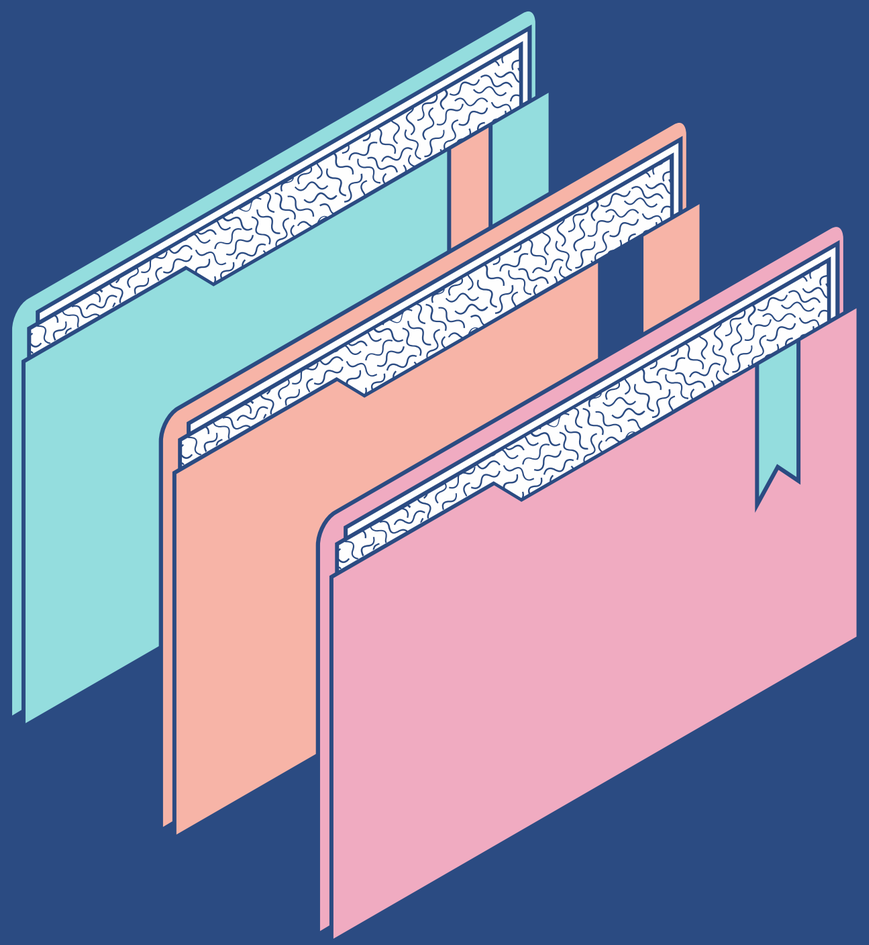
Klasifikasi Risiko Diabetes dengan Analisis Data Berbasis AI

Nama : Rizky Nanda Anggia



Daftar Isi

TOPIK UTAMA DALAM PRESENTASI INI

Project Overview

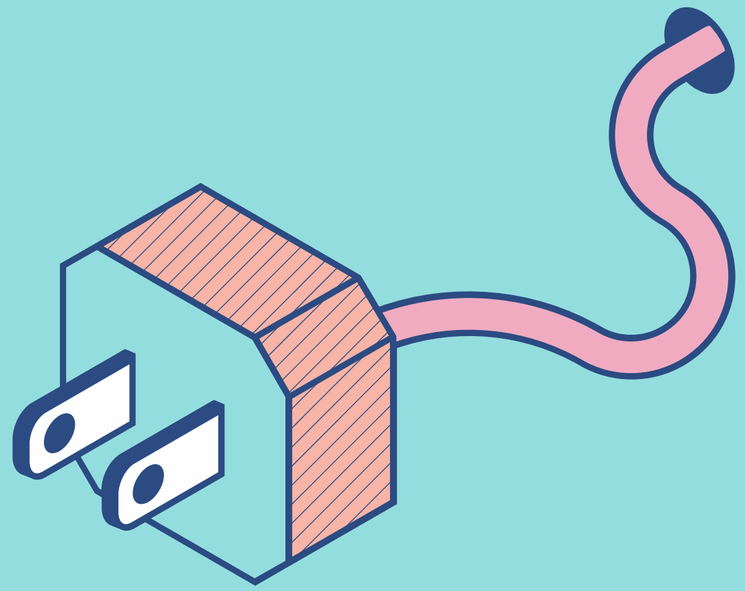
Latar Belakang:
Diabetes adalah penyakit kronis dengan prevalensi meningkat secara global. Prediksi dini sangat penting untuk pencegahan.

