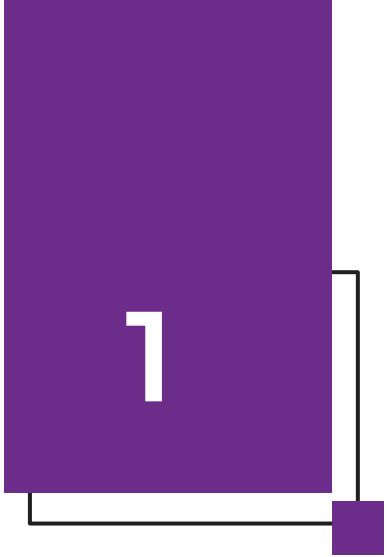
MANCHESTER 1824
The University of Manchester

HACE
Child Labour Risk

Bagus Pranata

Outline



Introduction

Introduction

Background & Problem Statement

HACE is a specialist data company committed to ending child labour in business supply chains. It assesses and analyzes businesses worldwide based on their exposure to child labor risk through the Child Labour Index (CLI).

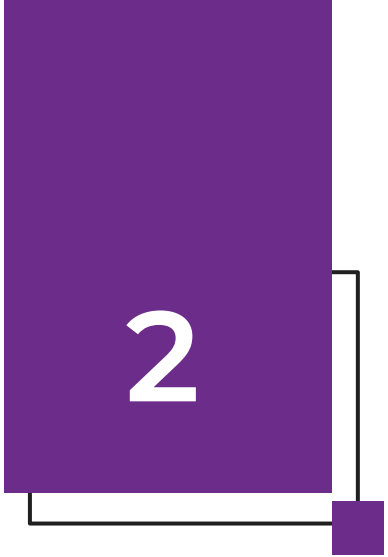
The task is to build a country-level algorithm around child labour issues that can aid HACE and its stakeholders in gaining a more comprehensive grasp of the associated risks and preventing the exploitation of vulnerable communities in the future

Objective

- To build a country-level algorithm or model that produces a risk/exposure level assessment for child labour.
- To visualise the model's insights through a dashboard.
- To incorporate any necessary weightings within the data and algorithm.

Significance

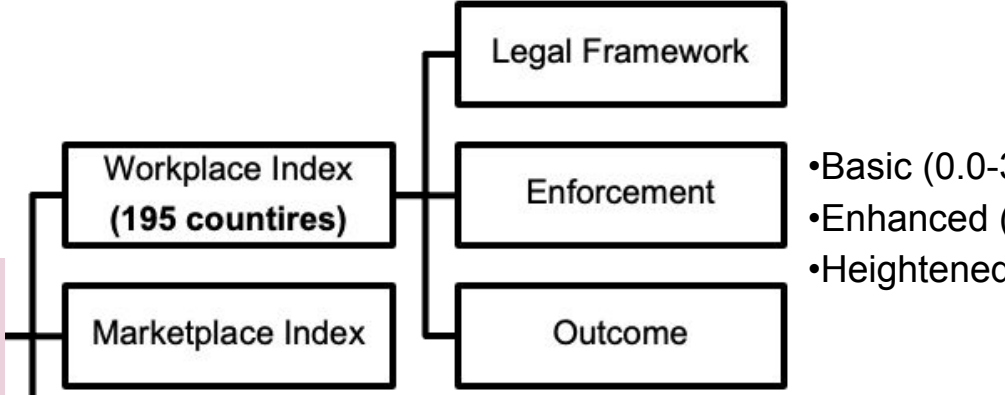
- Enhanced Monitoring
- Data-driven Decisions
- Customizable Strategies
- Stakeholder Engagement
- Scalability



