

# Machine learning and supervision

Developing tools for supervision

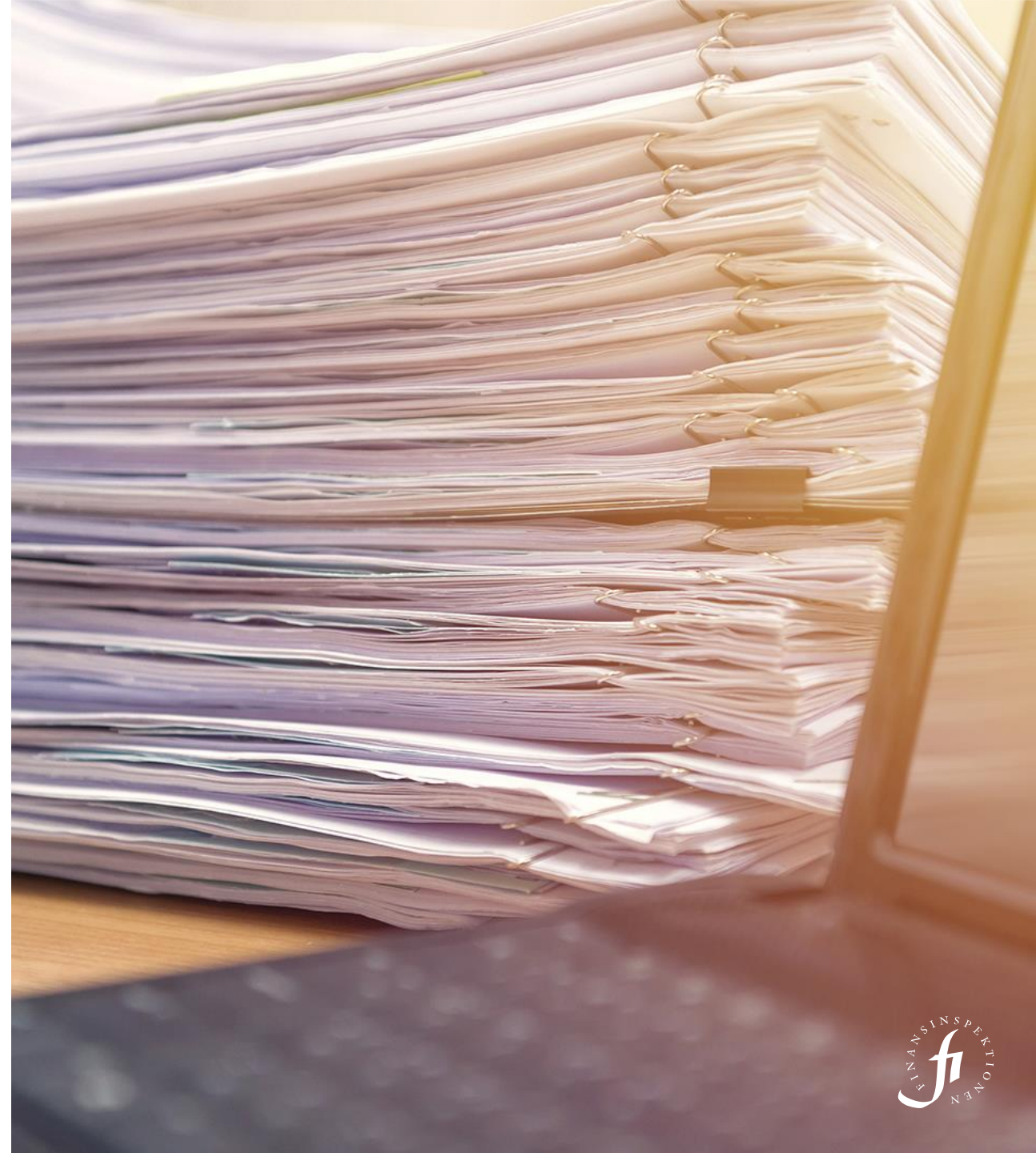


June 17 2024



## Background

- Supervision of the insurance market
  - Cover all companies
  - Riskbased
  - Quantitative vs. qualitative information
- DG Reform



# Qualitative data

- Three qualitative reports
  - Solvency and Financial Condition Report (SFCR)
  - Regular Supervisory Report (RSR)
  - Own Risk Self-Assessment (ORSA)
- Format PDF

|                   | 2018 | 2019 | 2020 |
|-------------------|------|------|------|
| SFCR              | 42   | 45   | 1    |
| RSR               | 22   | 22   | 1    |
| ORSA              | 1    | 16   | 3    |
| SFCR & RSR        | 12   | 13   | 0    |
| SFCR & ORSA       | 0    | 7    | 1    |
| RSR & ORSA        | 0    | 13   | 1    |
| SFCR & RSR & ORSA | 0    | 7    | 0    |

Table 1: Number of companies with submitted reports from each corresponding year and report type(s).