Tujuan:

- Membuat model untuk memprediksi risiko diabetes menggunakan dataset publik.
- Membandingkan dua algoritma: Logistic Regression & Random Forest.
- Memberikan insight & rekomendasi kesehatan.
- memanfaatkan AI (IBM Granite / Watson) untuk mempercepat analisis.
- dan terakhir untuk menyempurnakan project dari website model diabetes menggunakan random forest yaitu sebagai berikut:
  https://diabetesproject.pythonanywhere.com/

Dataset: Diabetes data publik yang saya ambil dari kaggle berjumlah 46 fitur dan 1879 baris data (diabetes_data.csv) dengan link dataset sebagai berikut:

- https://www.kaggle.com/datasets/rabieelkharoua/diabetes-health-dataset-analysis



Data Loading & Setup

```
import pandas as pd

df = pd.read_csv("diabetes_data.csv")
df.head()
```

Dataset diabetes_data.csv dibaca menggunakan pandas. Granite (LLM) dan Watson API disiapkan untuk mendukung analisis berbasis AI.

Exploratory Data Analysis (EDA)

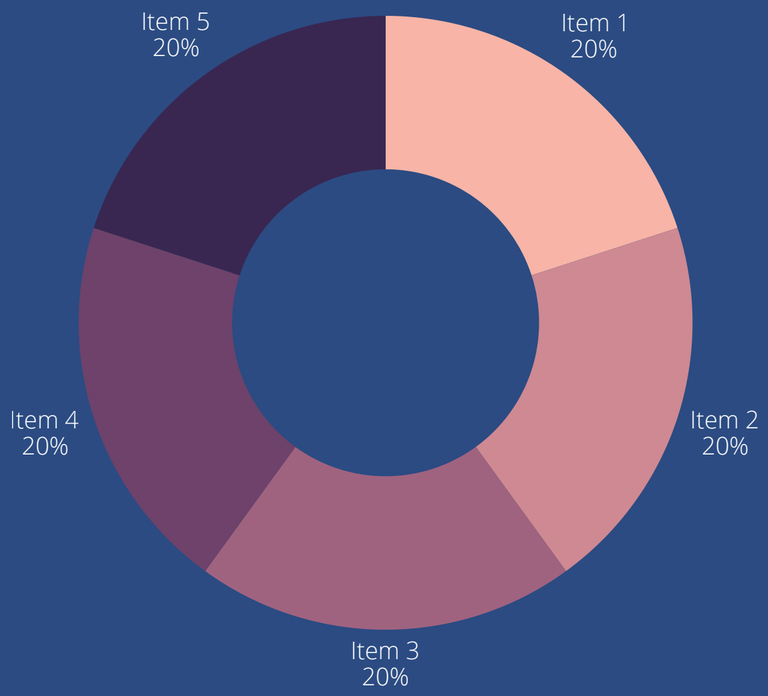
```
agent.invoke({"input": "Show the first 5 rows of the dataset along with the column names"})
agent.invoke({"input": "Provide summary statistics (mean, std, min, max) for all columns"})
agent.invoke({"input": "Check if there are any missing values in the dataset"})
agent.invoke({"input": "Visualize the distribution of the target variable (Outcome) using a bar chart"})
```