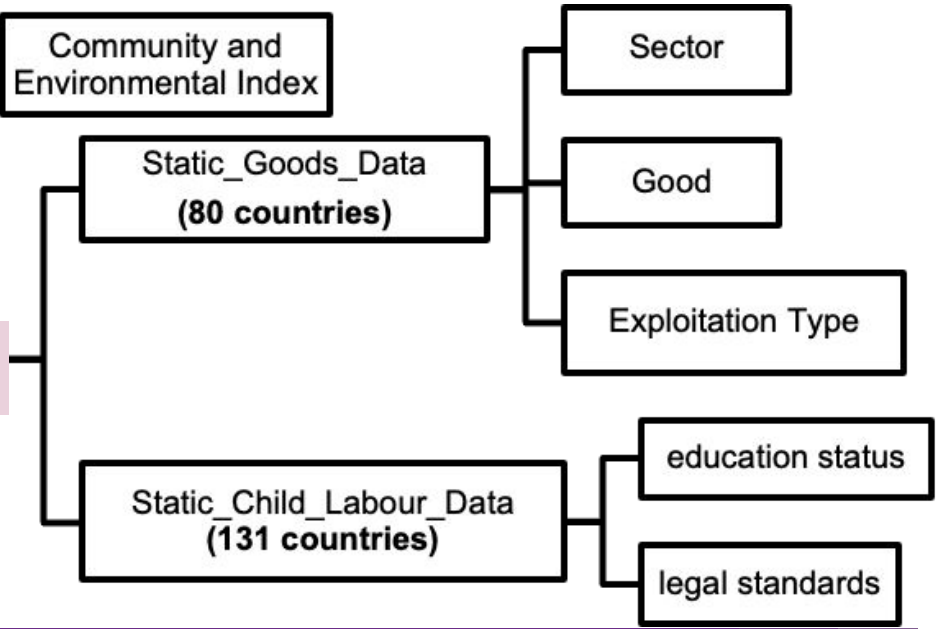
Datasets

Dataset Overview

Children's Rights and Business Atlas



Sweat and Toil API

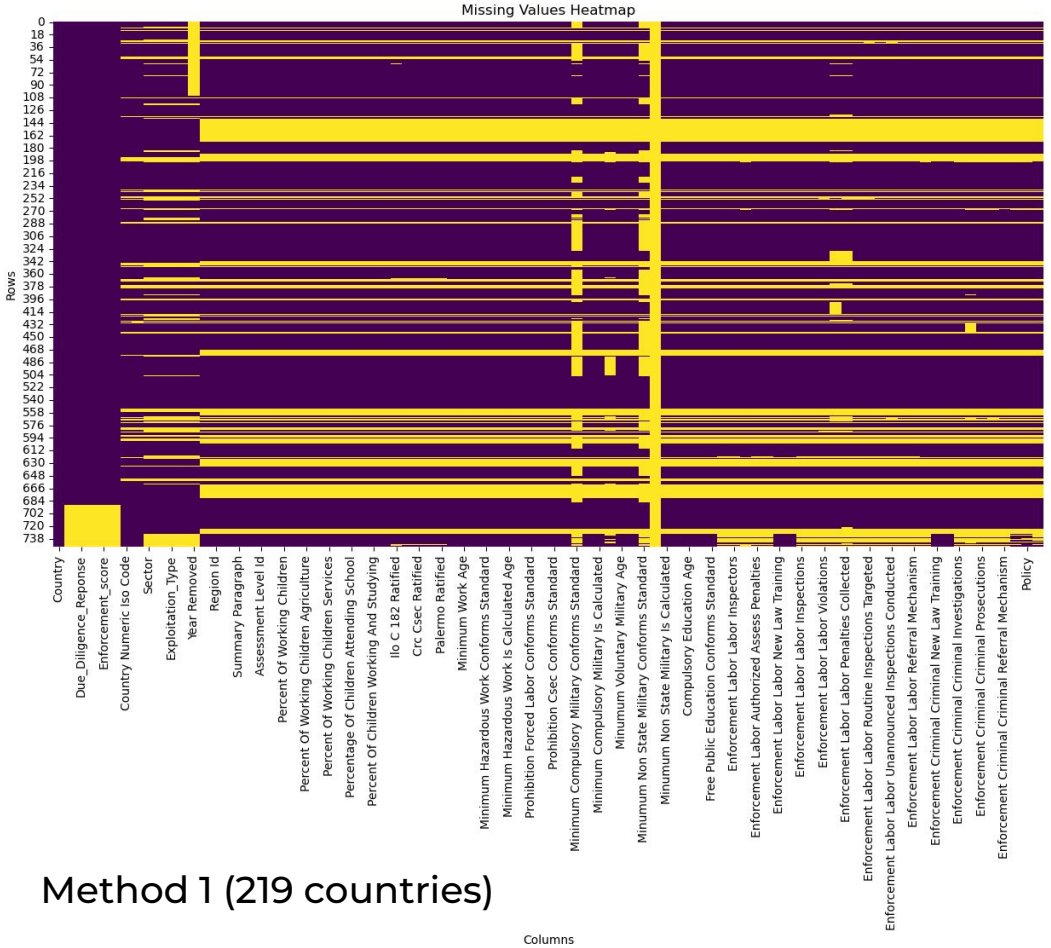


Data Integration

Filtering

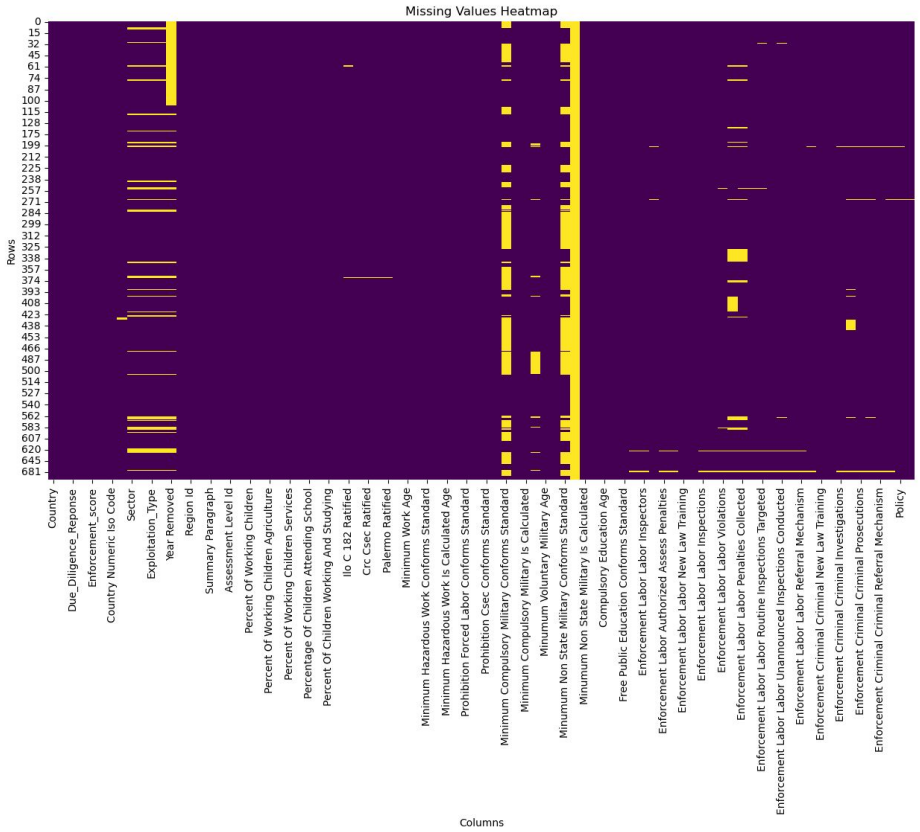
- Exploitation Type Filtering
- Temporal Data Filtering
- Filtering Based on Availability of Data Across **Countries**

