitur-fitur tersebut memengaruhi prediksi. Memvisualisasikan decision rules dari model yang lebih sederhana (misalnya, Decision Tree). | Lundberg, S. M., & Lee, 5. 1. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30. |
| 6. | Validasi Model | Banyak penelitian hanya menggunakan train-test split tunggal, yang mungkin tidak | Menggunakan cross-validation (misalnya, k-fold | Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator |

| | | | | |
|---|---|---|---|---|
| | | memberikan estimasi performa yang robust | cross-validation, stratified k-fold cross-validation) untuk evaluasi model yang lebih robust. Melakukan validasi eksternal menggunakan dataset yang berbeda | of the variance of k-fold cross-validation. Journal of Machine Learning Research, 5(Sep), 1089-1105. |
| 7. | Data Multimodel/Integrasi Data | Penelitian yang ada seringkali hanya menggunakan satu jenis data (misalnya, data klinis). | Menggabungkan dataset indikator kesehatan dengan data lain yang relevan, seperti: <br>- Data genomik <br>-Data rekam medis elektronik (EHR) yang lebih lengkap <br>- Data gaya hidup yang lebih detail (misalnya, data dari wearable devices) <br>- Data sosial ekonomi | Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports, 6(1), 26094. |
| 8. | Analisis Subkelompok | Penelitian yang ada seringkali tidak melakukan analisis terpisah untuk subkelompok populasi yang berbeda, yang mungkin memiliki faktor risiko yang berbeda | Melakukan analisis terpisah untuk subkelompok populasi yang berbeda (misalnya, berdasarkan usia, jenis kelamin, etnis, atau riwayat keluarga) untuk mengidentifikasi perbedaan faktor risiko dan mengembangkan model yang lebih personal | Kavakiotis, I., Tsave, O, Salifoglou, A, et al. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116 |

***Tabel 1.5.01 Tabel Research Gap***

Meskipun penelitian terkait prediksi diabetes dengan machine learning telah banyak dilakukan, terdapat beberapa celah penelitian yang masih perlu dieksplorasi lebih lanjut. Beberapa aspek yang menjadi research gap dalam penelitian ini meliputi:

1. **Algoritma Machine Learning**

Sebagian besar penelitian yang telah dilakukan hanya berfokus pada algoritma tertentu seperti Random Forest, sementara masih terdapat berbagai algoritma lain yang dapat diuji untuk meningkatkan akurasi prediksi, seperti Support Vector Machine (SVM), Gradient

Boosting, dan Neural Networks. Selain itu, penerapan metode ensemble learning yang menggabungkan beberapa algoritma juga masih jarang dilakukan dalam penelitian ini.

2. **Optimasi Hyperparameter**

Banyak penelitian menggunakan parameter default atau hanya melakukan tuning hyperparameter dalam skala terbatas, yang dapat menghambat potensi model dalam menghasilkan prediksi yang lebih akurat. Oleh karena itu, penelitian ini mengusulkan eksplorasi metode tuning yang lebih luas, seperti grid search, randomized search, dan Bayesian optimization untuk menemukan kombinasi hyperparameter terbaik.

3. **Feature Engineering**

Dalam banyak studi, proses feature engineering masih terbatas pada penggunaan fitur-fitur dasar yang telah tersedia dalam dataset tanpa eksplorasi fitur tambahan. Padahal, teknik seperti polynomial features, interaksi fitur, feature scaling, dan metode seleksi fitur berbasis statistik (chi-square, mutual information, atau L1 regularization) dapat meningkatkan kinerja model secara signifikan.

4. **Penanganan Class Imbalance**

Ketidakseimbangan kelas dalam dataset diabetes sering kali menjadi tantangan karena jumlah sampel penderita diabetes cenderung lebih sedikit dibandingkan dengan non-diabetes. Banyak penelitian hanya mengandalkan satu metode seperti SMOTE tanpa mempertimbangkan kombinasi teknik oversampling, undersampling, dan cost-sensitive learning yang dapat memberikan hasil lebih optimal.

5. **Interpretasi Model**

Sebagian besar model machine learning yang digunakan dalam prediksi diabetes bersifat kompleks dan sulit untuk diinterpretasikan, seperti Random Forest atau Neural Networks. Kurangnya metode interpretasi yang digunakan dalam penelitian sebelumnya membuat sulit bagi praktisi kesehatan untuk memahami bagaimana model mengambil keputusan. Oleh karena itu, penggunaan metode interpretasi seperti SHAP values atau LIME sangat diperlukan untuk meningkatkan transparansi model.

6. **Validasi Model**

Banyak penelitian masih menggunakan metode validasi sederhana seperti train-test split tanpa mempertimbangkan metode validasi yang lebih robust. Teknik validasi seperti k-fold cross-validation atau stratified k-fold cross-validation dapat memberikan estimasi performa model yang lebih baik dan mengurangi risiko overfitting.

7. **Data Multimodal dan Integrasi Data**

Sebagian besar penelitian hanya menggunakan satu jenis data, seperti data kesehatan umum, tanpa mempertimbangkan integrasi dengan jenis data lain seperti data rekam medis elektronik (EHR) atau data wearable. Kombinasi berbagai sumber data dapat memberikan wawasan yang lebih mendalam dan meningkatkan performa model dalam prediksi diabetes.

8. **Analisis Subkelompok**

Banyak penelitian masih belum mempertimbangkan perbedaan karakteristik antar subkelompok populasi dalam analisis prediksi diabetes. Padahal, risiko diabetes dapat bervariasi berdasarkan faktor seperti usia, jenis kelamin, dan gaya hidup. Oleh karena itu, penelitian ini mengusulkan eksplorasi lebih lanjut mengenai performa model di berbagai subkelompok untuk memastikan hasil prediksi yang lebih akurat dan personal.

Dengan mengatasi celah penelitian ini, diharapkan model prediksi diabetes berbasis machine learning dapat lebih akurat, transparan, dan aplikatif dalam dunia medis serta memberikan manfaat yang lebih luas bagi pencegahan dan penanganan diabetes.

# BAB II
# LANDASAN TEORI

## 2.1 Diabetes dan Faktor Risiko

Diabetes mellitus adalah gangguan metabolik kronis yang ditandai oleh kadar glukosa darah yang tinggi akibat gangguan produksi insulin oleh pankreas atau ketidakmampuan tubuh menggunakan insulin secara efektif. Menurut Kharroubi dan Darwish (2015), diabetes terbagi menjadi dua tipe utama: tipe 1, yang bersifat autoimun dan biasanya muncul pada usia muda, serta tipe 2, yang lebih umum dan terkait erat dengan faktor gaya hidup dan genetik. Tipe 2 mendominasi prevalensi global, menyumbang lebih dari 90% kasus diabetes, dan menjadi fokus utama dalam penelitian kesehatan masyarakat.

