Introduction

This report details the "Global Trust Engine," a Machine Learning for Development Integrity (ML4DI) model designed as an early warning "financial smoke detector" to identify high-risk environments for corruption. The goal is to flag issues before catastrophic financial diversion occurs, safeguarding development funds. Our model's theoretical framework is based on insights from major scandals, including the 1MDB affair in Malaysia and the "Hidden Debts" crisis in Mozambique. These cases revealed common quantitative precursors, such as the degradation of governance indicators. Our objective is to use these recurring patterns to build a predictive model that detects these shifts and provides actionable risk assessments.

Data Collection and Methodological Framework

To construct a robust predictive model, comprehensive and reliable data are paramount. Our primary data source is the World Bank's Public API, providing a wealth of governance and economic indicators across numerous countries and years. Data spanning from 2010 to 2023 were meticulously collected for two categories of indicators:

The six World Governance Indicators (WGI) were selected as the core for our corruption risk labeling strategy, given their direct correlation with institutional strength and transparency – factors demonstrably compromised in our case studies. These include:

- Voice and Accountability
- Political Stability and Absence of Violence/Terrorism
- Government Effectiveness
- Regulatory Quality
- Rule of Law
- Control of Corruption

Five additional economic indicators were collected to serve as potential predictive features in the subsequent model training phase. Crucially, these were *not* used in the initial labeling process to avoid circularity. These include metrics such as External Debt as a percentage of GNI and GDP Growth.

Establishing the Baseline Model and Labeling Strategy

The first step involved loading a baseline dataset (corruption_data_baseline.csv) comprising our three foundational countries: Malaysia and Mozambique (representing high-risk scenarios), and Canada (serving as a stable, low-risk control). Data for the period 2010-2023 were cleaned, with particular attention to handling missing values for economic indicators using a forward-fill method, a reasonable approach given the annual nature of these metrics.

A critical innovation of this project is the creation of a binary corruption_risk label (1 for high risk, 0 for low risk). Since "corruption" is not a directly measurable statistic, we devised a rules-based flagging system utilizing the WGI scores:

- A country-year is flagged if its Political Stability score falls below 0.50, or if any of the other five WGI indicators drop below 1.15.
- The ultimate corruption_risk label is assigned as High Risk (1) if a country-year triggers four or more of these six flags. Otherwise, it is designated Low Risk (0).

This labeling strategy was rigorously validated against the qualitative insights from the case studies:

- **Canada**: Consistently labeled as Low Risk (0) across all 14 years, reaffirming its status as a stable, low-corruption environment.
- **Malaysia**: Accurately flagged as High Risk (1) during the peak of the 1MDB scandal (2013-2015).
- **Mozambique**: Similarly, correctly identified as High Risk (1) during the "Hidden Debts" crisis (2013-2016).

This validation confirmed that our quantitative rule could effectively capture the periods of heightened corruption risk identified through qualitative historical analysis, thereby establishing a robust foundation for the model.

