DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DATA MINING AND KNOWLEDGE DISCOVERY IN EDUCATION

by

Ferda BALCI ÜNAL

May, 2022

İZMİR

DATA MINING AND KNOWLEDGE DISCOVERY IN EDUCATION

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computer Engineering

by
Ferda BALCI ÜNAL

May, 2022
İZMİR

Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "**DATA MINING AND KNOWLEDGE DISCOVERY IN EDUCATION**" completed by **FERDA BALCI ÜNAL** under supervision of **Assoc. Prof. Dr. Derya BİRANT** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Assoc. Prof. Dr. Derya BİRANT

Supervisor

Prof. Dr. Alp KUT

Thesis Committee Member

Asst. Prof. Dr. Pelin YILDIRIM TAŞER

Thesis Committee Member

Assoc. Prof. Dr. Tuğba ÖZACAR ÖZTÜRK

Examining Committee Member

Assoc. Prof. Dr. Aytuğ ONAN

Examining Committee Member

Prof. Dr. Okan FISTIKOĞLU

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my supervisor, Assoc. Prof. Dr. Derya BİRANT for her great mentorship, continuous support, encouragement, thoughtful comments, and recommendations during my Ph.D. study. Her insightful feedback has encouraged me all the time.

In addition, I would also like to thank my thesis committee; Prof. Dr. Alp KUT, and Asst. Prof. Dr. Pelin YILDIRIM TAŞER for their valuable advice for this study.

Finally, I am very much thankful to my husband and my son for their love, encouragement, and continuing support over the past few years. Without their tremendous understanding, it would be impossible for me to complete my study.

Ferda BALCI ÜNAL

DATA MINING AND KNOWLEDGE DISCOVERY IN EDUCATION

ABSTRACT

This thesis aims to increase knowledge discovery in the field of education by developing data mining methods. Considering that there is a large amount of ordinal and unlabeled data in the education area, this study focused on using semi-supervised classification and ordinal classification techniques.

Semi-supervised learning is a type of machine learning technique that constructs a classifier by learning from a small collection of labeled samples and a large collection of unlabeled ones. Although some progress has been made in this research area, the existing semi-supervised methods provide a nominal classification task. However, semi-supervised learning for ordinal classification is yet to be explored. To bridge the gap, two concepts, "semi-supervised learning" and "ordinal classification", were combined in this study for the categorical class labels for the first time and introduced a new concept of "semi-supervised ordinal classification".

Our study proposes a new method for semi-supervised learning that takes into account the relationships between the class labels, especially class orderings such as low, medium, and high. We performed an extensive empirical study that involved 10 benchmarks and 3 educational ordinal different quantities of labeled datasets with samples varying from 15% to 50% with an increment of 5%, aiming to evaluate the performance of our method by combining different base learners. The experiments showed that the proposed method improved the classification accuracy of the model compared to the existing semi-supervised method on ordinal data. We also developed a web application to provide the accessibility of our method.

Keywords: Data mining, education, semi-supervised learning, ordinal classification, machine learning

EĞİTİM ALANINDA VERİ MADENCİLİĞİ VE BİLGİ KEŞFİ

ÖZ

Tezin amacı, eğitim alanında veri madenciliği yöntemleri geliştirerek, bu alandaki bilgi keşfini arttırmaktır. Eğitim alanında çok sayıda sıralı ve etiketlenmemiş veri olduğu göz önüne alındığından, yarı denetimli sınıflandırma ve sıralı sınıflandırma tekniklerinin kullanılmasına odaklanılmıştır.

Yarı denetimli öğrenme, etiketli az sayıda örnek koleksiyonundan ve etiketlenmemiş çok sayıda örnek koleksiyonundan öğrenerek bir sınıflandırıcı oluşturan bir tür makine öğrenimi tekniğidir. Bu araştırma alanında bazı ilerlemeler kaydedilmesine rağmen, mevcut yarı denetimli yöntemler nominal bir sınıflandırma görevi sağlamaktadır. Ancak, sıralı sınıflandırma için yarı denetimli öğrenme henüz keşfedilmemiştir. Boşluğu kapatmak için bu çalışma, kategorik sınıf etiketleri için ilk kez "yarı denetimli öğrenme" ve "sıralı sınıflandırma" kavramlarını birleştirmekte ve yeni bir "yarı denetimli sıralı sınıflandırma" kavramını sunmaktadır.

Çalışmamız, özellikle düşük, orta ve yüksek gibi sınıf sıralamaları olmak üzere sınıf etiketleri arasındaki ilişkileri dikkate alan yarı denetimli öğrenme için yeni bir metot önermektedir. Farklı temel öğrenicileri birleştirerek yöntemimizin performansını değerlendirmeyi amaçlayan, %5' lik bir artışla %15 ila %50 arasında değişen farklı miktarlarda etiketli numunelere sahip 10 kıyaslama ve 3 eğitim sıralı veri setini içeren kapsamlı bir deneysel çalışma gerçekleştirdik. Deneyler, önerilen yöntemin, sıralı veriler üzerinde mevcut yarı denetimli yönteme kıyasla modelin sınıflandırma doğruluğunu iyileştirdiğini göstermiştir. Çalışmalarımıza ek olarak yöntemin erişilebilirliğini sağlamak için bir web uygulaması geliştirdik.

Anahtar kelimeler: Veri madenciliği, eğitim, yarı denetimli öğrenme, sıralı sınıflandırma, makine öğrenmesi

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 General

Machine learning builds a model to be able to predict an outcome for a new data instance based on prior data. The main types of machine learning are *supervised learning* and *unsupervised learning* which perform learning with labeled data and unlabeled data, respectively. A typical supervised learning task is a classification that finds the relationship between the objects in the training set and then assigns the unknown data to the classes previously determined. The classification has been known for the ability to give high accuracy to data estimation due to using labeled data. Nevertheless, obtaining labeled samples is a costly process, since it is usually based on expert experience. Therefore, this study focuses on *semi-supervised learning* (SSL) in which a small number of labeled data and a large number of unlabeled data are used together.

Ordinal classification is a special case of multiclass classification in which an inherent ordering among the classes exists. The ordinal classification aims to predict the class label of an unseen data instance by considering ranking relationships between the classes. For instance, an ordinal class attribute can have the five ranking values such as “strongly disagree”, “disagree”, “neutral”, “agree”, and “strongly agree” that are ordered from the worst situation to the best. Hence, the first-class label can be considered five times worse than the last one. Another example is that the size of an object can be one of the following labels: “large”, “medium”, and “small”. According to the literature (Frank et al., 2001), in case there is an order relationship among the class label, ignoring this relationship negatively affect classification performance. Therefore, this study focuses on applying ordinal classification to the field of education.

1.2 Purpose

The main motivations of this study are two-fold. (i) The existing SSL algorithms are capable of processing nominal or numerical class labels; however, we also need to capture the orderings among the categorical class labels. (ii) The limitation of the traditional ordinal classification is that it only uses labeled ordinal data to construct a classifier; nevertheless, we also need a learning paradigm dealing with the design of classifiers in the presence of both labeled ordinal data and unlabeled ordinal data.

The main novelty of this paper is that semi-supervised learning is used by ordinal classification together as a new approach. It proposes a new method, called *semi-supervised ordinal classification* (SSOC), which relies on semantic knowledge of the class label ordering and uses both labeled and unlabeled data together for classification. Our algorithm is different from the traditional regression and multiclass classification algorithms since, in the former, numeric target values are predicted based on a metric and, in the latter, there is no ordering between classes.

The first step of the proposed SSOC method is to train multiple binary classifiers on original labeled ordinal data. After that, the unlabeled data is labeled by the constructed classifiers via their predictions to generate additional labeled data, which is called pseudo-labeled data. Eventually, the final classifier is built by using both pseudo-labeled and originally labeled ordinal data.

1.3 Main Contributions of the Thesis

The main contributions of our study can be listed as follows: (i) This study combines “ordinal classification” and “semi-supervised learning” concepts for the categorical class labels for the first time and introduces a new concept of “semi-supervised ordinal classification”; (ii) This paper proposes a new method, called SSOC, for semi-supervised learning that takes into account the relationships between the class labels, especially class orderings such as bad, regular, and good; (iii) Our new method allows a standard base learner to be applied to an ordinal classification

task. Our study is original in comparing alternative base learners in conjunction with the proposed method, including decision tree (DT), support vector machines (SVM), k-nearest neighbors (KNN), random forest (RF) and neural network (NN); (iv) This is the first time that various proportions of labeled data (ranging between 15% and 50%) are investitaged for ordinal classification.

In the experiments, the efficiency of SSOC was verified on 10 benchmarks and 3 educational ordinal datasets by comparing it with the standard ordinal classification algorithm (Frank et al., 2001). In addition, it was also compared with YATSI (yet another two-stage idea) (Driessens et al., 2006) which is one of the well-known semi-supervised classification algorithms.

In the experiments, the SSOC method achieved high accuracy by the use of a set of labeled ordinal data and a set of unlabeled ordinal data. The Wilcoxon statistical test results showed that the differences in the performances of the proposed method and the existing method are statistically significant. As a result, the proposed SSOC method can be successfully used for the tasks in management decisions and assessments where both unlabeled and labeled data are available and the classes of the objects are discretely ordinal, instead of nominal or numerical.

1.4 Organization of the Thesis

The thesis consists of eight chapters. The remainder of this thesis is organized as follows:

Section 2 presents the previous works on educational data mining, semi-supervised learning as well as ordinal classification.

Chapter 3 explains the background information about this thesis, such as data mining methods, and educational data mining.

Chapter 4 explains the methods used in this thesis, such as semi-supervised learning and ordinal classification. Moreover, it introduces the proposed method and

highlights the main steps involved in it. Furthermore, the formal definition for the proposed SSOC method is also given in this chapter.

The first part of Chapter 5 is about the general description of SSOC. After that, it explains the datasets used in this study, presents the experimental studies, and gives results for the new method. Moreover, it presents a case study, which is conducted on educational data, and shows experimental results.

In Chapter 6, discussions about our study and main findings are given.

Chapter 7 presents the web application which was developed to make the proposed SSOC method accessible via a web interface.

Finally, concluding remarks and possible future works are given in the last chapter.

CHAPTER TWO
LITERATURE REVIEW

This chapter first gives a brief summary of previous works on educational data mining. In our study, two concepts of “ordinal classification” and “semi-supervised learning” were combined for the categorical class labels for the first time; we herein present a literature review of previous studies on both of them separately.

2.1 Literature Review for Educational Data Mining

Educational data mining is the process of extracting meaningful and useful information and patterns from large-scale educational datasets. The main subject of educational data mining is to predict students' academic performance (Fan et al., 2019; Guruler & Istanbullu, 2014). The increase of technological investments in the education area is observed as a result of the advancement of technology. E-Learning platforms (i.e., web-based online learning tools) have been developed along with technological developments. In addition, time and space constraints have been eliminated and learning costs have been minimized (Hu et al., 2014). Digital data in education is increasing as a result of availability of online training courses. Costa et al. (2017) reported that students at risk for failing courses can be predicted using advanced algorithms.

The estimation of performances of students becomes more difficult due to massive volume of edicational data (Shahiri & Husain, 2015). To provide descriptive information's basics of a given data, analysis of descriptive statistics can be used efficiently (Fernandes et al., 2019). Nevertheless, it is not always enough. Using estimation techniques, it is possible to identify students in risk in early stages to inform the instructors early. To provide descriptive information's basics of a given data, analysis of descriptive statistics can be used efficiently. (Marbouti et al., 2016). Managing the resources and increasing success rates is helpful to classify students to their academic performance potential properly (Miguéis et al., 2018). Nowadays, need to obtain meaningful information from the large volume of universities's

electronic data getting bigger rapidly (Asif et al., 2017). It is possible to increase the education processes quality, by data mining techniques applied to education data (Jayakumar et al., 2019).

