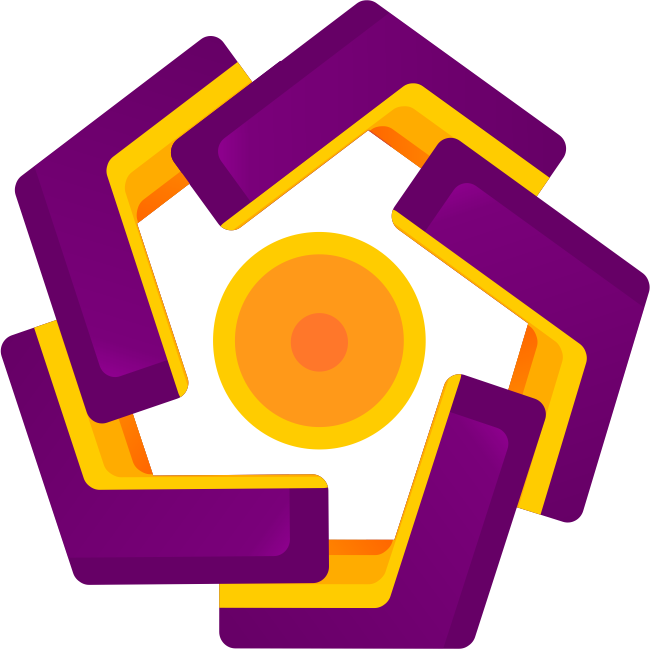
LAPORAN

UJIAN AKHIR SEMESTER

**BIG DATA & PREDICTIVE ANALYTICS LANJUT**

Dosen Pengampu :

Anna Baita, S.Kom., M.Kom



**Disusun oleh :**

Wahyu Setyaningtyas 22.11.4582

Satria Adhi Pramudita 22.11.4588

Muhammad Sachidhani 22.11.4576

**PROGRAM STUDI SARJANA INFORMATIKA**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS AMIKOM YOGYAKARTA**

**2025**

**Laporan Penggunaan Machine Learning untuk Prediksi Diabetes**

**1. Pemilihan Bidang dan Tujuan**

**Bidang yang Dipilih**:

Prediksi Diabetes Menggunakan Machine Learning

**Alasan Pemilihan**:

Diabetes adalah salah satu penyakit kronis yang prevalensinya terus meningkat di seluruh dunia, termasuk di Indonesia. Dengan meningkatnya angka penderita diabetes, prediksi dini sangat penting untuk pencegahan dan pengobatan yang lebih efektif. Oleh karena itu, kami memilih untuk menerapkan machine learning pada prediksi diabetes menggunakan dataset medis yang berisi informasi terkait faktor risiko diabetes.

**Tujuan**:

Dengan menggunakan machine learning, tujuan utama adalah untuk membangun model prediksi yang dapat membantu dalam identifikasi individu yang berisiko tinggi terhadap diabetes, sehingga langkah pencegahan dapat dilakukan lebih dini.

**2. Proses Mendapatkan Data dan Informasi Data**

**Sumber Data**:  
Dataset yang digunakan adalah **Diabetes Prediction Dataset** yang tersedia di Kaggle. Dataset ini berisi informasi medis yang dapat digunakan untuk memprediksi apakah seseorang berisiko menderita diabetes berdasarkan sejumlah fitur medis dan demografis.

**Waktu Pengambilan Data**: Dataset ini diunggah pada tahun 2020 dan berisi data terbaru yang diperoleh dalam rentang 1-4 tahun terakhir.