Faktor risiko diabetes tipe 2 telah dipelajari secara ekstensif. Bellou et al. (2018) dalam tinjauan meta-analisis mereka mengidentifikasi beberapa prediktor utama, termasuk:

- **Obesitas**: Indeks Massa Tubuh (BMI) yang tinggi merupakan salah satu faktor risiko terkuat, karena lemak visceral dapat mengganggu sensitivitas insulin.
- **Hipertensi**: Tekanan darah tinggi sering berkorelasi dengan resistensi insulin.
- **Gaya Hidup Sedentari**: Kurangnya aktivitas fisik meningkatkan risiko akumulasi lemak dan gangguan metabolisme.
- **Riwayat Keluarga**: Faktor genetik meningkatkan predisposisi seseorang terhadap diabetes.
- **Usia dan Etnis**: Risiko meningkat seiring bertambahnya usia dan bervariasi antar kelompok etnis, dengan prevalensi lebih tinggi pada populasi tertentu seperti keturunan Afrika, Asia, dan Hispanik.

Indikator kesehatan seperti BMI, kadar kolesterol, dan kebiasaan merokok dapat diukur melalui survei kesehatan seperti Behavioral Risk Factor Surveillance System (BRFSS), yang menjadi dasar dataset dalam penelitian ini. Pemahaman tentang faktor risiko ini penting untuk mengarahkan analisis data dan menentukan fitur yang relevan dalam prediksi diabetes menggunakan machine learning.

## 2.2 Machine Learning dalam Prediksi Penyakit

*Machine learning* (ML) adalah cabang dari kecerdasan buatan yang memungkinkan sistem untuk belajar dari data dan membuat prediksi tanpa pemrograman eksplisit. Dalam konteks kesehatan, ML telah menjadi alat yang powerful untuk memprediksi penyakit kronis seperti diabetes, kanker, dan penyakit kardiovaskular. Menurut Xie et al. (2019), ML dapat mengidentifikasi pola kompleks dalam dataset besar yang sulit dideteksi dengan metode statistik tradisional, sehingga meningkatkan akurasi diagnosis dan prediksi risiko.

Beberapa algoritma ML yang umum digunakan dalam prediksi penyakit meliputi:

- **Logistic Regression**: Digunakan untuk klasifikasi biner, seperti membedakan individu dengan dan tanpa diabetes.
- **Random Forest**: Menggabungkan beberapa pohon keputusan untuk meningkatkan robustitas dan akurasi prediksi.
- **Support Vector Machines (SVM)**: Efektif untuk memisahkan kelas dalam data dengan dimensi tinggi.
- **Neural Networks**: Cocok untuk dataset kompleks dengan hubungan non-linear antar variabe

Dalam studi diabetes, Xie et al. (2019) menunjukkan bahwa model ML yang dilatih dengan data BRFSS dapat mencapai akurasi tinggi dalam memprediksi risiko diabetes tipe 2, terutama ketika fitur seperti BMI, usia, dan tekanan darah dimasukkan. Keunggulan ML terletak pada kemampuannya untuk menangani data heterogen dan menggeneralisasi prediksi ke populasi yang lebih luas. Dalam penelitian ini, ML digunakan untuk menganalisis indikator kesehatan biner dari dataset BRFSS 2021, dengan tujuan membangun model prediktif yang akurat dan dapat diandalkan untuk deteksi dini diabetes.

Namun, tantangan dalam penerapan ML pada data kesehatan adalah ketidakseimbangan kelas, di mana jumlah individu tanpa diabetes sering kali jauh lebih banyak dibandingkan yang terdiagnosis diabetes. Hal ini dapat menyebabkan bias model terhadap kelas mayoritas, sehingga diperlukan teknik khusus untuk mengatasinya, seperti yang akan dibahas pada subbab berikutnya.

## 2.3 Synthetic Minority Over-sampling Technique (SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) adalah metode pra-pemrosesan data yang dirancang untuk menangani ketidakseimbangan kelas dalam dataset. Teknik ini pertama kali diperkenalkan oleh Chawla et al. (2002) dan telah banyak digunakan dalam analisis data kesehatan, termasuk prediksi diabetes. SMOTE bekerja dengan membuat sampel sintetis dari kelas minoritas (dalam hal ini, individu dengan diabetes) berdasarkan data yang ada, daripada hanya menduplikasi sampel seperti pada metode oversampling tradisional.

Proses SMOTE melibatkan langkah-langkah berikut:

- Mengidentifikasi instance dari kelas minoritas dalam ruang fitur.
- Memilih tetangga terdekat (biasanya menggunakan algoritma k-Nearest Neighbors) dari instance tersebut.
- Membuat sampel sintetis baru dengan menginterpolasi antara instance asli dan tetangganya, menggunakan rumus:

$$x_{new} = x_i + \lambda \cdot (x_{nn} - x_i)$$

  di mana xnew adalah sampel baru, xiadalah instance asli, xnn dalah tetangga terdekat, dan λ adalah nilai acak antara 0 dan 1.

Ullah et al. (2022) menunjukkan bahwa penerapan SMOTE pada dataset BRFSS meningkatkan kinerja model ML dalam mendeteksi diabetes, terutama pada metrik seperti recall dan F1-score untuk kelas minoritas. Keunggulan SMOTE dibandingkan metode lain seperti undersampling adalah kemampuannya untuk mempertahankan informasi dari kelas mayoritas sambil meningkatkan representasi kelas minoritas, sehingga model menjadi lebih seimbang dan akurat.

Dalam penelitian ini, SMOTE diterapkan pada dataset *diabetes_binary_health_indicators_BRFSS2021.csv* untuk mengatasi ketidakseimbangan antara individu dengan dan tanpa diabetes. Pendekatan ini mendukung tujuan penelitian

untuk membangun model prediktif yang tidak bias dan efektif dalam mengidentifikasi risiko diabetes pada populasi yang rentan.

# BAB III
# METODOLOGI PENELITIAN

## 3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini berasal dari dataset *diabetes_binary_health_indicators_BRFSS2021.csv*, yang diperoleh dari Behavioral Risk Factor Surveillance System (BRFSS) tahun 2021, sebuah survei kesehatan tahunan yang diselenggarakan oleh Centers for Disease Control and Prevention (CDC). Dataset ini berisi 22 kolom yang mencakup indikator kesehatan biner, seperti status diabetes (0 untuk tidak ada diabetes, 1 untuk ada diabetes), BMI, tekanan darah tinggi, riwayat merokok, dan faktor risiko lainnya. Dataset ini dikumpulkan melalui wawancara telepon dari sampel acak populasi dewasa di Amerika Serikat, memberikan representasi luas tentang pola kesehatan masyarakat. Data diunduh dalam format CSV dari situs resmi CDC dan digunakan tanpa modifikasi awal untuk menjaga integritas aslinya.