Common Identifier

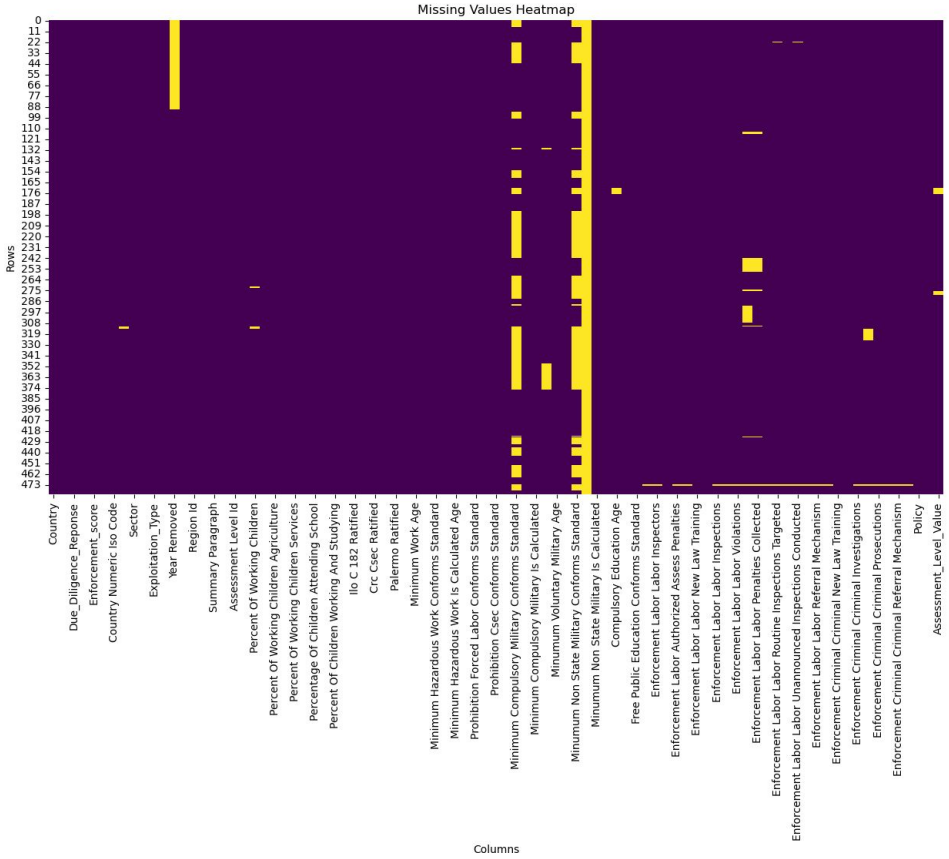


Method 1 (219 countries)

Data Integration



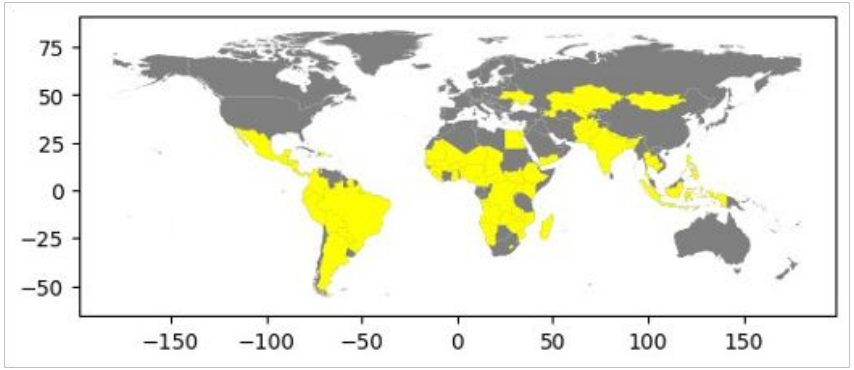
Method 2 (110 countries)



Method 3 (64 countries)

Merged Data EDA - Basics

- 64 countries
- in 5 regions of the world
- 140 unique Goods
- Main sectors: Agriculture, Mining, Manufacturing and Other.



Interesting insights

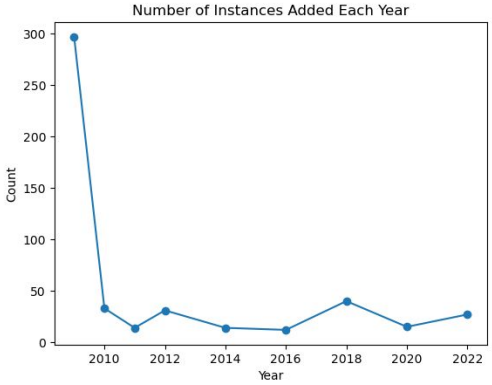
Scores generated by other Indexes

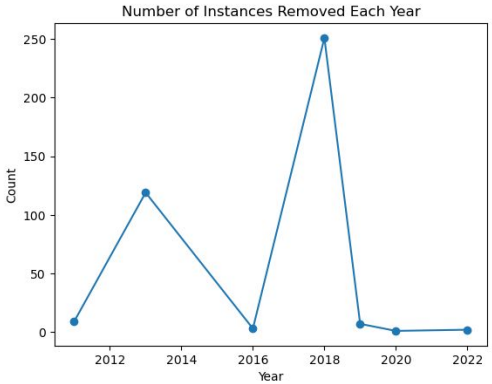
Sectors Goods

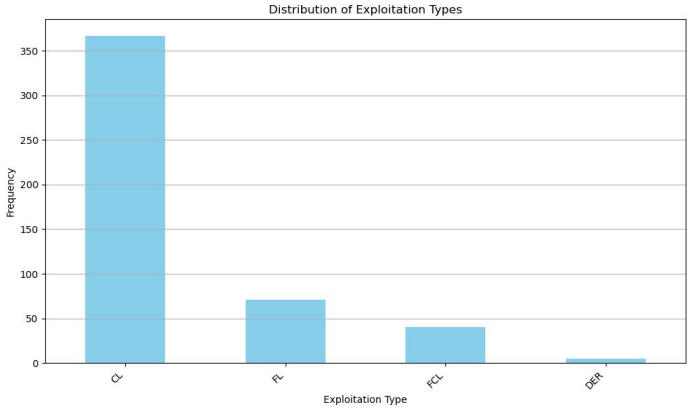
Education

Labour/ Work

Military







Missing Values Handling

- **'Unavailable' and 'unknown'**. To facilitate uniform treatment, these were replaced with NaN.
- Variables with missing value proportions exceeding **2%**.
- **Redundant** variable with **synonymous variables** that are easier for machine processing.
- Columns with only **one unique value**.
- Anomalous data, such as '13-Dec' in 'Compulsory Education Age'. Referring to other data, it was determined that this should be '13', likely indicating a recording error.
- **Numerical variables**, their **skewness** was assessed through distribution plots, and either mean or median imputation was applied based on specific circumstances.

Feature Engineering

- **ID-related data** was converted to integer type.
- **Binary variables** were transformed into boolean variables (0/1).
- **Age range variables** were encoded using the lower limit of the age range, for example, '5 to 14' mapped to 5, '6 to 14' mapped to 6, '7 to 14' mapped to 7, and '10 to 14' mapped to 10.
- **Variables representing different degrees**. For instance, in 'Due_Diligence_Reponse', 'Basic' was mapped to 1, 'Heightened' to 2, and 'Enhanced' to 3.
- **Categorical features** such as 'Sector' and 'Good' were directly subjected to one-hot encoding for machine readability.

3

Methodology

Feature Selection

The features were selected using a systematic approach. This approach used statistical tests that assessed all variables to see if there was a significant relationship between them and the target variable.