**Penjelasan Kolom Dataset**:

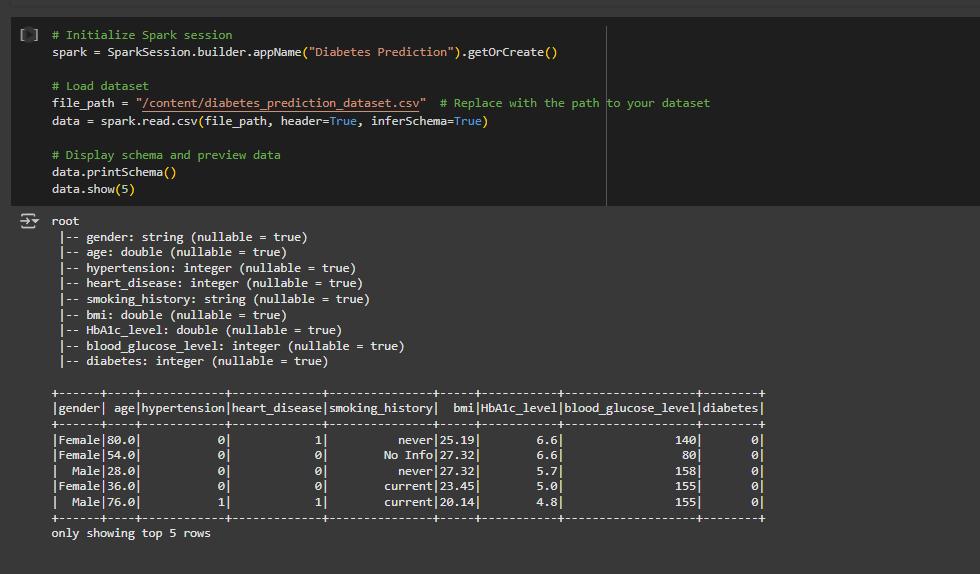
1. **gender**: Jenis kelamin (1 = male, 0 = female)
2. **age** : Usia pasien
3. **hypertension** : Apakah pasien memiliki hipertensi (1 = Yes, 0 = No)
4. **heart\_disease** : Apakah pasien memiliki penyakit jantung (1 = Yes, 0 = No)
5. **smoking\_history** : Riwayat merokok (never, formerly, currently)
6. **bmi** : Indeks Massa Tubuh (Body Mass Index)
7. **HbA1c\_level** : Tingkat HbA1c (indikator kontrol gula darah jangka panjang)
8. **blood\_glucose\_level** : Tingkat glukosa darah
9. **diabetes** : Status diabetes (1 = positive, 0 = negative)

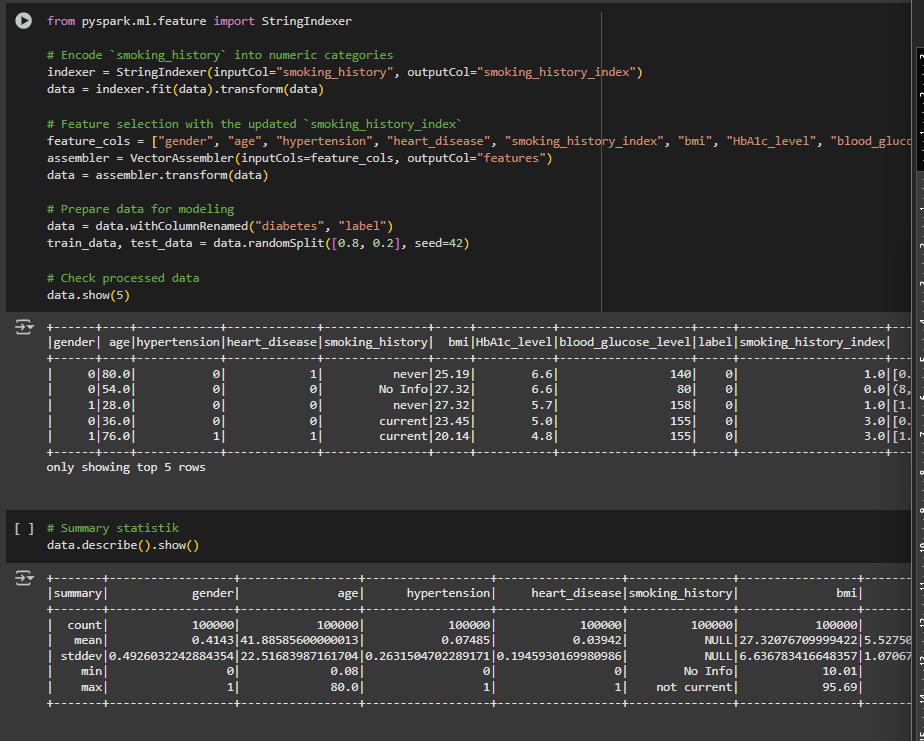
**Sumber**:  
<https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>