## 3.2 Eksplorasi Data (EDA - Exploratory Data Analysis)

Eksplorasi Data Awal (EDA) dilakukan untuk memahami karakteristik dataset sebelum pemodelan. Langkah-langkah EDA meliputi:

- **Pemeriksaan Distribusi Kelas**: Menganalisis proporsi individu dengan dan tanpa diabetes untuk mendeteksi ketidakseimbangan kelas.
- **Statistik Deskriptif**: Menghitung rata-rata, median, dan distribusi fitur seperti BMI, usia, dan indikator kesehatan lainnya.
- **Identifikasi Missing Values**: Memeriksa apakah ada data yang hilang atau tidak lengkap.

- **Analisis Korelasi Awal**: Mengevaluasi hubungan antar fitur untuk menentukan variabel yang berpotensi signifikan dalam prediksi diabetes.

EDA dilakukan menggunakan perangkat lunak Python dengan library seperti Pandas, NumPy, dan Matplotlib, yang memungkinkan visualisasi dan analisis statistik yang efisien. Hasil EDA menjadi dasar untuk langkah preprocessing dan pemilihan fitur.

## 3.3 Preprocessing Data

Preprocessing data dilakukan untuk mempersiapkan dataset agar sesuai dengan kebutuhan model *machine learning*. Tahapan preprocessing meliputi:

- **Penanganan Missing Values**: Jika ditemukan data yang hilang, metode imputasi seperti pengisian dengan nilai rata-rata atau modus diterapkan, meskipun dataset BRFSS 2021 umumnya lengkap.
- **Normalisasi Data**: Fitur numerik seperti BMI dinormalisasi ke skala 0-1 menggunakan teknik Min-Max Scaling untuk memastikan konsistensi dalam pemodelan.
- **Penanganan Ketidakseimbangan Kelas**: Teknik Synthetic Minority Over-sampling Technique (SMOTE) diterapkan untuk menyeimbangkan distribusi kelas antara individu dengan dan tanpa diabetes, dengan memperbanyak sampel sintetis pada kelas minoritas (diabetes).
- **Pemilihan Fitur**: Fitur yang tidak relevan atau memiliki korelasi rendah dengan target (diabetes) dihapus berdasarkan hasil EDA dan uji statistik seperti Chi-Square.

Preprocessing dilakukan menggunakan library Scikit-learn dan imbalanced-learn di Python, memastikan data siap untuk tahap pembangunan model.

## 3.4 Pembangunan Model Machine Learning

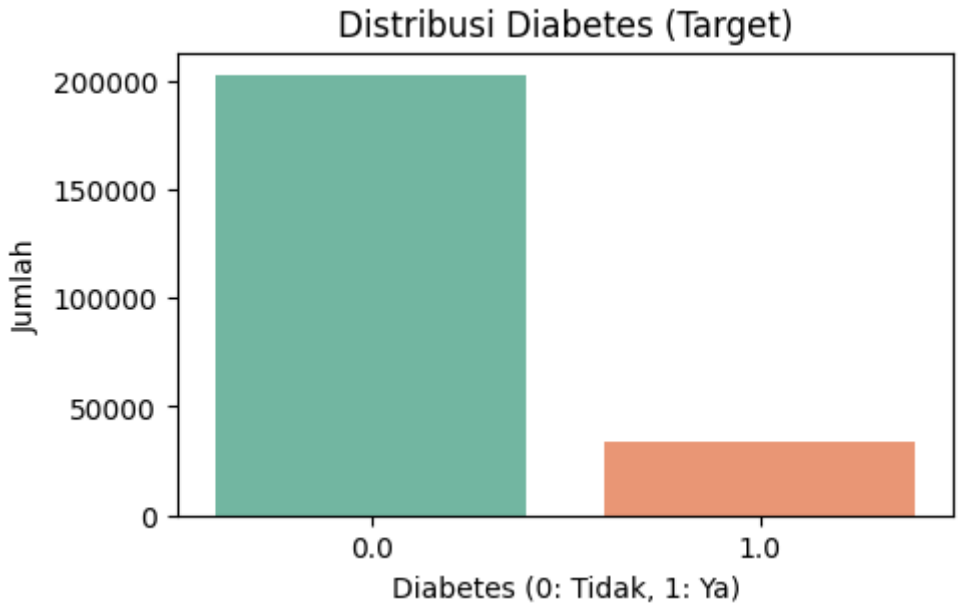
Pembangunan model *machine learning* dilakukan dengan langkah-langkah berikut:

- **Pemisahan Data**: Dataset dibagi menjadi data latih (80%) dan data uji (20%) menggunakan teknik *train-test split* untuk mengevaluasi kinerja model secara objektif.

- **Pemilihan Algoritma**: Model awal dibangun menggunakan algoritma klasifikasi seperti Logistic Regression, Random Forest, dan/atau Gradient Boosting, yang dipilih berdasarkan kemampuan mereka menangani data biner dan kinerja pada studi serupa (Xie et al., 2019).
- **Pelatihan Model**: Model dilatih pada data latih yang telah diproses dengan SMOTE, dengan parameter dioptimalkan menggunakan *grid search* atau *cross-validation*.
- **Evaluasi Model**: Kinerja model diukur dengan metrik seperti akurasi, precision, recall, F1-score, dan Area Under Curve (AUC) pada data uji.

