Data Quality Assessment and Recommendation of Feature Selection Algorithms: An Ontological Approach

Aparna Nayak*, Bojan Božić and and Luca Longo

SFI Centre for Research Training in Machine Learning, School of Computer Science, Technological University Dublin, Dublin, Republic of Ireland
E-mail: aparna.nayak@tudublin.ie; bojan.bozic@tudublin.ie; luca.longo@tudublin.ie
**Corresponding Author*

Received 17 October 2022; Accepted 10 January 2023;
Publication 14 April 2023

Abstract

Feature selection plays an important role in machine learning and data mining problems. Identifying the best feature selection algorithm that helps to remove irrelevant and redundant features is a complex task. This research tries to address it by recommending a feature selection algorithm based on dataset meta-features. The main contribution of the work is the use of Semantic Web principles to develop a recommendation model for the feature selection algorithm. As a result, dataset meta-features are modeled in a domain ontology, and a set of Semantic Web rule language (SWRL) predictive rules have been proposed to recommend a feature selection algorithm. The result of this research is a feature selection algorithm recommendation based on the data characteristics and quality (FSDCQ) ontology, which not only helps with recommendations but also finds the data points with data quality violations. An experiment is conducted on the classification datasets from the UCI repository to evaluate the proposed ontology. The usefulness

Journal of Web Engineering, Vol. 22_1, 175–196.
doi: 10.13052/jwe1540-9589.2219
© 2023 River Publishers

and effectiveness of the proposed method is evaluated by comparing it with the widely used method in the literature for the recommendation. Results show that the ontology-based recommendations are equally good as the widely used recommendation model, which is k-NN, with added benefits.

Keywords: Data quality, feature selection algorithm, meta-features, ontology, recommendation.

1 Introduction

The selection of feature subsets is an essential part in the fields of data mining and machine learning. An optimal feature subset helps to improve the performance of machine learning models by making them more generalizable and interpretable [8, 12]. A good feature selection algorithm can eliminate irrelevant and redundant features [11]. Applying candidate feature selection algorithms to the given dataset and selecting the most effective feature subset is one of the most practical methods of determining the best feature selection algorithm. However, this is a time-consuming task. One of the methods to tackle this problem is by identifying the relationship between the feature selection algorithms and dataset meta-features.

The existing literature demonstrates a positive correlation between the performance of a feature selection algorithm and the dataset's characteristics [1, 47]. To address this specific relationship, we propose a domain ontology that models both dataset meta-features and feature selection algorithms. Along with dataset meta-features, we also propose to include dataset quality as it contributes to machine learning model performances [23]. As a result, the proposed ontology is modeled with both data quality metrics and dataset meta-features. Hence, our ontology is named dataset characteristics and quality (DCQ) ontology. Feature selection algorithm recommendation using DCQ (FSDCQ) is modeled by adding rules to the domain ontology DCQ. The rules in Semantic Web Rule Language (SWRL) format helps to infer new knowledge from the existing ontology. Thus, it enhances the expressivity and completeness of the ontology [6]. The benefits of using an ontology to deliver such a recommendation include interoperability, potential reuse, and knowledge sharing [50].

Additionally, the FSDCQ ontology is intended to identify the data points with data quality issues. The quality of the dataset has a significant impact on the performance of machine learning tasks [19, 46]. Various techniques are available for data quality assessment [4, 7, 38]. However, they fail to identify the data points that have violated the data quality [29]. In this work, we attempt to identify and represent the data quality problems associated with the dataset using the ontology FSDCQ.

This study investigates the specific research question, “To what extent can a domain ontology facilitate machine learning tasks by recommending feature selection algorithms and analysing data quality issues?”. The work’s main objective is to adopt Semantic Web techniques to develop a novel model that can aid in feature selection algorithm recommendation. The use of rule language enables a better understanding of the role of each meta-feature, thereby increasing the model’s explainability [24, 55].

In our earlier work, we introduced the FSDCQ ontology [30], where the ontology is modeled and tested with a small number of datasets. The current work extends FSDCQ by adding data quality analysis and investigating the outcome of 100+ datasets, thus making the ontology more robust. The remainder of this article is structured as follows. Section 2 reviews related work on the existing approaches to automatically recommending feature selection algorithms and existing ontologies to describe the dataset quality and its characteristics. Section 3 describes the basic workflow of the recommendation of feature selection algorithm using ontology, followed by a detailed analysis of the datasets in Section 4. The implementation specifics are discussed in Section 5. The results of the experiment are presented and discussed in Section 6. Finally, Section 7 concludes the research work by providing directions for future work.

2 Related Work

This section briefly discusses the existing work on automatic feature selection recommendation methods and the application of ontologies related to data characteristics and quality.

2.1 Feature Selection

The two most commonly used methods for selecting a subset are (i) the filter approach and (ii) the wrapper approach. While various feature selection algorithms have been proposed, some of these outperform others in terms of performance (for example, classification accuracy) for a given dataset [57]. This leads to the emergence of a new research field associated with establishing intrinsic relationships between dataset characteristics and feature selection algorithms. In order to identify methods to recommend feature selection algorithms, a literature review was carried out. Dataset meta-features describe the properties of the dataset which are predictive for the performance of machine learning algorithms trained on them [27, 42].

