# Chapter 1

# Data

## 1.1 Introduction

The identification of code smells is critical for enhancing software maintainability and scalability in software engineering. This chapter presents the datasets used to analyze code smells in two Java projects: xerces-2.7.0 and fineract. The xerces-2.7.0 dataset, with 865 instances and 12 attributes, focuses on detecting Blob Class, Spaghetti Code, and Swiss Army Knife smells. The fineract dataset, referred to as code_smells_extended, contains 4059 instances and is evaluated in two configurations (57 and 16 attributes), targeting Spaghetti Code detection. Two classifiers, J48 and RandomForest, are assessed using 10-fold cross-validation to evaluate their effectiveness. The chapter also outlines plans to incorporate additional Java projects and results from SMURF and modern code smell detection tools for microservices, aligning with the objectives of this MSc AI capstone project.

## 1.2 Dataset Descriptions

### 1.2.1 Xerces-2.7.0 Dataset

The xerces-2.7.0 dataset, derived from WekaNose's analysis, initially contained 1000 instances, reduced to 865 after preprocessing to eliminate incomplete data. It includes 12 attributes, as detailed in Table 1.1, comprising nine numerical metrics and three binary labels for Blob Class, Spaghetti Code, and Swiss Army Knife.

The dataset is imbalanced, with 42 instances labeled "yes" (indicating at least one code smell) and 823 labeled "no," reflecting the rarity of code smells in software projects.

### 1.2.2 Fineract Dataset

The fineract dataset, code_smells_extended, contains 4059 instances and is analyzed in two configurations: one with 57 attributes and another with 16 attributes, both focusing on detecting Spaghetti Code (is_spaghetti). The 57-attribute set includes detailed metrics such as fan-in, fan-out, and quantities of specific code elements (e.g., loops, try-catch blocks), as listed in Table 1.2. The 16-attribute set is a subset, focusing on core metrics and smell indicators, as shown in Table 1.3.

Table 1.1: Xerces-2.7.0 Dataset Attributes

| Attribute | Type | Description |
|---|---|---|
| loc | Numeric | Lines of Code - measures class size |
| wmc | Numeric | Weighted Methods per Class - indicates class complexity |
| cbo | Numeric | Coupling Between Objects - measures class coupling |
| lcom | Numeric | Lack of Cohesion of Methods - assesses method relatedness |
| rfc | Numeric | Response For a Class - counts invocable methods |
| dit | Numeric | Depth of Inheritance Tree - measures inheritance depth |
| noc | Numeric | Number of Children - counts immediate subclasses |
| totalMethodsQty | Numeric | Total number of methods in the class |
| totalFieldsQty | Numeric | Total number of fields in the class |
| is_blob_class | {yes, no} | Indicates if the class is a Blob Class |
| is_spaghetti_code | {yes, no} | Indicates if the code is Spaghetti Code |
| is_swiss_army_knife | {yes, no} | Indicates if the class is a Swiss Army Knife |

The fineract dataset is highly imbalanced, with 17 instances labeled “yes” for Spaghetti Code and 4042 labeled “no” in both configurations, posing a significant challenge for detecting the minority class.

### 1.2.3 Additional Java Projects

Additional Java projects are planned for inclusion in the dataset to enhance the robustness and generalizability of the code smell detection analysis. These projects will be analyzed similarly, with results integrated once available, focusing on consistent metrics and classifiers to enable cross-project comparisons.

## 1.3 Experimental Setup

The J48 and RandomForest classifiers were implemented using Weka with 10-fold cross-validation. J48 was configured with a confidence factor of 0.25 and a minimum of 2 instances per leaf. RandomForest used 100 trees with a minimum of 1 instance per leaf and a variance threshold of 0.001. The classifiers were evaluated on the xerces-2.7.0 dataset (predicting any code smell) and the fineract dataset (predicting Spaghetti Code) in its 57- and 16-attribute configurations.

## 1.4 Results for Xerces-2.7.0 Dataset

### 1.4.1 J48 Classifier

The J48 classifier produced a pruned decision tree with 5 leaves, emphasizing class size (loc) and coupling (cbo):

loc <= 474: no (816.0)
loc > 474
| cbo <= 12
| | cbo <= 5: no (5.0)
| | cbo > 5
| | | loc <= 1021: yes (3.0)
| | | loc > 1021: no (2.0)
| cbo > 12: yes (39.0)

Performance metrics are summarized in Table 1.4, showing 99.19% accuracy.