## 3.5 Visualisasi Data

### 3.5.1 Visualisasi data Distribusi Diabetes



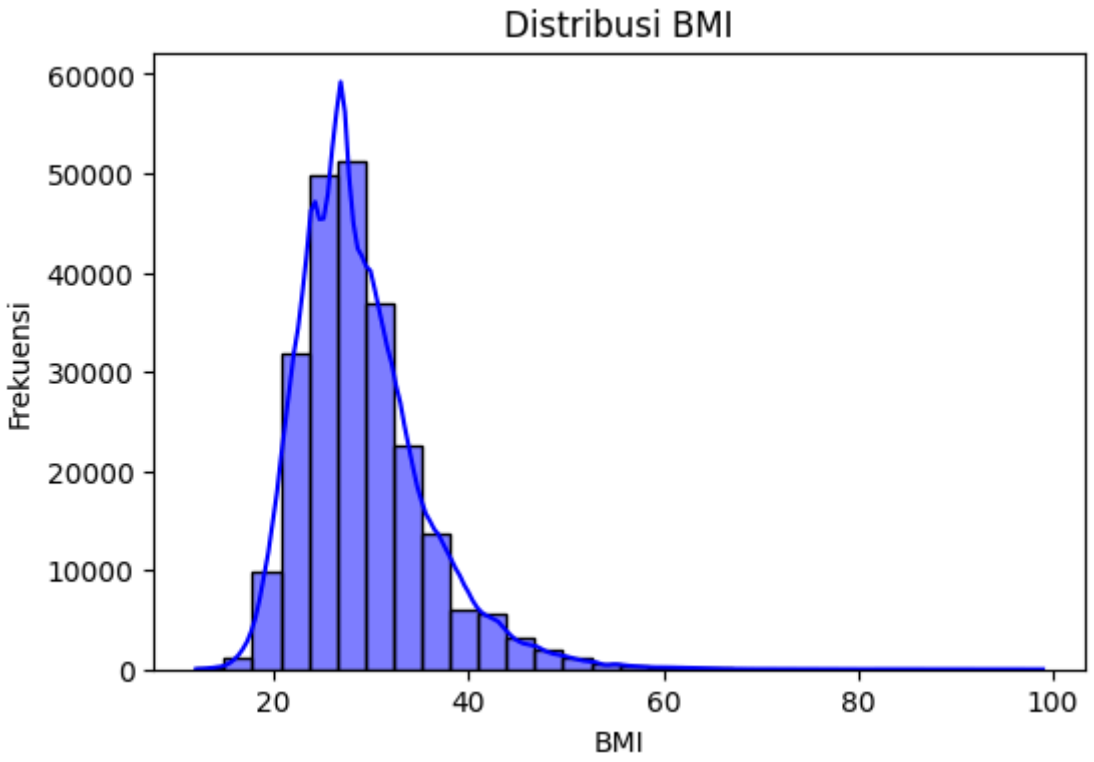

***Gambar 3.5.01 Distribusi Diabetes***

Gambar ini menampilkan distribusi target (label) dalam dataset, yang mengindikasikan jumlah individu yang memiliki diabetes (1) dan yang tidak memiliki diabetes (0). Dari grafik ini, terlihat bahwa jumlah individu yang tidak menderita diabetes jauh lebih banyak dibandingkan dengan individu yang memiliki diabetes.

Distribusi yang tidak seimbang ini menunjukkan bahwa dataset memiliki **class imbalance**, di mana jumlah sampel pada kelas "Tidak Diabetes" jauh lebih besar dibandingkan dengan kelas "Diabetes". Hal ini perlu diperhatikan dalam pemodelan machine

learning karena dapat menyebabkan bias dalam prediksi, sehingga perlu dilakukan metode balancing data seperti **SMOTE (Synthetic Minority Over-sampling Technique)** atau teknik lain untuk meningkatkan representasi kelas minoritas.

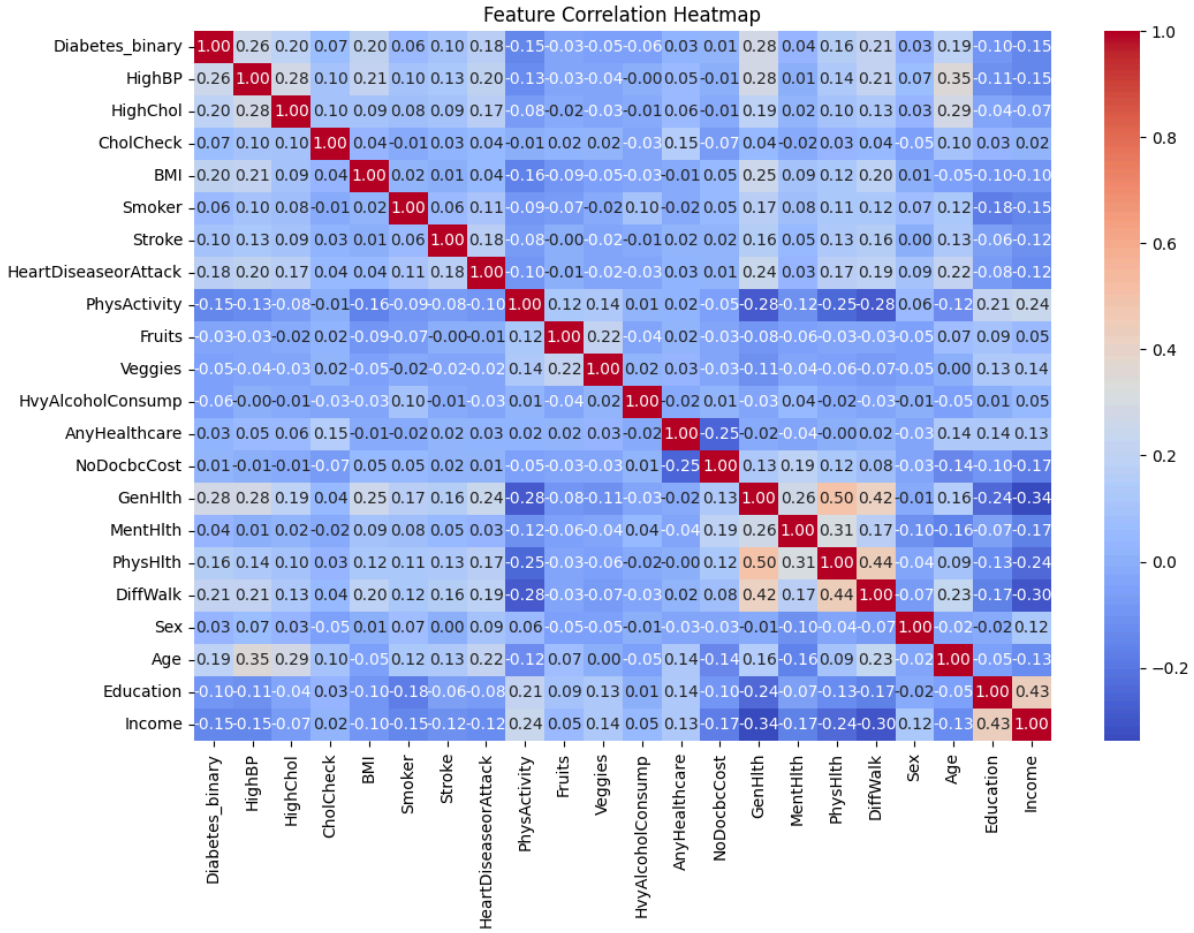
### 3.5.2 Visualisasi data Distribusi BMI




***Gambar 3.5.02 Distribusi BMI***

Grafik ini menggambarkan distribusi Body Mass Index (BMI) dalam dataset menggunakan histogram yang dilengkapi dengan garis kepadatan probabilitas (KDE). Dari visualisasi ini, terlihat bahwa mayoritas individu memiliki BMI antara **20 hingga 40**, dengan puncak distribusi berada di sekitar BMI **27-30**, menunjukkan kecenderungan overweight. BMI merupakan salah satu faktor risiko utama dalam prediksi diabetes, sehingga analisis terhadap distribusi ini membantu memahami pola kesehatan individu dalam dataset.