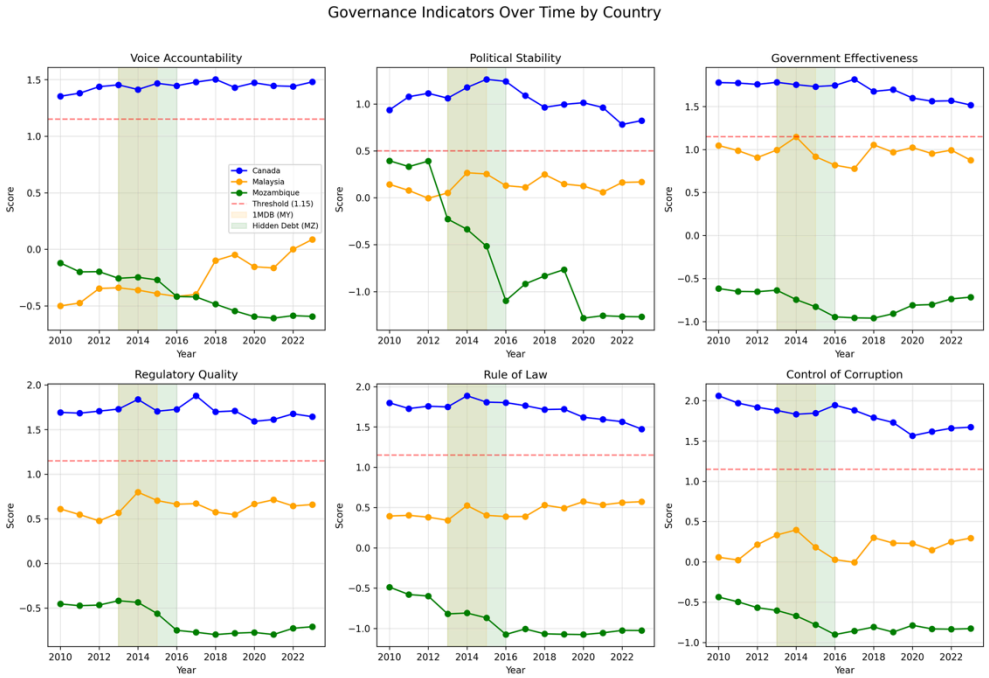

Figure 1: Governance Indicators Over Time by Country. This six-panel chart tracks the governance indicators for our baseline countries. It visually validates our hypothesis by showing Canada's scores (blue) remaining stable, while Malaysia's (orange) and Mozambique's (green) scores clearly deteriorate during the documented scandal periods (shaded areas), often crossing the risk thresholds (red lines).

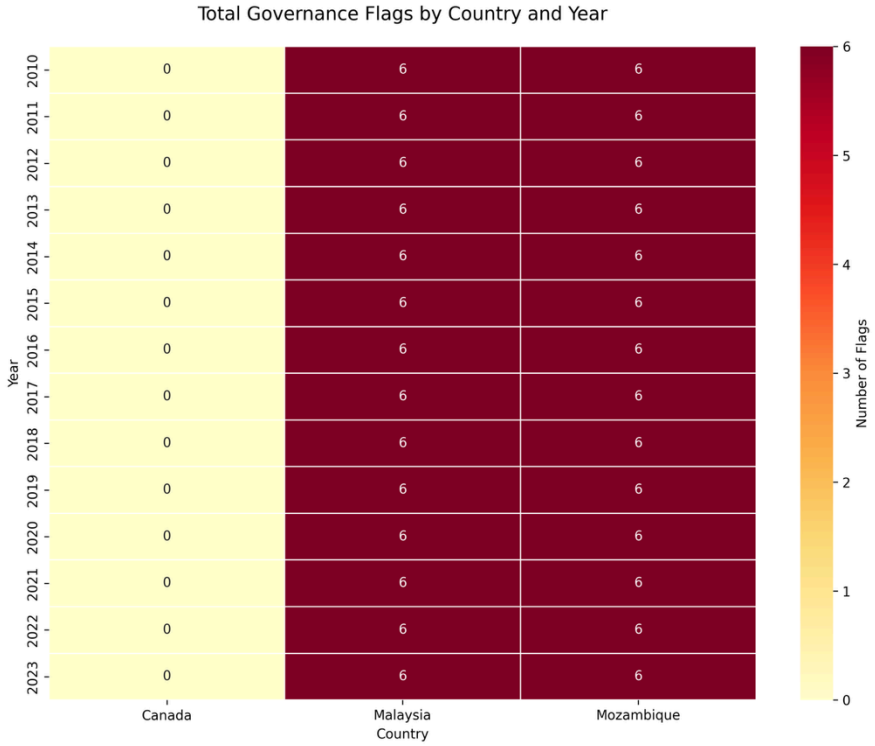

Figure 2: Total Governance Flags by Country and Year. This heatmap shows the annual count of governance threshold violations (0-6). It validates our case studies by showing Canada consistently at 0 flags, while Malaysia and Mozambique consistently trigger 6, indicating systemic governance failures.

Figure 3: Corruption Risk Labels by Country and Year. This heatmap applies our '4 of 6' flagging rule to generate the final binary corruption_risk label. It confirms Canada as Low Risk (0) and Malaysia/Mozambique as High Risk (1), creating the target variable (y) for our model.

