Predicting Diabetic Readmission: A Databricks Lakehouse Solution

End-to-End Medallion Architecture for Risk Stratfication in Healthcare

databricks

CODE BASICS

INDIAN DATA CLUB

INTRODUCTION

Suchorita Das (She/Her)

Data Analyst Intern — AtliQ Technologies

Core Expertise

- Data Visualization
- Statistical Problem Solving

Technical Skills

- Python
- SQL
- Microsoft Excel
- Power BI

Professional Affiliations

- AtliQ Technologies
- Ivy Professional School

Location

- Greater Kolkata Area

Build with Databricks: A Hands-On Challenge

A hands-on project challenge designed to test and build your skills.

Sponsored by

The official data and AI platform partner for this challenge.

Powered by

INDIAN DATA CLUB

The community driving and supporting this initiative.

Organized by

CODE BASICS

The educational platform leading and presenting the event.

Project Objective & Success Criteria

Primary Objective

Build an end-to-end data pipeline that ingests raw hospital records and outputs a binary prediction: Will this patient be readmitted within 30 days?

Scalability

Must use industry-standard Databricks Medallion architecture to handle volume.

Auditability

Data lineage must be preserved from ingestion to inference.

Interpretability

The model must be a “Glass Box”. Clinicians need to understand why a patient is flagged to trust the system.

Breaking the Cycle: The Diabetic Readmission Challenge

The Business Problem: The Cost of Readmission

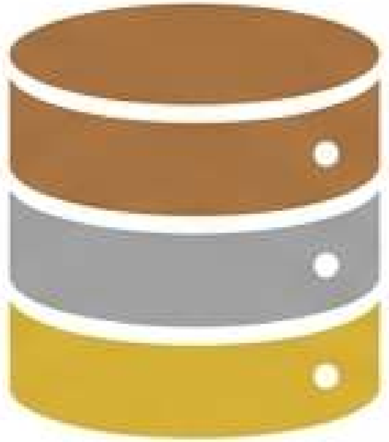
- **Context:** 30-day readmission rates are a critical quality metric influencing insurance reimbursements.
- **The Gap:** Current methods are reactive. Hospitals struggle to distinguish between patients likely to recover and those likely to return.
- **Goal:** Shift from reactive treatment to proactive risk management using historical operational data.

Executive Summary

The Challenge

Hospitals face financial penalties and operational strain due to high 30-day readmission rates among diabetic patients.

The Solution



A scalable Lakehouse architecture processing 10 years of clinical data (1999–2008) to feed an interpretable Logistic Regression model.

The Outcome

A transparent risk-scoring engine that identifies key drivers—such as prior inpatient visits—allowing clinicians to intervene before discharge.

~100k
Patient Encounters

50+
Variables Analyzed

White Box
Transparency

Engineering Directive: A Scalable, Audit-Ready Pipeline

Objective

Build a full data-to-decision workflow, not just a standalone model.