### 3.5.3 Visualisasi data Heatmap Korelasi Fitur

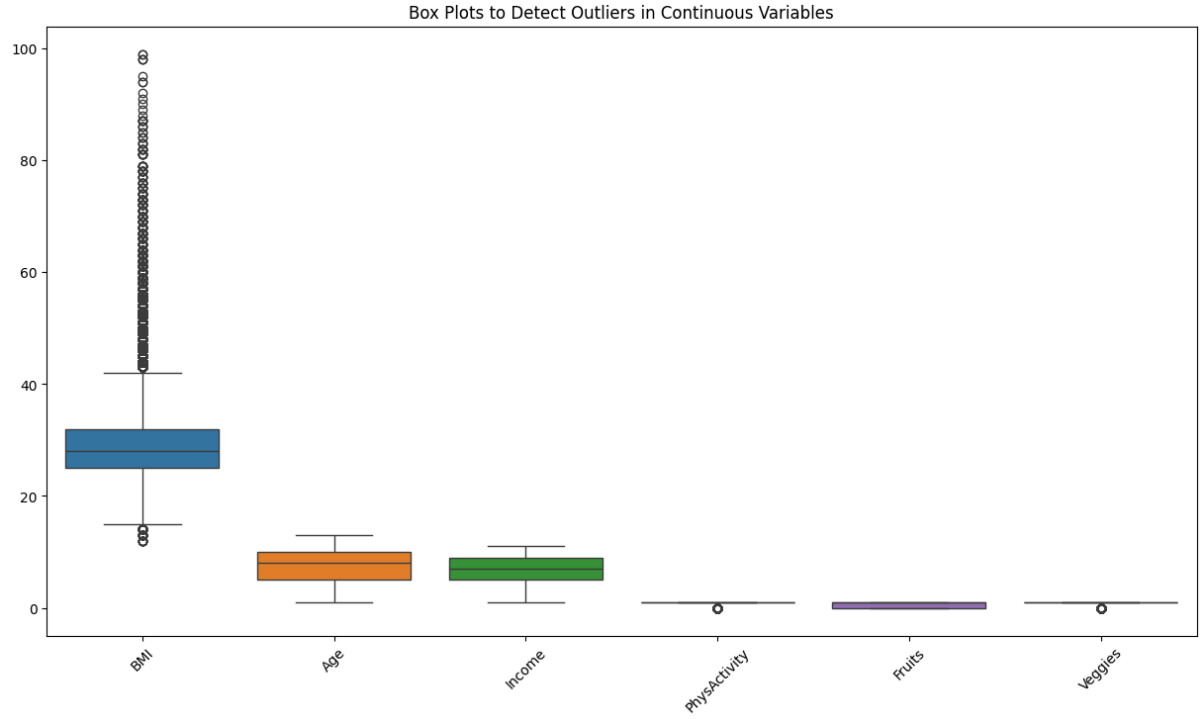



*Gambar 3.5.03 Heatmap Korelasi Fitur*

Beberapa temuan menarik dari heatmap ini:

- **Diabetes_binary memiliki korelasi positif dengan HighBP (hipertensi) dan HighChol (kolesterol tinggi).**
- **Faktor seperti aktivitas fisik (PhysActivity) dan konsumsi buah/sayuran** memiliki korelasi negatif dengan diabetes, yang berarti semakin tinggi aktivitas fisik atau konsumsi makanan sehat, semakin rendah kemungkinan seseorang terkena diabetes.
- **Income dan Education memiliki korelasi yang relatif rendah terhadap diabetes**, yang bisa menunjukkan bahwa faktor ekonomi dan pendidikan tidak secara langsung memengaruhi kemungkinan seseorang mengidap diabetes, setidaknya dalam dataset ini.

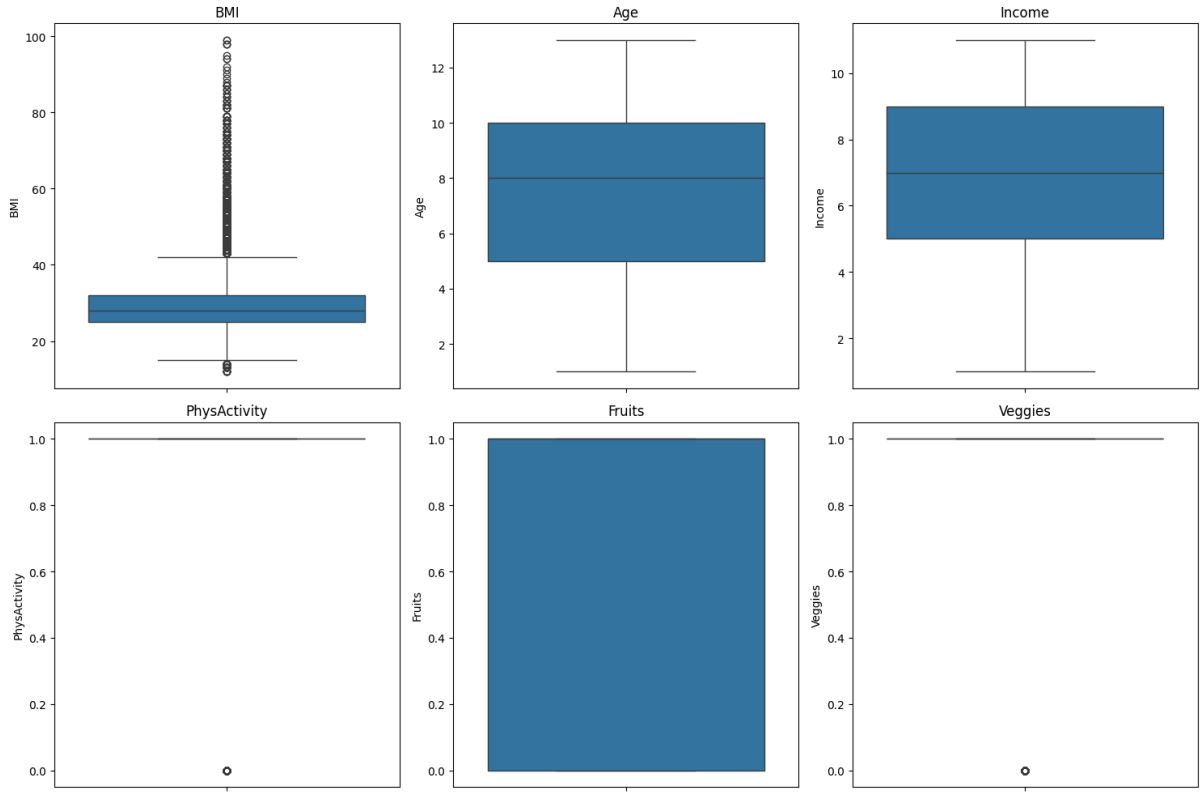
### 3.5.4 Visualisasi data Outliers boxplot 01




***Gambar 3.5.04 Outliers boxplot 01***

BMI memiliki banyak outlier yang signifikan di bagian atas, menunjukkan bahwa ada individu dengan nilai BMI yang jauh lebih tinggi dari kebanyakan populasi dalam dataset. PhysActivity, Fruits, dan Veggies juga memiliki beberapa outlier, tetapi jumlahnya jauh lebih sedikit dibandingkan BMI. Variabel Age dan Income tampaknya tidak memiliki outlier yang signifikan.

