MEDALLION ARCHITECTURE

DIABETES RISK ANALYSIS

Built a production-level data pipeline on Databricks using PySpark

MEDALLION ARCHITECTURE

1

BRONZE

Read and save raw data

2

SILVER

Change columm names,
Handle missing values,
Handle outliers,
Calculate BMI and add its column,
Save the data as a table

3

GOLD

Develop custom features in the Gold Layer to capture complex health-risk patterns that raw data alone cannot represent

GOLD LAYER

FEATURE ENGINEERING HIGHLIGHTS

BMI_GROUP	IS_ OVERWORKED	WORK_METABOLIC_ BENEFIT	LEISURE_VITALITY_ SCORE
Grouped BMI into 'Obese', 'Overweight', and 'Normal'. This transforms non-linear physiological risks into interpretable segments for the model and business reporting	Flagged users working >40 hours/week. This acts as a proxy for stress-related metabolic risks and lifestyle imbalance.	Quantified the "sweet spot" of physical labor. I assigned positive weights ($+1.5$) to moderate intensity but applied a penalty (-0.5) to high-intensity work to account for potential stress-related health trade-offs.	Weighted high-intensity leisure activities (2.0) more heavily than moderate ones (1.0). This emphasizes the quality of physical recovery outside of working hours as a key preventative factor.

AD-HOC (EDA)

While high-intensity at work often failed to show benefits due to occupational stress, high-intensity in leisure leads to pure metabolic gain. It effectively lowers blood sugar because it is voluntary and enjoyable.

This protective effect exists independent of BMI. This means that even without significant weight loss, the physiological act of engaging in vigorous fun directly improves your insulin response.

If you have limited time, "vigorous leisure" (High Intensity) offers the highest "Return on Investment" for diabetes prevention. One day of intense play is more valuable than one day of moderate effort.

1 **"High Intensity is Joy, Not Stress"**

2 **The "Active Fun" Shield**

3 **Efficiency Comparison (ROI)**