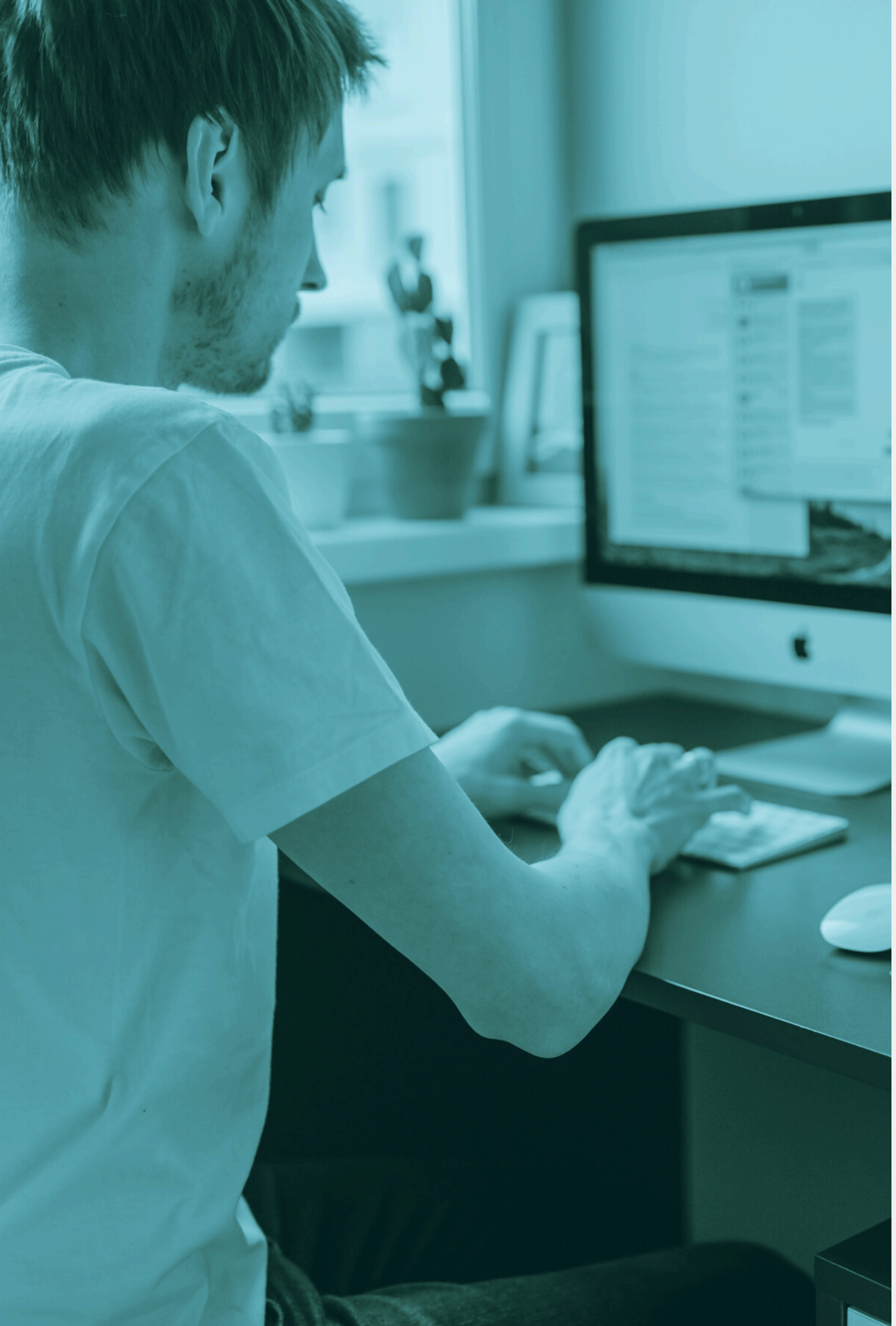
EDA digunakan untuk memahami struktur data, distribusi target, korelasi antar fitur, dan kondisi missing values.

Data Cleaning

AGENT.INVOKE({"INPUT": "FOR EACH NUMERIC COLUMN IN THE DATAFRAME, FILL MISSING VALUES WITH THE COLUMN MEAN. DO NOT APPLY THIS TO NON-NUMERIC COLUMNS."})

Missing values pada fitur numerik diisi dengan nilai rata-rata (mean) agar dataset tetap lengkap tanpa menghapus baris.





Modeling

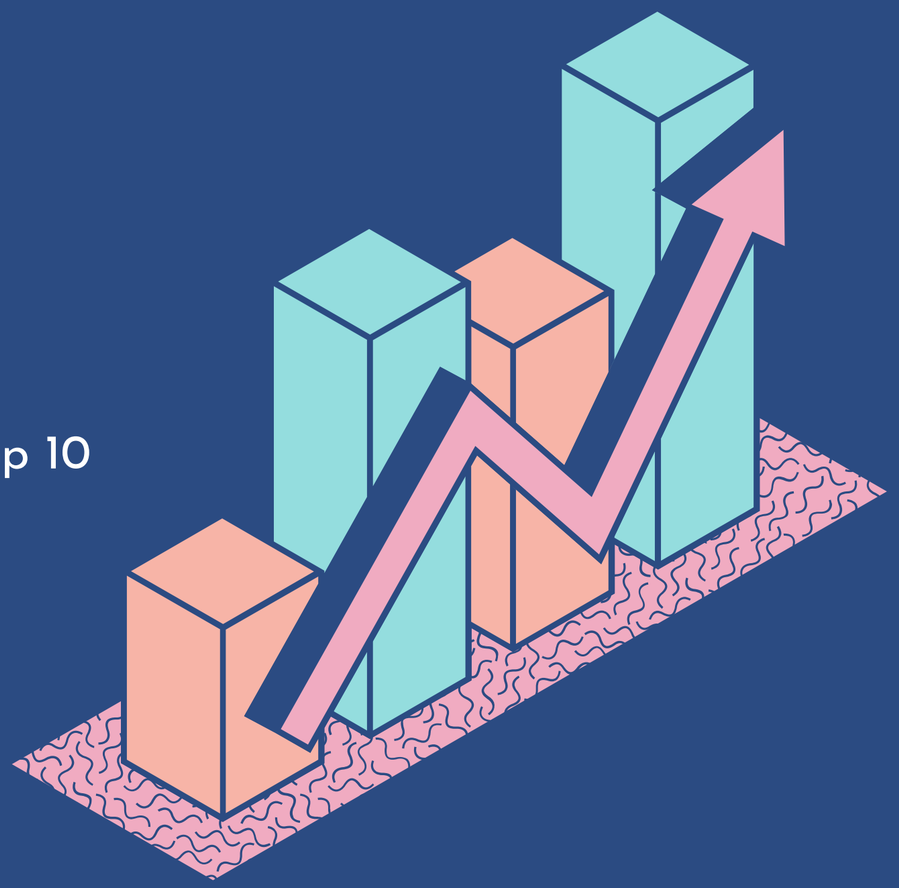
```
AGENT.INVOKE({"INPUT": "TRAIN A LOGISTIC REGRESSION MODEL TO PREDICT DIABETES USING ALL FEATURES"})
AGENT.INVOKE({"INPUT": "TRAIN A RANDOM FOREST MODEL AND COMPARE ITS PERFORMANCE WITH LOGISTIC REGRESSION"})
AGENT.INVOKE({"INPUT": "EVALUATE THE MODEL USING ACCURACY, PRECISION, RECALL, AND F1-SCORE"})
```