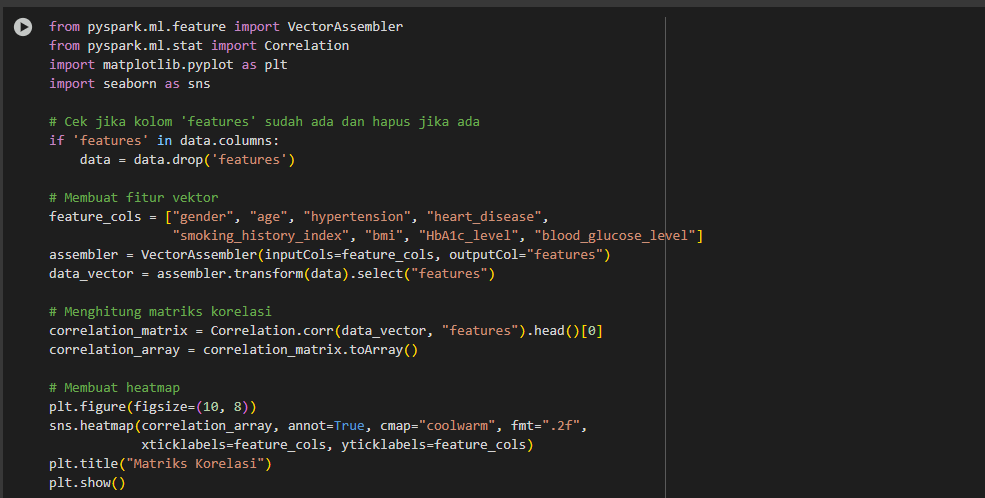
**3. Pre-processing Data**

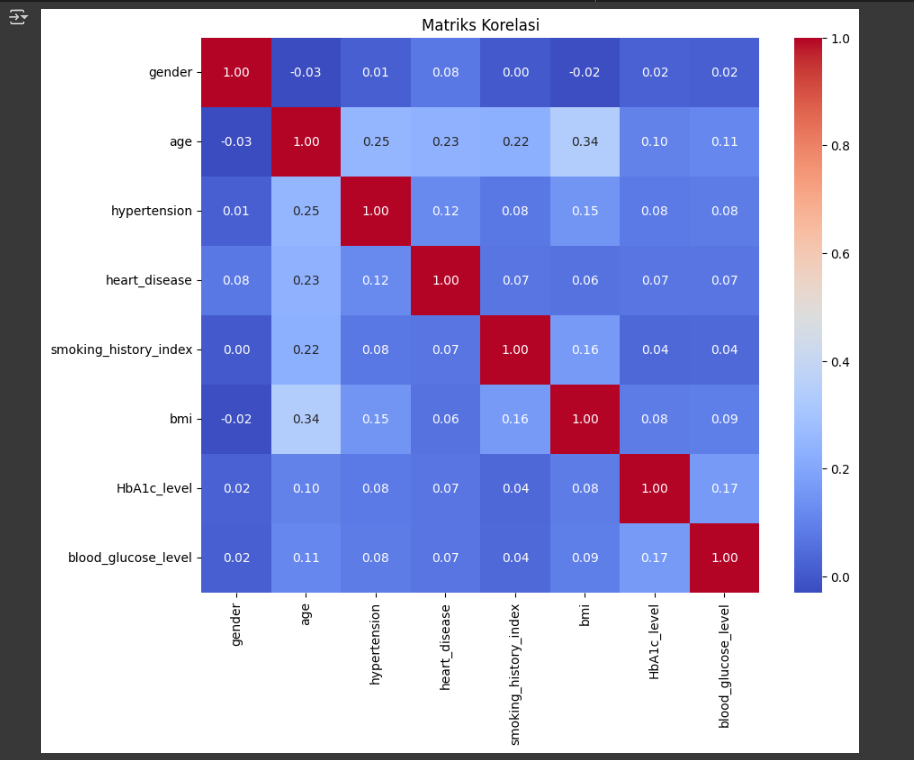
Pada tahap pre-processing, beberapa langkah dilakukan untuk menyiapkan data agar dapat digunakan dalam model machine learning:

1. Pemeriksaan Tipe Data: Memeriksa tipe data setiap kolom untuk memastikan bahwa tipe data sesuai dengan analisis yang akan dilakukan.
2. Mengganti Nama Kolom: Beberapa kolom diubah namanya untuk mempermudah pemahaman dan penggunaan.
3. Pemeriksaan Nilai Null: Memastikan tidak ada nilai yang hilang atau null yang dapat mengganggu proses modeling.
4. Mengubah Tipe Data: Mengubah tipe data kolom yang diperlukan menjadi tipe numerik.
5. Menampilkan Summary: Menampilkan statistik ringkasan untuk melihat distribusi data.
6. Matriks Korelasi: Menghitung matriks korelasi antar fitur untuk mengetahui hubungan antar variabel.