Detailed accuracy by class is shown in Table 1.5, with a confusion matrix in Table 1.6.

### 1.4.2 RandomForest Classifier

RandomForest, with 100 trees, achieved identical accuracy (99.19%) to J48, as shown in Table 1.7. It outperformed J48 in ROC Area (0.999 vs. 0.963) and PRC Area for the "yes" class (0.985 vs. 0.880), as detailed in Table 1.8.

## 1.5 Results for Fineract Dataset

### 1.5.1 J48 Classifier (57 Attributes)

For the 57-attribute configuration, J48 produced a simple tree with 3 leaves, relying on is_blob and totalmethodsqty:

is_blob = yes
| totalmethodsqty <= 22: no (2.0)
| totalmethodsqty > 22: yes (17.0)
is_blob = no: no (4040.0)

The classifier achieved 99.9261% accuracy, as shown in Table 1.10, with a high Kappa statistic (0.9139).

Detailed accuracy is shown in Table 1.11, with a confusion matrix in Table 1.12.

### 1.5.2 J48 Classifier (16 Attributes)

The 16-attribute configuration produced an identical J48 tree, achieving a slightly higher accuracy of 99.9507%, as shown in Table 1.13. The perfect recall (1.000) for the "yes" class is notable, as seen in Table 1.14.

### 1.5.3 RandomForest Classifier (57 Attributes)

RandomForest on the 57-attribute set achieved 99.9015% accuracy, slightly lower than J48, with a lower Kappa statistic (0.8745), as shown in Table 1.16. The recall for the “yes” class (0.824) is lower than J48, indicating missed Spaghetti Code instances (Table 1.17).

### 1.5.4 RandomForest Classifier (16 Attributes)

RandomForest on the 16-attribute set matched J48’s accuracy (99.9261%) and Kappa statistic (0.9139), as shown in Table 1.19. It achieved high recall (0.941) for the “yes” class, as detailed in Table 1.20.

## 1.6 Comparison of Classifiers and Datasets

### 1.6.1 Xerces-2.7.0 Dataset

Both J48 and RandomForest achieved 99.19% accuracy on the xerces-2.7.0 dataset, with identical confusion matrices (39 TP, 3 FN, 4 FP, 819 TN). RandomForest outperformed J48 in ROC Area (0.999 vs. 0.963) and PRC Area for the “yes” class (0.985 vs. 0.880), indicating better ranking performance for code smells. J48’s simpler tree (5 leaves) offers interpretability, highlighting loc and cbo as key predictors.

### 1.6.2 Fineract Dataset

The fineract dataset’s larger size (4059 instances) and focus on Spaghetti Code present a more challenging task due to greater imbalance (17 “yes” vs. 4042 “no”). Key observations include:

- J48 (16 Attributes): Achieved the highest accuracy (99.9507%) and perfect recall (1.000) for the “yes” class, with only 2 misclassifications. The tree structure, relying on is_blob and totalmethodsqty, suggests a strong correlation between Blob Class and Spaghetti Code.
- J48 (57 Attributes): Slightly lower accuracy (99.9261%) with 3 misclassifications, but an identical tree structure, indicating that the additional 41 attributes did not enhance decision-making.
- RandomForest (16 Attributes): Matched J48’s accuracy (99.9261%) and Kappa (0.9139), with strong recall (0.941) for the “yes” class and superior PRC Area (0.990 vs. 0.975).
- RandomForest (57 Attributes): Lower accuracy (99.9015%) and recall (0.824) for the “yes” class, with 4 misclassifications, suggesting that the additional attributes introduced noise or complexity that reduced performance.

The identical J48 trees across both attribute sets indicate that is_blob and totalmethodsqty are dominant predictors for Spaghetti Code in fineract. The 16-attribute configuration generally outperformed the 57-attribute set, suggesting that a focused feature set improves classification performance for this imbalanced dataset.

### 1.6.3 Cross-Project Comparison

The xerces-2.7.0 dataset, with a smaller size (865 instances) and broader smell detection (three smells), shows lower accuracy (99.19%) than the fineract dataset (up to 99.9507%) but a higher number of positive instances (42 vs. 17). The fineract dataset's extreme imbalance makes perfect recall more critical, which J48 (16 attributes) achieves. RandomForest's superior ranking metrics (ROC and PRC) in both datasets highlight its robustness for imbalanced data, though J48's interpretability is valuable for understanding key predictors like loc, cbo, and is_blob.