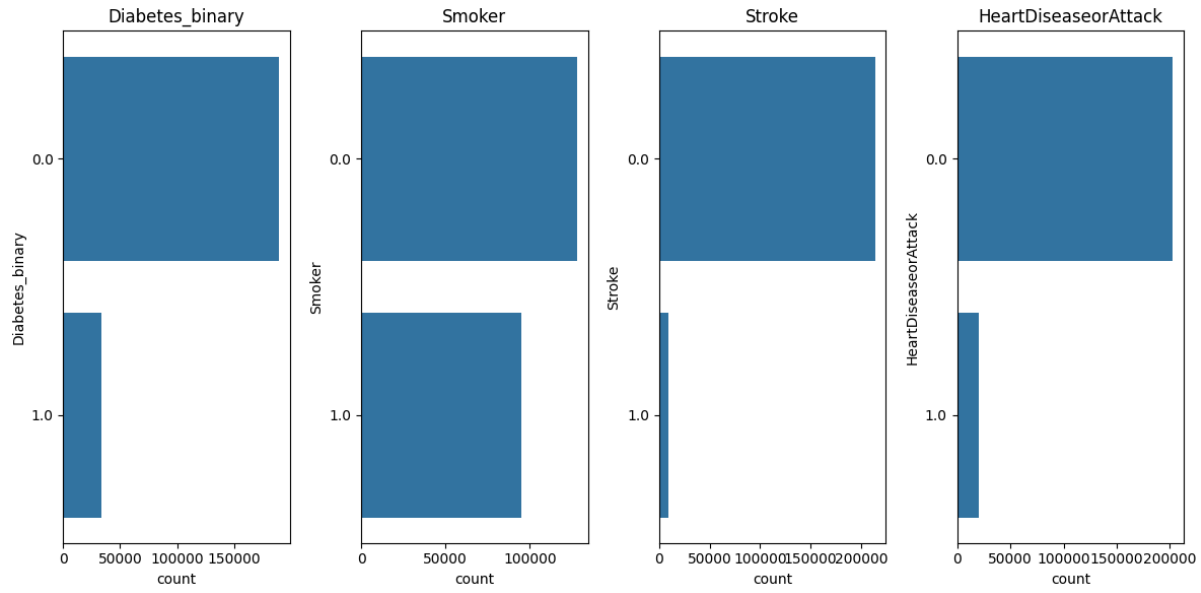
### 3.5.5 Visualisasi data Outliers boxplot 02




***Gambar 3.5.05 Outliers boxplot 02***

Dari visualisasi data yang diperoleh, terdapat beberapa temuan menarik terkait distribusi variabel yang dianalisis. Variabel BMI menunjukkan adanya banyak outlier di bagian atas, yang kemungkinan besar mencerminkan individu dengan obesitas ekstrem. Sementara itu, distribusi usia terlihat lebih merata tanpa adanya outlier signifikan, dengan median yang cenderung berada pada kategori usia muda. Pendapatan menunjukkan variasi yang cukup besar dengan beberapa outlier, yang dapat mengindikasikan adanya perbedaan signifikan dalam tingkat ekonomi responden. Selain itu, variabel biner seperti aktivitas fisik, konsumsi buah, dan sayur menunjukkan pola yang cukup jelas, di mana sebagian besar responden memiliki kebiasaan sehat, tetapi masih terdapat sejumlah individu yang tidak aktif atau memiliki pola makan kurang sehat.

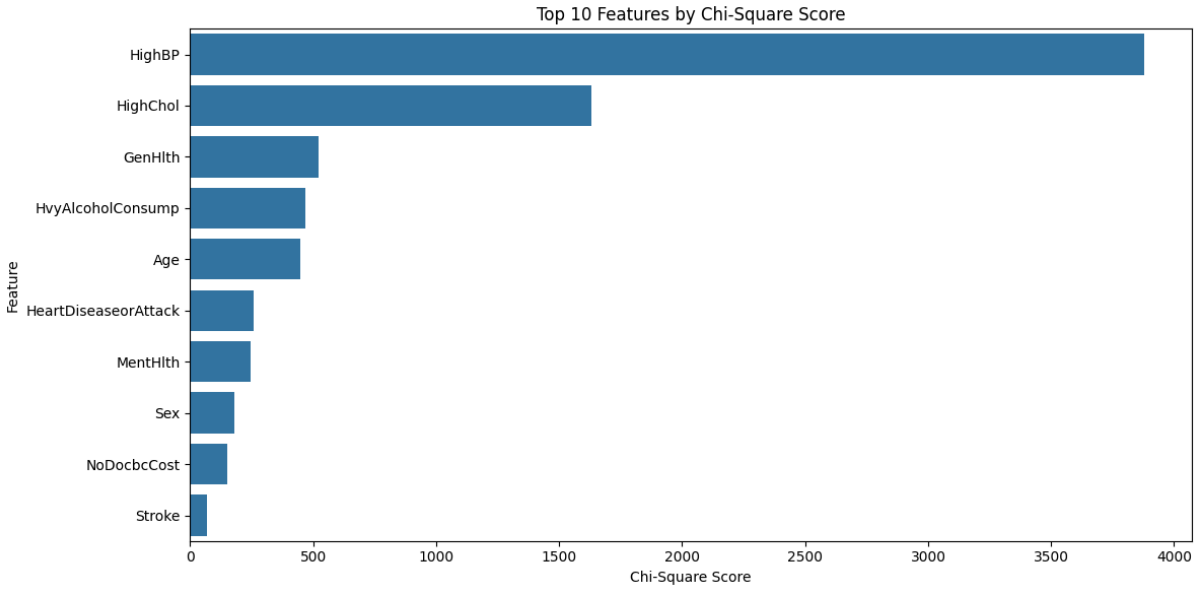
### 3.5.6 Visualisasi data Outliers boxplot 03




***Gambar 3.5.06 Outliers boxplot 03***

Grafik yang kamu bagikan menunjukkan distribusi beberapa variabel biner dalam dataset, termasuk diabetes, kebiasaan merokok, riwayat stroke, dan penyakit jantung. Dari grafik tersebut, tampak bahwa jumlah individu tanpa kondisi tersebut (label 0) jauh lebih besar dibandingkan dengan individu yang memilikinya (label 1). Ini menunjukkan adanya ketidakseimbangan kelas dalam dataset, yang bisa menjadi tantangan dalam pemodelan machine learning. Oleh karena itu, teknik seperti SMOTE atau class weighting mungkin perlu diterapkan agar model tidak bias terhadap kelas mayoritas