The Dataset

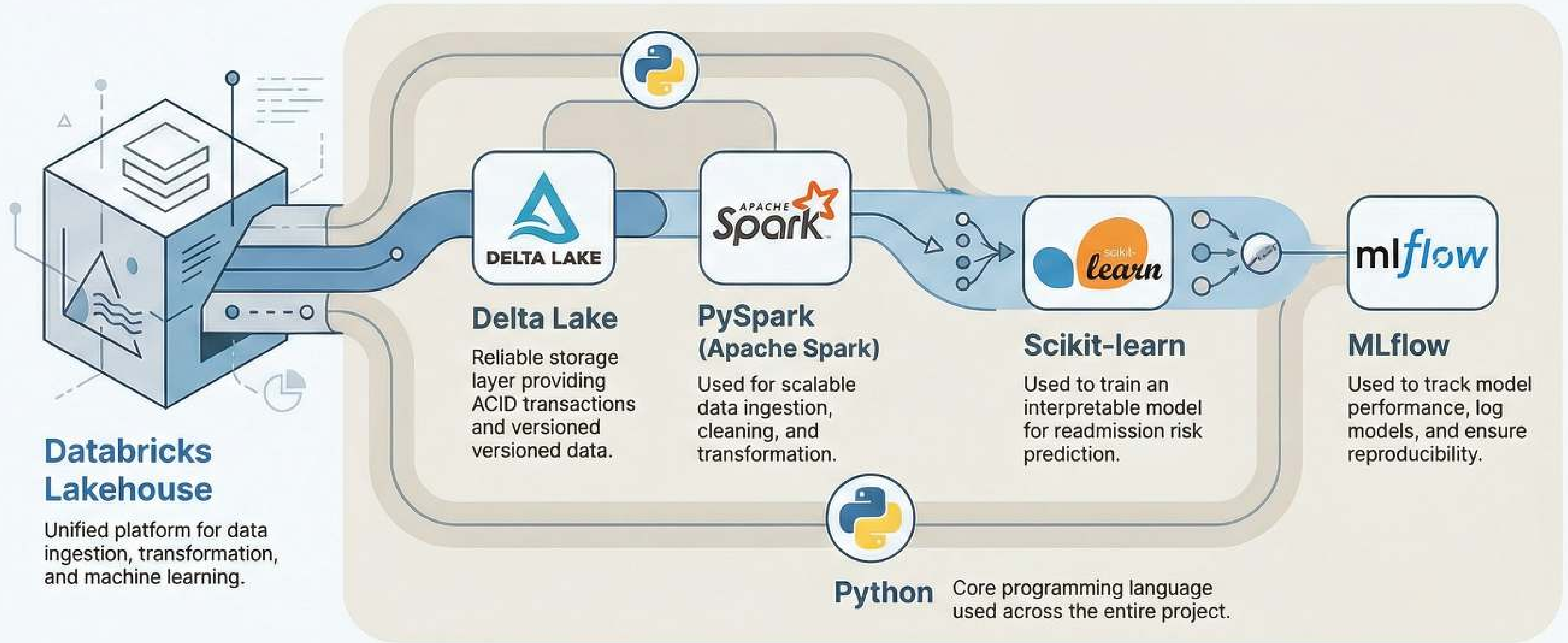
US	~100,000	50+
Hospitals (1999–2008)	Patient Encounters	Features (Demographics, Labs, Meds)

Key Constraints

- **Auditability:** Must preserve data lineage for healthcare compliance.
- **Interpretability:** No “black boxes”—clinicians must understand risk drivers.
- **Reproducibility:** Experiments must be tracked and versioned.