# Converting PDF to text

- Relation between number of pages and number of extracted sections
- Many different PDF producers

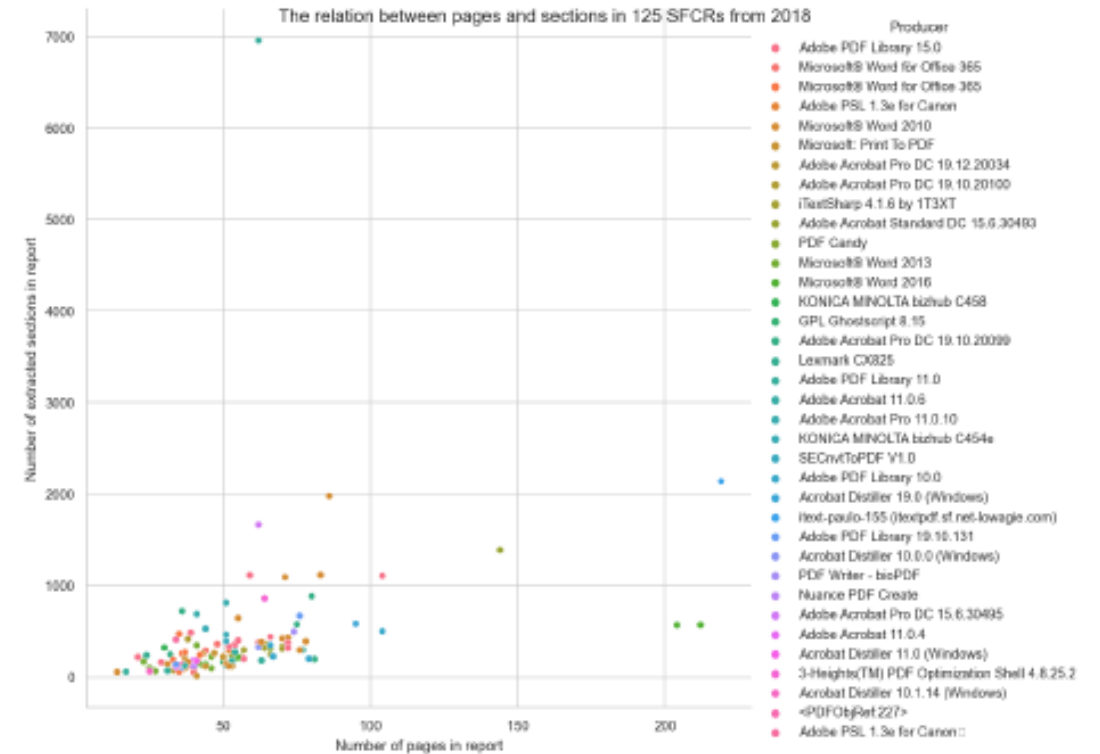


Figure 2: The relation between the number of pages and section in SFCRs. Each dot represents a single SFCR.

# Classification

- Risk matrix
- Classifier
  - Predefined classes – true label
  - Classifier output: Known risk category