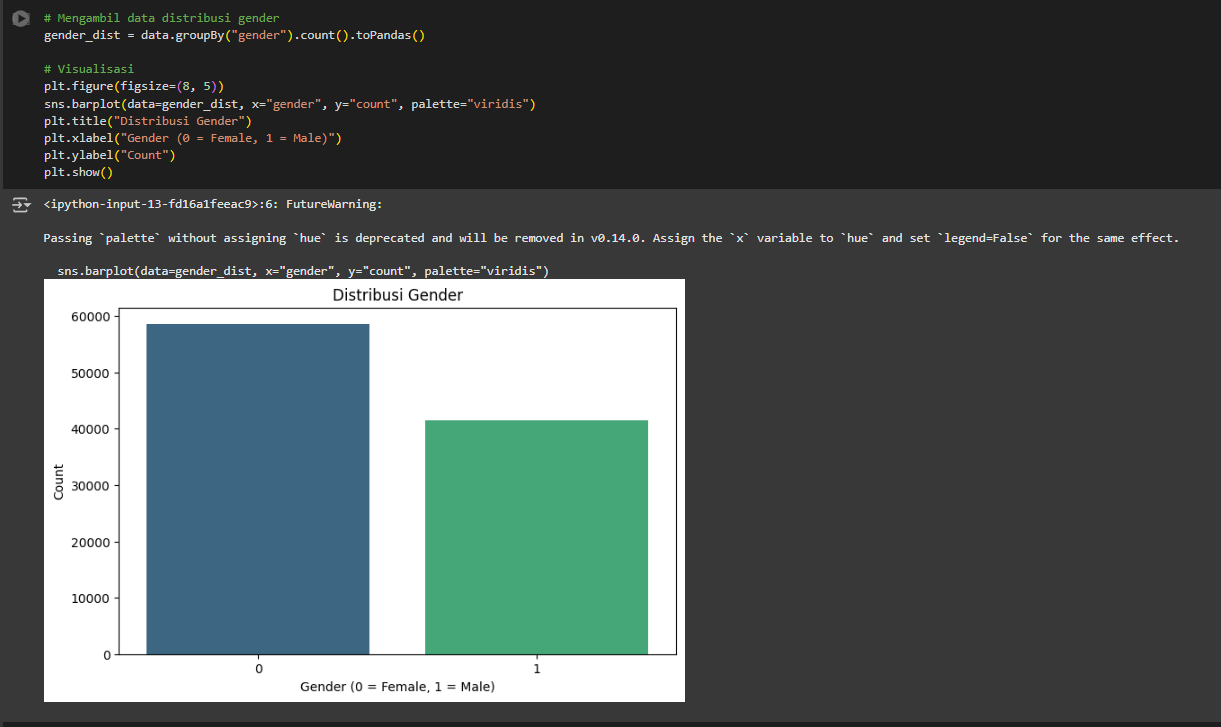


**4. Exploratory Data Analysis (EDA)**

Pada tahap EDA, kami melakukan analisis visual untuk mendapatkan wawasan dari dataset. Berikut adalah beberapa visualisasi yang digunakan:

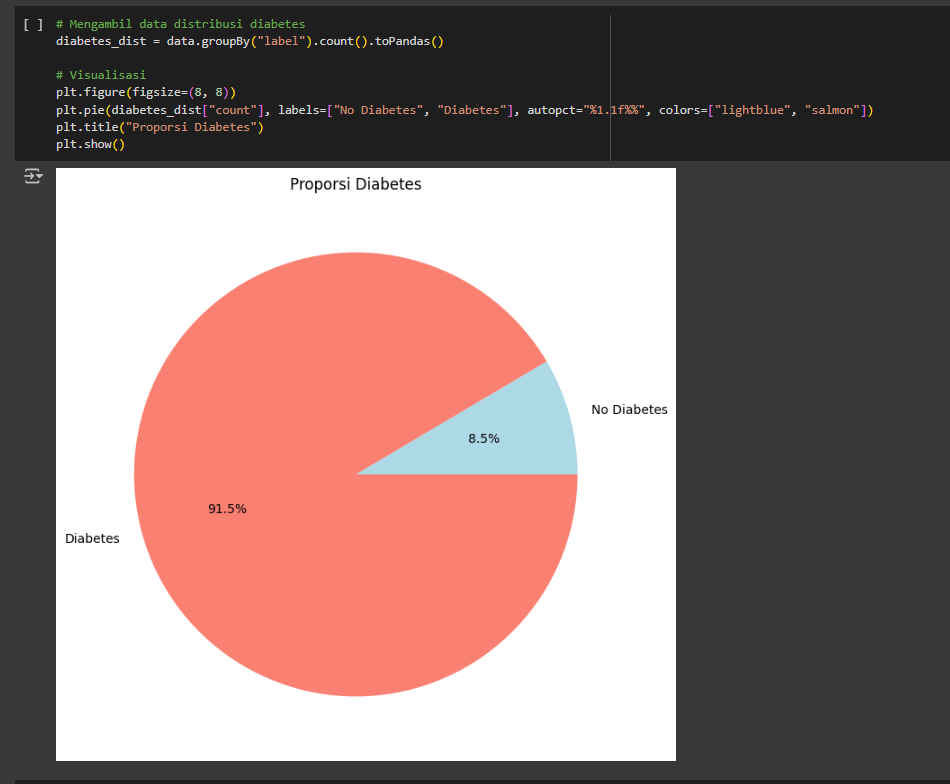
1. Bar Chart: Untuk melihat distribusi jenis kelamin.
2. Pie Chart: Untuk melihat proporsi penderita diabetes dalam dataset.
3. Scatter Plot: Untuk melihat hubungan antara age dan bmi.
4. Histogram: Untuk melihat distribusi HbA1c\_level.