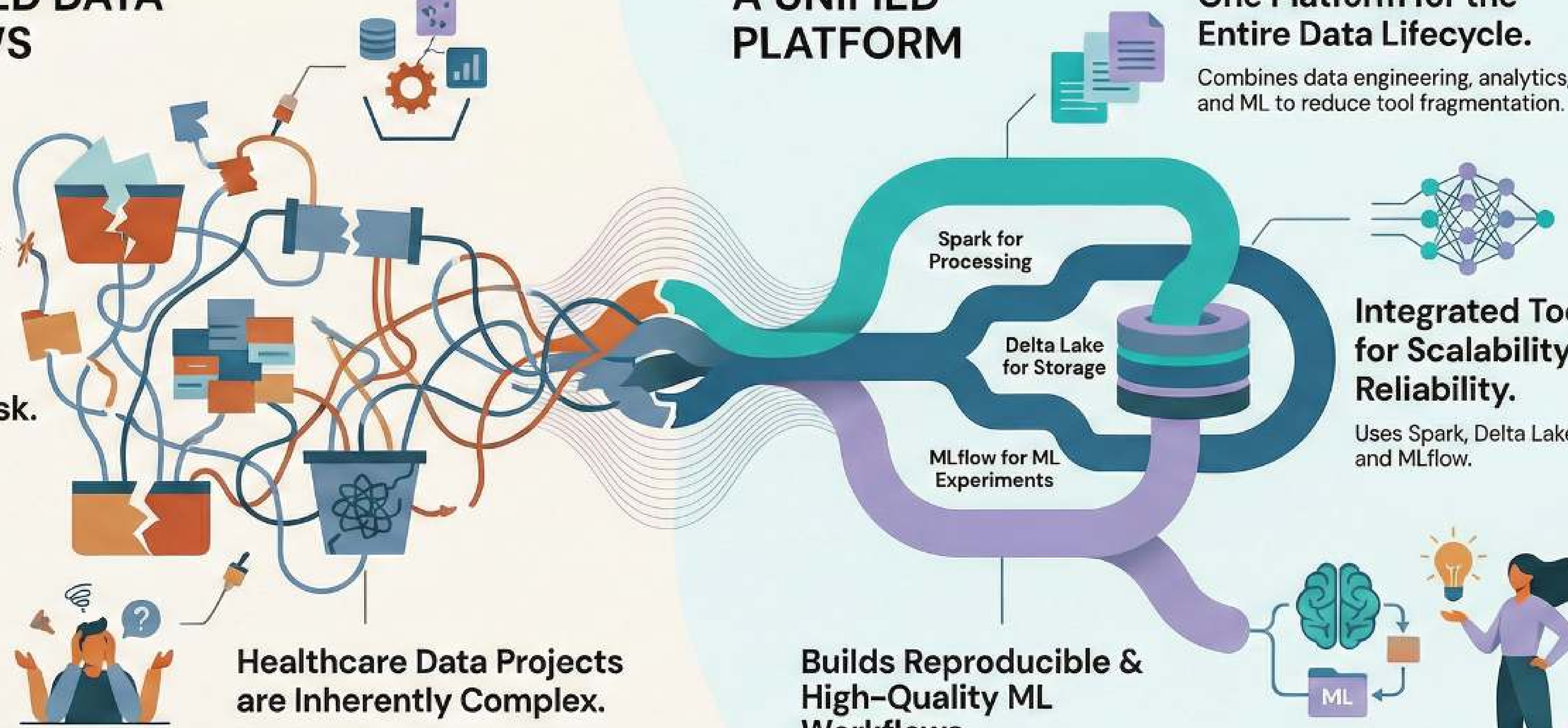
Tools & Technologies Used

Key technologies used in a Databricks machine **learning project**, from data to ML tracking.



Unifying Healthcare Analytics: The Databricks Advantage

THE CHALLENGE: FRAGMENTED DATA WORKFLOWS



Disconnected Tools Increase Complexity & Risk.

Managing data stages separately leads to errors and high maintenance coveronde overhead.

Healthcare Data Projects are Inherently Complex.

They involve large datasets, multi-stage transformations, and strict governance requirements.

THE SOLUTION: A UNIFIED PLATFORM

One Platform for the Entire Data Lifecycle.

Combines data engineering, analytics, and ML to reduce tool fragmentation.

Integrated Tools for Scalability & Reliability.

Uses Spark, Delta Lake, and MLflow.

Builds Reproducible & High-Quality ML Workflows.

Enables real enterprise architecture while ensuring data quality and traceability.

From Guesswork to Guidance: Using Machine Learning to Predict Diabetes Readmissions

THE PROBLEM:
LIMITATIONS OF TRADITIONAL ANALYSIS

Predicting 30-day diabetes readmission is difficult due to multiple interacting factors like patient history and hospital stay patterns. Traditional analysis methods are insufficient, creating a need for a more advanced, transparent, and scalable solution to support clinical decision-making.

Too Complex for Simple Rules

Traditional methods struggle to capture the complex relationships between numerous patient factors.

Fails to Scale & Adapt

Manual analysis cannot efficiently process large datasets or adapt to new incoming data.

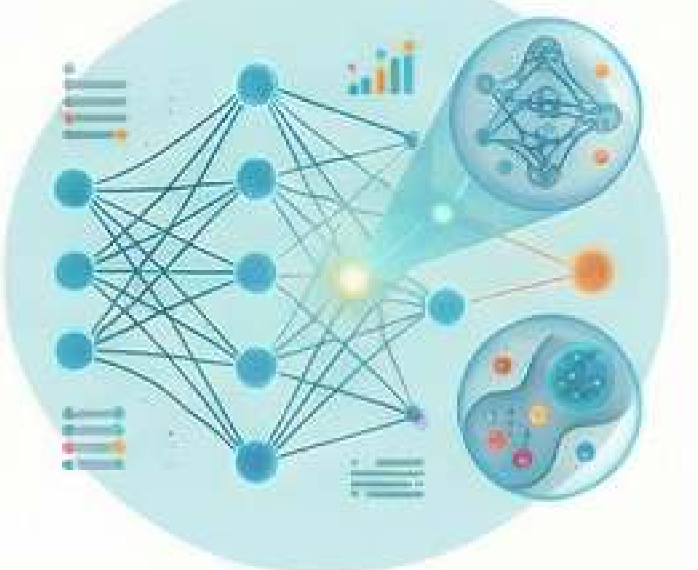
Offers Only Binary Decisions

Lacks the nuance of a personalized risk score, providing simple yes/no answers.