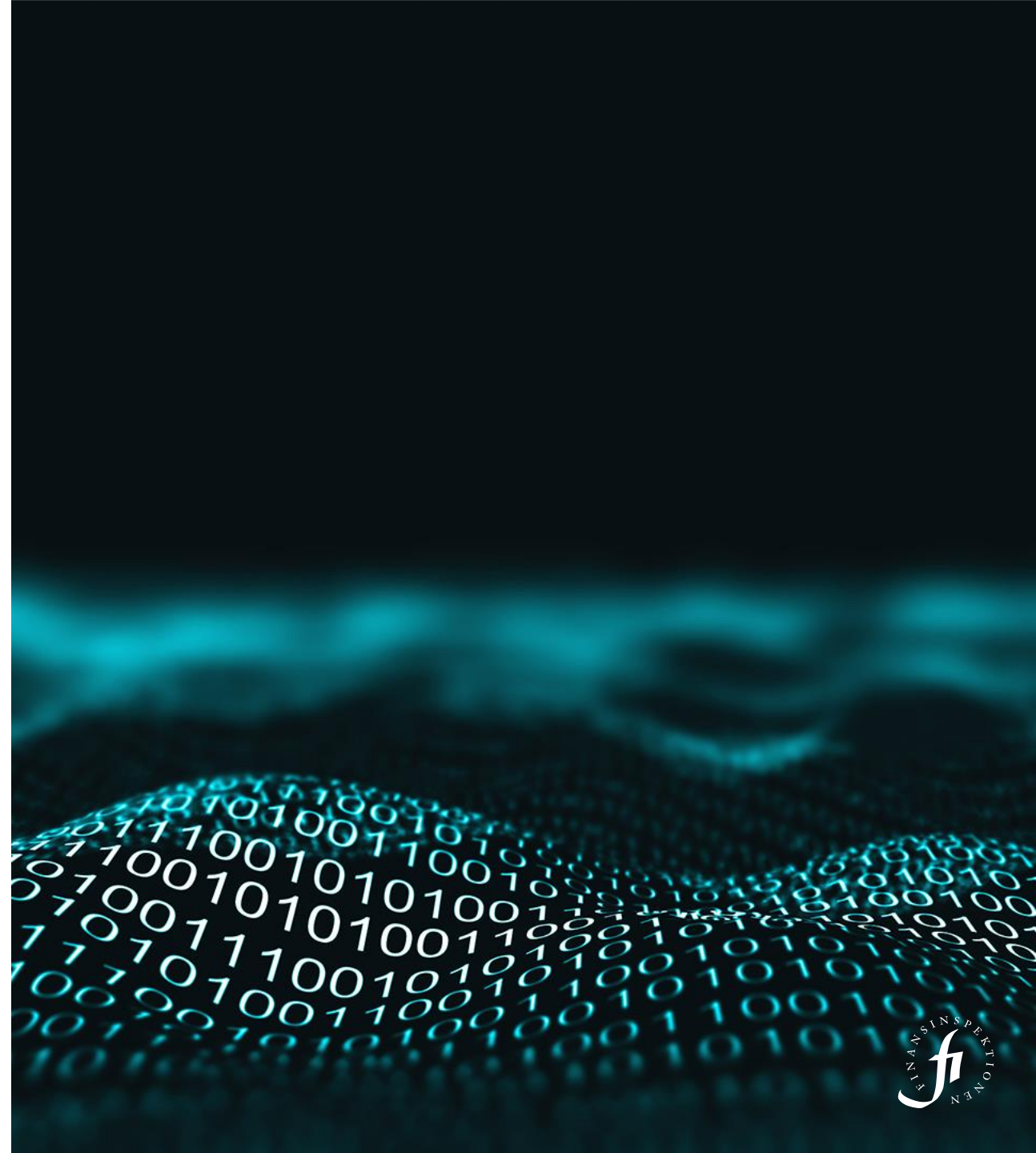
| REK TOTAL   | SOLO      |
|-------------|-----------|
| 2023 (2022) | Category  |
| Red         |           |
| Orange      |           |
| Yellow      |           |
| Green       |           |
| Total       | 156 (158) |

| Name     | Number of Classes | Classes     |        |           |     |
|----------|-------------------|-------------|--------|-----------|-----|
| Original | 4                 | Green       | Yellow | Orange    | Red |
| Ternary  | 3                 | GreenYellow |        | Orange    | Red |
| Binary   | 2                 | GreenYellow |        | OrangeRed |     |

Table 2: Illustration of the categorization schemes (Original, Ternary, and Binary).

# Experiments

- Text classification model
  - TF-IDF for numerical vector
    - Dummy
    - Logistic regression
    - Multinomial Naive Bayes classifier
    - Complement; different version of Multinomial Naive Bayes
    - Linear Support Vector classifier



## Result clustering PDF

- Overall clustering of SFCR
- Also clustered per mandatory section
- Distance between different reports