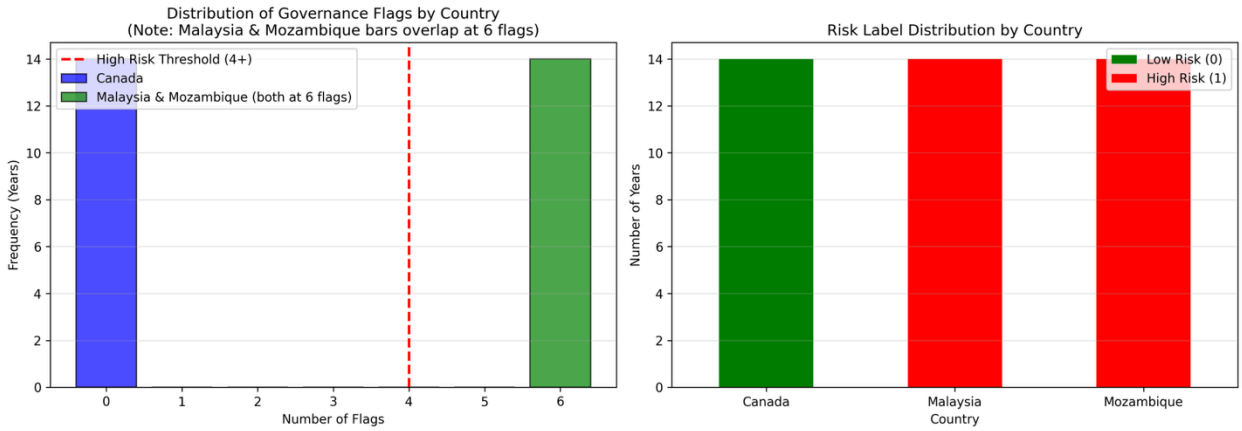

Figure 4: Distribution of Governance Flags and Risk Labels The left panel's histogram validates our threshold (red line), showing a clear bimodal split: Canada's 14 observations cluster at 0 flags, while Malaysia's and Mozambique's cluster at 6. The right panel confirms this, showing the final label distribution by country.

Dataset Expansion and Final Labeling

With the labeling strategy validated, Phase 2 focused on scaling the dataset to provide sufficient data for robust machine learning model training. Building upon the baseline, data for an additional 16 countries were collected from the World Bank API, ensuring a diverse representation of global risk profiles. These countries were initially categorized to ensure a balanced dataset:

- **High-Risk (5 countries):** e.g., Angola, Venezuela, Zimbabwe.
- **Medium-Risk (4 countries):** e.g., Brazil, India, South Africa.
- **Low-Risk (7 countries):** e.g., Norway, Denmark, Switzerland.

This expansion resulted in a comprehensive, unlabeled dataset of 266 country-years (19 countries × 14 years). The identical 4-of-6 WGI flag strategy was then applied to this expanded 19-country dataset. This process generated the final corruption_risk labels for all 266 observations, resulting in a dataset with 112 Low-Risk (0) instances (42.1%) and 154 High-Risk (1) instances (57.9%).

It is important to note that countries initially classified as "Medium-Risk" by our descriptive categorization often received a "High Risk (1)" label from our rules-based system. This outcome is not contradictory but rather an affirmation of the model's sensitivity: even "medium" risk environments, when assessed against our strict governance thresholds, may exhibit sufficient indicators to trigger an early warning. The binary label reflects the system's operational output rather than a broader geopolitical categorization.

The culmination of this phase is the corruption_data_expanded_labeled.csv dataset. This comprehensive file, containing all governance and economic indicators along with the definitive corruption_risk label, is now fully prepared for the subsequent machine learning training phase.