**Visualisasi Bar Chart** (Distribusi Jenis Kelamin):



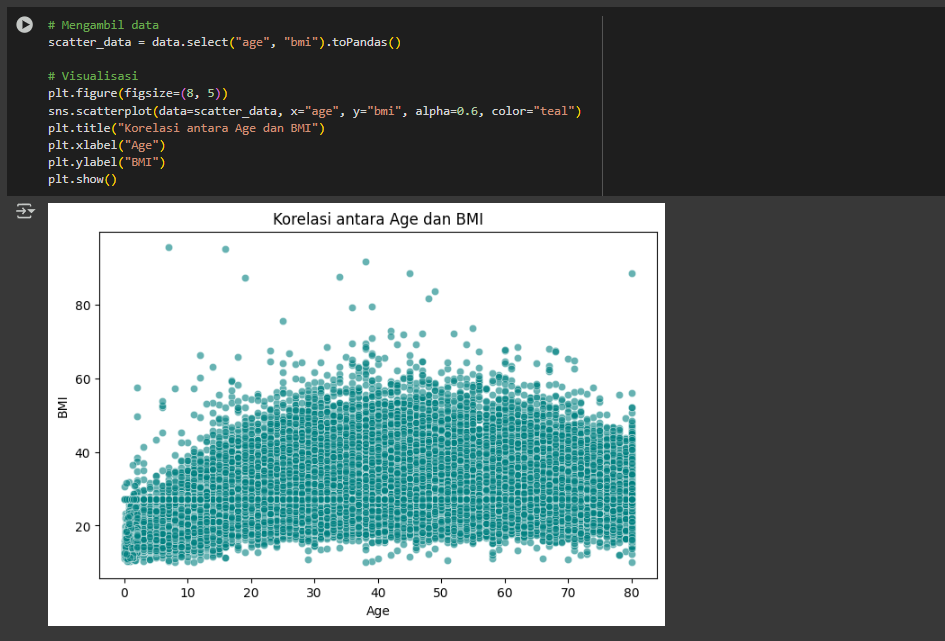
Penjelasan: Pada **Visualisasi Bar Chart** kami menemukan ada hampir 60.000 orang dari gender **Perempuan** yang terdampak diabetes, dan ada kurang dari 45.000 orang dari gender **Laki-laki** yang terkena diabetes selama 2 tahun terakhir.

**Visualisasi Pie Chart** (Proporsi Diabetes):



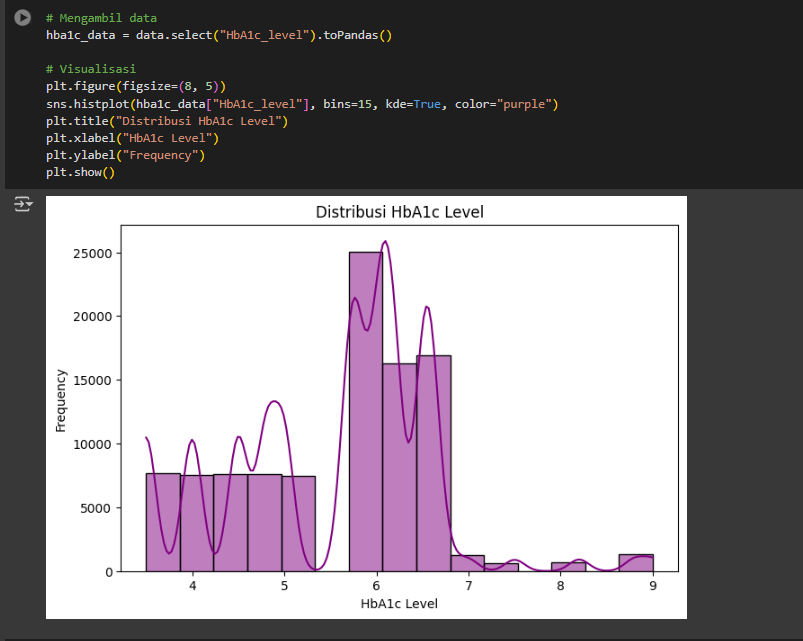
Penjelasan: Dari dataset yang kami dapat, terindikasi bahwa 91,5% terkena penyakit diabetes, dan hanya 8,5% nya masih sehat.