THE SOLUTION:
EXPLAINABLE MACHINE LEARNING

Identifies Hidden Patterns

ML analyzes high-dimensional healthcare data to find patterns not visible to humans.



Generates Nuanced Risk Scores

Estimates the probability of readmission, allowing clinicians to rank patients by risk.

Builds Trust Through Transparency

Explainable models show *why* a patient is flagged, supporting—not automating—clinical decisions.

The Lakehouse Architecture

A unified platform for data engineering and data science, moving from raw chaos to refined insights.

Bronze Layer: Preserving the Source of Truth

Core Concept: Data is ingested "as-is" without transformation. This ensures full lineage and auditability—critical for healthcare compliance.

```python
# The Bronze layer stores the raw dataset exactly as received

raw_df = (spark.read.option('header', 'true')
    .option('inferSchema', 'true')
    .csv('/Volumes/diabetes_readmission/healthcare/diabetic_data.csv'))
```

```python
dbutils.fs.ls('.../healthcare/')
```

Code Example: Raw Data Ingestion & Verification

Silver Layer: Refining Signal from Noise

Real-world medical data is messy. In this layer, we **standardize** features and handle missing values to prepare for analysis.

Handling Nulls

```
silver_df =
silver_df.replace('?', None)
    None
```

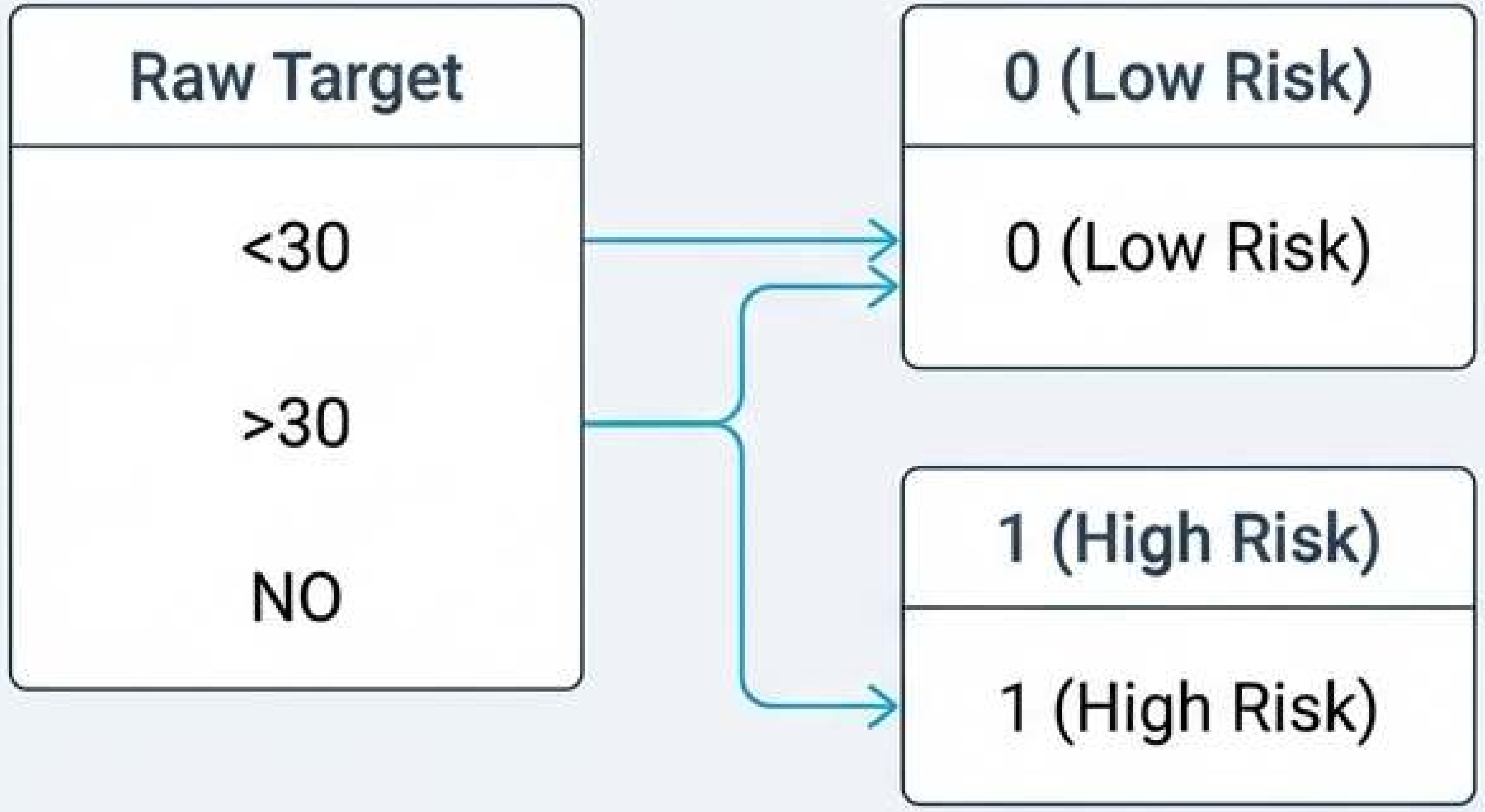
Schema Validation

Converting raw string columns into proper statistical data types.

Dropping Identifiers

Removing 'encounter_id' and 'patient_nbr' to prevent model leakage.