## 1.7 Future Work

This chapter will be expanded to include additional Java projects to enhance the generalizability of the findings. These projects will be analyzed using consistent metrics and classifiers to enable cross-project comparisons. Additionally, results from SMURF, an SVM-based method for uncovering refactoring opportunities, and modern code smell detection tools for microservices (e.g., leveraging Convolutional Neural Networks or Neutrosophic logic [1]) will be integrated. These additions will provide a comprehensive evaluation of code smell detection across traditional and modern software architectures.

## 1.8 Conclusion

The J48 and RandomForest classifiers demonstrate high accuracy in detecting code smells in the xerces-2.7.0 (99.19%) and fineract (up to 99.9507%) datasets. For xerces-2.7.0, RandomForest's superior ROC and PRC metrics make it preferable for ranking code smells, while J48's simplicity aids interpretability. For fineract, J48 with 16 attributes achieves the best performance, with perfect recall for Spaghetti Code, highlighting the effectiveness of a focused feature set. The inclusion of additional Java projects and tools like SMURF will further strengthen this capstone project's contribution to software quality assessment.

Table 1.2: Fineract Dataset Attributes (57 Attributes)

| Attribute | Type | Description |
|---|---|---|
| cbo | Numeric | Coupling Between Objects |
| cbomodified | Numeric | Modified Coupling Between Objects |
| fanin | Numeric | Number of classes calling this class |
| fanout | Numeric | Number of classes called by this class |
| wmc | Numeric | Weighted Methods per Class |
| dit | Numeric | Depth of Inheritance Tree |
| noc | Numeric | Number of Children |
| rfc | Numeric | Response For a Class |
| lcom | Numeric | Lack of Cohesion of Methods |
| lcom* | Numeric | Normalized Lack of Cohesion |
| tcc | Numeric | Tight Class Cohesion |
| lcc | Numeric | Loose Class Cohesion |
| totalmethodsqty | Numeric | Total number of methods |
| staticmethodsqty | Numeric | Number of static methods |
| publicmethodsqty | Numeric | Number of public methods |
| privatemethodsqty | Numeric | Number of private methods |
| protectedmethodsqty | Numeric | Number of protected methods |
| defaultmethodsqty | Numeric | Number of default methods |
| visiblemethodsqty | Numeric | Number of visible methods |
| abstractmethodsqty | Numeric | Number of abstract methods |
| finalmethodsqty | Numeric | Number of final methods |
| synchronizedmethodsqty | Numeric | Number of synchronized methods |
| totalfieldsqty | Numeric | Total number of fields |
| staticfieldsqty | Numeric | Number of static fields |
| publicfieldsqty | Numeric | Number of public fields |
| privatefieldsqty | Numeric | Number of private fields |
| protectedfieldsqty | Numeric | Number of protected fields |
| defaultfieldsqty | Numeric | Number of default fields |
| finalfieldsqty | Numeric | Number of final fields |
| synchronizedfieldsqty | Numeric | Number of synchronized fields |
| nosi | Numeric | Number of statements in the class |
| loc | Numeric | Lines of Code |
| returnqty | Numeric | Number of return statements |
| loopqty | Numeric | Number of loops |
| comparisonsqty | Numeric | Number of comparison operations |
| trycatchqty | Numeric | Number of try-catch blocks |
| parenthesizedexpsqty | Numeric | Number of parenthesized expressions |
| stringliteralsqty | Numeric | Number of string literals |
| numbersqty | Numeric | Number of numeric literals |
| assignmentsqty | Numeric | Number of assignments |
| mathoperationsqty | Numeric | Number of mathematical operations |
| variablesqty | Numeric | Number of variables |
| maxnestedblocksqty | Numeric | Maximum number of nested blocks |
| anonymousclassesqty | Numeric | Number of anonymous classes |
| innerclassesqty | Numeric | Number of inner classes |
| lambdasqty | Numeric | Number of lambda expressions |
| uniquewordsqty | Numeric | Number of unique words |
| modifiers | Numeric | Modifier flags |
| logstatementsqty | Numeric | Number of log statements |
| is_functional_decomposition | {yes, no} | Indicates Functional Decomposition smell |