**Visualisasi Scatter Plot** (Korelasi antara age dan bmi):



Penjelasan: Rentang tersebut mencakup multiple klasifikasi BMI menurut WHO, dari normal weight (18.5-24.9 kg/m²) hingga obesity class II (35.0-39.9 kg/m²). Tidak terdeteksi adanya korelasi signifikan antara variabel usia dan BMI (r ≈ 0)

**Visualisasi Histogram** (Distribusi HbA1c\_level):



Penjelasan: Dari visualisasi distribusi HbA1c Level yang ditampilkan. Puncak tertinggi berada pada rentang 6-7 dengan frekuensi sekitar 25.000 kasus, yang mengindikasikan zona pre-diabetes hingga diabetes. Puncak kedua yang lebih rendah berada pada rentang 4-5, menunjukkan level HbA1c normal.

**5. Pemilihan Kolom/Fitur yang Relevan**

Berdasarkan analisis yang dilakukan, kami memilih beberapa kolom yang relevan untuk memprediksi diabetes, yaitu:

1. Age: Usia adalah faktor penting dalam risiko diabetes.
2. BMI: Indeks Massa Tubuh tinggi dapat meningkatkan risiko diabetes.
3. Hypertension: Hipertensi adalah faktor risiko utama untuk diabetes.
4. Blood Glucose Level: Kadar glukosa darah yang tinggi berhubungan langsung dengan diabetes.
5. HbA1c Level: Mengukur kontrol gula darah jangka panjang, yang dapat mengindikasikan risiko diabetes.

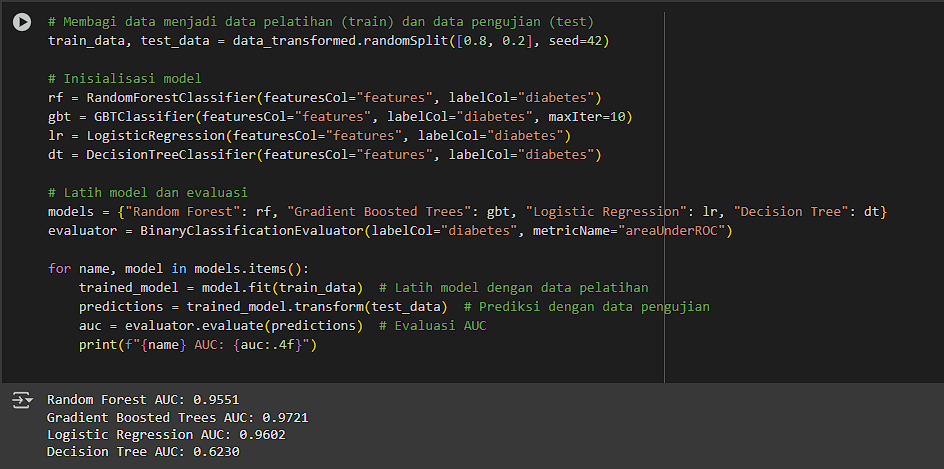
* **Relevansi Fitur dengan Diabetes**: Anda telah menghubungkan setiap fitur dengan risiko diabetes secara logis (misalnya, usia sebagai faktor risiko, BMI yang tinggi, dan relevansi kadar glukosa). Hal ini menunjukkan pemahaman yang baik tentang faktor-faktor penyebab diabetes.
* **Korelasi Data**:Anda menyebutkan bahwa hasil korelasi digunakan untuk mendukung pemilihan fitur. Namun, bisa lebih jelas dengan menambahkan hasil spesifik, seperti nilai korelasi atau insight dari matriks korelasi untuk setiap fitur yang dipilih.
* **Fokus pada Prediktor Utama**: Anda bisa menegaskan bahwa fitur yang dipilih adalah yang paling relevan dan signifikan dari dataset, sehingga dapat meningkatkan akurasi model prediksi.

