AI-Powered Supplier Sustainability Evaluation

A MACHINE LEARNING APPROACH TO ESG RISK CLASSIFICATION

SAMAN NAGHSHI | APRIL 10, 2025

Global Context & Motivation

- Coffee is a globally consumed commodity with significant ESG implications.
- Deforestation, water use, and labor rights are major concerns.
- Sustainable sourcing is essential for companies and regulators.
- ESG data is fragmented and hard to evaluate.
- Procurement teams lack objective risk assessments.
- Can we use machine learning to predict supplier risk?

The Solution

- Used Random Forest to classify suppliers as Low, Medium, or High Risk.
- Features: ESG Score, Certifications, Emissions, Cost, Distance, Violations.
- Built a visualization map to locate risks globally.

How We Created the Dataset

- We created a **synthetic dataset** of 100 coffee suppliers using controlled randomness and domain knowledge.
- Each supplier is randomly assigned to a major coffee-producing country (e.g., Brazil, Ethiopia, Indonesia).
- Features were generated using realistic distributions:
- **ESG Score** ~ Normal(7.5, 1.2), clipped between 4 and 10
- **Coffee Quality Score** ~ Normal(80, 5), clipped between 70 and 95
- **Emissions**, **Distance**, and **Cost** generated with variability
- **Certifications** added using binomial (yes/no) probabilities
- **Latitude and Longitude** are randomly assigned **within the actual boundaries** of the supplier's country → makes mapping possible!
- **Sustainability Risk** is assigned using ESG and violation data plus controlled **random noise** to simulate real-world messiness.

Table 1: Data Structure

Feature	Type	Example/Source
Supplier Name	Identifier	"Coop_Coffee_Rwanda", "Fair_Bean_Colombia"
Country of Origin	Categorical	Brazil, Colombia, Ethiopia, Vietnam, Rwanda
Coffee Quality Score	Numerical (0-100)	Specialty Coffee Scores (Q-Grader)
ESG Score	Numerical (0-10)	Synthetic or derived from industry reports
Certification (Fairtrade/Organic)	Binary (0 or 1)	Fairtrade Intl., Rainforest Alliance
Distance to Market (km)	Numerical	Synthetic (realistic)
Emissions per shipment (kg CO_2)	Numerical	Estimated by shipping route/distance
Cost per shipment (USD)	Numerical	Synthetic, realistic ranges
Historical ESG violations	Numerical (count)	Simulated data
Risk/Sustainability Class (Target)	Categorical	Low, Medium, High

Why We Chose Random Forest

Real-world ESG data is **messy, nonlinear, and contains both numeric and categorical variables.** Random Forests are:

- Ensemble models — combine many weak learners (trees) into a strong predictor
- Robust to noise & overfitting
- Handle feature interactions and missing information well
- Offer interpretability via feature importance & tree inspection

Supply Risk Model

- 100 decision trees trained on different data and features.
- Each tree votes on classification.
- Final output: majority vote of all trees.