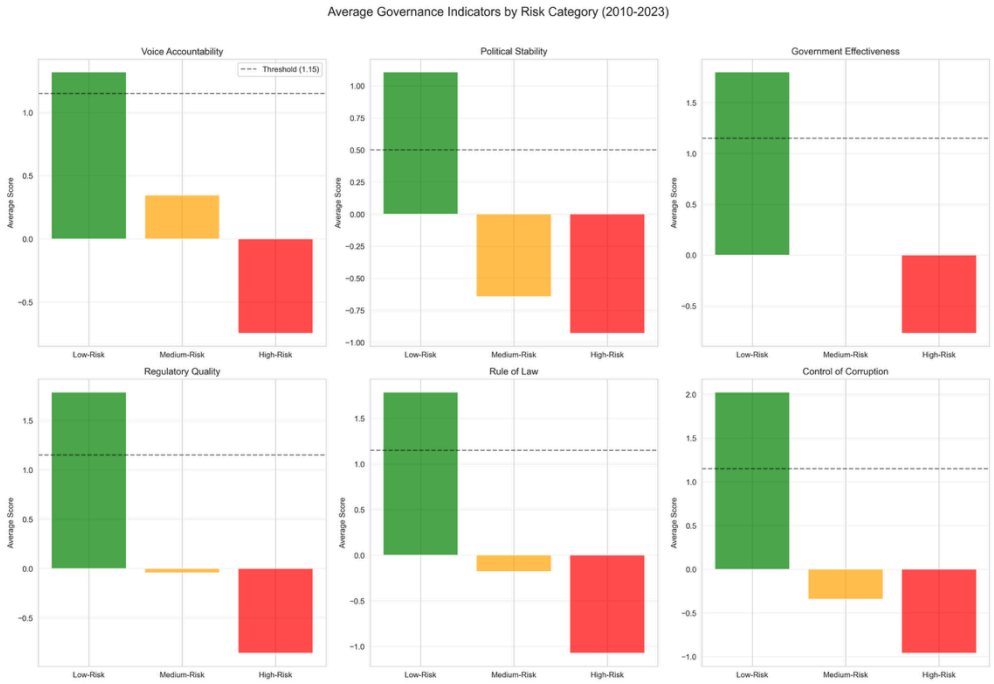

Figure 5: Average Governance Indicators by Risk Category (2010-2023). This multi-panel bar chart shows the distribution of governance scores across the six WGI for Low, Medium, and High-Risk countries. It illustrates a clear separation, with Low-Risk countries scoring consistently above the thresholds (dashed lines) and High/Medium-Risk countries falling below, validating our categorization approach.

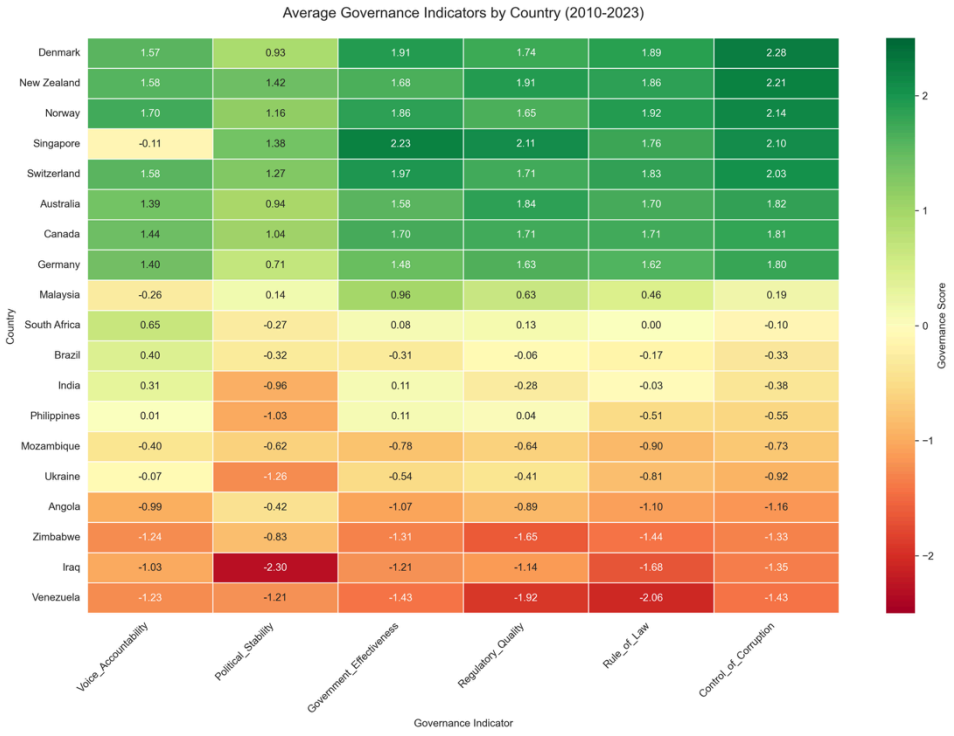

Figure 6: Average Governance Indicators by Country (2010-2023). This heatmap displays the 14-year average governance scores for all 19 countries, sorted by 'Control of Corruption'. It reveals a clear gradient from high-governance (green) to low-governance (red) environments, illustrating the wide variation in institutional quality across our dataset.

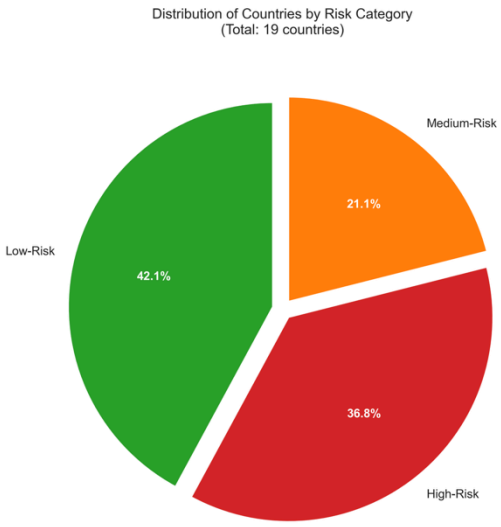

Figure 7: Distribution of Countries by Risk Category. This pie chart shows the composition of our 19-country dataset by descriptive risk category: 42.1% Low-Risk (8 countries), 36.8% High-Risk (7 countries), and 21.1% Medium-Risk (4 countries). This balanced distribution provides adequate representation for model training.