Dataset characteristics are the description of a dataset, representing its structural, statistical, and other properties. Most of the literature focuses on three distinct sets of measures of dataset characteristics: (i) simple, statistical, and information-theoretical features, (ii) model-based features, and (iii) landmarking features [54]. Simple properties are those taken directly from the attribute value table of the dataset. Statistical properties represent the correlation and symmetry of attributes. Information-theoretical properties seek to characterise the nominal attributes and their relationship with the class attribute. Model-based properties adopt machine learning methods to represent dataset features. Landmarking properties illustrate the performance achieved by simple classification algorithms. Table 1 summarises the approaches that have used meta-features to build recommendation models to automatically select algorithms for machine learning tasks.

2.2 Ontology

A methodology for constructing an ontology from conception to completion is discussed in Methontology [14] where a set of activities conforming the ontology development process is presented. Following best practices in ontology development, the data characteristics and quality (DCQ) ontology reuses appropriate classes from a set of ontologies that are designed for data quality and data mining applications. An extensive literature review has been conducted to understand existing vocabularies to support meta-features, and a vocabulary of terms have been composed for DCQ.

Meta-features are usually described as a part of data mining (DM) ontologies. “OntoDM” is a general data mining ontology designed to provide a unified framework for data mining research. It makes an attempt to encompass the entirety of the data mining cycle [33]. “Expose” is an ontology for standardizing the description of machine learning experiments. This ontology is used to express and share metadata about experiments [53]. To represent the relationship between data mining tasks and dataset characteristics, multiple ontologies have been designed. “OntoDM-KDD” [34], “OntoDT” [35], and “CRISP-DM” [49] are some of the additional ontologies that are based on data mining related concepts. “DMOP” is a data mining optimization ontology that supports various stages of the data mining process [21]. A class

Table 1 Literature review and comparison of advisory functions used for recommendations

						Dataset Characteristic			
Source	Advisory Function	Number of Datasets	Number of Classification Techniques	Number of Feature Selection algorithms	Evaluation Metrics	Simple, Statistical	Information Theoretical	Model Based	Landmarking
[20]	Ranking based on McNemar test	1082*	5	8	Accuracy	✓	✓	✗	✗
[26]	SVM	156	–	7	Accuracy	✓	✓	✗	✗
[28]	k-NN	58	–	–	F1 score				
[32]	C5.0 decision tree	128	5	3	Accuracy, time complexity	✓	✓	✗	✗
[36]	Ranking based on MCPM	213	5	5	Learning time, percentage of selected attributes, error rate	✓	✓	✓	✓
[37]	k-NN	47	–	10	Spearman's rank correlation	✓	✓	✗	✓
[39]	k-NN	38	–	9	Accuracy	✓	✓	✗	✗
[40]	Regression	123	–	5	Correlation	✓	✓	✓	✗
[41]	Regression	54	–	9	Accuracy	✓	✓	✓	✓
[47]	J4.8 decision trees	26	4	3	Accuracy	✓	✗	✗	✗
[48]	k-NN	84	–	–	Accuracy, execution time	✓	✗	✗	✗
[57]	k-NN	115	22	5	Recommendation hit ration based on accuracy	✓	✓	✗	✗
[59]	Variance, LIBSVM	84	–	3	Accuracy	✓	✓	✓	✓

*Includes artificial dataset.

hierarchy established in DMOP between datasets and their attributes is reused in DCQ.

Data quality is one of the essential components while describing a dataset. Data quality management (DQM) is an ontology that refers to the conceptualization of the data quality domain, the establishment of cleaning standards, and the reporting of data quality problems [15]. Data cleaning ontology (DCO) refines and extends data cleaning operations which directly assesses data quality [3]. Reasoning violations ontology (RVO) describes the reasoning errors of RDF and OWL [5]. The World Wide Web Consortium (W3C)[1] recommends a set of standard vocabularies data quality vocabulary (DQV), which covers most of the aspects of data quality [2]. None of the

[1]https://www.w3.org/TR/vocab-dqv/

aforementioned ontologies discuss the analysis of data quality assessment. However, in the case of linked data, analysis of data quality assessment is discussed in [29, 51] by identifying the erroneous triples based on metric execution failure.

In detail, an advisory function refers to a method that aims to recommend an algorithm from an existing knowledge base. The proposed work aims to use ontology as advisory function. Some of the applications that use ontology as advisory methods/recommendation are product recommendation based on text [43], health-care [9, 10] and higher education [31]. Therefore, it is a novel approach to solve recommendation of a feature selection algorithm using ontology. To the best of our knowledge, no research has focused on considering data quality as a characteristic of a dataset for the task under investigation. In this article, beside the aforementioned simple, statistical, information, and quality-based measures we propose an additional category to characterise datasets, which includes quality-based measures.