- The target variable we defined came from encoding the Exploitation Type which has FL (0), CL (1), FCL (2).

The chosen tests were:

➢ **One-way ANOVA:** to test the significance of continuous features.
➢ **Chi-squared:** to test the significance of categorical features.

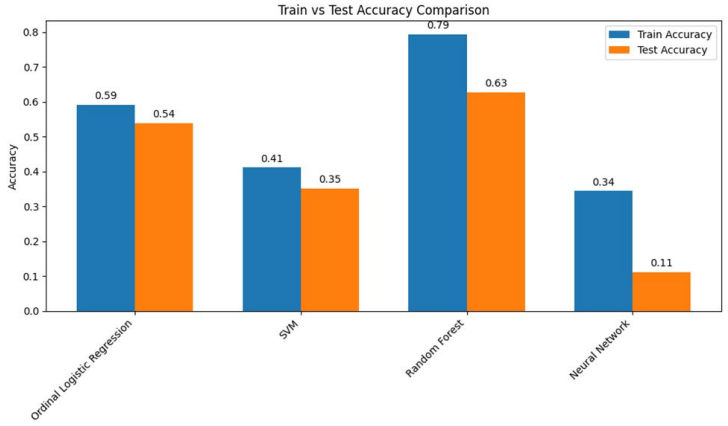
- The tests were conducted at the 10% Significance level, resulting in a total of 23 features being selected.

Model Selection

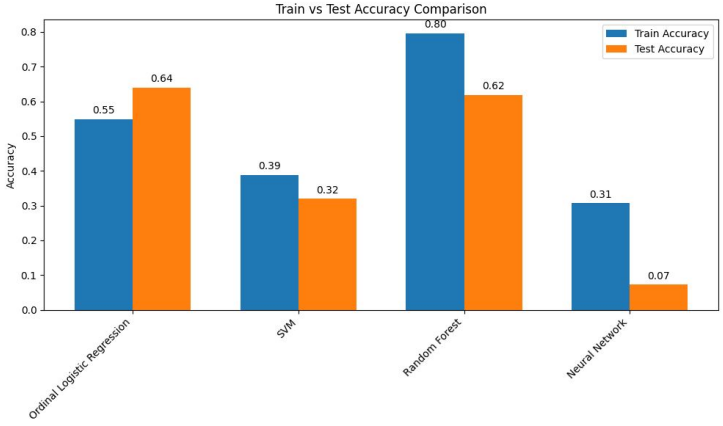
- Ordinal Logistic Regression (OLR)
- Support Vector Machine (SVM)
- Random Forest
- Neural Network

OLR is the most stable, while keeping good scores and interpretability

Results by splitting the training and test sets by 70/30% with a random state = 42



Results by splitting the training and test sets by 80/20% with a random state = 42



Metrics

- **F1 Scores**:
  - Ordinal Logistic Regression: ~0.67 — Good score, good interpretability
  - SVM: ~0.40
  - Random Forest: ~0.66 — (tends to overfit)
  - MLP: ~0.69 — Good score, low interpretability, imbalanced

- **Accuracy scores:**
  - Ordinal Logistic Regression: ~ 0.54 - 0.64
  - SVM: ~ 0.24 - 0.37
  - Random Forest: ~ 0.60 - 0.72
  - MLP: 0.11 - 0.68

Quick insights (top 5 coefficients)

- **Class 0 (DER/FL):**
  - Enforcement_score: 0.10327314
  - Index_score: -0.06283733
  - Region Id: -0.0567223
  - Prohibition Child Trafficking Conforms Standard: -0.04535533
  - Minimum Voluntary Military Age: -0.04273657

Sum of the coefficients: 0.31 from the total 0.39

- **Class 1 (CL):**
  - Minimum Work Age: 0.18593251
  - Region Id : 0.11394711
  - Prohibition Child Trafficking Conforms Standard: 0.08319089
  - Minimum Voluntary Military Age: 0.05554535
  - Prohibition Illicit Activities Conforms Standard: 0.04698918

Sum of the coefficients: 0.49 from the total 0.57

- **Class 2 (FCL):**
  - Minimum Work Age : -0.15097449
  - Enforcement_score : 0. 10776667
  - Index_score : 0.07056057
  - Region Id: -0.0572248
  - Prohibition Illicit Activities Conforms Standard: -0.05090132

Sum of the coefficients: 0.44 from the total 0.55

About Logistic Regression

$$y_k(\beta, x) = \beta_{0k} + \beta_{1k} x_1 + \beta_{2k} x_2 + \ldots + \beta_{pk} x_p = \beta_{0k} + X_i W_k$$

Class 0 (Forced Labor): $y_0(\beta, x) = \beta_{0,0} + \beta_{1,0} x_1 + \beta_{2,0} x_2 + \ldots + \beta_{p,0} x_p$

Class 1 (Child Labor): $y_1(\beta, x) = \beta_{0,1} + \beta_{1,1} x_1 + \beta_{2,1} x_2 + \ldots + \beta_{p,1} x_p$

Class 2 (Forced Child Labor): $y_2(\beta, x) = \beta_{0,2} + \beta_{1,2} x_1 + \beta_{2,2} x_2 + \ldots + \beta_{p,2} x_p$

Then, we can calculate the probabilities for each class like:

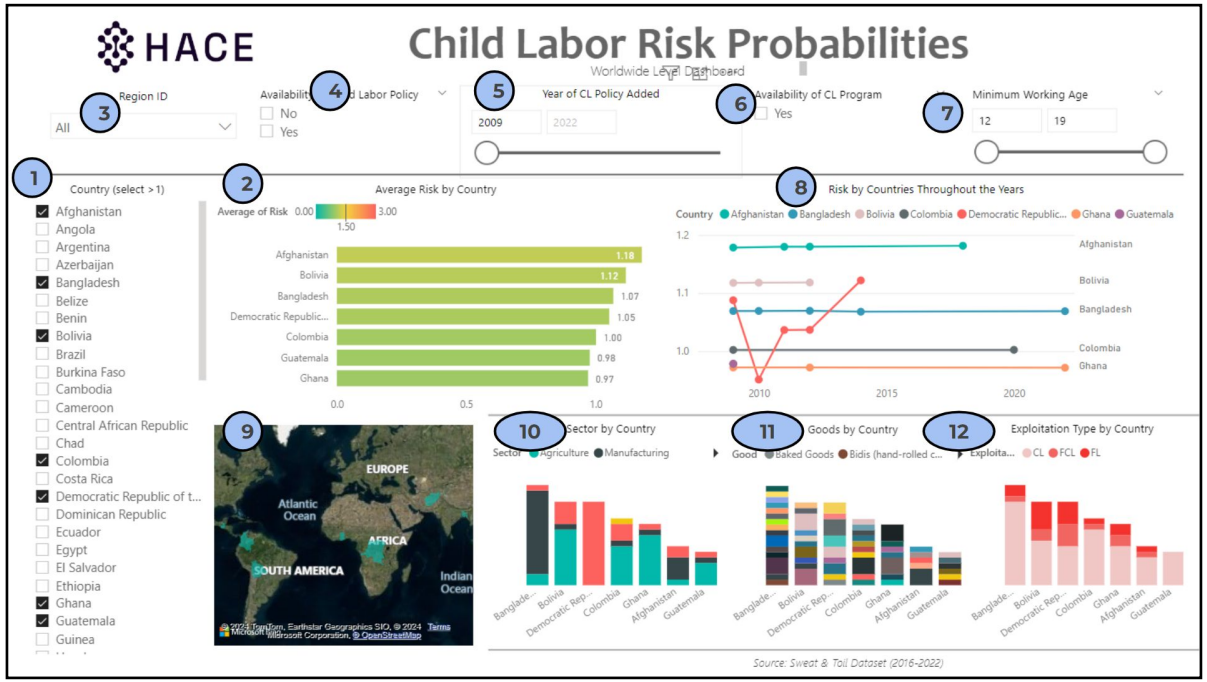
$$\widehat{p_k}(x_i) = \frac{exp(\beta_{0,k} + X_i W_k)}{\sum_{j=0}^{K-1} exp(\beta_{0,j} + X_i W_j)}$$

Risk Index Calculation

- A risk index that measures the risk of exposure to child labour was derived to give insights into the data.
- The purpose of creating an index is to create a summary of the analysis found through our model.
- The logistic regression model yields a probability class of the possible outcomes.
- The risk index can be calculated using the mathematical expected value of the discrete random variables

$$\textit{Risk Index} = 0 \times \textit{Probability}(0) + 1 \times \textit{Probability}(1) + 2 \times \textit{Probability}(2)$$