Feature Engineering: Defining the Target




```
silver_df =
silver_df.withColumn('readmitted_30',
   when(col('readmitted') == '<30',
   1).otherwise(0))
```

Outcome: A clear binary flag where 1 indicates immediate action required.

Gold Layer: Refining for Machine Learning

Feature Categories

- Demographics
- Medications
- Procedures
- Hospital Stay Details

```
feature_cols = ["race", "gender",
                "time_in_hospital",
                "num_lab_procedures",
                "insulin", ...]

gold_ready_df = gold_df.select(feature_cols
    + ["readmitted_30"])

# Polished Silver #AA0A0A0
gold_ready_df.write.saveAsTable(
   "...gold_diabetes_ready")
```

Curating specific clinical signals.

Final hand-off to Data Science.

Modeling Strategy: Trust over Complexity

Strategy

Model Choice: Logistic Regression.

Rationale: In healthcare, explainability is paramount. Clinicians need to know why a patient is flagged. While neural networks may offer marginal accuracy gains, they lack the transparency required for this use case.

Preparation Code

```
for col in X.columns:
    if X[col].dtype == 'object':
        X[col] = LabelEncoder().fit_transform(X[col])

scaler = StandardScaler()
```

Clinical Engineering

Production-Ready MLOps with MLflow

```python
with mlflow.start_run(run_name
="Logistic_Regression_Baseline"):
    mlflow.log_metric("AUC", auc)

    mlflow.sklearn.log_model(lr_scaled,
"logistic_regression_model")

```

Tracking Area Under Curve (AUC) to handle class imbalance.