Pemilihan fitur dilakukan berdasarkan keterkaitan langsung antara faktor-faktor tersebut dengan diabetes, serta hasil korelasi yang diperoleh dari matriks korelasi.

**6. Penggunaan 4 Model Machine Learning**

Pada tahap ini, kita akan menggunakan empat model machine learning untuk memecahkan masalah klasifikasi yang bertujuan memprediksi apakah seseorang berisiko terkena diabetes atau tidak. Model-model yang dipilih adalah:

1. Random Forest Classifier  
   Random Forest adalah metode ensemble learning yang membangun banyak pohon keputusan dan menggabungkan hasil dari setiap pohon untuk menghasilkan prediksi yang lebih baik. Random Forest sangat baik untuk menangani data dengan fitur yang memiliki hubungan non-linear dan dapat menangani dataset yang besar dengan baik. Model ini secara umum mampu menangani overfitting dan memberikan hasil yang stabil.
2. Gradient Boosted Trees (GBT) Classifier  
   Gradient Boosting adalah teknik boosting yang membangun pohon keputusan secara berurutan, di mana setiap pohon baru berusaha untuk memperbaiki kesalahan yang dibuat oleh pohon sebelumnya. GBT sering digunakan untuk dataset yang lebih kompleks karena dapat menangani hubungan non-linear antar fitur dengan sangat baik.
3. Logistic Regression  
   Logistic Regression adalah model klasik yang digunakan untuk klasifikasi biner. Model ini memodelkan probabilitas dari hasil yang mungkin (misalnya, kemungkinan diabetes) dan menggunakan fungsi logistik untuk mengklasifikasikan hasil. Walaupun lebih sederhana dan sering dianggap tidak sekuat model lain dalam menangani hubungan non-linear, logistic regression sering kali memberikan hasil yang baik pada masalah klasifikasi sederhana.
4. Naive Bayes Classifier  
   Naive Bayes adalah model probabilistik yang menggunakan Teorema Bayes untuk klasifikasi. Model ini berasumsi bahwa fitur-fitur dalam dataset bersifat independen, meskipun kenyataannya jarang terjadi demikian. Namun, meskipun asumsi independen ini sering kali tidak terpenuhi, Naive Bayes sering kali memberikan hasil yang cukup baik dalam berbagai masalah klasifikasi, terutama untuk dataset dengan banyak fitur kategorikal.



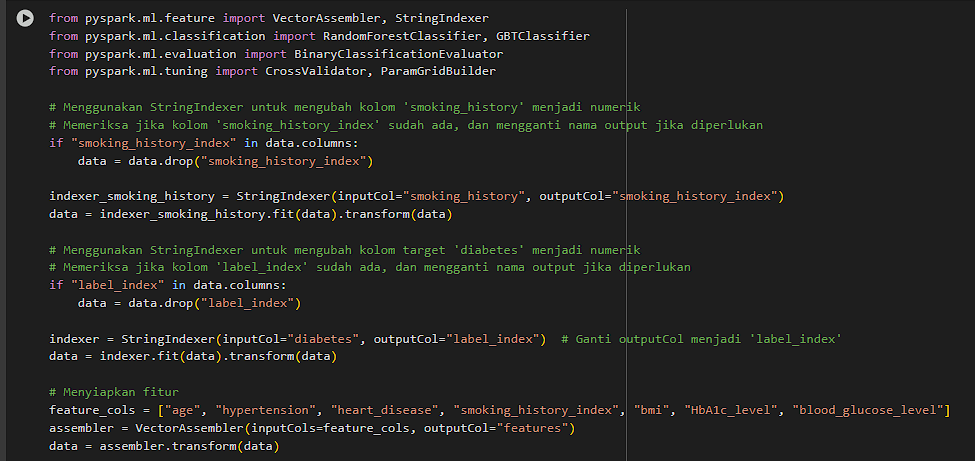
Metrik Evaluasi: Untuk membandingkan kinerja keempat model tersebut, kami akan menggunakan beberapa metrik evaluasi klasifikasi, yaitu:

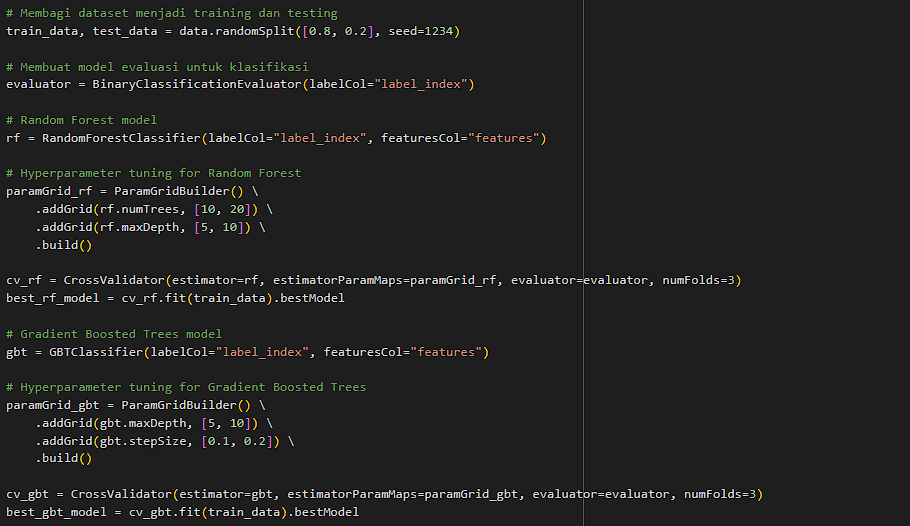
* AUC (Area Under the Curve): Mengukur kualitas prediksi model dengan cara menghitung luas di bawah kurva ROC (Receiver Operating Characteristic). AUC yang lebih tinggi menunjukkan kemampuan model dalam membedakan antara kelas positif dan negatif.
* Akurasi: Persentase prediksi yang benar di antara seluruh data. Akurasi adalah metrik yang paling sederhana untuk mengevaluasi model, tetapi tidak selalu mencerminkan performa model, terutama jika data sangat tidak seimbang.
* F1 Score: Harmonik rata-rata antara precision dan recall. F1 score memberikan gambaran yang lebih jelas tentang keseimbangan antara keduanya, terutama ketika data sangat tidak seimbang.
* Precision: Mengukur seberapa banyak dari prediksi positif yang benar-benar positif.
* Recall: Mengukur seberapa banyak dari data yang sebenarnya positif yang berhasil diprediksi sebagai positif.

Berikut adalah contoh kode untuk evaluasi model dengan menggunakan PySpark:

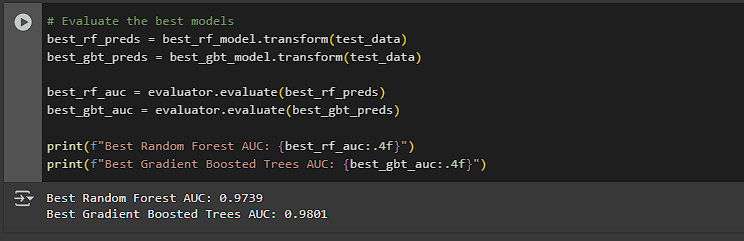
**7. Penggunaan 4 Model Machine Learning**

Di tahap ini, kami memilih dua model dengan performa yang paling baik yaitu **Random Forest Classifier** dan **Gradient Boosted Trees (GBT) Classifier**





**8. Karakteristik Model Terbaik**



Penjelasan: Hasil evaluasi menunjukkan bahwa model **Gradient Boosted Trees memiliki AUC sebesar 0.9801**, sedikit lebih unggul dibandingkan dengan **Random Forest yang memiliki AUC sebesar 0.9739**. Berdasarkan hasil ini, **Gradient Boosted Trees** dapat dipilih sebagai **model yang lebih baik** untuk tahap hyperparameter tuning dan analisis lebih lanjut.

**Gradient Boosted Trees (GBT)** adalah model yang sangat baik untuk data dengan pola yang rumit dan tidak langsung. Model ini cocok untuk data yang memiliki hubungan kompleks antar kolom atau fitur, tanpa perlu mengubah skala data terlebih dahulu. **GBT** juga bisa menangani data yang memiliki nilai kosong dengan baik dan membantu kita memahami fitur mana yang paling berpengaruh. Namun, jika data terlalu banyak noise (data yang tidak relevan), model ini bisa menjadi terlalu "pas" dengan data tersebut, sehingga perlu diatur dengan hati-hati. Keunggulan inilah yang membuat **GBT** bekerja lebih baik pada data dengan pola yang sulit ditemukan, seperti yang terlihat dari hasil evaluasi sebelumnya.