In the field of education such as English Language education (Lu, 2017), physical education (Zhu, 2019), and engineering education (BuenañoFernandez et al., 2019), data mining algorithms have been used so far. Some studies have relevance to higher education (Abu et al., 2019), some of them have interested in high school students (Márquez-Vera et al., 2016). On the other hand, some studies have focused on instructor performance (Agaoglu, 2016), the other studies have interested in the prediction of student performance (Deshmukh et al., 2018).

2.2 Literature Review for Semi-Supervised Learning

Semi-supervised learning (SSL) is a type of machine learning technique that constructs a classifier by learning from a small collection of labeled samples and a large collection of unlabeled ones. In the field of SSL, some new methods have been developed, as well as some existing supervised methods have been adapted to the problem. Until now, semi-supervised learning methods have been successfully applied to problems in many different areas such as health (Lang et al., 2020), security (Li et al., 2020), education (Livieris et al., 2019), geology (Xu et al., 2019), biology (Stanescu & Caragea, 2019), real estate (Yang et al., 2016), and energy (Wang et al., 2019). A comprehensive survey on the SSL topic was conducted by Van & Hoos (2020). They give a broad overview of the SSL methods by presenting a new taxonomy, explaining recent advances, and providing basic assumptions underlying SSL.

Semi-supervised learning can be grouped into two different categories (Van et al., 2020): inductive learning and transductive learning. The goal of *inductive learning* (Dornaika et al., 2017) is to produce a prediction function that is defined on the whole input space. On the other hand, the idea of *transductive learning* (Xu et al., 2019; Rahangdale & Raut, 2019) is to directly perform predictions only for the

unlabeled data. In other words, given a dataset consisting of labeled (X_L) and unlabeled (X_U) samples such that X_L, $X_U \subseteq X$, with labels $Y_L \in Y$ for the labeled samples, the inductive methods consider both X_L and X_U to yield a model f : $X \rightarrow Y$, whereas the output of the transductive methods predicted labels $\hat{Y}_U$ for the given unlabeled samples (X_U) only. Our proposed algorithm (SSOC) is based on the inductive learning paradigm.

The semi-supervised learning methods can be mainly classified into generative models (Van et al., 2020), co-training (Stanescu & Caragea, 2019; Yang et al., 2016), self-training (Livieris et al., 2019), semi-supervised support vector machines (i.e., semi-SVM, S3VM) (Wang et al., 2019), disagreement-based methods (Li et al., 2020), and graph-based methods (Lang et al., 2020; Dornaika et al., 2017). Generative models (Van et al., 2020) build a joint probability model (i.e., Gaussian mixture model) depending on a distribution assumption, and the decision boundary is determined by using both unlabeled data samples and labeled data samples in the probabilistic framework such as expectation-maximization, maximum probability likelihood, or Bayes. The co-training methods (Stanescu & Caragea, 2019; Yang et al., 2016) iteratively build two or more classifiers by using multiple different views of data, in which the most confident predictions of a classifier on the unlabeled samples are utilized as the labeled training samples by another classifier in each iteration. Self-training (Livieris et al., 2019) is one of the widely used wrapper approaches, where a classifier firstly is trained with the initial labeled data for the purpose of classifying unlabeled samples, and then it is retrained by adding its predictions to the labeled data. The Semi-SVM method (Wang et al., 2019) is an extended version of the traditional SVM, which is used both labeled and unlabeled data to iteratively find a boundary that maximizes the margins between classes. The disagreement-based methods (Li et al., 2020) train multiple learners and exploit the disagreements among the learners during the SSL process. The graph-based methods (Lang et al., 2020; Dornaika et al., 2017) firstly build a weighted graph, in which each vertex corresponds to a data sample and the edge between two vertices refers to the pairwise similarity of data samples and then the methods use the graph to assign

class labels to the unlabeled data samples. In our study, a new approach based on the self-training method is proposed to address the ordinal classification.

No technique has been discovered for semi-supervised learning yet to determine prior information in which method is well-suited for a certain problem. Each method has some advantages and disadvantages, as well as each problem, even each dataset, has its characteristics. Therefore, a combination of empirical evaluation and theoretical analysis should be used to determine a method that is best suited to the given problem. For instance, Livieris et al. (2019) compared four different semi-supervised learning algorithms (co-training, self-training, tri-training, and YATSI) to determine the best one for the student performance prediction problem. Similarly, Uylas Sati (2020) tested many different SSL methods on real-world datasets.

A semi-supervised learning problem is designed to reflect an assumption that the algorithm builds on. The most commonly known assumptions are cluster assumption, manifold assumption, smoothness assumption, and low-density separation assumption (Chen & Wang, 2010). The cluster assumption considers that samples belonging to the same group are likely to be of the same class. The manifold assumption states that data samples on the same low-dimensional manifold have a great probability of having the same class value. The smoothness assumption means that if samples are close to each other in a high-density space, these samples may have the same label. The low-density separation assumption states that the class-decision boundary lies in the region where the data density is low in the input space. Since our proposed algorithm deals with the ordinal classification task, it is based on the assumption that if two samples are close, then, may have similar class rankings (orders).

Although most of the research on SSL is centered on semi-supervised classification, the remaining methods such as semi-supervised clustering (Rahangdale & Raut, 2019) and semi-supervised regression (Yang et al., 2016) have also been studied. The former focuses on building a model using labeled data and

unlabeled data together to make a prediction based on continuous variables, whereas the latter aims to improve clustering results with the help of labeled data.

Since the aforementioned SSL algorithms focus on nominal classification, they are out of the scope of this paper. They do not capture and reflect the orderings among the class labels; hence, they may lead to the construction of inefficient models in terms of accuracy in the case of ordinal data. Unlike the previous studies, this paper proposes a new paradigm for semi-supervised learning where ordinal data is used in the model construction process. More specifically, semi-supervised learning is extended by considering the relationships between class labels.

2.3 Literature Review for Ordinal Classification

Ordinal classification (OC) considers the problems where the class labels of the target attribute in the dataset follow a given order, such as very hot, hot, warm, cold, and very cold labels in a weather prediction problem. Recently, the ordinal classification has received much attention in the machine learning field with the development of a growing number of real-world applications, such as customer segmentation (i.e., gold, silver, and bronze), sentiment analysis (i.e., happy, natural, and sad), credit scoring (i.e., high, medium, and low risk levels), human age estimation (i.e., old, adult, and young), medical diagnosis (i.e., initiation, proliferation, and progression stages of cancer), and survey analysis (i.e., disagree, neutral, and agree). In these contexts, the ability to capture the natural order of the labels is crucial to improving the classification performance of the model. The importance of taking into account inter-classes relations has already been proven in (Frank et al., 2001).

The Ordinal Classification approaches can be grouped under three categories: threshold approaches, naive approaches, and ordinal binary decomposition approaches (Gutiérrez et al., 2015). The *threshold approaches* (Perez-Ortiź et al., 2013) obtain a set of thresholds by dividing the target variables into successive intervals, where each class label belongs to an interval limited by these thresholds.

The *naive approaches* (Sánchez-Monedero et al., 2013) use an appropriate simplifying assumption on the class labels to treat OC problems as if they were standard classification problems. For instance, one possible solution is to use different weights for different class labels, or another alternative solution is to map the class labels into numeric values and then implement a standard regression algorithm such as the support vector regression. The *ordinal binary decomposition (OBD) approaches* (Frank et al., 2001) transform the original ordinal classification mission into a set of binary classification missions. Each sub-task is separately solved by a binary classification algorithm, and then the final class labels are determined by ultimately combining the binary outputs into one label. In this study, we propose a new method based on the OBD approach.

The supervised ordinal classification approaches aforementioned above present limitations when only a small number of labeled ordinal data is available. The reason is that data instances are labeled by a user is important problem of ordinal classification, and this process could be difficult, expensive, or time-consuming, especially when the number of class labels is high (i.e., more than three). To overcome this limitation, this article proposes a new paradigm for ordinal classification, where both labeled ordinal data and unlabeled data are utilized during the model construction process.

Semi-supervised ordinal regression (SSOR) has a very rare field in the literature, only a few detailed analyses (Pérez-Ortiz et al., 2016; Srijith et al., 2013) have been made. The comparison of our SSOC study against existing SSOR studies is shown in Table 2.1. In three respects, our method differs from other existing methods. First, they performed the prediction of numeric target values, while we focus on the classification of categorical target values. Second, they used different methods such as kernel discriminant learning (Pérez-Ortiz et al., 2016), empirical risk minimization (Tsuchiya et al., 2021), max-coupled learning (Cardoso & Domingues, 2011), and Gaussian processes (Srijith et al., 2013); whereas we utilized the ordinal binary decomposition method. Third, since their target values are numeric they used different evaluation metrics such as mean absolute error (MAE), mean squared error

(MSE), and mean zero-one error (MZE), whereas we tested the performance with accuracy and F-score metrics. To the best of our knowledge, a semi-supervised ordinal classification that will consider the ordinal categorical class labels has not been studied until now. To bridge this gap, in this study, a new algorithm for semi-supervised learning is developed in the context of ordinal classification.

Cardoso and Domingues (2011) proposed a new learning paradigm, called max-coupled learning, and three different methodologies for breast cancer application. Our study differs from their study in many aspects. First, they focused on the prediction of numeric target values (breast imaging reporting and data system - BIRADS scores such as 1,2,3), whereas we consider the categorical class labels where the metrical distances between adjacent categories are commonly unknown. For instance, the question of whether the difference between classes c_1 and c_2 is the same as between classes c_2 and c_3 is unknown. We propose a completely different approach that enables unlabeled data via the smoothness assumption (i.e., close samples are likely to have similar categorical class labels). Second, their method was proposed for handling data with two views (x and z), however, no restrictions are encountered when applying our method to ordinal data. Third, they presented their solution for a specific problem (breast cancer) by generating synthetic data, whereas we demonstrated the generalization ability of our method on real-world datasets obtained from various domains. Fourth, they used a different method (max-coupling) which finds the maximum of the values obtained from the two "views" in prediction, while we utilized the ordinal binary decomposition method by considering upward and downward unions of classes. Fifth, to evaluate the performance, they used the MAE metric since their target values are numeric like a regression problem, whereas we assess the percentage of errors since the class labels are categorical.

Table 2.1 Comparison of our study with the existing studies.

Ref	Methods	Algorithm	Target class attribute		Evaluation metrics	Application domain
			Numeric	Categorical		
Pérez-Ortiz et al., 2016	Kernel discriminant learning (KDL)	Reduced empirical feature space semi-supervised KDL for ordinal regres- sion (ES-DL)	✓		MAE STD	Various domains
Tsuchiya et al., 2021	Empirical risk minimization	Semi-supervised ordinal re- gression with Gaussian kernel (SEMI-Kernel)	✓		MAE MSE, MZE	Various domains
Cardoso & Domingues, 2011	Max-coupled learning	Semi-supervised max-coupling algorithm	✓		MAE	Health
Srijith et al., 2013	Gaussian pro- cesses	Semi-supervised Gaussian process ordinal regression (SSGPOR)	✓		MAE, MZE	Various domains
Proposed approach	Ordinal binary decomposition	DT, SVM, KNN, RF, and NN		✓	Accuracy F-Score	Various domains

CHAPTER THREE
BACKGROUND INFORMATION

3.1 Data Mining

The increase in digitalization enables us to have plenty of data in many different fields. Having too much data is getting worth it if we know how to use it. Data mining aims to discover knowledge from large-scale datasets using advanced algorithms. With data mining, it becomes possible to extract the relationships within the historical data and make accurate predictions for the future.

One of the application areas of data mining is education. Data mining in education allows us to make predictions by analyzing the data obtained so far in the field of education by using advanced algorithms. There are basically three data mining methods: classification, clustering, and association rule mining. In this thesis, we focused on the classification task due to its popularity.

3.2 Data Mining Methods

The data mining methods may differ depending on the data nature we have and the field of study. The main data mining methods used for classification are described below.