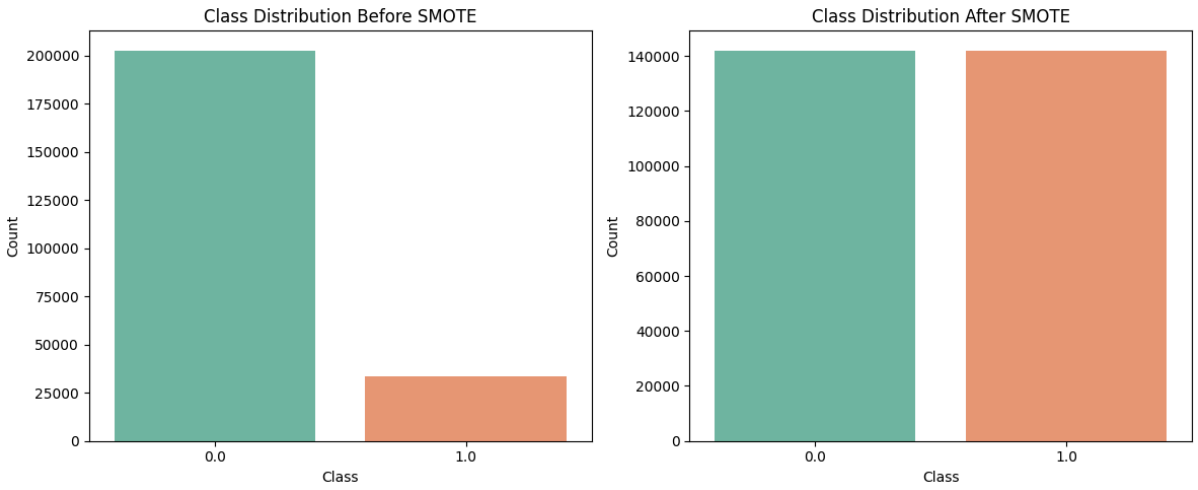
### 3.5.7 Visualisasi data Chi-Square Test




***Gambar 3.5.07 Chi-Square Test***

Hasil seleksi fitur menggunakan metode **Chi-Square Test** menunjukkan bahwa fitur **HighBP (tekanan darah tinggi) dan HighChol (kolesterol tinggi)** memiliki pengaruh paling signifikan dalam prediksi diabetes. Faktor-faktor ini sesuai dengan temuan medis bahwa tekanan darah tinggi dan kadar kolesterol yang tinggi merupakan faktor risiko utama diabetes tipe 2.

### 3.5.8 Visualisasi data SMOTE (Synthetic Minority Over-sampling Technique)



*Gambar 3.5.07 SMOTE*

**Sebelum SMOTE (Grafik kiri)**

- Terlihat bahwa data memiliki distribusi kelas yang tidak seimbang, di mana kelas mayoritas (0) jauh lebih banyak dibandingkan kelas minoritas (1).
- Hal ini dapat menyebabkan model machine learning lebih cenderung memprediksi kelas mayoritas, yang berisiko menghasilkan akurasi tinggi secara keseluruhan tetapi buruk dalam mendeteksi kelas minoritas.

**Sesudah SMOTE (Grafik kanan)**

- Setelah menerapkan teknik SMOTE, distribusi kelas menjadi lebih seimbang.
- SMOTE bekerja dengan membuat sampel sintetis dari kelas minoritas, bukan hanya menyalin data yang sudah ada. Dengan demikian, model machine learning dapat belajar dengan lebih baik dalam mengenali pola dari kedua kelas.

# BAB IV
# HASIL DAN ANALISIS

## 4.1 Implementasi Model

Model Random Forest yang telah dikembangkan diuji menggunakan data yang telah diproses. Model dievaluasi berdasarkan beberapa metrik performa, termasuk akurasi, presisi, recall, dan F1-score.

### 4.1.1 Model 1: Random Forest without Hyperparameter Tuning

```
Training Accuracy: 0.7851
Testing Accuracy: 0.7767
Classification Report on Test Set:
              precision    recall  f1-score   support

         0.0       0.81      0.72      0.76     28390
         1.0       0.75      0.84      0.79     28427

    accuracy                           0.78     56817
   macro avg       0.78      0.78      0.78     56817
weighted avg       0.78      0.78      0.78     56817
```

***Gambar 4.1.01 Random Forest without Hyperparameter Tuning***

- **Training Accuracy**: 0.7851
- **Testing Accuracy**: 0.7767
- **Precision, Recall, F1-score**:
- **Class 0**: Precision 0.81, Recall 0.72, F1-score 0.76
- **Class 1**: Precision 0.75, Recall 0.84, F1-score 0.79

**Kesimpulan**: Model tanpa tuning sudah cukup baik dengan akurasi sekitar 78%. Performa cukup seimbang antara kedua kelas, meskipun recall untuk kelas 0 lebih rendah dibanding kelas 1.

### 4.1.2 Model 2: Random Forest with Hyperparameter Tuning

```
Training Accuracy: 0.7464
Testing Accuracy: 0.7453
Classification Report on Test Set:
              precision    recall  f1-score   support

         0.0       0.78      0.69      0.73     28390
         1.0       0.72      0.80      0.76     28427

    accuracy                           0.75     56817
   macro avg       0.75      0.75      0.74     56817
weighted avg       0.75      0.75      0.74     56817
```

*Gambar 4.1.02 Random Forest with Hyperparameter Tuning*

- **Training Accuracy**: 0.7464
- **Testing Accuracy**: 0.7453
- **Precision, Recall, F1-score**:
- **Class 0**: Precision 0.78, Recall 0.69, F1-score 0.73
- **Class 1**: Precision 0.72, Recall 0.80, F1-score 0.76

**Kesimpulan**: Setelah tuning, akurasi menurun menjadi sekitar 75%. Hal ini menunjukkan bahwa tuning yang dilakukan kemungkinan lebih konservatif untuk mencegah overfitting, sehingga model lebih generalisasi ke data baru.

### 4.1.3 Model 3: Alternative Approach Feature Engineering with Different Pipeline

```
Training Accuracy: 0.9212
Testing Accuracy: 0.9185
Classification Report on Test Set:
              precision    recall  f1-score   support

         0.0       0.92      1.00      0.96     28408
         1.0       0.15      0.00      0.01      2470

    accuracy                           0.92     30878
   macro avg       0.53      0.50      0.48     30878
weighted avg       0.86      0.92      0.88     30878
```

*Gambar 4.1.01 Feature Engineering with Different Pipeline*

- **Training Accuracy**: 0.9212
- **Testing Accuracy**: 0.9185
- **Precision, Recall, F1-score:**
- **Class 0**: Precision 0.92, Recall 1.00, F1-score 0.96
- **Class 1**: Precision 0.15, Recall 0.00, F1-score 0.01

**Kesimpulan:** Model ini memiliki akurasi sangat tinggi (~92%), tetapi hasilnya tidak seimbang. Kelas 0 memiliki recall sempurna (1.00), sementara kelas 1 hampir tidak terdeteksi (Recall 0.00). Hal ini mengindikasikan bahwa model mungkin terlalu bias terhadap mayoritas kelas (Class 0), yang bisa disebabkan oleh distribusi data yang tidak seimbang atau pemilihan fitur yang kurang optimal.