3 Methodology

The basic workflow of the proposed method is depicted in Figure 1. It consists of three phases: dataset meta-feature extraction, target feature extraction, and

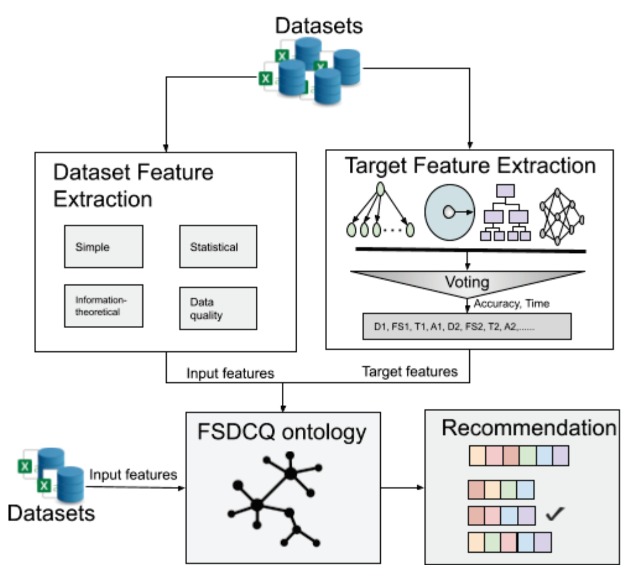

Figure 1 Recommendation model for feature selection algorithms.

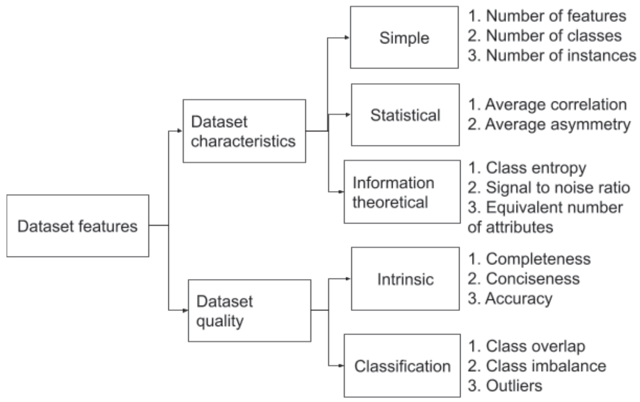

Figure 2 Dataset meta-features [30].

ontology modelling. Each phase contributes to the final ontology, which is finally combined to construct a recommendation model.

3.1 Dataset Feature Extraction

Dataset features represent the characteristics and quality of the dataset. A total of 14 characteristics have been extracted from each dataset. Figure 2 lists all the features that have been extracted from the datasets. The majority of features are extracted from raw datasets, as characteristics represent the nature of the dataset prior to preprocessing. A minimal preprocessing is applied before extracting some of the features, which is mentioned in Section 5.

3.2 Target Feature Extraction

The objective of target feature extraction is to find a feature selection algorithm that performs better on the considered dataset. In this step, the ensemble classifier is used with each feature selection algorithm to find out which one works best. Four classifiers, i.e, instance (k-NN) [18], symbolic (C4.5) [44], statistical (Naive Bayes) [52], and connectionist (SVM) are used as base classifiers in ensemble classification. The advantage of using an ensemble classifier is that it accounts for the bias of different machine learning algorithms as well as boosts the performance of a single model's predictions by training numerous models and integrating their results [13,45]. Results of the base classifier predictions are aggregated by a soft voting method [16, 22]. Finally, the performance of a feature selection algorithm is measured by the classification accuracy and the time required to select features by feature selection algorithms.

3.3 Ontology for Recommendation and Data Quality Assessment

The proposed ontology is modeled by considering existing domain ontologies from data mining and machine learning. Some of the classes/properties are modified based on the requirements of FSDCQ. Dataset meta-features will be either object properties or data properties in the ontology. Each dataset is paired with a feature selection algorithm selected based on its performance. Thus, in the proposed ontology, each dataset is associated with its meta-features and feature selection algorithm.

Apart from feature selection algorithm recommendation, the ontology is also modeled to identify data quality issues. Typically, when evaluating dataset quality, data points with quality violations are not specified. In the proposed method, each data quality metric is associated with a number of data points with quality violations. These data points with quality violations can be considered for quality enhancement using either predefined procedures or a human in the loop. The FSDCQ ontology is modeled to locate quality-violating data points in the dataset. Thus, the user can query the ontology to identify data points that have violated the quality.

The ontology is populated with datasets, their features, and a feature selection algorithm. This ontology serves as a recommendation model, capable of recommending an optimal algorithm for feature selection by computing dataset meta-features.

4 Datasets

We analysed all the datasets from the machine learning repository of the University of California, Irvine (UCI), a popular data source in the classification literature [25, 56]. There are 599 datasets in the UCI collection, of which 466 are acceptable for classification tasks.[2] Initially, 128 datasets were eliminated for lacking textual content. An additional 251 datasets were excluded for various reasons, including the following: (1) data having image features, text features, time series, molecular information, and geospatial features; (2) the presence of duplicate datasets; (3) datasets with empty files;

[2] Searched in March 2022.

and (4) extremely small datasets. Furthermore, datasets were excluded to reduce the amount of data cleaning. The majority of exclusionary criteria only excluded one or two datasets: (1) datasets that span multiple sheets within a single file (two datasets); (2) datasets with labels in a separate file (one dataset); and (3) datasets with multiple delimiters (one dataset). Finally, we considered 82 datasets for the complete analysis.