Naive Bayes. Naive Bayes classifiers use events that occur previously to find a new event possible, based on Bayes' Theorem. The algorithm has a common principle but according to the theorem each feature of data is independent.

Random Forest. Random forest is a typical ensemble learning method. The algorithm contains many decision trees generated randomly. The random forest algorithm aims to get stable predictions and high accuracy rates by using multiple decision trees combined. The decision trees group trained with the bagging method and constituted the forest form.

Decision Tree. Decision trees consist of branches and nodes which are lined up in a row. The trees are similar to flowcharts but not non-cyclic. There is a root node that symbolizes the entire dataset at the top of the tree. Entropy and information gain are calculated when building trees.

Neural Network. A neural network tries to recognize fundamental relationships by mimicking the way the human brain works. Neural networks can adapt to changing input without needing to redesign the system.

3.3 Educational Data Mining

Educational data mining is the subfield of data mining that analyzes large-scale educational data to extract meaningful and useful information within the education context. With the increase in education data, the development of specific data mining methods in this field has also increased. Measuring and evaluating the effect of a newly released algorithm on the results is an important research topic in this field.

The data mining model for classification shown in Figure 3.1 as a workflow. The first step is related to applying feature selection algorithms to educational data. In the following step, interested in creating a good model that finds desired outputs by mapping inputs. The following step is the model evaluation phase intends to improve the performance of classification by supplying feedback to first step feature selection and data mining step. After the model creation process is completed, the prediction phase is started and many predictions can be produced with the same model.

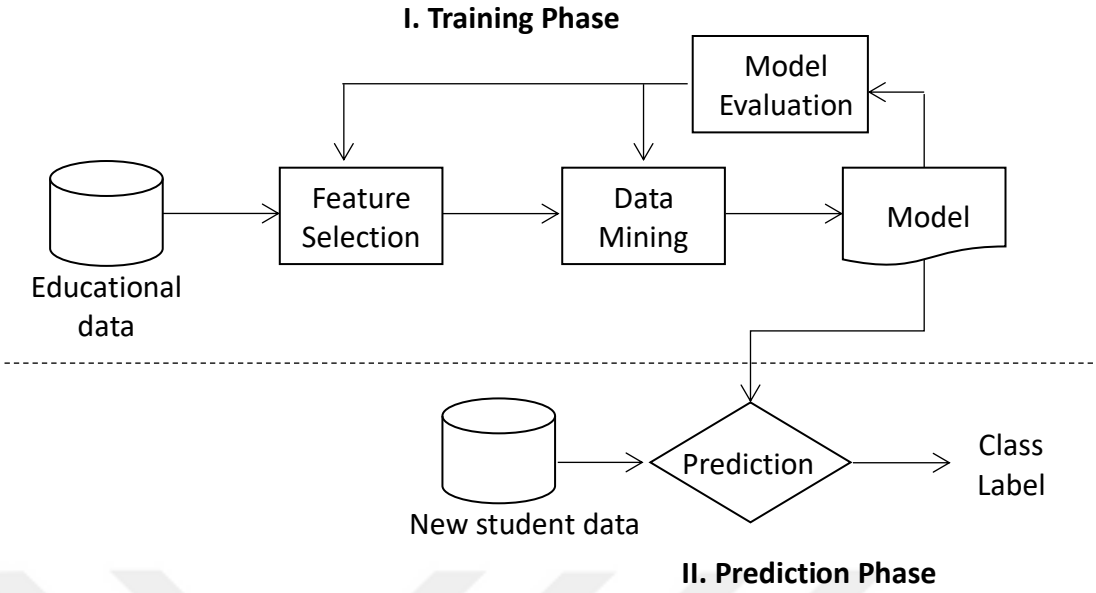


Figure 3.1 Educational data mining prediction process

CHAPTER FOUR
MATERIALS AND METHODS

4.1 Semi-Supervised Learning

Semi-Supervised Learning (SSL) is an important and useful type of machine learning concerned with using both labeled and unlabeled data to perform a particular learning task. It is conceptually situated between unsupervised learning and supervised learning and implies a more complex problem than both of them. The semi-supervised learning topic has received much attention in many areas ranging from health (Lang et al., 2020) to education (Livieris et al., 2019), where it is cheaper and easier to obtain unlabeled data rather than labeled data since it requires less expertise, effort, and time consumption.

Semi-Supervised Classification (SSC) is a classification type between supervised learning and unsupervised learning. SSC is a method that can be used in problematic situations where a small amount of labeled data and a large amount of unlabeled data are available. Many approaches have been developed for semi-supervised learning. Among them, the self-training approach is used in this study due to its popularity and efficiency.

The existing SSL algorithms are capable of processing nominal data; however, they do not capture and reflect the orderings among the class labels. Hence, they may lead to building inefficient models in the case of ordinal data. To address this limitation of the existing SSL algorithms, in this article, we propose a new SSL algorithm to exploit the presence of the orders among class labels in ordinal data. Our study aims to investigate how combining labeled and unlabeled ordinal data may improve the classification performance, and develop an algorithm that takes advantage of such a combination.

4.2 Ordinal Classification

Ordinal classification (OC) is a special kind of supervised multi-class classification task that addresses the class attributes whose labels exhibit a form of ordering. For example, an ordinal class attribute can represent four wind-level categories in a wind speed prediction problem: very high, high, moderate, and low. There exists an order among the class labels such that very high > high > moderate > low, where > represents that the former class label is better than the latter class label.

The standard classification algorithms ignore the orders among class labels, however, in many cases the class values display a natural order. Ordinal classification is a supervised learning technique used for situations where the class labels are ordered in a categorical format. For example, five ordinal values related to object physical characteristics such as “extra small” < “small” < “medium” < “large” < “extra-large” can be considered.

The conventional classification algorithms disregard the ordering information in class labels; however, it usually leads to a significant loss of performance. In an ordinal classification problem, a sample from the lowest class is remarkably different from a sample from the highest class, whereas two samples, from the lowest and middle classes, are relatively close to each other. Therefore, taking into account such ordering information aims to improve the performances of classifiers.

The limitation of classical ordinal classification is the use of labeled ordinal data alone to build a classifier. However, in this study, we propose a novel learning paradigm dealing with the design of classifiers in the presence of both labeled ordinal data and unlabeled ordinal data.

4.3 The Proposed Approach: Semi-Supervised Ordinal Classification (SSOC)

In this study, we propose a new approach “*semi-supervised ordinal classification*” (SSOC), which considers the hierarchical relationship among the target variables

when using both labeled and unlabeled ordinal data for classification. In other words, we present a new classification strategy for incorporating ordinal data information into semi-supervised learning.

There are many cases where unlabeled data in addition to labeled ordinal data can help in building or improving a classifier. Consider, for instance, the case of movie classification, where we want to assign a rating to a collection of movies such as very good, good, average, poor, and terrible ratings. However, most users might not have an interest in rating movies, thus labeled ordinal data is scarce or difficult to obtain; on the other hand, a huge amount of unlabeled data is available. This is where semi-supervised ordinal classification comes in. The proposed SSOC method is able to use unlabeled data along with the labeled ordinal data to construct more efficient models in terms of accuracy compared to the standard ordinal classification methods.

A simple solution for the SSOC can be converting class labels into real values, e.g., {1,2,3,4,5}, and then solving the problem as a standard regression problem. However, this solution may lead to building unreliable models, since the metrical distances between adjacent categories are commonly unknown. For instance, the question is whether the difference between saying “strongly agree” and “agree” is the same as between saying “neutral” and “disagree”. The answer is that it is not possible to ensure this assumption. Therefore, in this paper, we propose a completely different approach. We extended the ordinal classification algorithm presented in (Frank et al., 2001) to enable unlabeled data via the smoothness assumption (i.e., close samples are likely to have similar categorical class labels).

The proposed SSOC method is based on the self-training principle (also known as self-learning or self-labeling). Firstly, an ordinal classifier is trained in a standard way with an initially small number of labeled samples on an ordinal scale. After that, the constructed ordinal classifier labels the unlabeled data by its predictions. Lastly, a purely ordinal classification algorithm builds the final inductive classifier by using both pseudo-labeled and original labeled ordinal data together.

4.3.1 The Formal Definition of the Proposed Method

Formally let X be an input space with d-dimensional features $X \subseteq \mathbb{R}^d$, and, $Y = \{c_1 , c_2 , ..., c_k \}$ be a finite set of predetermined ordered classes such that $c_1 < c_2 < ... < c_k$. We denote two sample sets X_L and X_U as the collection of input instances for the ordinal labeled and unlabeled samples, respectively, where $X_L , X_U \in X$ and $X = X_L \cup X_U$. Given a set of l labeled instances $\{(x_i, y_i)\}_{i=1}^{l}$ and u unlabeled instances $\{x_j\}_{(l+1)}^{(l+u)}$ we denote labeled data as $D_L = \{(x_1, y_1), (x_2, y_2), , (x_l, y_l)\}$ and unlabeled data as $D_U = \{x_{l+1}, x_{l+2}, , x_{l+u}\}$. Each data point (x_i, y_i) in the labeled data D_L consists of an instance x_i from a given input space X, such that $x_i \in X$, and has an associated label y_i, where y_i is one of the ordinal categories such that $y_i \in Y$. The two datasets D_L and D_U form the training set D together with n instances, where $n = l + u$. We are especially interested in cases where $u >> l$ since an unlabeled instance is abundant and easy to be obtained while labeling the instance is expensive and it usually requires expert knowledge.

Definition 1. Semi-supervised ordinal classification (SSOC) refers to the problem that utilizes both ordinal labeled data D_L and unlabeled data D_U and aims to find a decision function $f : X_{L+U} \rightarrow Y$ that can predict y^* class labels correctly of some unseen inputs x^*.

The proposed SSOC method consists of two main steps. In the first step, the SSOC method applies an ordinal classification algorithm to the labeled ordinal instances in X_L to build an ordinal classifier and then produces pseudo-labels $\hat{y} = \{y_{l+1} , y_{l+2} , , y_{l+u}\}$ for the unlabeled data instances in X_U by using the predictions of the resulting classifier. In the second step, the ordinal classifier is then retrained on the newly obtained pseudo-labeled ordinal data in addition to the originally labeled ordinal data.

Definition 2. Ordinal decomposition assume that the output space is defined as $y = \{c_1, c_2 ,, c_k\}$ and the labels are ordered according to the ranking structure $c_k > ... > c_2 > c_1$, where > denote this order information. Ordinal decomposition is to

decompose the original ordinal classification problem involving *k* classes into *k-1* binary classification problems to encode the ordering information among the class labels. The *i*-th binary problem is defined by separating the classes from 1 to *i*, denoted by $y_- = \{c_1, c_2 ,, c_i\}$, and the classes from $i + 1$ to k , denoted by $y_+ = \{c_{i+1}, c_{i+2} ,, c_k\}$.

Rather than the standard class probability $P(label = c_i)$, the *i*-th problem estimates the binary composition probability $P(label > c_i)$, which is the probability of a sample having a class greater than c_i in the ordinal scale. In this way, it takes into account the order among the classes since the upward and downward unions of classes are progressively considered during the creation of the binary datasets. The final output is then predicted by multiple models (*k*-1 models), which are trained in conjunction with a base learner, i.e., using a decision tree method.