Random Forest General Idea

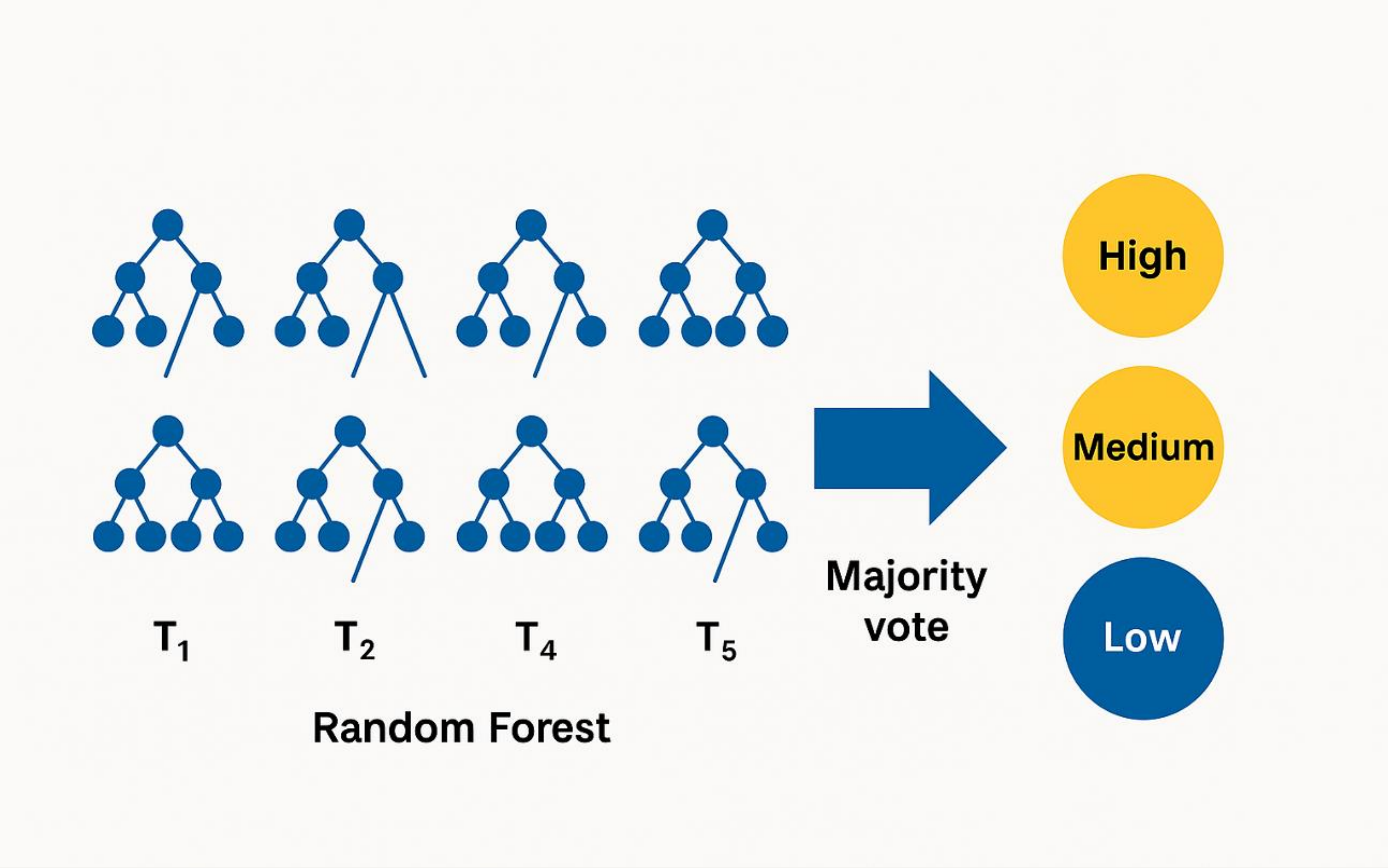

Figure 3: Overview of the Model

Example of Tree Logic

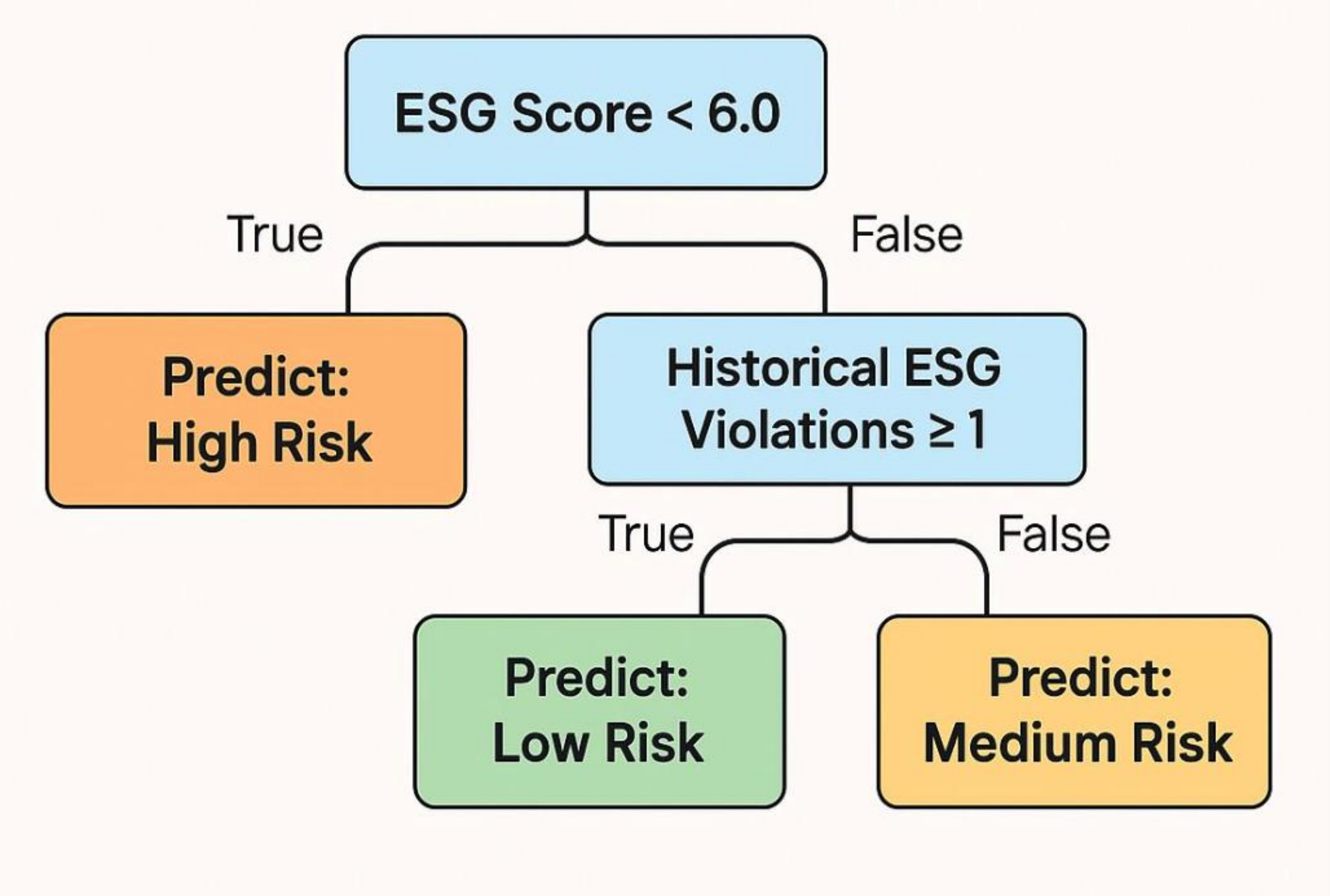

Figure 1: Subset of a Decision Tree

Example of Tree Logic

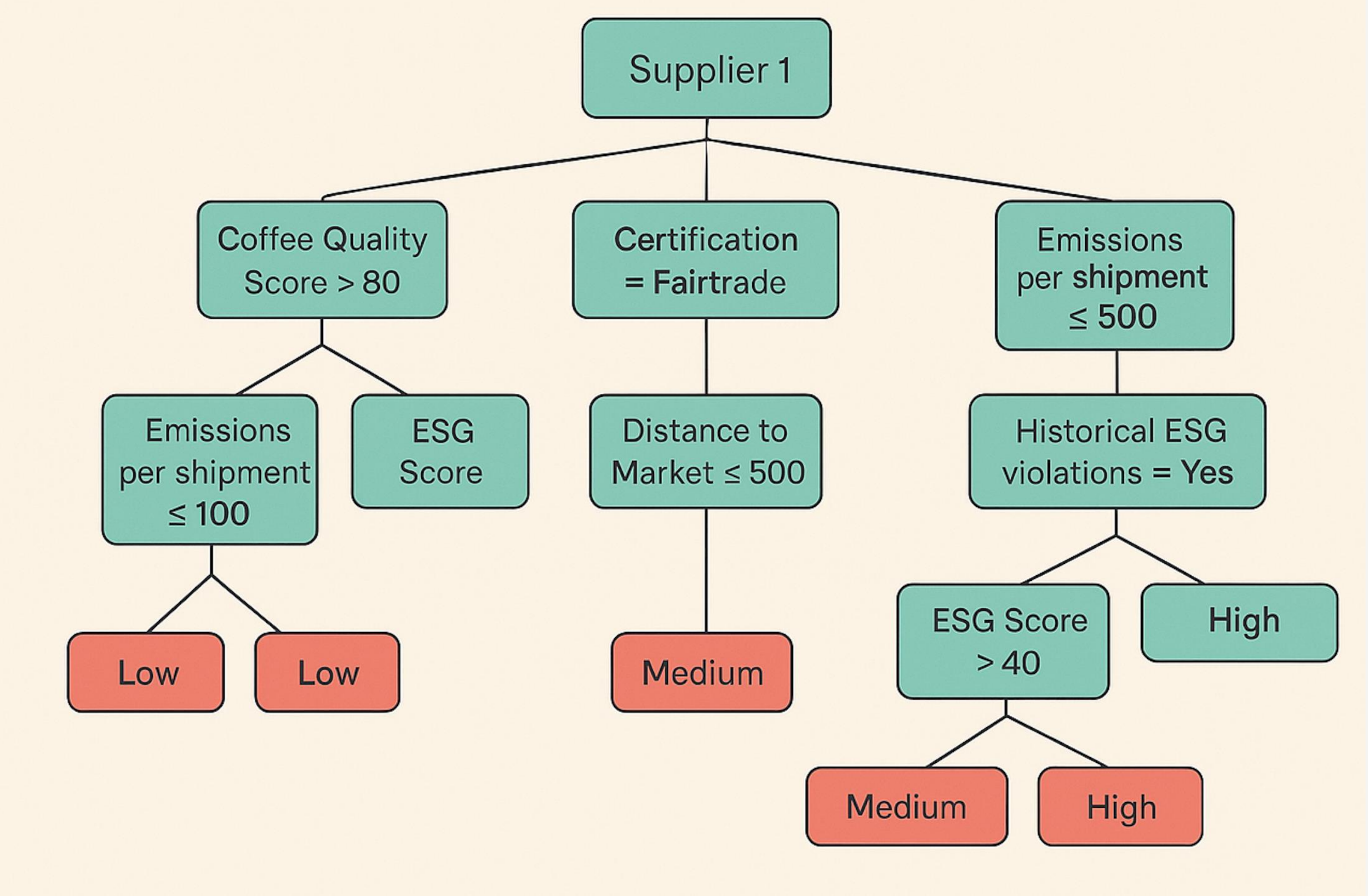

Figure 2: Decision Tree

Table 2: The Importance of Variables

Feature	Importance	What It Tells Us
ESG_Score	0.308	The **strongest** signal — higher ESG → lower risk
Historical_ESG_Violations	0.162	Repeated violations are a major red flag
Coffee_Quality_Score	0.089	High quality correlates with lower risk — surprising but insightful
Longitude / Latitude	~0.15 total	Geography matters — region-specific ESG patterns
Emissions_kg_CO2	0.073	Higher emissions → higher risk
Distance_km	0.063	Long shipping distances factor into sustainability
Cost_USD	0.060	Cost may reflect operational scale or region — indirectly meaningful