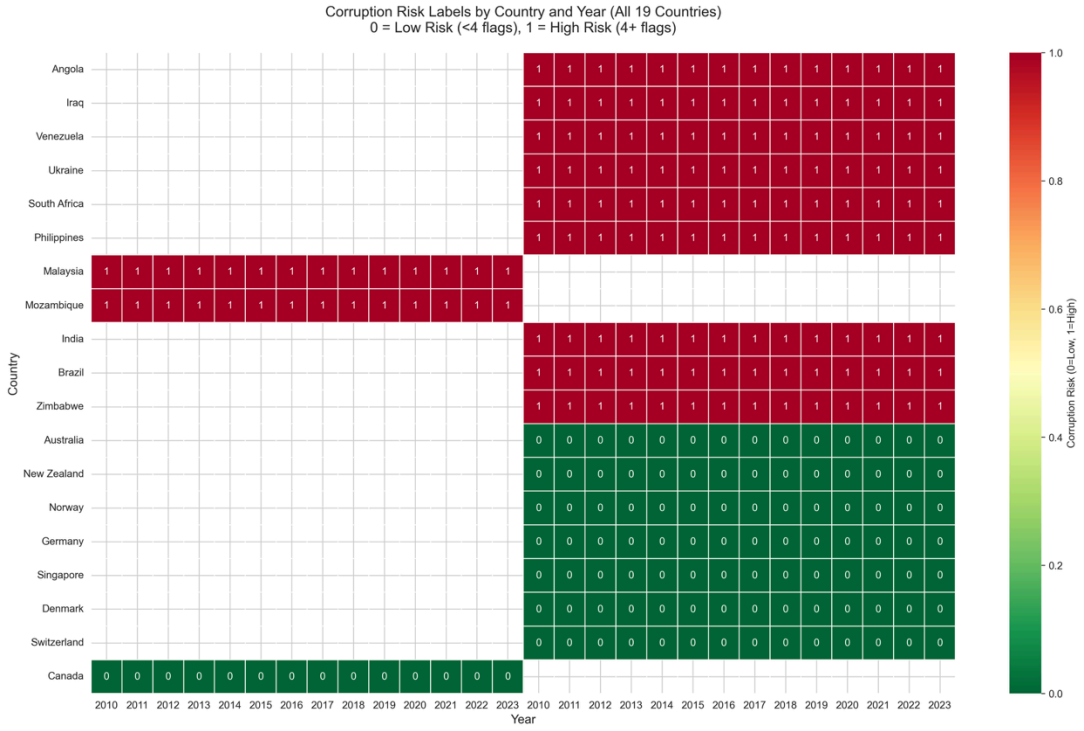

Figure 8: Corruption Risk Labels by Country and Year (All 19 Countries). This heatmap is the primary result, showing the binary corruption_risk labels for all 266 country-year observations. Red cells (1) indicate High Risk (4+ flags), while green cells (0) indicate Low Risk (<4 flags). Countries are sorted by risk, showing 11 nations with consistent high-risk patterns (red) and 8 with low-risk status (green). These labels serve as the final training data for our model.

Next Step: Model Training and Deployment

A crucial next step involves incorporating qualitative data to enrich the model's predictive power. As outlined in the proposal, we will perform sentiment analysis on corruption-related news headlines for the countries in our dataset. Utilizing Natural Language Processing (NLP) tools such as TextBlob or VADER, sentiment scores will be extracted and integrated as an additional predictive feature. This will test the hypothesis that shifts in public sentiment, as reflected in media coverage, can serve as an early qualitative warning sign alongside quantitative governance indicators.

The expanded and labeled corruption_data_expanded_labeled.csv dataset will be systematically divided into training and testing sets. A Decision Tree Classifier, chosen for its interpretability and robust performance on structured data, will be trained on the combined set of governance and economic indicators (and subsequently, sentiment scores). The model's performance will then be rigorously evaluated using key metrics such as accuracy, precision, and recall to ensure its reliability in identifying high-risk environments.

A core deliverable for this phase will be the visualization of the trained Decision Tree. This graphical representation will provide a clear, intuitive flowchart of the model's decision-making process, illustrating precisely how different indicator thresholds lead to a "corruption risk" classification. Such transparency is vital for stakeholder understanding and trust in an early warning system designed for development integrity.

The successful completion of these remaining steps will bring the Global Trust Engine to fruition, providing a powerful, data-driven tool to proactively combat corruption and safeguard international development efforts.