Definition 3. Binary composition probability let M_i for *i*=1, 2, , *k-1* , denotes the model constructed for the ordinal classification problem. Given a search instance *x*, an estimation $M_i(x)$ is considered as a prediction of the probability $P(L_x > c_i)$ that the class label of *x* is interpreted in Y_+, where L_x denotes the class label of *x*. Binary composition probabilities on *Y* are formulated as follows:

$$\begin{gathered} P(c_i) = 1\text{-}P(L_x > c_i) \\ P(c_i) = P(L_x > c_{i-1}) x (1\text{-}P(L_x > c_i)) \quad 1 > i > k \\ P(c_k) = P(L_x > c_{k-1}) \end{gathered} \tag{4.1}$$

To assign a class label (*L*) to a new sample *x*, we need to calculate the probabilities of the k ordinal classes using *k-1* binary model. The prediction of the probabilities for the first and last classes depends on a single model. The probability of an example being of the first class $P(c_1)$ is given by $1 - P(L_x > c_1)$ since we will only consider the first model that discriminates between c_1 and the rest. Similarly, the probability of the last class $P(c_k)$ is computed from $P(L_x > c_{k-1})$ in accordance with the one-vs-followers strategy. For the intermediate classes $1 < i < k$, the probability depends on a pair of models since the ordinal structure of a middle-class

c_i can be reflected by two indicators: greater than its previous class $P(L_x > c_{i-1})$ and not greater than itself $(1 - P(L_x > c_i))$. In this way, the method considers the order in the frontiers. For instance, when the model is queried with a sample *x* for the class c_2, a vote for the higher classes c_3 and a vote for the lower classes c_1 should be considered. Therefore, the upward and downward unions of classes are progressively considered. In other words, adjacent classes are grouped to encode the ordering among classes. For example, assume that there are five ordered classes: very sad, sad, natural, happy, and very happy; then the probability of each class is calculated as follows:

Ordinal decomposition	Binary composition probability
{verysad}{sad, natural,happy,veryhappy}	P(verysad)=1-P(label>verysad)
{verysad}{sad}{natural,happy,veryhappy}	P(sad)=P(label>verysad)×(1-P(label>sad))
{verysad,sad}{natural}{happy,veryhappy}	P(natural)=P(label>sad)×(1-P(label>natural))
{verysad,sad,natural}{happy}{veryhappy}	P(happy)=P(label>natural)×(1-P(label>happy))
{verysad,sad,natural,happy}{veryhappy}	P(veryhappy)= P(label>happy)

At the end of the binary decomposition, the label that has the highest probability is predicted as the final class for a given instance. This probability estimation is meaningful for ordinal classification since it typically assigns higher scores to the instances from higher classes. As an example, an instance *x* of class c_i will be in a positive class (Y_+) of all *i*+1 binary classifiers, and therefore the model M_i should return a high score. Thereby, the model M_i yields a reasonable probability estimation that is consistent in the sense of $M_i(x) \geq M_{i+1}(x)$ Essentially, the mapping $i \rightarrow M_i(x)$ is similar to a decumulative distribution function.

Coming back to the problem of semi-supervised ordinal classification, an ordinal classifier model is constructed by using labeled data until now. In the next step, the pseudo-labeling step, the resulting ordinal classifier is used to predict labels for the previously unlabeled instances. Since the SSOC method takes into account the relationships among the class labels, an unlabeled sample will be assigned to a pseudo-label by the maximum ranking probability to which it is associated. The

pseudo-labeled data is appended to the initial labeled data. Finally, the SSOC method retrains the underlying ordinal classifier with this augmented data.

4.3.2 The Algorithm of the Proposed Method

Figure 4.1 shows the algorithm of the SSOC proposed method. The algorithm can be summarized in 5 steps.

Step 1. In the first step (lines 9-18), called ordinal binary decomposition, the method converts the original labeled dataset D_L into a set of binary datasets $\{D_i\}_{i=1}^{k}$ to encode the ranking relation among the classes. In this process, the label y_i associated with the instance x_j is replaced with $y_j = 0$, $\forall\, y_j \leq c_i$, and, $y_j = 1$, $y_j > c_i$. In other words, if we consider the class c_i, the class values higher than c_i are labeled as 1, and the others are labeled as 0. In this way, the algorithm transforms the ordinal classification problem involving *k* classes into *k-1* binary classification problems. For every $i \in \{1, 2, ..., k\text{-}1\}$, the dataset D_i is generated using classes $\{c_1, c_2, ..., c_i\}$ against $\{c_{i+1}, ..., c_k\}$.

Step 2. In the second step (lines 19–23), a separate model M_i is trained for each binary dataset D_i by using a base learner. The first model M_1 is built to estimate what is the probability of a given sample belonging to any of the classes that are located higher than class c_1 (class $> c_1$), the second model M_2 is constructed to estimate the probability of belonging to any of the classes that are located higher than c_2 (class $> c_2$), and so on. Finally, the last model M_k is trained to estimate the probability of belonging of a sample to the highest class c_k (class $> c_k$). In the training phase, a standard classification algorithm (i.e., DT, SVM, and NN) can be used as a base learner.

Step 3. In the third step (lines 24-33), a query instance $x_i \in D_u$ is submitted to all models M^*, and their estimations ($M^*(x_i)$) are compared, and the class label with the highest probability is assigned to x_i as a pseudo-label. For each instance x_i in the unlabeled dataset D_u, this process is repeated In the for-each-loop, the probability of

belonging the query instance x_i to each class is estimated by using Equation (4.1). While the probability of the first-class $P(c_1)$ is computed (line 27), the upward union of classes $P(L_x > c_1)$ is subtracted from 1. While the probability of the last class $P(c_k)$ is computed (line 31), the upward union of classes P ($L_x > c_{k-1}$) is directly considered. On the other hand, the probabilities for the intermediate classes $2 \leq j \leq k-1$ are computed in the for-loop (lines 28-30) by considering both upward and downward unions of classes. In the end, the label that has the highest (MAX) probability is predicted as the final class (y) for the query instance x_i.

Algorithm - SSOC: Semi-Supervised Ordinal Classification.

1. **Inputs:**
2. D_L: The labeled ordinal dataset $D = (x_1, y_1), (x_2, y_2), ., (x_l, y_l)$ with l intances
3. D_U: the unlabeled dataset $D = x_{l+1}, x_{l+2}, ., x_{l+u}$ with u intances
4. Y: ordinal class labels $y \in \{c_1, c_2, ., c_k\}$ with an order $c_k \succ ... \succ c_2 \succ c_1$
5. k: the number of classes
6. **Output:**
7. M^* : semi-supervised ordinal classification model
8. **Begin:**
9. // Step 1 - Construction of binary datasets from the labeled data, D_L
10. **for** $i \leftarrow 1$ **to** k-1 **do**
11. **foreach** (x_j, y_j) **in** D_L **do**
12. **if** $(y_j \geq c_i)$ **then**
13. D_i.Add$(x_j, 0)$ // The class values smaller than or eqaul to c_i are labeled as 0
14. **else**
15. D_i.Add$(x_j, 1)$ // The class values higher than c_i are labeled as 1
16. **end**
17. **end**
18. **end**
19. // Step 2 - Construction of unified binary models
20. **for** $i \leftarrow 1$ **to** k-1 **do**
21. $M_i = Train(D_i)$ // Building a classifier on the training set using a base learner
22. $M^* = M^* \cup M_i$
23. **end**
24. // Step 3 - Pseudo-labeling unlabeled data, D_U
25. **foreach** x_i **in** D_U **do**
26. $y = M^*(x_i)$ =MAX (// The label with max probability is predicted as the final class
27. $P(c_1) = 1 - P(L_i > c_1)$ // The probability of belonging to the first class
28. **for** $j \leftarrow 2$ **to** $k - 1$ **do**
29. $P(c_j) = P(L_i > c_{j-1})) \times (1 - P(L_i > c_j))$//The probability of belonging to a middle class
30. **end**
31. $P(c_k) = P(L_i > c_{k-1}))$ // The probability of belonging to the last class
32. D_L.Add(x_i, y)
33. **end**
34. // Step 4 - Construction of binary datasets from the labeled and pseudo-labeled data
35. Repeat step 1
36. // Step 5 - Construction of semi-supervised ordinal classification model
37. Repeat step 2
38. **End Algorithm**

Figure 4.1 Algorithm of the proposed SSOC method

Step 4. In the fourth step (lines 34-35), the binary datasets are generated from the combined (labeled and pseudo-labeled) data, similar to the first step. In other words, the algorithm repeats Step 1 by simply using both the pseudo-labeled and original labeled ordinal data as if it was regular labeled ordinal data.

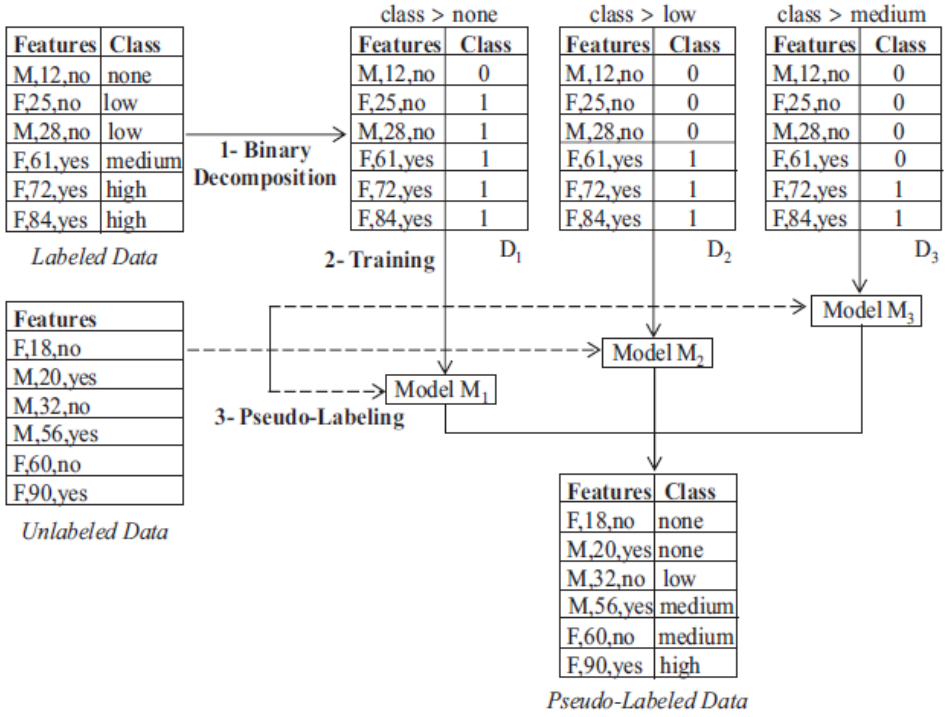
Step 5. In the fifth step (lines 36-37), the ordinal classifier is re-trained by using the newly obtained binary datasets, similar to the second step. In other words, the ordinal classifier is re-trained with its estimationsby enlarging its initial labeled set.

The SSOC method builds *k-1* model on the training set, where *k* is the number of classes. When we use a base learner with complexity *T*, the overall time complexity of the method for training is given by $O((k-1).T(l) + u + T(l+u))$, where *l* and *u* are the numbers of labeled and unlabeled ordinal instances, respectively.