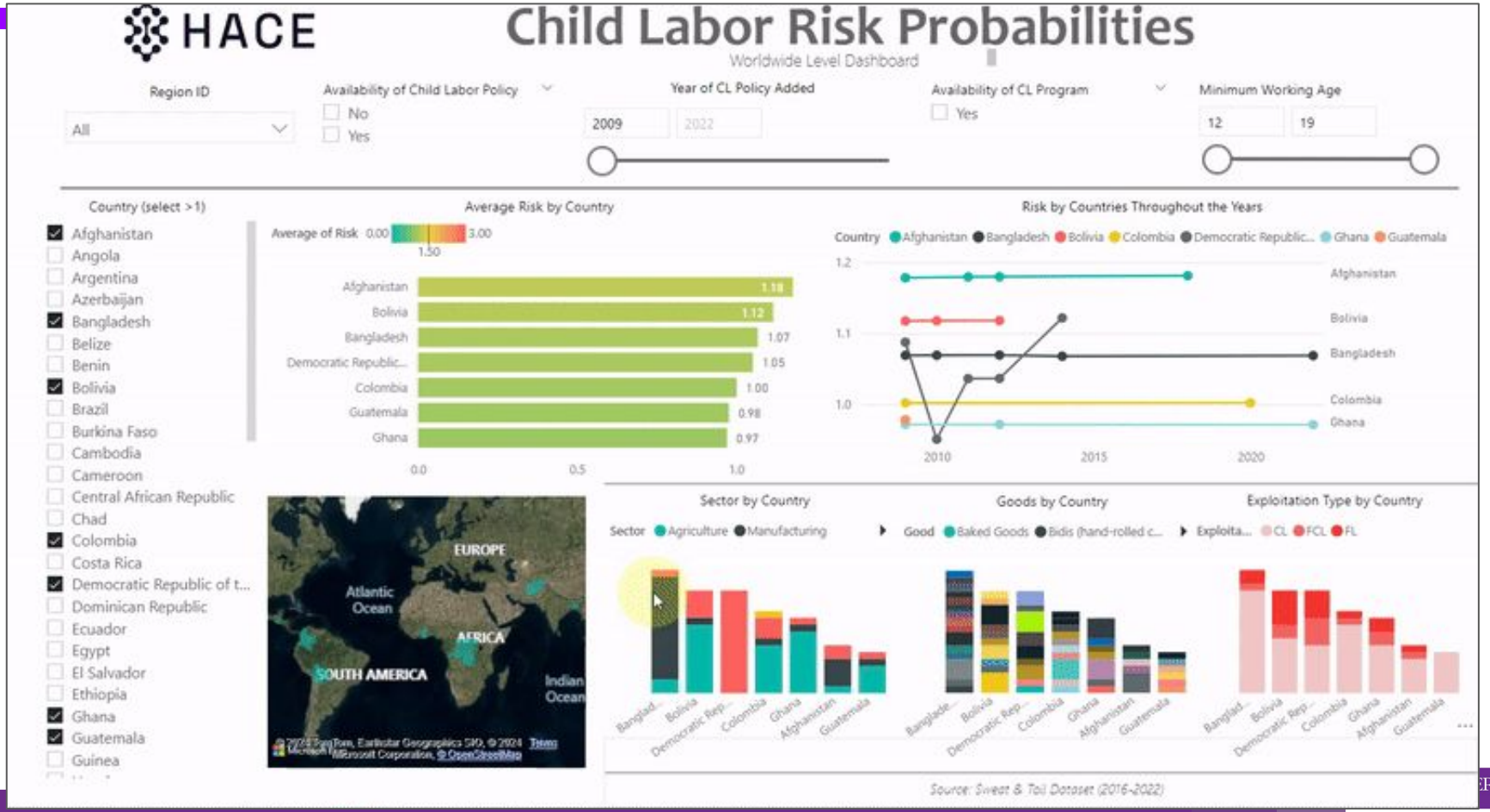
Worldwide Level Dashboard



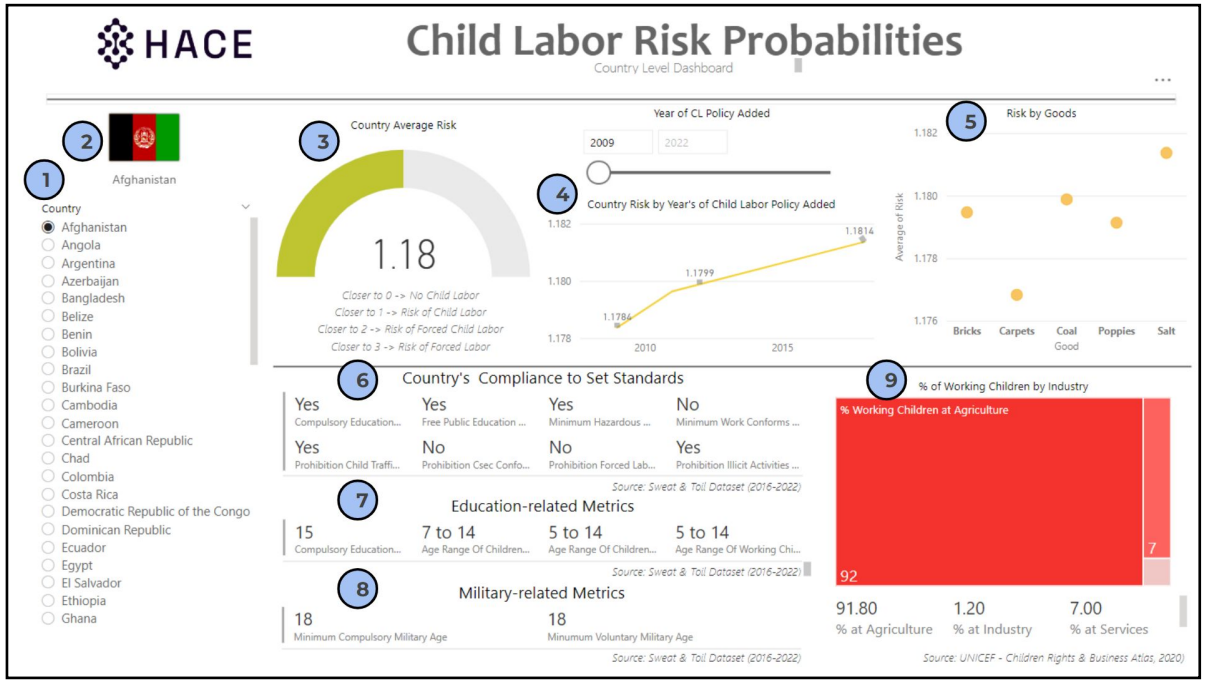
Interface of worldwide level dashboard

- The **(1) multi-selection country** enables customization of the study of the desired comparison nations and displays the average risk on the **(2) country matrix** risk as well as other **chosen criteria (3) through (6)**.
- **(8) A line chart is used to visualise the Risk Comparison**, providing information on how different nations compare over time with regard to their policies regarding child labour risk.
- The **stacked bar chart** titled **"Index by Country" and the filled map titled "Countries Filled Map" (9) and (10)** respectively show the frequency and severity of child labour in different geographic regions, providing a clear visual differentiation between those with different levels of concern.
- Detailed stacked column charts on **(11) Sector, (12) Goods, and (13) Exploitation Type** by Country illustrate which sectors, kinds of goods, and **types of exploitation** are more common in particular nations.