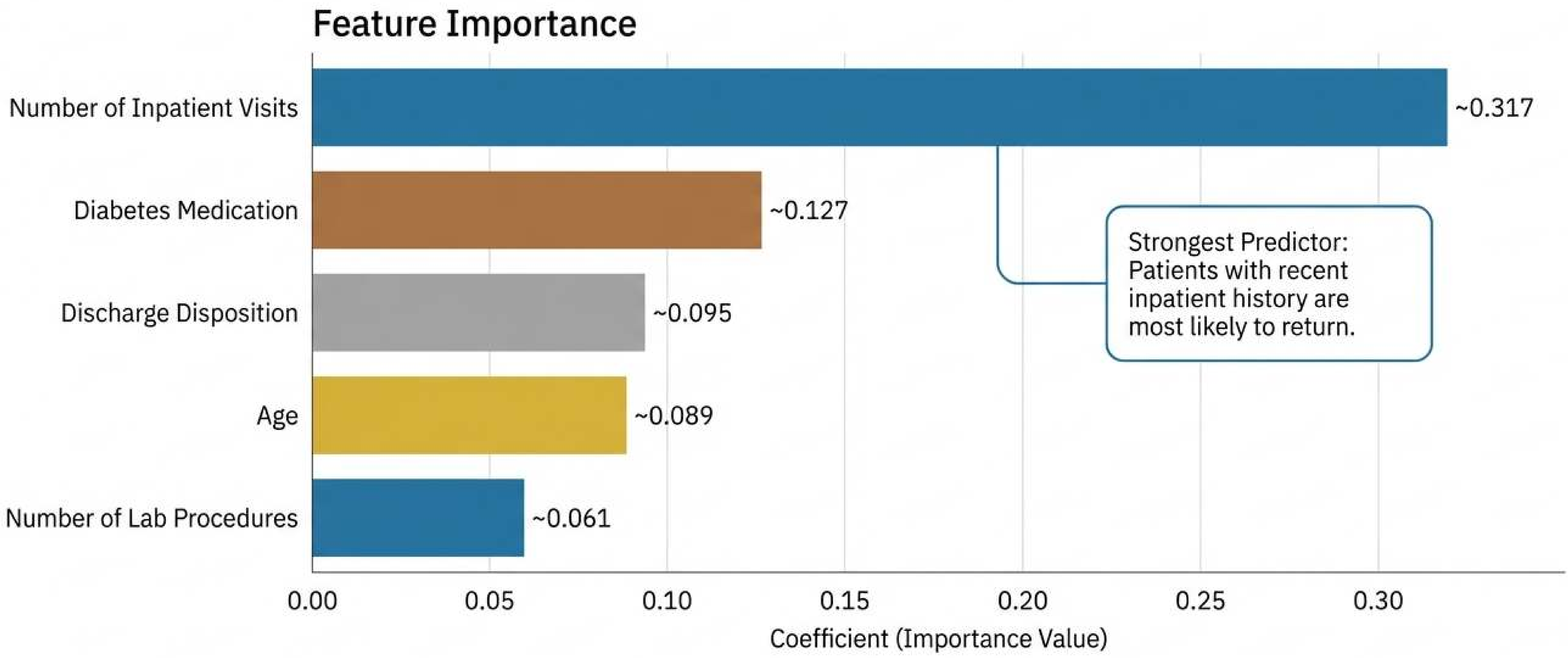
Model Performance: A Realistic Baseline

AUC Score: 0.64

This represents a valid ‘better than random’ baseline on **highly noisy real-world data**. It proves the pipeline functions end-to-end and provides a benchmark for future model iteration.

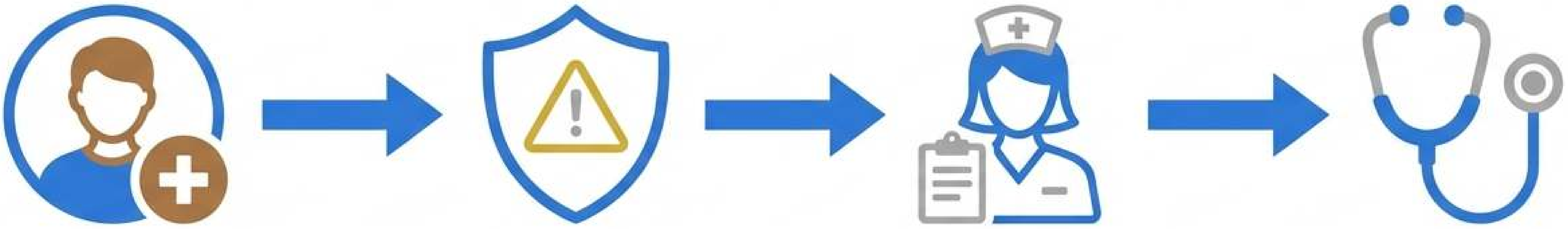
True Negative (Correctly Safe)	False Positive (False Alarm)
False Negative (Missed Risk)	True Positive (Correctly Flagged)

Decoding Risk: Key Drivers of Readmission



Clinical Engineering

From Prediction to Intervention



Pipeline Processes Patient Data → Model Flags ‘High Risk’ (> Threshold) → Nurse Schedules 72hr Follow-up → Doctor Reviews Meds

Impact: Proactive care reduces readmission rates, improves patient health, and avoids penalties.

Predicting 30-Day Hospital Readmissions: A Project Impact Summary

THE PROBLEM

High Hospital Readmission Rates
Diabetic patients were frequently being readmitted to the hospital shortly after discharge.

Lack of Proactive Identification
There was no effective way to identify which patients were at high risk.

OUR SOLUTION

End-to-End Data Pipeline
A structured pipeline was built in Databricks to process patient data efficiently.

Explainable Machine Learning Model
We used a transparent AI model to predict risk without being a "black box".

Clear Risk Flag
The model produces a simple, binary flag to indicate a patient's 30-day readmission risk.

KEY OUTCOMES

Early Identification of At-Risk Patients
Healthcare teams can now see which patients need extra attention before they leave.

Actionable Insights for Care Teams
The risk flag enables targeted interventions and better care planning.

