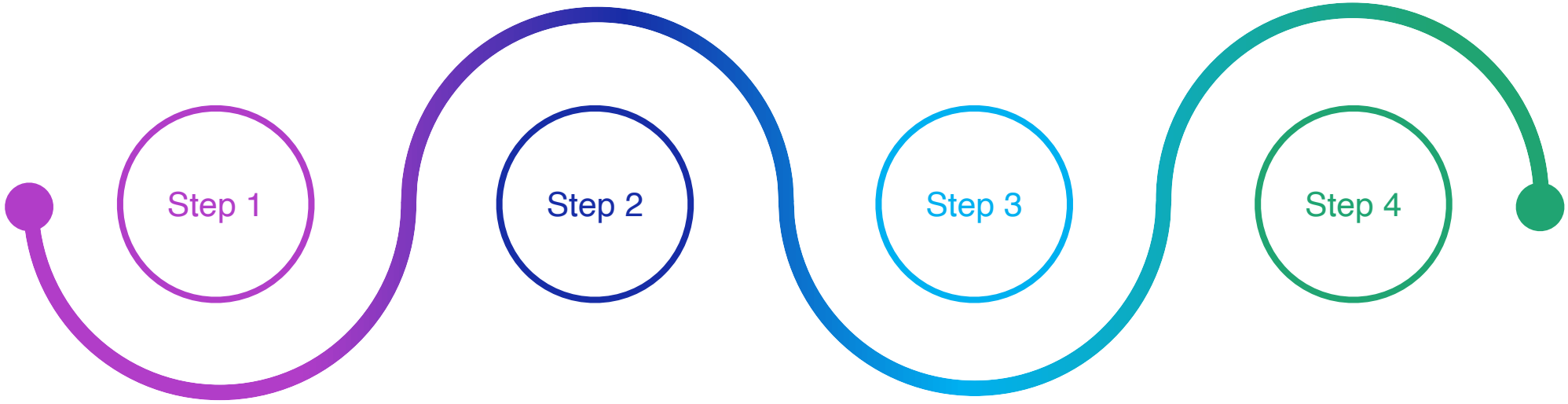
SCHNEIDER ELECTRIC HACKATHON

PIPELINE & STRATEGIES



Data Reading

- Management and integration all the different data sources
- Csv, API and PDF reading

Data Processing

Binding and standardizing rows and columns for a smooth modelling

Modelling

- We decided to use a xgboost with an extensive grid

Predictions

- F1 score was chosen as our main metric

DATA READING & PROCESSING

- All different sources of data were integrated in the data frame
- The API connection was made by the httr and jsonlite R libraries
- PDF files reading was automated using pdftools package
- Reg expressions to extract all the info from the strings
- Nonimportance columns (based on correlation with the response var) were dropped

MODELLING

- **Feature Selection of predictors:** Correlation-based feature selection with 0.01 as threshold. Chosen predictors are: “country", "sector_name", “main_activity_label”.
- **Model used for the problem: XGBoost** - fast, efficient & less prone to overfitting than RandomForest. Supports multi-classification problems.
- **Hyperparameter tuning:** 5-fold Cross-Validation with random selection of 200 combinations from the expanded grid. F1 score computed on the test-set for every iteration.
- **Final hyper-parameters:** Best combination selected based on out-of-sample F1 score.