Table 1.3: Fineract Dataset Attributes (16 Attributes)

| Attribute | Type | Description |
|---|---|---|
| loc | Numeric | Lines of Code |
| wmc | Numeric | Weighted Methods per Class |
| cbo | Numeric | Coupling Between Objects |
| lcom | Numeric | Lack of Cohesion of Methods |
| rfc | Numeric | Response For a Class |
| noc | Numeric | Number of Children |
| totalmethodsqty | Numeric | Total number of methods |
| totalfieldsqty | Numeric | Total number of fields |
| is_functional_decomposition | {yes, no} | Indicates Functional Decomposition smell |
| is_data_class | {yes, no} | Indicates Data Class smell |
| is_lazy_class | {yes, no} | Indicates Lazy Class smell |
| is_spaghetti | {yes, no} | Indicates Spaghetti Code smell |
| is_swiss$_{a}rmy_knife$ | {yes, no} | Indicates Swiss Army Knife smell |
| is_blob | {yes, no} | Indicates Blob Class smell |
| has_long_method | {yes, no} | Indicates Long Method smell |
| has_long_parameter_list | {yes, no} | Indicates Long Parameter List smell |

Table 1.4: J48 Performance Metrics (Xerces-2.7.0 Dataset)

| Metric | Value |
|---|---|
| Correctly Classified Instances | 858 (99.1908%) |
| Incorrectly Classified Instances | 7 (0.8092%) |
| Kappa Statistic | 0.9134 |
| Mean Absolute Error (MAE) | 0.0078 |
| Root Mean Squared Error (RMSE) | 0.086 |
| Relative Absolute Error (RAE) | 8.3432% |
| Root Relative Squared Error (RRSE) | 40.0059% |

Table 1.5: J48 Detailed Accuracy by Class (Xerces-2.7.0 Dataset)

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Are |
|---|---|---|---|---|---|---|---|---|
| yes | 0.929 | 0.005 | 0.907 | 0.929 | 0.918 | 0.913 | 0.963 | 0.88 |
| no | 0.995 | 0.071 | 0.996 | 0.995 | 0.996 | 0.913 | 0.963 | 0.99 |
| Weighted Avg. | 0.992 | 0.068 | 0.992 | 0.992 | 0.992 | 0.913 | 0.963 | 0.99 |

Table 1.6: J48 Confusion Matrix (Xerces-2.7.0 Dataset)

| | Predicted Yes | Predicted No |
|---|---|---|
| Actual Yes | 39 | 3 |
| Actual No | 4 | 819 |

Table 1.7: RandomForest Performance Metrics (Xerces-2.7.0 Dataset)

| Metric | Value |
|---|---|
| Correctly Classified Instances | 858 (99.1908%) |
| Incorrectly Classified Instances | 7 (0.8092%) |
| Kappa Statistic | 0.9134 |
| Mean Absolute Error (MAE) | 0.0134 |
| Root Mean Squared Error (RMSE) | 0.0733 |
| Relative Absolute Error (RAE) | 14.3392% |
| Root Relative Squared Error (RRSE) | 34.0957% |

Table 1.8: RandomForest Detailed Accuracy by Class (Xerces-2.7.0 Dataset)

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Are |
|---|---|---|---|---|---|---|---|---|
| yes | 0.929 | 0.005 | 0.907 | 0.929 | 0.918 | 0.913 | 0.999 | 0.98 |
| no | 0.995 | 0.071 | 0.996 | 0.995 | 0.996 | 0.913 | 0.999 | 1.00 |
| Weighted Avg. | 0.992 | 0.068 | 0.992 | 0.992 | 0.992 | 0.913 | 0.999 | 0.99 |

Table 1.9: RandomForest Confusion Matrix (Xerces-2.7.0 Dataset)

| | Predicted Yes | Predicted No |
|---|---|---|
| Actual Yes | 39 | 3 |
| Actual No | 4 | 819 |

Table 1.10: J48 Performance Metrics (Fineract Dataset, 57 Attributes)

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4056 (99.9261%) |
| Incorrectly Classified Instances | 3 (0.0739%) |
| Kappa Statistic | 0.9139 |
| Mean Absolute Error (MAE) | 0.0008 |
| Root Mean Squared Error (RMSE) | 0.0261 |
| Relative Absolute Error (RAE) | 8.9412% |
| Root Relative Squared Error (RRSE) | 40.4741% |

Table 1.11: J48 Detailed Accuracy by Class (Fineract Dataset, 57 Attributes)