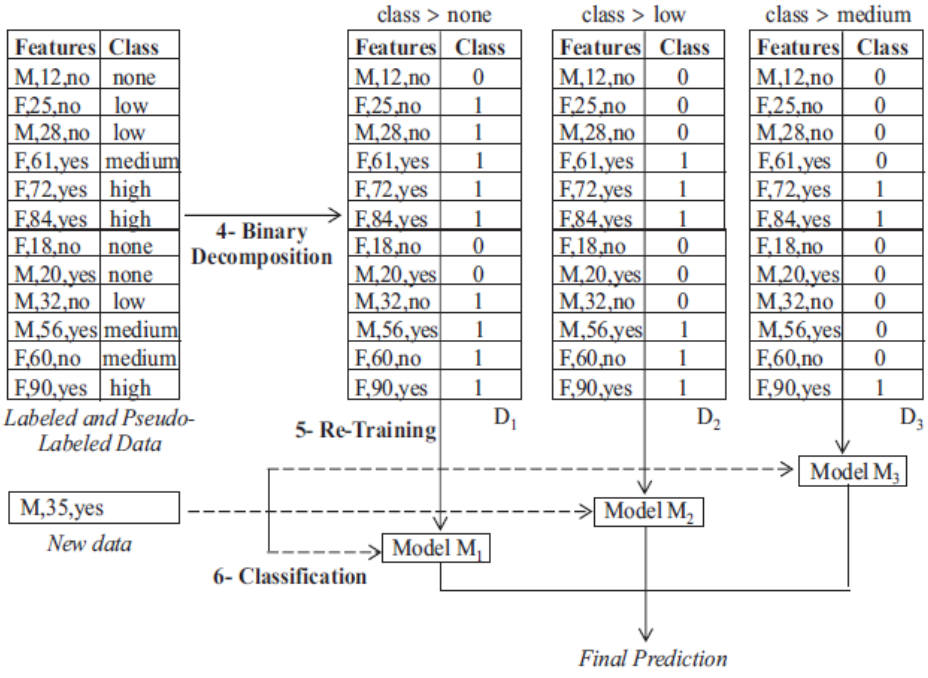
4.3.3 An Example of the Proposed Method

Figure 4.2 shows a concrete example of the proposed approach: semi-supervised ordinal classification (SSOC). Assume that each sample in the given dataset has three features (gender, age, and status) and is classified into one of the predefined four categories on an ordinal scale (none < low < medium < high). The SSOC method consists of six consecutive steps grouped under two main stages. The first stage illustrated in Figure 4.2 (a) is to train ordinal classifiers on the existing labeled ordinal data, and after that, use the predictions of these classifiers on the previously unlabeled data to generate additional labeled data indicate as pseudo-labeled data. This process provides to introduce the unlabeled data to the training process efficiently and straightforwardly. In the binary decomposition step, the ordinal classification problem involving *k* classes ($k = 4$ in this example) is converted into ($k - 1$) binary classification problems to encode the ordering among the class labels such that *P(class > none)*, *P(class > low)*, and *P(class > medium)*. In the second main stage illustrated in Figure 4.2 (b), the ordinal classifiers are retrained on the newly obtained pseudo-labeled data in addition to the originally labeled data. The

method simply passes both the originally labeled and pseudo-labeled ordinal data to a supervised base classifier as if they were regular labeled ordinal instances.



(a)



(b)

Figure 4.2 An example of the proposed SSOC method

4.3.4 The Proposed Educational Data Mining Process

Figure 4.3 shows the general steps of the educational data mining process proposed in this study. In the first step, education data is collected via an information system. It contains some demographic, behavioral, and performance features about students. In the second step, called data preprocessing, a discretization method is applied to make raw education data ordinal. In the third step, both labeled and unlabeled ordinal data are used to train a learner for a semi-supervised learning task. Finally, the model is evaluated by using the *k*-fold cross-validation technique.

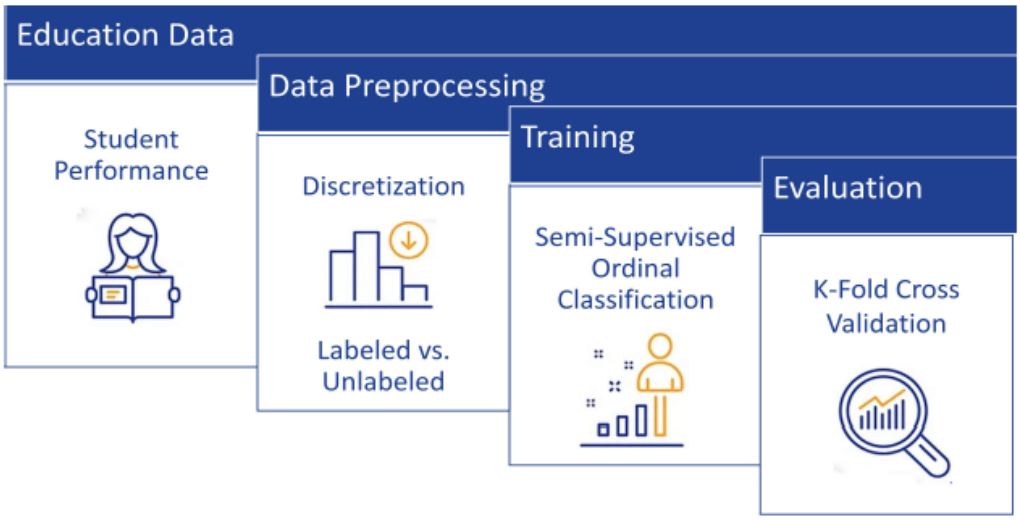


Figure 4.3 The proposed educational data mining process

CHAPTER FIVE
EXPERIMENTAL STUDIES AND RESULTS

5.1 Dataset Description

To demonstrate the validity and effectiveness of the proposed SSOC method, the experiments were carried out on 10 ordinal datasets available in the UCI (the University of California at Irvine) and OpenML repositories. Table 5.1 gives the names and main characteristics of the datasets, including the number of features, classes, instances, the number of instances per class, and the source repository.

Table 5.1 Summary description of the datasets

ID	Dataset Name	#Features	#Instances	#Classes	Class Distribution	Source
1	Automobile	26	205	7	(0, 3, 22, 67, 54, 32, 27)	UCI
2	Car Evaluation	7	1728	4	(384, 69, 1210, 65)	UCI
3	Employee Selection (ESL)	5	488	9	(2, 12, 38, 100, 116, 135, 62, 19, 4)	OpenML
4	Eucalyptus	20	736	5	(180, 107, 130, 214, 105)	OpenML
5	Nursery	9	12960	5	(4320, 2, 328, 4266, 4044)	UCI
6	Squash-Stored	25	52	3	(8, 21, 23)	OpenML
7	User knowledge model. (UKM)	6	403	5	(50, 129, 122, 102)	UCI
8	Volcanoes on venus (B6)	4	10130	5	(88, 86, 58, 68, 2952)	UCI
9	Wine quality-red	12	1599	6	(10, 53, 681, 638, 199, 18)	UCI
10	Wine quality-white	12	4898	7	(20, 163, 1457, 2198, 880, 175, 5)	UCI

Each dataset is divided into two parts: labeled and unlabeled sets because the main purpose of the SSOC method is to utilize unlabeled ordinal data to improve learning performance. We used a random selection of instances that were marked as labeled samples, and the class labels of the rest of the samples were removed. To investigate the effect of the amount of labeled ordinal data, we considered different percentage values when dividing the dataset, varying from 15% to 50% with an

increment of 5%. For example, assuming a dataset that contains 1000 instances, when the labeled data rate is 10%, 100 samples were put into the labeled dataset D_L with their class labels, while the remaining 900 samples were placed into the unlabeled dataset D_U without their class labels.

5.2 Experimental Studies

We conducted five experiments on 10 ordinal datasets for illustrating the efficiency and validity of the proposed SSOC algorithm.

Experiment 1 - We compared the SSOC method with the existing semi-supervised learning algorithm (YATSI) presented in (Driessens et al., 2006) to show the superiority of our algorithm on ordinal data.

Experiment 2 - We investigated the performance of the SSOC method with different ratios of labeled data (from 15% to 50%) to determine its impact on the method.

Experiment 3 - We evaluated the performance of the SSOC method with different base classifier combinations (DT, SVM, KNN, RF, and NN).

Experiment 4 - We compared the SSOC method with the standard ordinal classification algorithm presented in Frank et al., (2001) to show the effectiveness of our method.

Experiment 5 - We investigated that after retraining the classifier with new pseudo-labels, how much the performance is increased compared to the performance with initial training labels.

In these experiments, the performances of the classification models were evaluated according to the accuracy metric, which is the percentage of correctly predicted samples. Accuracy is calculated by the formula: *Accuracy=*

(TP+TN)/(TP+TN+FP+FN), where true positive (*TP*) and true negative (*TN*) indicate the correct predictions for the positive samples and negative samples, respectively; whereas false-positive (*FP*) and false-negative (*FN*) represent the misclassified positive and negative samples, respectively. The SSOC and YATSI methods were compared by using the original dataset for testing. Ordinal classification accuracies were obtained by using the 10-fold cross-validation technique, in which the data is randomly divided into ten disjoint and equal-sized partitions, and one of the partitions is kept for the testing process, while the remaining partitions are utilized for the training process.

We implemented the proposed method in the Java programming language by using the WEKA machine learning library (Witten et al., 2005). We also compared our method with the existing methods (Frank et al., 2001; Driessens et al., 2006) which are available as packages for the WEKA tool. The comparison results were evaluated by using Wilcoxon statistical test to ensure the significance of the performance results obtained by the methods on the ordinal datasets.

In each experiment, all the input parameters of the algorithms were left as default values, except the KNN algorithm. The number of neighbors (the parameter *k*) was selected as $log_2(n)$ based on the previous study (Jiang et al., 2014), where *n* is the number of objects in the respective dataset. The default value of this parameter is 1; however, it is too small and it generally does not make sense to choose the parameter *k* so small when a large number of instances are available in the dataset.

5.3 Experimental Results

In this study, the proposed SSOC approach was tested with different base classifiers (i.e., DT, SVM, KNN, RF, and NN) on the benchmark datasets. In other words, SSOC with each different base classifier is built as an independent classifier. From here onwards, the abbreviation of the method followed by the abbreviation of the base classifier technique is used to refer to the related approach. For example, SSOC-SVM refers to the SSOC method with the SVM base classifier.

5.3.1 The Results of Experiment 1

In this experiment, we compared the proposed method (SSOC) with the existing method (YATSI) (Driessens et al., 2006) on various ordinal datasets in terms of classification accuracy (%). YATSI (yet another two-stage idea) is a collective classifier for semi-supervised learning and it can be used in the WEKA machine learning tool (Witten et al., 2005) with an additional Collective Classification package. Table 5.2 shows the comparison results. The selection of initial labeled data was repeated five times for the proposed method and the average results were reported here. It can be observed from Table 5.2 that the proposed SSOC method resulted in performance improvement over the existing YATSI method on 10 benchmark ordinal datasets. For example; for the "UKM" dataset, the SSOC-DT algorithm (90.07%) is significantly better than the YATSI-DT algorithm (83.13%). The predictive performance difference between SSOC and YATSI is considerable for almost all datasets, for different percentages of labeled data (i.e., 25%, 50%, and 75%), and for all machine learning algorithms (DT, SVM, KNN, RF, and NN). Hence, the results indicated that semi-supervised based ordinal classification could generally produce a robust model with high prediction accuracy.

Although the SSOC method usually outperforms YATSI, in some cases the performance of the YATSI is better than SSOC, especially for the Car Evaluation and Nursery datasets. This is probably true when the data is noisy or some classes are similar to each other. The reason is probably that YATSI is a collective classification algorithm that uses an off-the-shelf classifier in the first step and benefits from the weighted nearest neighbor approach. The main idea behind collective classification is that the predicted class label of a test sample is influenced by the predictions made by related samples. However, if the user does not determine the optimal number of neighbors as well as the optimal weights, YATSI may lose this advantage, especially when the data is noisy.

