1. Environment Setup

1.1. Prerequisites

To run the Datathon_2025.ipynb file and reproduce the results, ensure you have the required versions of Python and necessary libraries installed.

Python Version: 3.12.1

Required Libraries:

- pandas (v2.2.3) - Data manipulation
- numpy (v2.2.2) - Numerical operations
- scikit-learn (v1.6.1) - Machine learning models and evaluation metrics
- matplotlib (v3.10.0) - Visualization
- seaborn (v0.13.2) - Statistical data visualization
- scipy (v1.15.1) - Statistical transformations

2. Instructions to Run the Notebook

1. Ensure that the dataset file _.csv is in the same directory as the Datathon_2025.ipynb file.
2. Ensure that required libraries are up to date to prevent any errors.
3. Uncomment the pip install lines if you have missing dependencies.
4. Run the cells in numerical order to prevent any errors.

3. Process

3.1. Data Characteristics

- The dataset contains a mix of financial, structural, and operational characteristics.
- Only the following columns were used in the model, and all other columns were dropped due to high missingness and redundancy:
    - SIC Code
    - 8-Digit SIC Code
    - Year Found
    - Ownership Type
    - Employees (Single Site)
    - Employees (Domestic Ultimate Total)
    - Employees (Global Ultimate Total)

    - Sales (Domestic Ultimate Total USD)
    - Sales (Global Ultimate Total USD)
    - Import/Export Status
    - Is Domestic Ultimate (Target Variable)
    - Is Global Ultimate (Target Variable)
- Some variables, like Employees (Single Site), had many missing values and required imputation.
- Categorical variables, such as Ownership Type, were encoded using Label Encoding or Frequency Encoding for further analysis.
- Outliers were handled using SD method, where values that deviates more than 3 standard deviations away from mean were removed (only upper bound values were removed as it was necessary to keep lower bound values due to employees column).

3.2. Data Splitting and Modelling

- The cleaned data was split into 70% Training, 10% Validation and 20% Test data.
- A Random Forests model from the sklearn.ensemble library was utilised for the classification task. Random Forests, which utilise an aggregation of decision trees, is said to have high accuracy and robustness to noise in data. Its non-parametric nature aids in organizing the various skewed columns of data as well.

4. Key Insights and Findings

4.1. Model Performance

The model was evaluated using test data and the results are as follows:

Model + Classification task	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Random Forest (Is Domestic Ultimate)	93.3%	89.6%	96.2%	92.8%	0.968
Random Forest (Is Global Ultimate)	95.1%	89.6%	96.2%	92.8%	0.984

4.2. Key Findings

As initially hypothesised when deciding on a machine learning model for the classification task, Random Forests was able to perform well in predicting Is Domestic Ultimate and Is Global Ultimate classifications with the available data.

- By analysing the feature importance across both models, Single Site Employee numbers were the most important feature. For the former task, Domestic Ultimate and Global Ultimate Employee numbers were the next most relevant features. For the latter, Global Ultimate and Domestic Sales were the next most relevant features, but with a larger gap from Single Site Employees comparatively.
- Recall is about 7-8% higher compared to Precision in both models, which suggests a notable chance for false positives in identification. This may affect the effectiveness of the models in real-world scenarios depending on the costs of wrongly classifying a company as Domestic or Global Ultimate.

5. Conclusion

This Python Notebook has trained a Random Forests machine learning model through data preprocessing, feature engineering and model evaluation and achieved high performance metrics in predicting companies as "Domestic Ultimate" or "Global Ultimate". More testing could be performed with other models such as Logistic Regression or Gradient Boosting Machines to compare performance and weigh the trade-offs between Precision and Accuracy.

Additionally, more analyses could be administered on columns unused in training this model, for example word embeddings on textual columns like Company Descriptions to inspect if some keywords were linked to Domestic Ultimate and Global Ultimate status. Furthermore, if incomplete columns such as Square Footage or Fiscal Year End were able to be scraped for data, it could potentially enhance the predictive capabilities of machine learning models for this task.