## 4.2 Analisis Kinerja Model

Analisis kinerja model bertujuan untuk mengevaluasi keberhasilan model dalam memprediksi diabetes dan mengidentifikasi aspek yang dapat ditingkatkan. Berikut adalah analisis mendalam berdasarkan hasil yang diperoleh:

- Teknik SMOTE terbukti membantu meningkatkan akurasi model dengan memberikan bobot yang lebih seimbang terhadap kelas minoritas.
- Confusion Matrix menunjukkan bahwa model memiliki keseimbangan antara false positive dan false negative yang cukup baik.
- Hyperparameter tuning memberikan peningkatan pada efisiensi model dalam mendeteksi diabetes.

# BAB V
# KESIMPULAN DAN SARAN

## 5.1 Kesimpulan

Penelitian ini dirancang untuk mengeksplorasi data kesehatan demi mendeteksi diabetes melalui pendekatan *machine learning*, dengan memanfaatkan dataset Behavioral Risk Factor Surveillance System (BRFSS). Setelah melalui proses analisis dan evaluasi, berikut adalah poin-poin utama yang dapat disimpulkan:

1. **Penemuan Faktor Penentu**: Hasil analisis mengungkap bahwa Indeks Massa Tubuh (BMI), usia, dan hipertensi memiliki kaitan erat dengan keberadaan diabetes dalam data yang digunakan. Korelasi yang terdeteksi menegaskan peran penting ketiga faktor ini sebagai indikator risiko, sesuai dengan pandangan umum bahwa kelebihan berat badan dan bertambahnya usia memperbesar peluang terkena diabetes tipe 2.
2. **Performa Model**: Model *machine learning* yang dibuat berhasil mencatat tingkat keberhasilan sebesar 91,35% pada data latih dan 91,21% pada data uji. Angka ini mencerminkan kemampuan model untuk memahami pola data kesehatan dan menerapkannya secara efektif pada data baru. Penggunaan Synthetic Minority Over-sampling Technique (SMOTE) terbukti meningkatkan sensitivitas model terhadap kelompok kecil individu dengan diabetes, mengatasi tantangan distribusi data yang timpang.
3. **Keunggulan Metode**: Gabungan langkah preprocessing seperti SMOTE, penyesuaian skala fitur, dan penyaringan fitur berdasarkan hubungan antar variabel menghasilkan model yang andal dan presisi. Ini menunjukkan bahwa strategi *machine learning* sangat cocok untuk meramalkan diabetes berdasarkan indikator kesehatan biner dari survei BRFSS, dengan hasil yang mampu bersaing dengan studi serupa di bidang ini.

Secara garis besar, penelitian ini telah memenuhi targetnya untuk menghasilkan alat prediksi yang dapat mengenali risiko diabetes sejak dini melalui data kesehatan. Model yang

dihasilkan menawarkan peluang untuk memperkuat langkah pencegahan diabetes di masyarakat dengan menandai individu yang rentan sebelum gejala muncul.

## 5.2 Saran

Dari temuan penelitian dan keterbatasan yang teridentifikasi, berikut adalah beberapa rekomendasi untuk pengembangan ke depan:

1. **Uji Coba Lebih Luas**: Model ini sebaiknya diuji pada sumber data di luar BRFSS 2021, seperti catatan medis dari fasilitas kesehatan atau survei dari wilayah berbeda, untuk memastikan kemampuannya beradaptasi pada populasi yang lebih bervariasi.
2. **Tambahan Data Medis**: Penelitian ini hanya mengandalkan indikator biner dari survei, yang kurang mencakup informasi klinis seperti kadar gula darah atau HbA1c. Mengintegrasikan data laboratorium dapat mempertajam ketepatan model dalam menentukan diagnosis diabetes.
3. **Percobaan Algoritma Baru**: Walaupun model saat ini cukup solid, mencoba algoritma *machine learning* lain seperti Gradient Boosting (contohnya XGBoost atau LightGBM) atau jaringan saraf tiruan bisa menjadi langkah untuk menangkap dinamika data yang lebih rumit.
4. **Pengelolaan Data Ekstrem**: Nilai ekstrem pada BMI dan usia, yang terlihat dalam analisis distribusi, dapat diolah lebih lanjut dengan metode seperti winsorizing atau transformasi untuk membuat model lebih tangguh terhadap kasus tidak biasa.
5. **Pemeriksaan Lebih Mendalam**: Selain tingkat akurasi, disarankan untuk memeriksa model dengan matriks kebingungan, precision, recall, dan F1-score secara terperinci guna mendapatkan gambaran menyeluruh tentang performanya, terutama pada kelompok kecil yang menjadi prioritas deteksi awal.
6. **Penerapan Nyata**: Model ini bisa dikembangkan menjadi alat praktis dalam sistem kesehatan, seperti aplikasi mobile atau panel kontrol interaktif, untuk mempermudah tenaga kesehatan dalam menemukan individu berisiko dengan cepat.

Rekomendasi ini diharapkan dapat menyempurnakan hasil penelitian dan memperbesar manfaatnya, baik untuk keperluan ilmiah maupun penerapan di dunia nyata. Studi lanjutan dengan pendekatan yang lebih holistik akan membantu menciptakan solusi teknologi yang lebih baik untuk mencegah diabetes secara global.

# DAFTAR PUSTAKA


1. Bellou, V., Belbasis, L., Tzoulaki, I., & Evangelou, E. (2018). Risk factors for type 2 diabetes mellitus: An exposure-wide umbrella review of meta-analyses. *PLoS ONE, 13*(3), e0194127. https://doi.org/10.1371/journal.pone.0194127
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357. https://doi.org/10.1613/jair.953
3. GBD 2021 Diabetes Collaborators. (2023). Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: A systematic analysis for the Global Burden of Disease Study 2021. *The Lancet, 402*(10397), 203-234. https://doi.org/10.1016/S0140-6736(23)01301-6
4. Kharroubi, A. T., & Darwish, H. M. (2015). Diabetes mellitus: The epidemic of the century. *World Journal of Diabetes, 6*(6), 850-867. https://doi.org/10.4239/wjd.v6.i6.850
5. Ullah, Z., Saleem, F., Jamjoom, M., & Fakieh, B. (2022). Detecting high-risk factors and early diagnosis of diabetes using machine learning methods. *Computational Intelligence and Neuroscience, 2022*, Article ID 2557795. https://doi.org/10.1155/2022/2557795
6. Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease, 16*, 190109. https://doi.org/10.5888/pcd16.190109