On the one hand, some datasets contain multiple files, each of which represents a distinct dataset. One example is wine quality, represented by two datasets containing samples of red and white wine. On the other hand, multiple files containing the sub-strings "train" and "test" that represent a single dataset are merged into a single file and treated as a single dataset. This enables the experiment to treat all datasets in a consistent way. As a result, we now have a total of 104 datasets, which are summarised in appendix Table 4 in the appendix. The first column indicates the name of the dataset. The second column denotes whether or not a dataset contains multiple files. The third column preprocessing indicates if any preprocessing is required and its type is mentioned in the last column.

Each dataset in our work contains samples between 31 and 49999, with features ranging from 4 to 242, and labels ranging from 2 to 10. We do not perform extensive data preprocessing or data transformation. The reasons for this are as follows: (1) our objective is not to achieve the state of the art performance for each dataset but to determine which feature selection algorithm performs best on the dataset regardless of its domain; (2) to avoid bias introduced by preprocessing, it may be prudent to use an unprocessed original dataset or a dataset that has undergone minimal preprocessing; (3) classification results could be improved further by applying dataset specific preprocessing, which requires domain knowledge and which is outside the scope of the paper.

4.1 Label Identification

Identifying the column that corresponds to label information is important for our approach. The following assumptions are made to correctly determine the label column within the dataset; (1) when both the first and the last columns contain categorical data, a number of distinct labels are identified. The column with the fewest unique values is considered as the label column; (2) when both the first and last columns contain the same number of unique values, the last column is given priority; (3) when neither the first nor the last column contains a categorical value, the columns in the dataset are scanned

Table 2 Datasets that are impacted by manual preprocessing

Method	Total Datasets
Changed header	1
Encoding UTF-16	1
File extension modification	2
Changed filename	2
Deleted metadata file	4
Removed additional header	5
No preprocessing	71

from the beginning to find categorical values. If the number of unique values in any column is less than 10, the column is considered the label. The above-mentioned assumptions are verified across all datasets. Eight datasets that did not meet the criteria are handled individually.

4.2 Manual Preprocessing

The minimal manual preprocessing steps that are applied to datasets are covered in this subsection. This preprocessing step assists in standardising the format of all datasets. The majority of datasets required no form of preprocessing. Among all of the considered datasets, fifteen datasets required manual preprocessing with the following steps: (1) in most cases, text files (txt extension) represent additional information related to datasets, therefore, text files representing datasets are converted; (2) remove additional headers in xlsx; (3) delete metadata presented in the same sheet; (4) rename the file to combine train and test datasets; (5) change the encoding format from UTF-16 to UTF-8; and (6) remove additional headers. Table 2 shows the number of datasets affected by manual preprocessing. The proposed model processes datasets with the file extensions ARFF, XLS, XLSX, CSV, and DATA. The system also handles datasets that are compressed (zip, rar). When multiple datasets with the same file content are represented by different file extensions, the CSV, ARFF, and XLSX file extensions are prioritized in order.

5 Implementation Specifics

This section describes implementation specifics and the experiments carried out to validate the proposed methodology. Experiments are conducted on a machine running Linux Mint 19.3 Cinnamon and powered by an Intel(R) Core(TM) i7-9750H CPU running at 2.60 GHz with 16 GB of RAM. The datasets that are considered for the experiment are tabulated in Table 4 in the appendix.

The majority of the meta-features are extracted before the preprocessing steps are applied. However, the presence of non-integer data prevents the extraction of certain characterization measures. Some features necessitated the following preprocessing on the datasets: (1) missing values are either substituted with zeros or excluded from the analysis; (2) sklearn's label encoder is used to convert qualitative nominal values to integer values.

Dataset features are extracted from each dataset to model the ontology FSDCQ, as shown in Figure 2. A supporting document is made available in the git repository that explains the formulas/algorithms used to compute all the meta-features.[3] Dataset characteristics are broadly classified into three dimensions as described in Section 2. The proposed research takes into account the characteristics of the dataset identified as significant in [36].

Meta-features related to data quality are classified into two dimensions. The intrinsic dimension represents the metrics that are independent of the user's context [58]. A classification dimension represents the metrics that are important for a machine learning classification algorithm [17].

The target feature is constructed by adopting filter based feature selection algorithms that assess the features using various evaluation methods. Filter based feature selection algorithms are mainly based on the evaluation metrics dependency, distance, and consistency. The experiments are based on the following six feature selection algorithms: (1) mutual information (MI); (2) gain ratio (GR); (3) fast correlation based filter (FCBF); (4) minimum redundancy maximum relevance (mRMR); (5) Relief; (6) ReliefF. Each feature selection algorithm is evaluated by passing it through an ensemble classifier.