Table 5.2 Proposed method (SSOC) and the existing method (YATSI) comparison in terms of accuracy (%)

Dataset No	Labeled data (%)	SSOC					YATSI				
		DT	SVM	KNN	RF	NN	DT	SVM	KNN	RF	NN
1	25	56.78	63.02	48.39	**71.71**	66.15	48.78	46.34	50.24	47.32	47.80
	50	73.56	74.54	58.83	**87.80**	81.95	57.07	57.56	50.73	56.10	55.12
	75	76.88	80.20	67.41	**96.10**	91.12	60.49	64.39	59.02	61.95	63.41
2	25	81.42	82.13	75.21	89.28	**95.30**	83.39	89.12	76.33	84.38	91.78
	50	86.79	80.31	82.35	96.06	**98.62**	91.32	92.94	82.99	92.71	94.91
	75	91.63	81.31	91.86	98.34	**99.76**	91.90	92.01	88.54	92.82	93.23
3	25	62.83	58.98	59.92	68.93	**71.15**	60.25	51.23	60.45	61.68	63.11
	50	68.57	63.81	66.68	**75.33**	74.18	67.21	58.40	65.16	68.44	68.03
	75	72.75	65.53	69.47	**79.67**	76.43	64.96	59.22	65.16	64.55	65.16
4	25	61.90	65.76	47.93	67.42	**68.13**	52.04	53.53	48.23	51.63	50.82
	50	66.85	70.22	57.04	**81.20**	79.08	54.35	56.39	54.21	56.25	55.98
	75	72.80	72.55	61.82	**90.35**	85.90	50.00	54.08	55.03	55.43	54.76
5	25	92.78	91.88	91.87	96.07	**99.58**	93.74	94.05	92.35	95.00	96.25
	50	96.30	92.64	95.84	98.74	**99.98**	96.99	95.44	95.81	97.38	98.74
	75	97.43	93.09	97.26	99.60	**99.98**	96.63	94.93	96.03	97.38	97.57
6	25	64.62	65.00	58.08	**70.00**	67.69	38.46	40.38	42.31	42.31	40.38
	50	69.23	77.31	62.31	81.54	**82.69**	57.69	48.08	53.85	55.77	51.92
	75	78.46	87.31	68.46	**91.15**	90.38	57.69	51.92	44.23	51.92	50.00
7	25	90.07	81.84	72.95	90.82	**95.53**	83.13	77.42	72.95	83.13	83.37
	50	95.53	83.13	84.42	**96.92**	96.87	85.36	83.37	84.37	84.86	86.10
	75	96.97	91.41	87.39	**98.86**	98.21	84.62	82.63	82.88	84.12	84.12
8	25	96.20	96.21	96.23	**97.11**	96.48	96.47	96.20	96.23	96.41	96.42
	50	96.25	96.21	96.42	**98.17**	96.62	96.54	96.20	96.43	96.54	96.58
	75	96.30	96.21	96.53	**98.89**	96.67	96.52	96.21	96.51	96.54	96.60
9	25	61.79	57.66	57.36	**69.79**	60.38	57.22	57.16	57.47	58.04	57.29
	50	68.09	58.09	61.71	**81.78**	63.63	58.16	58.16	59.16	58.85	57.72
	75	77.24	58.41	63.09	**91.56**	67.03	57.79	57.35	60.54	59.47	60.23
10	25	56.57	45.87	53.44	**67.92**	53.84	52.10	46.96	53.63	54.23	51.12
	50	65.66	51.67	57.76	**81.76**	56.61	55.59	53.74	56.57	56.94	55.96
	75	73.44	50.32	59.58	**91.19**	57.92	55.45	53.12	56.98	56.17	54.02

Although accuracy is a widely used evaluation metric, sometimes it is not enough alone due to its strong bias against the rare (minority) class, especially when the data is imbalanced. Since even considering all examples as the majority class, high accuracy can be achieved. Since some of the datasets used in this study is imbalanced, we applied the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) and we used the F-Score measure to evaluate the results. Options for setting hyperparameters were left as default values from the WEKA software package, except for the percentage of SMOTE instances to be created for the minority class. It was set according to the ratio of instances between the majority and minority classes in the particular dataset. Table 5.3 shows the comparison results in terms of F-Score which is the harmonic mean of precision and recall. The selection of initial labeled data was repeated five times and the average results were reported here. The value of the F-Score metric ranges between 0 and 1, where 1 is the best value. As seen in Table 5.3; the F-score values obtained by the SSOC-RF algorithm are close to 1, especially when 75% of the data was considered labeled. The F-score values slightly increased as the number of labeled data samples increased. As a result, it can be concluded that it is possible to achieve high F-score values by the proposed method for the ordinal data.

Even though the proposed SSOC method has higher accuracy than the YATSI method on average, the results were evaluated by using a formal statistical test to ensure the differences in performance are statistically significant. Table 5.4 demonstrates the p-values computed through the Wilcoxon test by pairwise comparisons between the methods. Based on the reported p-values, it can be noted that the results are statistically significant since almost all the obtained p-values are very smaller than the significance level ($\alpha = 0.05$). Hence, the statistical test strongly demonstrated the existence of significant differences between the methods.

Table 5.3 Comparison of the methods in terms of F-Score

Dataset	Labeled data (%)	SSOC				
		DT	SVM	KNN	RF	NN
Automobile	25	0.5206	0.5967	0.3481	0.6837	0.6227
	50	0.7708	0.7315	0.5584	0.9098	0.8448
	75	0.8402	0.7414	0.6222	**0.9785**	0.9458
Car Evaluation	25	0.5679	0.5359	0.4350	0.7488	0.8857
	50	0.6494	0.5038	0.4899	0.9106	0.9718
	75	0.8580	0.5074	0.7885	0.9737	**0.9945**
ESL	25	0.5451	0.3863	0.4882	0.5980	0.5942
	50	0.5823	0.4355	0.5616	0.6817	0.6377
	75	0.6316	0.4806	0.5731	**0.8035**	0.6699
Eucalyptus	25	0.5971	0.6359	0.4645	0.6582	0.6660
	50	0.6501	0.6821	0.5455	0.8016	0.7818
	75	0.7173	0.7060	0.5931	**0.8989**	0.8535
Nursery	25	0.7079	0.7283	0.6976	0.8127	0.8907
	50	0.8003	0.7740	0.7339	0.8837	0.9125
	75	0.8229	0.7856	0.7785	0.9003	**0.9125**
Squash-Stored	25	0.6107	0.6715	0.5325	0.709	0.6976
	50	0.6302	0.7769	0.5729	0.8081	0.8232
	75	0.806	0.8703	0.6899	**0.9083**	0.9011
UKM	25	0.9043	0.7772	0.7324	0.9112	0.9539
	50	0.9543	0.8053	0.8369	0.969	0.9699
	75	0.968	0.9093	0.8702	**0.9891**	0.9844
Volcanoes	25	0.4388	0.5810	0.43815	0.5207	0.5482
	50	0.4673	0.5810	0.5174	0.7300	0.5536
	75	0.5648	0.5810	0.5295	**0.9035**	0.7099
WQ-Red	25	0.3973	0.3305	0.3523	0.5676	0.4160
	50	0.4659	0.3325	0.3938	0.7639	0.4697
	75	0.5321	0.3341	0.3935	**0.9265**	0.5210
WQ-White	25	0.3565	0.2365	0.3164	0.5372	0.3446
	50	0.4487	0.2729	0.3386	0.7606	0.3800
	75	0.5152	0.2704	0.3638	**0.8928**	0.4116

Table 5.4 Wilcoxon test results.

Compared Algorithms	P-value	Significance Level
SSOC-DT vs. YATSI-DT	0.0000971	very strong
SSOC-SVM vs. YATSI-SVM	0.0349977	strong
SSOC-KNN vs. YATSI-KNN	0.0028530	very strong
SSOC-RF vs. YATSI-RF	0.0000017	very strong
SSOC-NN vs. YATSI-NN	0.0000017	very strong

5.3.2 The Results of Experiment 2

In this experiment, we considered different ratios of labeled data since the classification performance can be varied with the amount of labeled data. Figure 5.1 shows the accuracy results obtained by the proposed SSOC method with ranging labeled data ratios from 15% to 50% with an increment of 5%. The results indicated that when the ratio of labeled data increase, our method provide significant improvement. The best accuracy on average (79.35%) was achieved when 50% of the data was considered as labeled. The second-best accuracy on average (78.12%) was obtained under the circumstance: 45 % labeled and 55% unlabeled data. The percentages of 40% and 35% followed them with the average accuracy values of 76.98 % and 75.54 %, respectively. Hence, it can be concluded that, in ordinal classification tasks, a set of labeled ordinal data can be employed by the proposed SSOC algorithm to consistently improve performance. This result indicates that the initial small number of labeled ordinal data is not enough sufficient learn the model, and therefore, to improve the classification performance we need additional labeled data.

As has been observed in the Nursery, UKM, and Volcanoes datasets, it is possible for the ordinal classification problem, applying a few numbers of labeled data and a large number of unlabeled data can use for the ability of generalization. This is probably because of the following reasons. The Nursery and Volcanoes datasets are the largest datasets among others, hence the initial labeled ordinal data is enough to learn the model sufficiently. This makes the model more robust, and therefore, unlabeled data can be labeled correctly by using this initial model. The Volcanoes dataset has the least number of features among the other datasets, and therefore, it is free from irrelevant and redundant features. The presence of irrelevant or redundant features may mislead the algorithm to produce irrelevant patterns, leading to classification errors. The UKM dataset is a relatively balanced dataset, therefore, providing a representative investigation and is able to reliably classify all classes. When the dataset is imbalanced, the instances from the minority class may not exist in the small labeled data enough for learning. Furthermore, the model may fail to converge to a correct solution and cannot give good results, if labeled data does not represent patterns in the data. Performance declines could be observed when there was a mismatch between the classes available in the labeled ordinal data and the classes present in the unlabeled data. Therefore, the selection of optimal labeled data ratio is restricted so that at least one sample per class should be labeled. In multiclass classification, the number of labeled data samples can be increased linearly with respect to the number of classes.

It experimentally confirmed that the characteristics of the ordinal data and their partitioning may have an important impact on the classification performance of different machine learning algorithms. To provide a realistic evaluation, researchers should run alternative algorithms on available datasets with different ratios of labeled and unlabeled ordinal data. For example, in this study, it was observed from Figure 5.1 that the SSOC-RF and SSCO-NN algorithms were usually worked well even if a small number of labeled data is available.

The results obtained from the Nursery and Volcanoes datasets are quite stable, no matter what ratio of labeled data is used. This is probably because of the fact that they are the largest datasets and so the initial labeled data were supplied sufficiently, and therefore, the classification accuracy did not change so much by adding more unlabeled data. On the other hand, the labeling process is quite important for other datasets. For example, for the Automobile dataset, the classification accuracy of the SSOC-RF method increased from 84.88% to 87.80% after adding 5% labeled data.

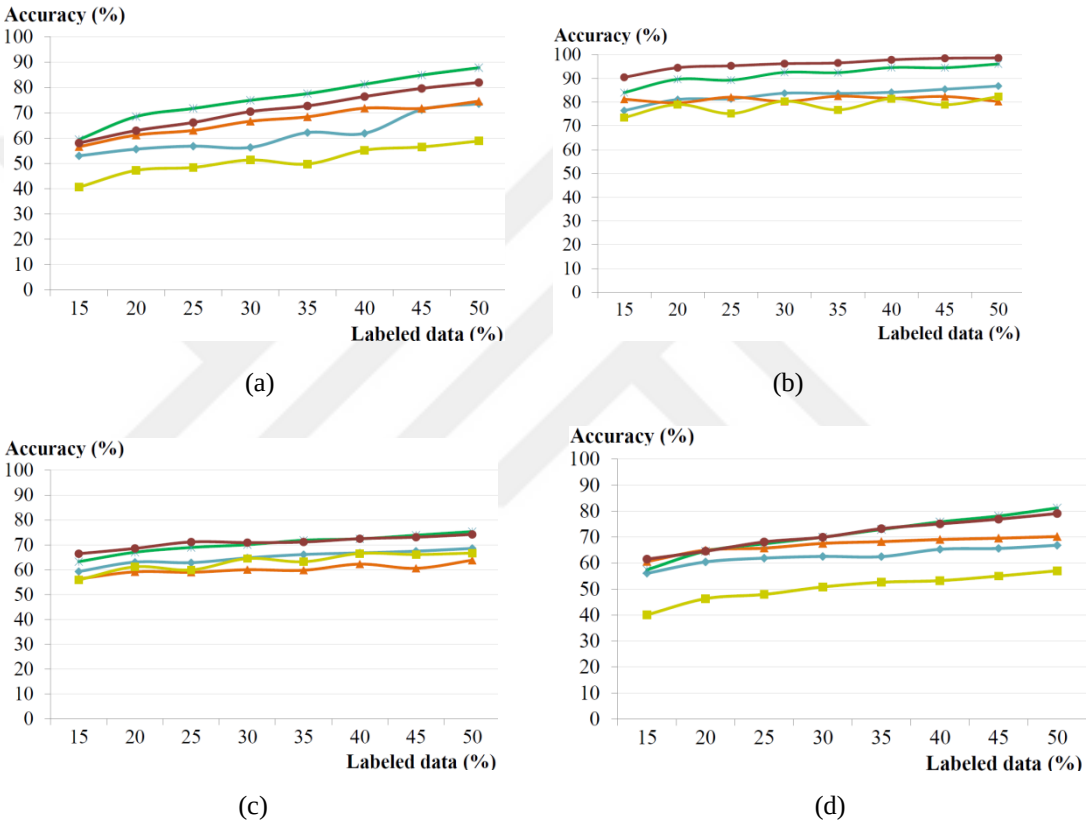


Figure 5.1 Accuracy Results were obtained with different ratios of labeled data. (a) Automobile Dataset; (b) Car Evaluation Dataset; (c) ESL Dataset; (d) Eucalyptus Dataset; (e) Nursery Dataset; (f) Squash-Stored; (g) UKM Dataset; (h) Volcanoes Dataset; (i) WQ-Red Dataset; (j) WQ-White Dataset.