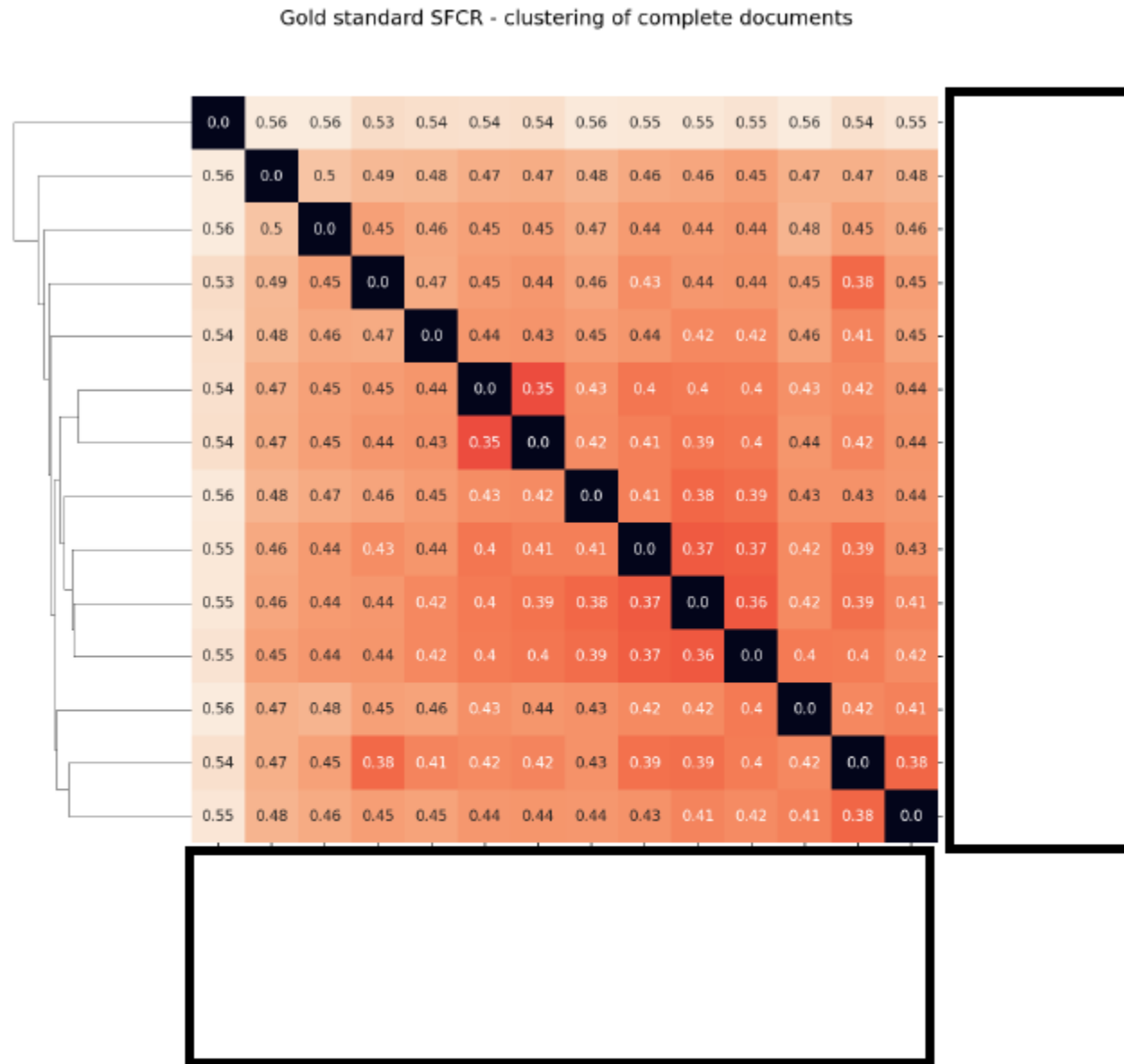


Figure 6: A cluster map indicating illustrating the pairwise distances between the SFCRs in the gold standard data set.

# Results

- Modelling effects
  - Limited SFCR
  - One-hot encoded

| Dataset    | Grouping | Size | Classifier |           |             |                  |                  |
|------------|----------|------|------------|-----------|-------------|------------------|------------------|
|            |          |      | Dummy      | Logistic  | Multinomial | Complement       | SVM              |
| ORSA       | Liv      | 7    | 71.4       | 64.7±18.3 | 64.7±18.3   | 60.9±23.9        | 60.4±23.5        |
|            | Skade    | 9    | 55.6       | 52.1±18.9 | 47.2±19.0   | 47.1±19.2        | 51.1±19.1        |
| RSR        | Liv      | 9    | 66.7       | 66.7±0.0  | 63.1±15.0   | 51.3±23.5        | 58.9±20.7        |
|            | Skade    | 11   | 54.5       | 55.9±15.7 | 59.3±17.5   | 66.2±18.3        | 62.5±19.8        |
| SFCR       | Liv      | 17   | 52.9       | 57.8±17.7 | 56.5±15.3   | 60.3±16.9        | 67.7±19.0        |
|            | Skade    | 25   | 60.0       | 64.7±9.5  | 60.5±4.0    | <b>73.4±12.5</b> | <b>85.7±10.3</b> |
| SFCR & RSR | Liv      | 7    | 57.1       | 46.6±17.2 | 46.6±17.9   | 45.4±30.1        | 41.1±24.8        |
|            | Skade    | 6    | 100.0      | N/A       | N/A         | N/A              | N/A              |

Table 10: Classification results (mean accuracy and standard deviation) of different classifiers on the three different types of reports, as well as the combination of SFCR and RSR, when split over company domain (Liv & Skade). Bold highlights the average accuracy that lies outside one standard deviation from the accuracy of the Dummy classifier. The N/A symbol means that no results were obtained.





# Lessons learned...

---



Thank you for your time!

Questions?

Contact:

**Jimmy Hollén**

**[jimmy.hollen@fi.se](mailto:jimmy.hollen@fi.se)**