A robust recommendation model has to be evaluated by considering multiple metrics. Hence, the final target feature is selected based on the accuracy of the ensemble classifier and the time required by each feature selection algorithm to select features. The extracted meta-features are populated in the proposed ontology using the owlready python package.[4]

SWRL works on the principle of unification. It is challenging to obtain datasets with identical characteristics in the real world. We have thus normalised every value in the dataset. Each value is encoded as either zero or one, depending on whether it falls within the column's normalised range. The ontology is populated with the normalised values of the dataset features. This

[3]https://github.com/aparnanayakn/onto-DCQ-FS
[4]https://owlready2.readthedocs.io/en/latest/index.html

Table 3 Evaluation comparison of the experiment

Dataset	k-NN	FSDCQ (Proposed)	Actual
Secondary data.csv	MI	[MI; relief]	MI
Cryotherapy.xlsx	GR	[relief; GR]	GR
Tuandromd.csv	relief	[MI; relief]	relief
Online shoppers intention.csv	relief	[mRMR; relief; GR]	relief
Wine.data	mRMR	mRMR	mRMR
Somerville Happiness Survey.csv	mRMR	[mRMR; relief; GR]	mRMR
Transfusion.data	MI	[FCBF; MI; relief]	MI
Spambase.data	relief	[GR; reliefF]	reliefF
Audit risk.csv	MI	[mRMR; GR]	MI
Divorce.csv	GR	[mRMR; relief; GR]	GR

populated ontology acts as a recommender model. The SWRL rule helps to recommend a feature selection algorithm if a dataset is not associated with one. We can also query the FSDCQ ontology with dataset features to get a better feature selection algorithm.

6 Results and Discussion

The experiment is evaluated by comparing the proposed rule-based method with most commonly used advisory function (refer to Table 1). Datasets considered for model evaluation are tabulated in Table 3. Ten datasets are randomly selected for evaluation, while the remaining datasets are used for training.

6.1 Results

Table 3 represents the comparison of the proposed method with k-NN on the test datasets. The actual and recommended/predicted (k-NN; FSDCQ) feature selection algorithms for each dataset are listed. Findings suggest that the FSDCQ performs similarly to the k-NN. However, FSDCQ provides multiple recommendations for most datasets, allowing the user to narrow down the number of candidate feature selection algorithms. In the case of multiple recommendations, it is remarkable that one of the recommendations is correct.

6.2 Discussion

The findings show that the rule-based method performs as good as the more prevalent advisory method. However, instead of recommending one outperforming feature selection algorithm, FSDCQ recommends multiple feature selection algorithms. One of the reasons for this could be SWRL rule unifies the testing dataset with multiple training data points. When the experiment was exhaustively carried out using different testing datasets, some datasets lacked recommendations. We assume that having more training data points might not lead to this problem.

To determine the impact of data quality metrics on the recommendation model, the experiment is conducted by eliminating dataset quality metrics. However, the recommendation performed poorly. We assume this is because SWRL rules receives only a limited number of attributes for unification.

Data points with quality violations are successfully stored in the ontology. The user can write SPARQL queries to identify data points and comprehend metrics that violate data quality. This allows users to improve data quality in the future without analysing the entire dataset.

7 Conclusion and Future Work

In this research work, we have presented the FSDCQ ontology. It provides a conceptual framework for meta learning and the relationships between meta-features to enable the recommendation of feature selection algorithms. The methodology proposed for recommending feature selection algorithms establishes relationships between ontology individuals and unifies them to recommend feature selection algorithms. Additionally, FSDCQ associates data quality metrics with data points that violate the metric definition.

In future study, we will strengthen the FSDCQ ontology by making it self-explainable. FSDCQ should be able to provide a reason for the recommendation. Another interesting extension would be clustering the datasets based on their domain, and tailor feature selection (recommendation) to the domain under consideration.