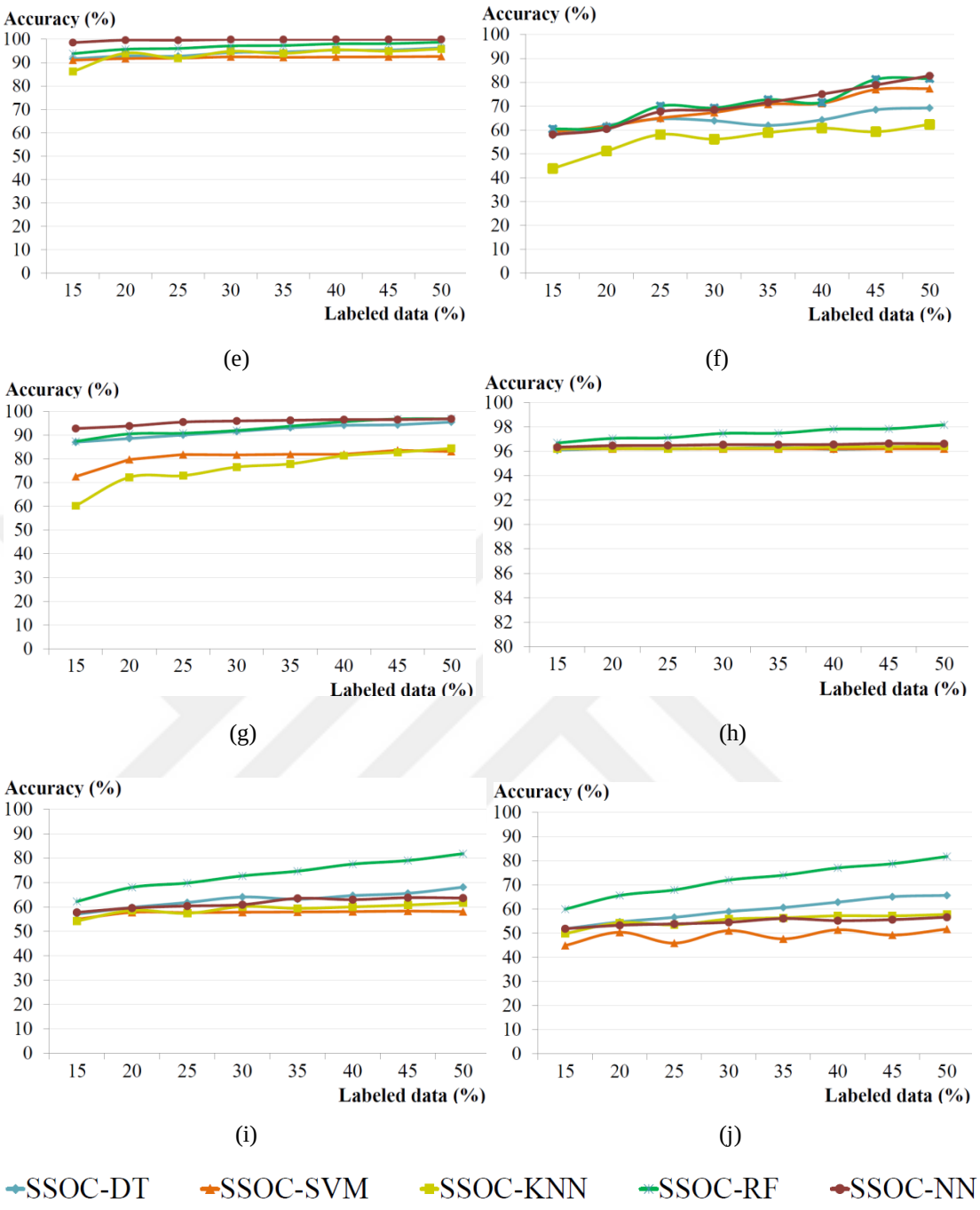


Figure 5.1 Continues

5.3.3 The Results of Experiment 3

A key property of our SSOC method is that it can be applied to any ordinal data in combination with any given base learner. The performance of the method is dependent on the selection of the appropriate base learning algorithm. Therefore, in this experiment, the proposed approach was tested on various datasets in

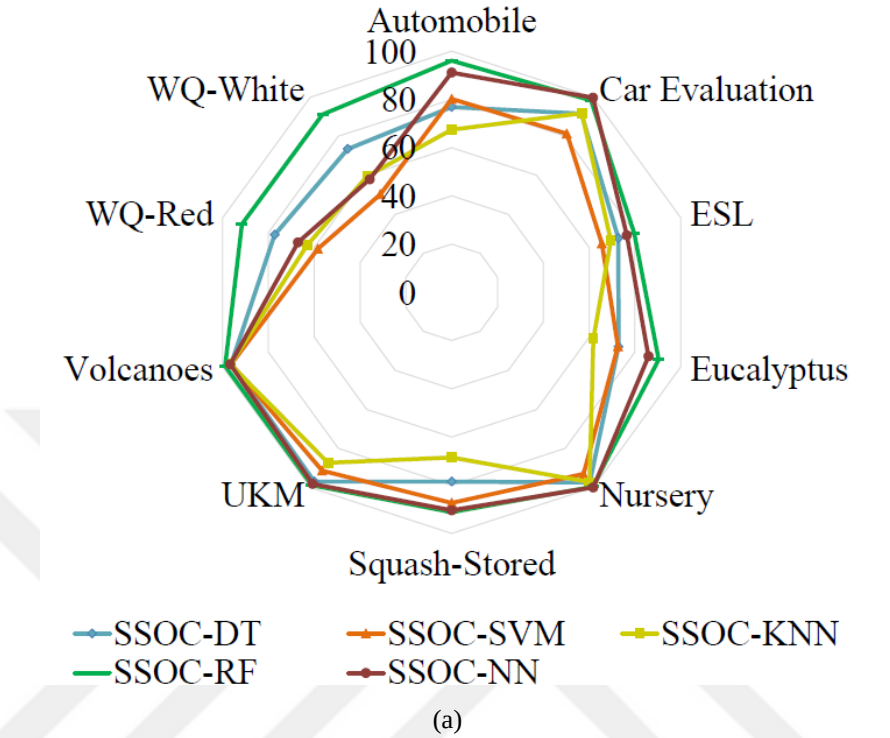
combination with five popular classification algorithms, including DT, SVM, KNN, RF, and NN.

Figure 5.2 shows the radar and bar charts that represent the accuracy values obtained from each algorithm for the ordinal datasets considered when using a 25%-75% unlabeled-labeled ratio. We selected this ratio since the accuracy of prediction increased as the number of labeled data samples increased, and the characteristics underlying the data cannot be suitable for training with a small ratio of labeled data due to many different reasons, i.e., small data size, imbalanced data, noisy data, the presence of similar classes, and a large number of classes. The radar chart illustrates the classification performance as the distance from the center; thus, a higher area indicates a better classification performance. It helps us to comparatively visualize the accuracy of the methods for each dataset separately. Figure 5.2 (b) shows the same results in a different way aiming to compare the accuracy values on all the datasets.

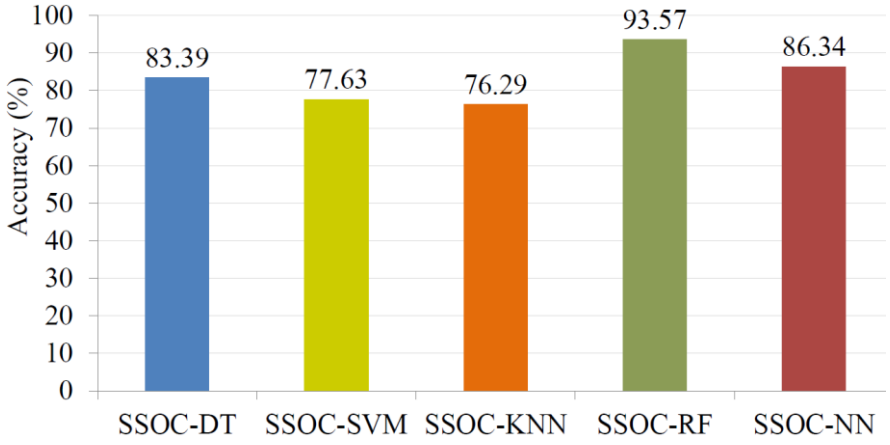
According to the results, SSOC-RF achieved slightly better accuracy (93.57%) than the rest. This is probably because RF is an ensemble learning method that constructs many decision trees to improve classification performance. The SSOC-NN and SSOC-DT methods follow the SSOC-RF method with accuracy values of 86.34% and 83.39%, respectively. It seems that KNN is not a good choice as a base learner for SSOC since their combinations showed the worst performance (76.29%) on almost all datasets. Furthermore, it was observed that SSOC-SVM provided better results than SSOC-KNN, although the differences between them are less remarkable than in the cases of other methods adopted as base learners.

As has been empirically observed, the characteristics of the ordinal data may have an important impact on the classification performance of different base learners. While the best average accuracy (96.92%) was achieved for the Volcanoes dataset, the worst average values were generally obtained for the WQ-red and WQ- white datasets. Some base learners may work well on particular types of datasets and may perform poorly on others. Therefore, a combination of empirical evaluation and

theoretical analysis should be used to determine the best base learner for the given problem.



(a)



(b)

Figure 5.2 Accuracy results were obtained with different ratios of labeled data. (a) Radar chart; (b) Two-dimensional column chart.

In this study, we considered different ratios of labeled data since the classification performance can be varied with the amount of labeled data. For example, when the 25%–75% labeled-unlabeled ratio was considered, the results are given in Table 5.2

and Figure 5.1 showed that the SSOC-DT, SSOC-RF, and SSOC-NN methods achieved high accuracy values (≥80%) on some datasets such as the Car Evaluation, Nursery, UKM, and Volcanoes datasets. However, a small ratio of labeled data could sometimes be insufficient enough to train the model. This is especially true in the following cases: (i) When the dataset size is small, the insufficient labeled data may not represent patterns in the data; (ii) when the dataset is imbalanced, the instances from the minority class may not exist in the small labeled data enough for learning; (iii) when the data is noisy, a small portion of labeled data may not tolerate the noise itself; (iv) when some classes are similar to each other, a small portion of labeled data may not contain useful information for distinguishing them; (v) when the number of classes is large, a small number of labeled data may not include at least one sample per class.