Worldwide Level Dashboard - Sample Video



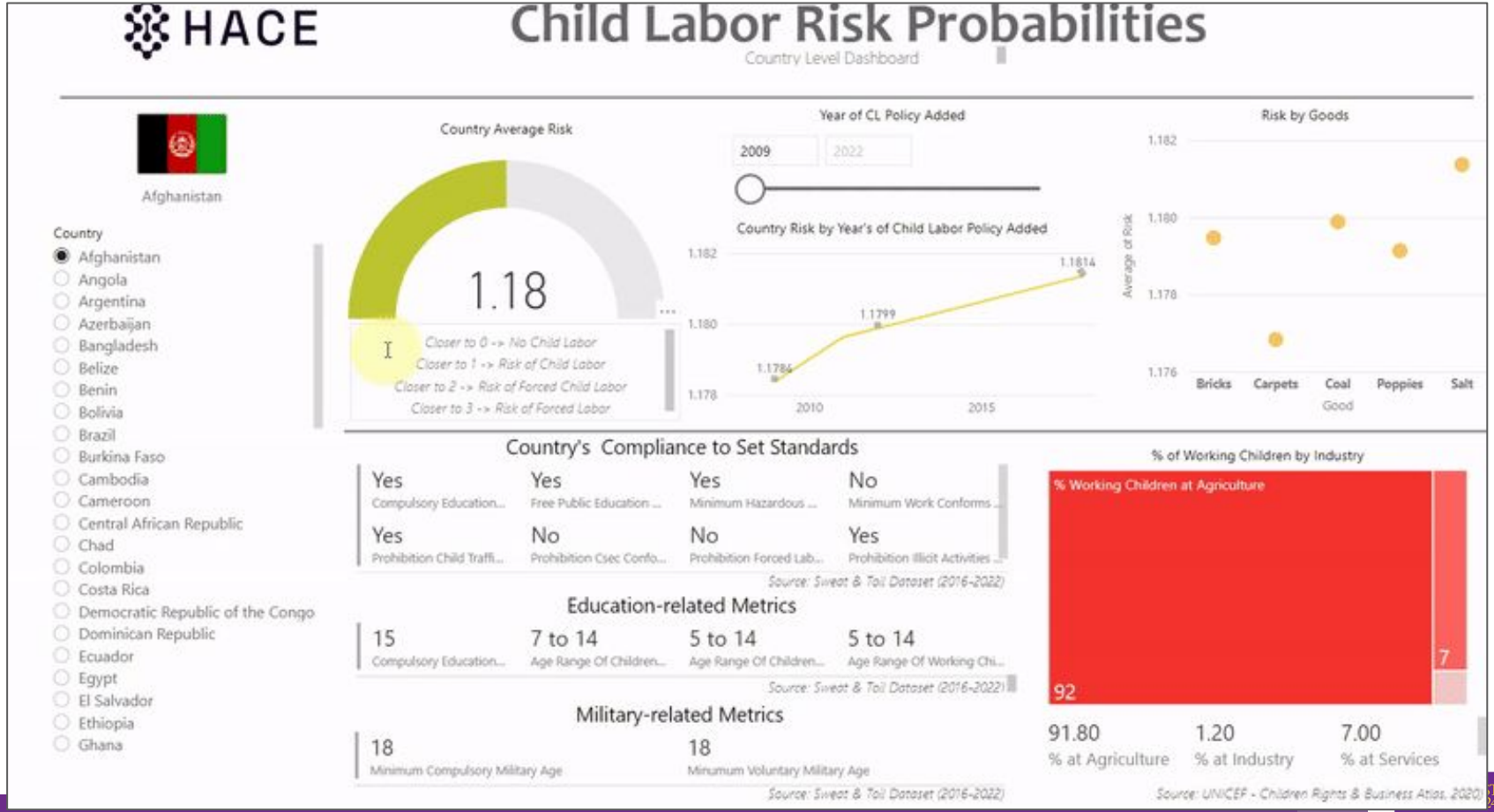
Country Level Dashboard



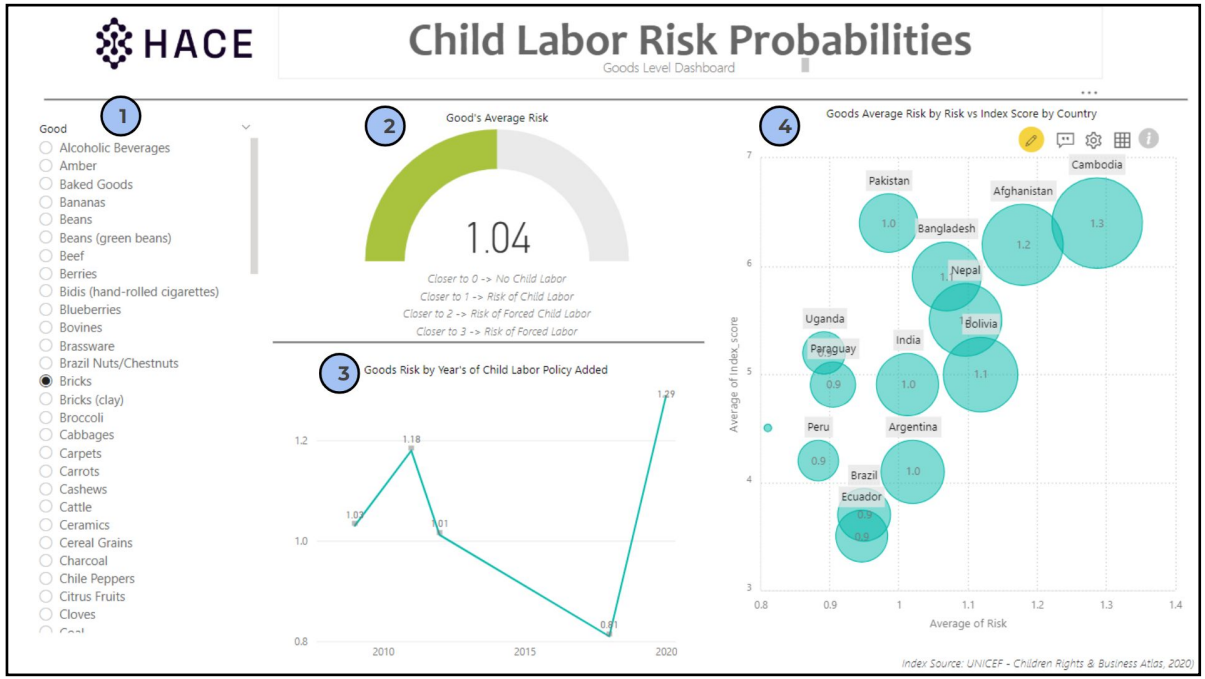
Interface of country level dashboard

- The user can choose a single country using the **(1) Country Slicer** on the dashboard, and then the selected country is visually anchored by the **(2) Country Image.**
- An immediate visual summary of the current risk level related to child labour in the chosen nation is provided by a **(3) nation Risk Gauge.**
- The **(4) Country Risk by Year of Child Labor** Policy Added Line Chart links the years that child labour policies were implemented to changes in risk level over time.
- **(5) Nation Risk by Goods** and **(6) Nation's Adherence to Specified Guidelines Multi-Row** Cards shed light on particular problem areas and the degree to which the nation complies with international child labour regulations.
- In-depth sections on the nation's child labour statistics pertaining to **(7) education** and **(8) the military** show how these two crucial aspects of children's life are impacted by child work.
- Finally, the distribution of child labour across the nation's many industries is shown graphically in the **(9) Percentages of Child Working Industry Treemap.**