Acknowledgements

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Appendix

A. Datasets

Table 4 Datasets

Dataset	Multiple files	Preprocessing	Type
Accelerometer Data Set	No	No	
Algerian Forest Fires Dataset Data Set	No	Yes	Removed additional header
Audit dataset	Yes	No	
Autism Screening Adult Data Set	No	No	
Autistic Spectrum Disorder Screening Data for Adolescent Data Set	No	No	
Autistic Spectrum Disorder Screening Data for Children Data Set	No	No	
Balance Scale Data Set	No	No	
Bank marketing	Yes	No	
Banknote authentication Data Set	No	No	
Blood Transfusion Service Center Data Set	No	No	
Bone marrow transplant: children Data Set	No	No	
Breast Cancer Coimbra Data Set	No	No	
Breast Cancer Wisconsin (Diagnostic) Data Set	No	No	
Burst Header Packet (BHP) flooding attack on Optical Burst Switching (OBS) Network Data Set	No	No	
Caesarian Section Classification Dataset Data Set	No	No	
Car Evaluation Data Set	No	No	
Cargo 2000 Freight Tracking and Tracing Data Set	No	No	
Cervical cancer (Risk Factors) Data Set	No	No	
Cervical Cancer Behavior Risk	No	No	
Chemical Composition of Ceramic Samples Data Set	No	No	
Chess	No	No	
Climate Model Simulation Crashes Data Set	No	No	
Congressional Voting Records Data Set	No	No	
Cryotherapy Dataset Data Set	No	Yes	Deleted metadata
Default of credit card clients Data Set	No	Yes	Removed additional header
Divorce Predictors data set Data Set	No	No	
Drug consumption (quantified) Data Set	No	No	
Dry Bean Dataset Data Set	No	No	
Ecoli	No	Yes	Changed header line
Electrical Grid Stability Simulated Data Data Set	No	No	
Estimation of obesity levels based on eating habits and physical condition Data Set	No	No	
Extention of Z-Alizadeh sani dataset Data Set	No	Yes	Deleted metadata
Fertility Data Set	No	No	
First-order theorem proving Data Set	No	Yes	Changed file name
Glass Identification Data Set	No	No	
Hayes-Roth Data Set	No	Yes	File extension changed from txt to data
HCC Survival Data Set	No	No	
HCV data Data Set	No	No	
Heart failure clinical records Data Set	No	No	
Hepatitis C Virus (HCV) for Egyptian patients Data Set	No	Yes	Deleted metadata
Higher Education Students Performance Evaluation Dataset Data Set	No	No	
HTRU2	No	No	
ILPD (Indian Liver Patient Dataset) Data Set	No	No	
Immunotherapy Dataset Data Set	No	Yes	Deleted metadata
Iris Dataset	No	No	
Las Vegas Strip Data Set	No	No	
Lung cancer	No	No	
Lymphography Data Set	No	No	
Mammographic Mass Data Set	No	No	
MONK's Problems Data Set	Yes	Yes	Removed additional header
Mushrooms	No	No	
Myocardial infarction complications Data Set	No	Yes	Removed additional header
Non verbal tourists data Data Set	No	No	
Nursery Data Set	No	No	
Online Shoppers Purchasing Intention Dataset Data Set	No	No	
Parkinsons Data Set	No	No	
Phishing Websites Data Set	No	No	
Polish companies bankruptcy data Data Set	Yes	No	
Primary Tumor Data Set	No	No	
QSAR biodegradation Data Set	No	No	
Raisin Dataset Data Set	No	No	
Risk Factor prediction of Chronic Kidney Disease Data Set	No	Yes	Removed additional header
Secondary Mushroom Dataset Data Set	Yes	No	
Seeds Data Set	No	Yes	File extension changed from txt to data
Seismic-bumps Data Set	No	No	
Sepsis survival minimal clinical records Data Set	Yes	No	
Somerville Happiness Survey Data Set	No	Yes	Encoding 16
South German Credit (UPDATE) Data Set	No	No	
Soybean	Yes	No	
Spambase Data Set	No	No	
SPECTF Heart Data Set	Yes	Yes	Changed file name
SUSY Data Set	No	No	
Tennis Major Tournament Match Statistics Data Set	Yes	No	
Thoracic Surgery Data Data Set	No	No	
Tic-Tac-Toe Endgame Data Set	No	No	
TUANDROMD (Tezpur University Android Malware Dataset) Data Set	No	No	
Turkish Music Emotion Dataset Data Set	No	No	
Vertebral Column Data Set	Yes	No	
Website Phishing Data Set	No	No	
Wholesale customers Data Set	No	No	
Wine Data Set	No	No	
Wine Quality Data Set	Yes	No	
Yeast	No	No	

References

[1] Robert Aduviri, Daniel Matos, and Edwin Villanueva. Feature selection algorithm recommendation for gene expression data through gradient boosting and neural network metamodels. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2726–2728, 2018.

[2] Riccardo Albertoni and Antoine Isaac. Introducing the data quality vocabulary (DQV). *Semantic Web*, 12(1):81–97, 2021.

[3] Ricardo Almeida, Paulo Maio, Paulo Oliveira, and João Barroso. An ontology-based methodology for reusing data cleaning knowledge. In *KEOD 2015 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, pages 202–211. SciTePress, 2015.

[4] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3), jul 2009.

[5] Bojan Bozic, Rob Brennan, Kevin Feeney, and Gavin Mendel-Gleason. Describing reasoning results with rvo, the reasoning violations ontology. In *MEPDaW and LDQ co-located with ESWC*, volume 1585 of *CEUR Workshop Proceedings*, pages 62–69, 2016.

[6] Qiushi Cao, Ahmed Samet, Cecilia Zanni-Merk, François de Bertrand de Beuvron, and Christoph Reich. An ontology-based approach for failure classification in predictive maintenance using fuzzy c-means and swrl rules. *Procedia Computer Science*, 159:630–639, 2019. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019.

[7] Cinzia Cappiello, Chiara Francalanci, and Barbara Pernici. Data quality assessment from the user's perspective. In *Proceedings of the 2004 International Workshop on Information Quality in Information Systems*, IQIS '04, page 68–73, New York, NY, USA, 2004. Association for Computing Machinery.

[8] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[9] Jianguo Chen, Kenli Li, Huigui Rong, Kashif Bilal, Nan Yang, and Keqin Li. A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. *Information Sciences*, 435:124–149, 2018.

[10] Rung-Ching Chen, Yun-Hou Huang, Cho-Tsan Bau, and Shyi-Ming Chen. A recommendation system based on domain ontology and swrl for anti-diabetic drugs selection. *Expert Systems with Applications*, 39(4):3995–4006, 2012.

[11] Padraig Cunningham, Bahavathy Kathirgamanathan, and Sarah Jane Delany. Feature selection tutorial with python examples, 2021.

[12] Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156, 1997.

[13] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers Comput. Sci.*, 14(2):241–258, 2020.

[14] Mariano Fernández-López, Asunción Gómez-Pérez, and Natalia Juristo. Methontology: from ontological art towards ontological engineering. 1997.