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Are |
|---|---|---|---|---|---|---|---|---|
| yes | 0.941 | 0.000 | 0.889 | 0.941 | 0.914 | 0.914 | 0.971 | 0.91 |
| no | 1.000 | 0.059 | 1.000 | 1.000 | 1.000 | 0.914 | 0.971 | 1.00 |
| Weighted Avg. | 0.999 | 0.059 | 0.999 | 0.999 | 0.999 | 0.914 | 0.971 | 0.99 |

Table 1.12: J48 Confusion Matrix (Fineract Dataset, 57 Attributes)

| | Predicted Yes | Predicted No |
|---|---|---|
| Actual Yes | 16 | 1 |
| Actual No | 2 | 4040 |

Table 1.13: J48 Performance Metrics (Fineract Dataset, 16 Attributes)

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4057 (99.9507%) |
| Incorrectly Classified Instances | 2 (0.0493%) |
| Kappa Statistic | 0.9442 |
| Mean Absolute Error (MAE) | 0.0005 |
| Root Mean Squared Error (RMSE) | 0.0209 |
| Relative Absolute Error (RAE) | 6.08% |
| Root Relative Squared Error (RRSE) | 32.365% |

Table 1.14: J48 Detailed Accuracy by Class (Fineract Dataset, 16 Attributes)

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Are |
|---|---|---|---|---|---|---|---|---|
| yes | 1.000 | 0.000 | 0.895 | 1.000 | 0.944 | 0.946 | 1.000 | 0.97 |
| no | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 0.946 | 1.000 | 1.00 |
| Weighted Avg. | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 0.946 | 1.000 | 1.00 |

Table 1.15: J48 Confusion Matrix (Fineract Dataset, 16 Attributes)

| | Predicted Yes | Predicted No |
|---|---|---|
| Actual Yes | 17 | 0 |
| Actual No | 2 | 4040 |

Table 1.16: RandomForest Performance Metrics (Fineract Dataset, 57 Attributes)

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4055 (99.9015%) |
| Incorrectly Classified Instances | 4 (0.0985%) |
| Kappa Statistic | 0.8745 |
| Mean Absolute Error (MAE) | 0.0021 |
| Root Mean Squared Error (RMSE) | 0.0264 |
| Relative Absolute Error (RAE) | 23.9332% |
| Root Relative Squared Error (RRSE) | 40.8019% |

Table 1.17: RandomForest Detailed Accuracy by Class (Fineract Dataset, 57 Attributes)

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Are |
|---|---|---|---|---|---|---|---|---|
| yes | 0.824 | 0.000 | 0.933 | 0.824 | 0.875 | 0.876 | 1.000 | 0.97 |
| no | 1.000 | 0.176 | 0.999 | 1.000 | 1.000 | 0.876 | 1.000 | 1.00 |
| Weighted Avg. | 0.999 | 0.176 | 0.999 | 0.999 | 0.999 | 0.876 | 1.000 | 1.00 |

Table 1.18: RandomForest Confusion Matrix (Fineract Dataset, 57 Attributes)

| | Predicted Yes | Predicted No |
|---|---|---|
| Actual Yes | 14 | 3 |
| Actual No | 1 | 4041 |

Table 1.19: RandomForest Performance Metrics (Fineract Dataset, 16 Attributes)

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4056 (99.9261%) |
| Incorrectly Classified Instances | 3 (0.0739%) |
| Kappa Statistic | 0.9139 |
| Mean Absolute Error (MAE) | 0.0014 |
| Root Mean Squared Error (RMSE) | 0.0212 |
| Relative Absolute Error (RAE) | 15.7365% |
| Root Relative Squared Error (RRSE) | 32.7691% |

Table 1.20: RandomForest Detailed Accuracy by Class (Fineract Dataset, 16 Attributes)

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Are |
|---|---|---|---|---|---|---|---|---|
| yes | 0.941 | 0.000 | 0.889 | 0.941 | 0.914 | 0.914 | 1.000 | 0.99 |
| no | 1.000 | 0.059 | 1.000 | 1.000 | 1.000 | 0.914 | 1.000 | 1.00 |
| Weighted Avg. | 0.999 | 0.059 | 0.999 | 0.999 | 0.999 | 0.914 | 1.000 | 1.00 |

Table 1.21: RandomForest Confusion Matrix (Fineract Dataset, 16 Attributes)

| | Predicted Yes | Predicted No |
|---|---|---|
| Actual Yes | 16 | 1 |
| Actual No | 2 | 4040 |

# Bibliography

[1] Cerny, T., et al. (2023). On Code Smells and Microservices: A Systematic Literature Review. Software: Practice and Experience.