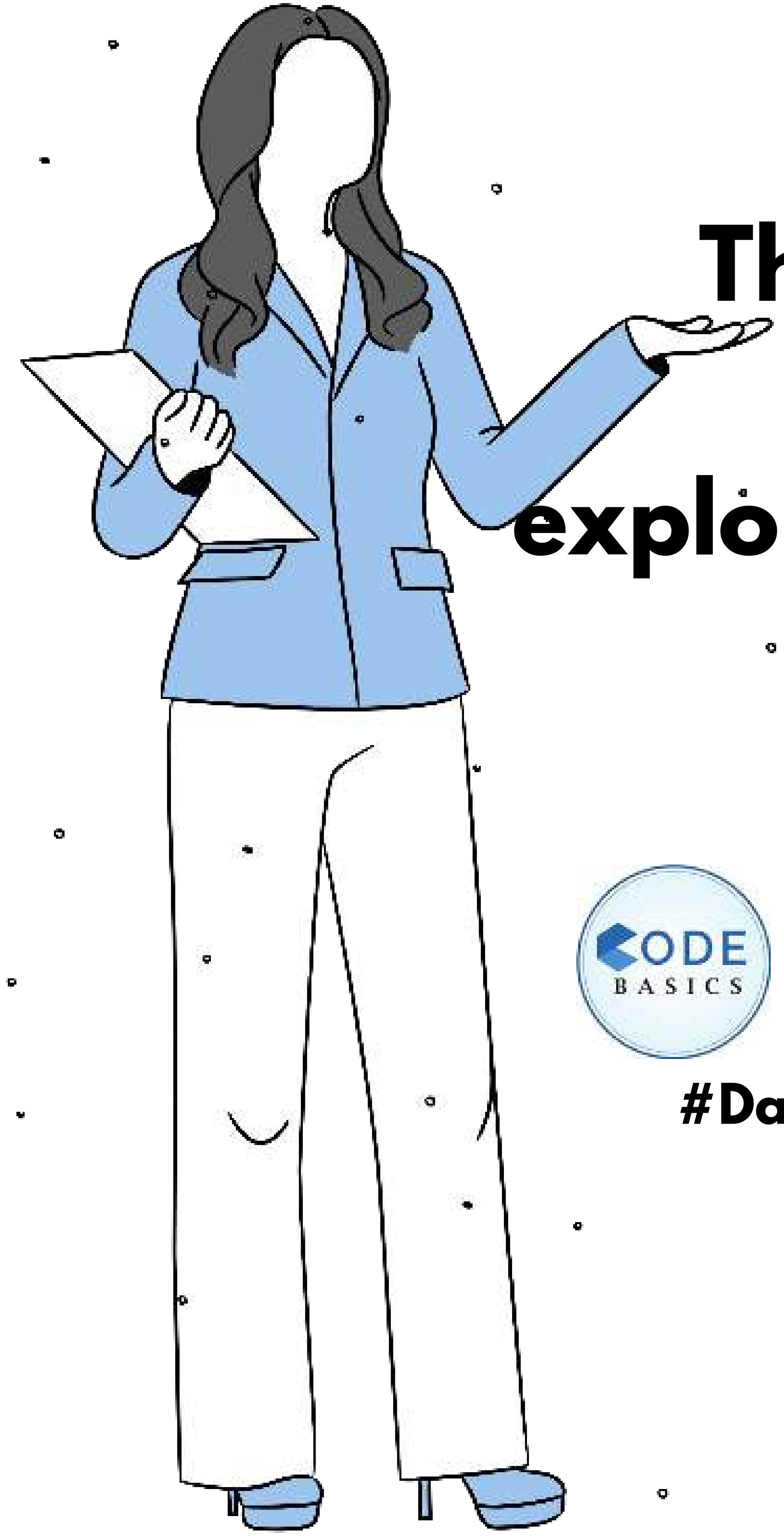
Reduced Readmission Risk
Proactive care helps lower the chances of patients returning to the hospital.

WHY IT MATTERS

Better Patient Outcomes
Patients receive more personalized care, leading to improved health and well-being.

Lower Healthcare Costs
Reducing readmissions saves significant money for both patients and providers.

Trustworthy AI in Healthcare
Demonstrates the value of explainable AI for making critical healthcare decisions.



Thank you for exploring with me

Suchorita Das

#DatabricksWithIDC