[15] Christian Fürber and Martin Hepp. Towards a vocabulary for data quality management in semantic web architectures. In *Proceedings of the 2011 EDBT/ICDT Workshop on Linked Web Data Management*, pages 1–8. ACM, 2011.

[16] Isha Gandhi and Mrinal Pandey. Hybrid ensemble of classifiers using voting. In *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pages 399–404, 2015.

[17] Nitin Gupta, Shashank Mujumdar, Hima Patel, Satoshi Masuda, Naveen Panwar, Sambaran Bandyopadhyay, Sameep Mehta, Shanmukha Guttula, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. Data quality for machine learning tasks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 4040–4041, New York, NY, USA, 2021. Association for Computing Machinery.

[18] Iris Hendrickx and Antal van den Bosch. Hybrid algorithms with instance-based classification. In *Machine Learning: ECML 2005, 16th European Conference on Machine Learning*, volume 3720 of *Lecture Notes in Computer Science*, pages 158–169. Springer, 2005.

[19] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3561–3562, New York, NY, USA, 2020. Association for Computing Machinery.

[20] Alexandros Kalousis and Melanie Hilario. Feature selection for meta-learning. In *Knowledge Discovery and Data Mining – PAKDD*, volume 2035 of *Lecture Notes in Computer Science*, pages 222–233. Springer, 2001.

[21] C. Maria Keet, Agnieszka Lawrynowicz, Claudia d'Amato, Alexandros Kalousis, Phong Nguyen, Raúl Palma, Robert Stevens, and Melanie Hilario. The data mining optimization ontology. *Journal of web semantics*, 32:43–53, 2015.

[22] Saloni Kumari, Deepika Kumar, and Mamta Mittal. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, 2:40–46, 2021.

[23] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. Cleanml: A study for evaluating the impact of data cleaning on ml classification tasks. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 13–24, 2021.

[24] Luca Longo, Randy Goebel, Freddy Lecue, Peter Kieseberg, and Andreas Holzinger. Explainable artificial intelligence: Concepts, applications, research challenges and visions. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 1–16. Springer, 2020.

[25] Núria Macià and Ester Bernadó-Mansilla. Towards UCI+: A mindful repository design. *Information Sciences*, 261:237–262, 2014.

[26] Rafael Gomes Mantovani, André L. D. Rossi, Edesio Alcobaça, Joaquin Vanschoren, and André C. P. L. F. de Carvalho. A meta-learning recommender system for hyperparameter tuning: Predicting when tuning improves SVM classifiers. *Information Sciences*, 501:193–221, 2019.

[27] L.C. Molina, L. Belanche, and A. Nebot. Feature selection algorithms: a survey and experimental evaluation. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 306–313, 2002.

[28] Munehiro Nakamura, Atsushi Otsuka, and Haruhiko Kimura. Automatic selection of classification algorithms for non-experts using meta-features. *China-USA Business Review*, 13(3), 2014.

[29] Aparna Nayak, Bojan Božić, and Luca Longo. Data quality assessment of comma separated values using linked data approach. In Witold Abramowicz, Sören Auer, and Milena Stróżyna, editors, *Business Information Systems Workshops*, pages 240–250, Cham, 2022. Springer International Publishing.

[30] Aparna Nayak, Bojan Božić, and Luca Longo. An ontological approach for recommending a feature selection algorithm. In *Web Engineering*, pages 300–314, Cham, 2022. Springer International Publishing.

[31] Charbel Obeid, Inaya Lahoud, Hicham El Khoury, and Pierre-Antoine Champin. Ontology-based recommender system in higher education. In *Companion Proceedings of the The Web Conference 2018*, pages 1031–1034, 2018.

[32] Dijana Oreski, Stjepan Oreski, and Bozidar Klicek. Effects of dataset characteristics on the performance of feature selection techniques. *Applied Soft Computing*, 52:109–119, 2017.

[33] Pance Panov, Saso Dzeroski, and Larisa N. Soldatova. Ontodm: An ontology of data mining. In *Workshops Proceedings of the 8th IEEE International Conference on Data Mining*, pages 752–760. IEEE Computer Society, 2008.

[34] Pance Panov, Larisa N. Soldatova, and Saso Dzeroski. Ontodm-kdd: Ontology for representing the knowledge discovery process. In *Discovery Science - 16th International Conference, DS*, volume 8140 of *Lecture Notes in Computer Science*, pages 126–140. Springer, 2013.

[35] Pance Panov, Larisa N. Soldatova, and Saso Dzeroski. Generic ontology of datatypes. *Information Sciences*, 329:900–920, 2016.

[36] Antonio Rafael Sabino Parmezan, Huei Diana Lee, Newton Spolaôr, and Feng Chung Wu. Automatic recommendation of feature selection algorithms based on dataset characteristics. *Expert Systems with Applications*, 185:115589, 2021.

[37] Yonghong Peng, Peter A. Flach, Carlos Soares, and Pavel Brazdil. Improved dataset characterisation for meta-learning. In *Discovery Science, 5th International Conference*, volume 2534 of *Lecture Notes in Computer Science*, pages 141–152. Springer, 2002.

[38] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. Data quality assessment. *Commun. ACM*, 45(4):211–218, apr 2002.

[39] Nitin Pise and Parag Kulkarni. Algorithm selection for classification problems. In *SAI Computing Conference (SAI)*, pages 203–211. IEEE, 2016.