Country_Colombia	0.027	Specific country-level patterns are being picked up
Certification_Organic	0.019	Organic certification has **some** impact, less than expected

Supplier Risk Map

Figure , Sustainability risk of suppliers; low, medium, and high

Results

Evaluation Method:

- We used **5-fold cross-validation** to assess the model's generalization.
- This means the dataset was split into 5 parts:
  4 used for training, 1 for testing — repeated 5 times.
- Ensures the model’s performance isn't biased by any one particular split.

Performance Metrics Accuracy: 94% ± 5%

- High predictive power, even with noisy label generation.
- The original data had fewer 'High' risk suppliers.
- We used **upsampling** to balance the classes before training.
- This ensured the model didn't underperform on minority classes.

What This Means:

- The model learned the **underlying ESG risk logic**, not just memorized the data.
- It generalized well even with **realistic noise** and **feature overlap**.
- Shows that Random Forest is a strong choice for complex, real-world ESG problems.

Conclusion

What We Built

- A working ML prototype to classify coffee suppliers by **sustainability risk**
- Simulated realistic supply chain data across **7 countries**
- Used a **Random Forest** model to capture ESG patterns and make accurate predictions

What It Shows

- **ESG Scores** and **historical violations** are highly predictive — but geography, emissions, and certifications matter too
- The model successfully learned **nonlinear, fuzzy rules** — like those found in real-world ESG systems

The Big Picture

- This approach can make supply chains more **transparent**, **accountable**, and **data-driven**
- It offers a blueprint for scalable ESG risk tools for buyers, NGOs, and compliance teams

Q&A

I'm happy to answer your questions.