Model Logistic Regression digunakan sebagai baseline, lalu Random Forest dilatih sebagai model utama. Evaluasi dilakukan dengan akurasi, precision, recall, dan F1-score.

Feature Importance

agent.invoke({"input": "Calculate and return only the top 10 most important features for predicting diabetes using Random Forest"})

Random Forest menunjukkan fitur paling berpengaruh terhadap diabetes, dengan Glucose, BMI, Age, dan HbA1c sebagai top 4.



Insights & Explanation

AGENT.INVOKE({"INPUT": "EXPLAIN WHY RANDOM FOREST PERFORMS BETTER THAN LOGISTIC REGRESSION FOR THIS DATASET"})

Random Forest bekerja lebih baik dibanding Logistic Regression karena mampu menangani hubungan non-linear, lebih tahan outlier, dan memberikan feature importance.

Conclusion & Recommendations

```
agent.invoke({"input": "Summarize the overall findings from the analysis in bullet points"})
agent.invoke({"input": "Provide actionable recommendations based on the model results and feature importance"})
```

Random Forest terbukti lebih akurat (~90%) dibanding Logistic Regression (~70%).
Rekomendasi: lakukan pemeriksaan rutin gula darah & BMI, gunakan model ini sebagai alat bantu screening.

AI Support Explanation

- IBM GRANITE VIA REPLICATE: MENERJEMAHKAN INSTRUKSI BAHASA INGGRIS MENJADI KODE PYTHON UNTUK ANALISIS DATA.
- IBM WATSON ML API: MENDUKUNG PENGELOLAAN MODEL AGAR MUDAH DIKEMBANGKAN LEBIH LANJUT.
- DUKUNGAN AI MEMBUAT WORKFLOW LEBIH CEPAT, KONSISTEN, DAN EFISIEN.

AI Support Explanation
- IBM Granite via Replicate digunakan untuk mengeksekusi instruksi bahasa alami menjadi kode Python.
- IBM Watson ML API digunakan untuk mengelola dan mendukung model.
- Dengan dukungan AI, analisis menjadi lebih cepat, mudah direplikasi, dan terstruktur.

Closing & Limitations

- Proyek ini berhasil menunjukkan bahwa Random Forest lebih baik daripada Logistic Regression dalam memprediksi risiko diabetes.
- Keterbatasan:
- Dataset bersifat publik & terbatas → mungkin tidak merepresentasikan seluruh populasi.
- Belum ada validasi pada data real-world klinis.
- Next Step:
- Bisa ditambah dataset medis lain.
- Bisa diuji dengan model AI lebih canggih (misalnya Gradient Boosting atau Neural Networks).

Access to Notebook & Code

- Untuk melihat hasil lengkap analisis dan modeling, silakan akses notebook di Google Colab.
- 📎 Google Colab Link: https://colab.research.google.com/drive/1HeMsj0he7bpc4CJAStM_FjRqUbp1Pg0Y?usp=sharing

- 📎 GitHub Repository (opsional): https://github.com/AnggiAllied/Final-Project-Klasifikasi-Risiko-Diabetes-dengan-Analisis-Data-Berbasis-AI

Ada pertanyaan?

@anggiallied@gmail.com
Terimakasih! atas perhatiannya!