[40] Matthias Reif, Faisal Shafait, and Andreas Dengel. Prediction of classifier training time including parameter optimization. In *Advances in Artificial Intelligence*, volume 7006 of *Lecture Notes in Computer Science*, pages 260–271. Springer, 2011.

[41] Matthias Reif, Faisal Shafait, Markus Goldstein, Thomas M. Breuel, and Andreas Dengel. Automatic classifier selection for non-experts. *Pattern Analysis and Applications*, 17(1):83–96, 2014.

[42] Adriano Rivolli, Luís P.F. Garcia, Carlos Soares, Joaquin Vanschoren, and André C.P.L.F. de Carvalho. Meta-features for meta-learning. *Knowledge-Based Systems*, 240:108101, 2022.

[43] Renata Lopes Rosa, Gisele Maria Schwartz, Wilson Vicente Ruggiero, and Demóstenes Zegarra Rodríguez. A knowledge-based recommendation system that includes sentiment analysis and deep learning. *IEEE Transactions on Industrial Informatics*, 15(4):2124–2135, 2018.

[44] Salvatore Ruggieri. Efficient c4.5 [classification algorithm]. *IEEE transactions on knowledge and data engineering*, 14(2):438–444, 2002.

[45] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *WIREs Data Mining Knowl. Discov.*, 8(4), 2018.

[46] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 614–622, New York, NY, USA, 2008. Association for Computing Machinery.

[47] Samar Shilbayeh and Sunil Vadera. Feature selection in meta learning framework. In *Science and Information Conference*, pages 269–275. IEEE, 2014.

[48] Qinbao Song, Guangtao Wang, and Chao Wang. Automatic recommendation of classification algorithms based on dataset characteristics. *Pattern Recognition*, 45(7):2672–2689, 2012.

[49] Man Tianxing, Myo Myint, Wang Guan, Nataly Zhukova, and Nikolay Mustafin. A hierarchical data mining process ontology. In *28th Conference of Open Innovations Association (FRUCT)*, pages 465–471. IEEE, 2021.

[50] Mike Uschold and Michael Gruninger. Ontologies: Principles, methods and applications. *The knowledge engineering review*, 11(2):93–136, 1996.

[51] Ramneesh Vaidyambath, Jeremy Debattista, Neha Srivatsa, and Rob Brennan. An intelligent linked data quality dashboard. In *Proceedings for the 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*, volume 2563 of *CEUR Workshop Proceedings*, pages 341–352. CEUR-WS.org, 2019.

[52] Linda C. van der Gaag and Andrea Capotorti. Naive bayesian classifiers with extreme probability features. In *International Conference on Probabilistic Graphical Models*, volume 72 of *Proceedings of Machine Learning Research*, pages 499–510. PMLR, 2018.

[53] Joaquin Vanschoren and Larisa Soldatova. Exposé: An ontology for data mining experiments. In *International workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2010)*, pages 31–46, 2010.
[54] Ricardo Vilalta, Christophe G. Giraud-Carrier, Pavel Brazdil, and Carlos Soares. Using meta-learning to support data mining. *International Journal of Computer Science Applications*, 1(1):31–45, 2004.
[55] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.
[56] Kiri Wagstaff. Machine learning that matters. *arXiv*, 2012.
[57] Guangtao Wang, Qinbao Song, Heli Sun, Xueying Zhang, Baowen Xu, and Yuming Zhou. A feature subset selection algorithm automatic recommendation method. *Journal of Artificial Intelligence Research*, 47:1–34, 2013.
[58] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016.
[59] Yang Zhongguo, Li Hongqi, Sikandar Ali, and Ao Yile. Choosing classification algorithms and its optimum parameters based on data set characteristics. *Journal of Computers*, 28(5):26–38, 2017.

Biographies

Aparna Nayak received her M.Tech degree from Manipal Academy of Higher Education, India. She has more than seven years of teaching experience. She is currently pursuing her Ph.D. at the Technological University Dublin, specializing in knowledge graphs. Her current research interests include machine learning and knowledge graphs.

Bojan Božić is a Lecturer in Computer Science at TU Dublin. He has worked on European research projects such as SANY (Sensor Web Enablement), TaToo (Tagging Tools for Semantic Discovery), Europeana Creative (Cultural Inheritage), PELAGIOS (Linked Data), and C2-SENSE (Sensor Web and Interoperability). He also has contributed to the H2020 project ALIGNED, modelling data and software engineering processes through ontologies and annotations for the Dacura platform. His current research interests are Semantic Web, machine learning, and natural language processing.

Luca Longo is a curious individual deeply devoted to and highly passionate for science. He strives for excellence and contribution to knowledge. He received his doctoral degree in Artificial Intelligence at Trinity College Dublin after a bachelor and masters in Computer Science, Statistics and Health Informatics. He is actively engaged in dissemination of scientific material to the public as his TEDx talks demonstrate. He has received various awards both for his research work and for his teaching. With his team of doctoral and post-doctoral students, he conducts fundamental research in explainable artificial intelligence, defeasible reasoning, and non-monotonic argumentation. He also performs applied research in machine learning and predictive data analytics, mainly applied to the problem of mental workload modelling.