Country Level Dashboard - Sample Video



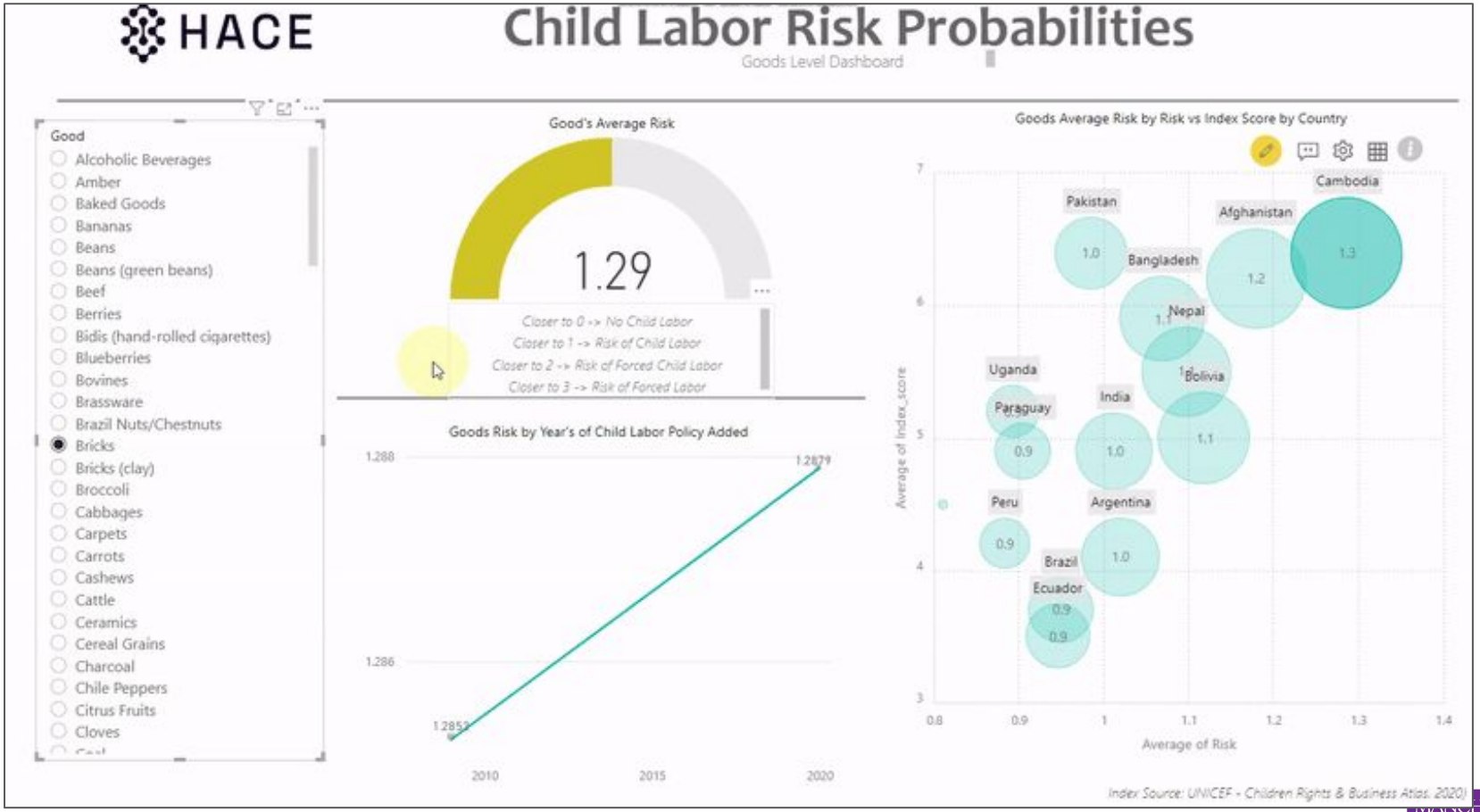
Goods Level Dashboard



Interface of goods level dashboard

- This dashboard's primary feature is a **(1) Good Slicer** for single selection, which enables users to select a particular good for in-depth examination.
- The **(2) Good's risk Gauge** shows the current danger level associated with the chosen good visually, conveying a sense of urgency or concern right away.
- A **(3) Good Risk by Child Policy Year Added** Line The selected good's historical trend of risk levels is depicted in the chart, showing how these risks have changed in correlation with the introduction of child labour laws.
- **(4) The bubble chart** provides a comparative perspective by displaying the relative rankings of various nations with respect to risk and index score for the chosen good.

Goods Level Dashboard - Sample Video



4

Outcomes & Discussion

Achievement of Objectives

1. **Build a country-level algorithm or model that produces a risk/exposure level assessment for child labour.**
   We trained an OLR model using 23 features that predicts the probabilities of different types of forced labour. The risk/exposure level are calculated through Risk Index = 0 x Probability(0) + 1 x Probability(1) + 2 x Probability(2)

1. **Visualise the model's insights through a dashboard.**
   We created a dashboard that brings the insights to life by making our model more understandable and actionable through interactive filters, maps, graphs, and the ability to focus on specific countries and/or goods.

1. **Incorporate any necessary weightings within the data and algorithm.**
   The weightings for each factor can be further refined or analysed to obtain more insights.
   In addition to the Ordinal Logistic Regression model, more complex models, such as Random Forest, can be used to deepen and enhance the analysis of the results. While the primary goal of creating a risk index is based on the OLR model, adding a Random Forest model could complement this analysis.

Challenges and Obstacles

1. **Data Quality**
   The proportion of missing values that made up each column of the dataset spanned from 0.20% to 99.8%. To solve this we discarded columns that had over a 2% missing value proportion and performed missing value imputation on the remaining columns.

1. **Class Imbalance**
   The number of classes available in the dataset was heavily imbalanced, which resulted in a bias towards the majority class (CL). To mitigate this, we resampled the data before training the logistic regression model.

1. **Limited access to resources:**
   Initially we explored the idea of utilising a dynamic dashboard to monitor the risk of child labour. However, we were severely limited by the length of the project, the complexity of Power BI, and the terms of our free student licence. As a result, we decided against these complex features and instead used a static dashboard that shows the current statistics

5

Industry Contribution Analysis

Impact on the Industry

Businesses

Short-term:

- Risk mitigation: quickly identify high-risk areas and take immediate action to mitigate risks.
- Supply chain efficiency: insight on labour practices that can optimise resource allocation.

Long-term:

- Reputation and brand Value: positive brand perception
- Innovation and adaptation: investing in solutions to address CL can drive innovation and be ahead of the curve.

Index Creators

Internally:

The data provided by the index and dashboard can aid in internal process optimization as large amounts of data are available through a very easy-to-access dashboard.

Externally:

- Market demand and revenue generation: from licensing agreements and consulting services
- Strategic partnerships and collaborations: Partnerships with NGOs, governments, and private businesses.

Project Reproducibility, Expandability and Benefits- SWOT ANALYSIS

Strengths

- Machine Learning (ML) model and Power bi Dashboard offer scalability
- ML's automation reduces time and effort for analysis, translating to cost savings.
- Model and Dashboard offer versatility, enabling changing priorities, contexts, and objectives, ensuring ongoing relevance.
- Static and historical data that will remain valuable even with the passing of time

Weaknesses

Dependency on external data sources

Opportunities

Expansion of Data Sources

Continuous improvement and Innovation

Threats

Discontinuation of one of the Datasets

6

Conclusion & Recommendation

Conclusion

This project has successfully developed and implemented a **strong algorithm that can accurately forecast** the dangers of child labour in a variety of global scenarios by applying machine learning models, feature engineering, and extensive data preparation. Then, the **dashboard implementation** offers HACE and its stakeholders with a structured and convenient means of interacting with data-driven insights for decision-making. Lastly, by carefully **utilising weightings in the data and algorithm**, the most crucial factors influencing the likelihood of child labour have been found, increasing the model's anticipated accuracy.

Recommendation

- **Continuous Data Enhancement**
- **Model Refinement and Iteration**
- **Engaging with Stakeholders**
- **Advanced Visualization Tools**
- **Policy Development and Advocacy**
- **Technology Integration**
- **Scalability and Replicability**

Thank you

MANCHESTER 1824
The University of Manchester