Since SSOC is a meta-algorithm, it allows any supervised base learner to be applied, such as DT, SVM, and NN. Since deep learning (DL) has been attracting attention in the machine learning community in recent years, our method can also be combined with it. Since deep learning has been proven to be a powerful ML technique in many studies, it can be used to improve the classification performance of our method. We experimented to investigate the performance of SSOC-DL by using the WekaDeeplearning4j package (Lang et al., 2019). The first experimental results showed that SSOC-DL outperformed the existing OC-DL on average, similar to other base learners. The optimal hyperparameters could be found and increased to build models with higher performance. However, we can have challenges of high computational time and high power consumption. Deep learning has generally been used to solve complex problems such as image classification, speech recognition, and natural language processing since it can automatically extract useful features (such as edges in images). The datasets used in this study contain simple transactional data and are not complex to be analyzed by deep learning. However, the proposed SSOC method can take advantage of deep learning when classifying ordinal image, voice, text, and video data.

5.3.4 The Results of Experiment 4

This experiment was conducted to evaluate semi-supervised learning in the case of ordinal classification. The central question in this experiment is: whether the introduction of unlabeled ordinal data yields a learner better than the traditional learner or not – shortly, be it semi-supervised or supervised. To answer this question, we compared the proposed SSOC algorithm with the existing ordinal classification algorithm presented in (Frank et al., 2001). Table 5.5 shows the comparison results obtained when using a 50% – 50% labeled-unlabeled ratio. It is clearly seen that the semi-supervised ordinal classification methods (SSOC-DT, SSOC-SVM, SSOC-RF, and SSOC- NN) were exceeded their supervised ordinal counterparts (OC-DT, OC-SVM, OC-RF, and OC-NN) in terms of accuracy on average. The results presented in Table 5.5 are promising since they indicate that, in ordinal classification, unlabeled ordinal data can be employed by the machine learning algorithms to consistently improve performance. This is because of the fact that when unlabeled data is available besides the labeled ordinal data, semi-supervised learning can better exploit and learn existing structures in the explanatory features than supervised learning would be. For instance, the SSOC-RF method (87.93%) remarkably outperformed the existing OC-RF algorithm (80.61%) on average. Similarly, the SSOC-NN method (83.02%) achieved better performance than the existing OC-NN algorithm (77.43%) on average. It experimentally confirmed that the introduction of unlabeled ordinal data can improve the generalization and so classification performances of any particular learning algorithm. Hence, our semi-supervised ordinal classification algorithms showed significant potential and competitiveness against supervised ordinal ones. Consequently, the experimental results provided good synergy for the ordinal version of semi-supervised learning.

The obtained results are promising indeed because, in real-world applications, unlabeled ordinal data are more available than labeled ordinal data, and labeling ordinal data is a difficult, expensive, or time-consuming process and it usually requires human efforts. Therefore, it is usually possible to construct a proper classification model by labeling at most half of the ordinal instances, instead of all.

Table 5.5 Proposed method (SSOC) and the existing method (OC) comparison in terms of accuracy (%)

Dataset	SSOC					OC				
	DT	SVM	KNN	RF	NN	DT	SVM	KNN	RF	NN
Automobile	73.56	74.54	58.83	**87.80**	81.95	66.34	66.83	61.95	85.85	72.68
Car evaluation	86.79	80.31	82.35	96.06	98.62	90.16	80.32	94.27	95.72	**99.42**
ESL	68.57	63.81	66.68	**75.33**	74.18	65.78	66.80	68.24	66.60	68.85
Eucalyptus	66.85	70.22	57.04	**81.20**	79.08	65.49	66.03	54.48	62.77	63.32
Nursery	96.30	92.64	95.84	98.74	**99.98**	97.04	93.13	97.77	99.14	99.95
Squash-stored	69.23	77.31	62.31	81.54	**82.69**	69.23	71.15	53.85	67.31	65.38
UKM	95.53	83.13	84.42	**96.92**	96.87	92.56	93.55	85.86	93.55	95.29
Volcanoes	96.25	96.21	96.42	**98.17**	96.62	96.08	96.18	96.49	96.42	96.61
WQ-red	68.09	58.09	61.71	**81.78**	63.63	61.60	58.41	58.04	69.48	59.29
WQ-white	65.66	51.67	57.76	**81.76**	56.61	57.62	51.72	54.10	69.27	53.51
Avg.	78.68	74.79	72.34	**87.93**	83.02	76.19	74.41	72.51	80.61	77.43

5.3.5 The Results of Experiment 5

The proposed SSOC method consists of two main steps. In the first step, a classifier is trained with the initial labeled data on an ordinal scale. In the second step, the constructed ordinal classifier labels the unlabeled data by its predictions, called pseudo-labels, and then the classifier is retrained with this augmented data. In this experiment, we investigated that after retraining the classifier with new pseudo-labels, how much the performance is increased compared to the performance with initial training labels. To answer this question, we compared the performances of the initial model and retrained model.

Figure 5.3 shows the average accuracy values obtained from the initially trained model and retrained model for all the ordinal datasets considered when using 25%-75%, 50%- 50%, and 75%- 25% labeled-unlabeled ratios. It can be seen from the results that the pseudo-labeled data can be employed by the algorithm to consistently improve performance. For example, when a 50%-50% labeled-unlabeled ratio was

considered, the SSOC-NN method achieved a higher accuracy value (83.02%) than the initial model (76.07%) on average. It experimentally confirmed that the introduction of unlabeled ordinal data can improve the generalization and so classification performances of any particular learning algorithm. This is because when unlabeled data is available besides the labeled ordinal data, the algorithm can better exploit and learn existing structures in the explanatory features.

Figure 5.3 Performance improvement provided by retraining compared to the initial training

Providing additional knowledge by the unlabeled data instances relevant for ordinal classification improves the generalization ability of the learning system successfully. The best improvement (14.81%) was observed for the SSOC-RF algorithm when the labeled data rate is 75% probably due to its ensemble structure. Based on the results, we can conclude that the introduction of unlabeled ordinal data yields a learner better than the initial learner. Unlabeled data in addition to labeled ordinal data can help in improving a classifier in terms of accuracy.

As a result of the five experiments aforementioned, it can be concluded that the proposed SSOC algorithm looks quite promising in achieving high accuracy for the ordinal version of semi-supervised learning. The experiments demonstrated that exploiting the order among class labels in semi-supervised learning could lead to building better models, compared to traditional nominal and supervised learning.

5.4 Case Study for Education

5.4.1 Description of Educational Datasets

It has been studied with three datasets suitable for ordinal classification related to the field of education. A brief description of the datasets is given in Table 5.6.

Table 5.6 Description of the education datasets

No	Name	Feature	Instance	Class	Class Distribution	Source
1	xAPIEdu	17	205	3	(211, 127, 142)	Kaggle
2	studentmat	33	649	5	(130, 103, 60, 40, 62)	UCI
3	studentpro	33	649	5	(201, 154, 112, 82, 100)	UCI

The first dataset, named xAPI-Edu, was obtained from the Kaggle data repository (https://www.kaggle.com/aljarah/xAPIEdu-Data). It contains some demographic / behavioral features and information about the academic achievements of the students. This data was collected from an e-Learning system, called Kalboard 360, using the Experience API Web service (XAPI) (Amrieh et al., 2016). The target attribute in the dataset involves ordinal class labels about student performances like low-level, middle level, and high level.

The second and third datasets, named student-mat and student-por, were obtained from the University of California Irvine (UCI) machine learning repository (https://archive.ics.uci.edu/ml/datasets/Student+Performance). These datasets also contain student academic performance information (Cortez & Silva, 2008). Two datasets were provided regarding the performances of students in two distinct courses: Mathematics (mat) and Portuguese language (por). In these datasets, the final score is ranged between 0 and 20, making prediction difficult since it is wide in scope. For this reason, some data preprocessing activities were performed on the datasets as described in (Unal, 2020). The final score was re-graded according to an ordinal structure such that the range 0-9 expressing F, 10-11 expressing D, 12-13 expressing C, 14-15 expressing B, and 16-20 expressing A.

5.4.2 Experimental Studies for Education

In the experiments, three different classification algorithms were used as base learners with SSOC, including Decision Tree (DT), Random Forest (RF), and Neural Network (NN). The SSOC method with each different base learner constructed an independent classifier. From here onwards, the related approach is named with the abbreviation of the method followed by the abbreviation of the base learner. For example, SSOC.DT denotes that the SSOC method is used with the DT base learner.

The performances of the models were assessed according to the accuracy and F-Score measures. Accuracy is the ratio of correctly predicted instances to the total number of instances and it is calculated by the formula: $Accuracy = (TN+TP)/(FN+FP+TN+TP)$, where false negative (*FN*) and false positive (*FP*) represent incorrectly classified instances; whereas, true negative (*TN*) and true positive (*TP*) are measures of correct predictions. F-Score is also a useful evaluation measure that considers three quantities: *TP*, *FP*, and *FN*. It is a summary performance metric since it is the harmonic mean of precision and recall.

The results were obtained by using the 10-fold cross-validation methodology, in which the whole data is randomly divided into ten disjoint and equally populated groups, and one of the groups is utilized as a testing set, while the remaining groups are used as the training set. Initial data was selected five times by using different random seed numbers (from 1 to 5), so the training process is repeated five times and the average accuracy results were reported here.

5.4.3 Experimental Results for Education

To investigate the effectiveness of SSOC on education data, several experiments had been carried out. Each dataset was considered as labeled and unlabeled at certain rates and the results obtained with various ratios of labeled data were compared. We

took into consideration different percentage values when partitioning the datasets, varying from 15% to 50% with an increment of 5%, and also 75% labeled data.

Figure 5.4 shows the accuracy values obtained by the application of the SSOC.DT, SSOC.RF, and SSOC.NN methods on the first dataset. Although there is not much difference in performances of SSOC.NN and SSOC.RF, it is possible to say that the SSOC.NN method is more likely to classify correctly many ratios of labeled data. As expected, the performances of all the methods improved as the number of labeled data instances increased. The best accuracy (92.29%) was obtained by SSOC.RF under the circumstance: 75% labeled and 25% unlabeled data. It seems that DT is not a good choice to be used with SSOC since their conjunction showed the worst performance for this dataset.

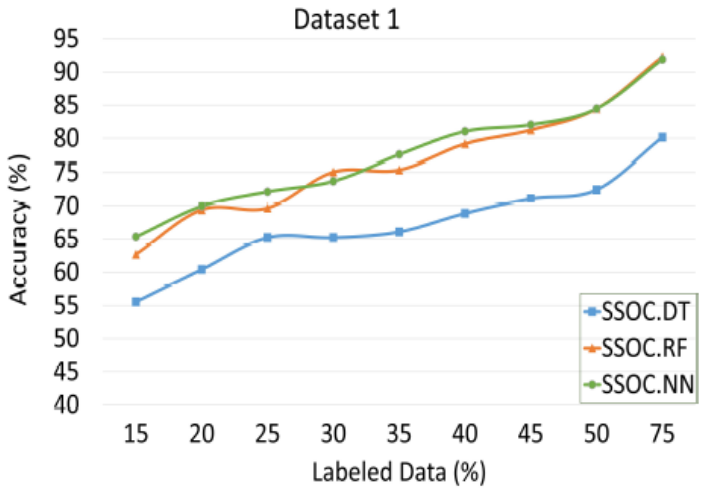


Figure 5.4 Comparison of methods in terms of accuracy on dataset 1 with different labeled data ratios

Figure 5.5 and Figure 5.6 show the comparison of the performances of the SSOC.DT, SSOC.RF, and SSOC.NN methods on the second and third datasets, respectively. According to the results, the Random Forest-based SSOC method always achieved the best results for both of the datasets. Another remarkable point here is that even with 15% labeled data, an accuracy of 71.74% was obtained in the third dataset. Thereby, it can be concluded that it is possible to achieve high accuracy values by SSOC.RF even if a very small amount of labeled ordinal data is available. The highest accuracy value (93%) was obtained by SSOC.RF for Dataset 3 under the

circumstance of 75% labeled data. The reason is probably that RF is an ensemble-styled learning algorithm and usually increases accuracy by combining the outputs of many decision trees.

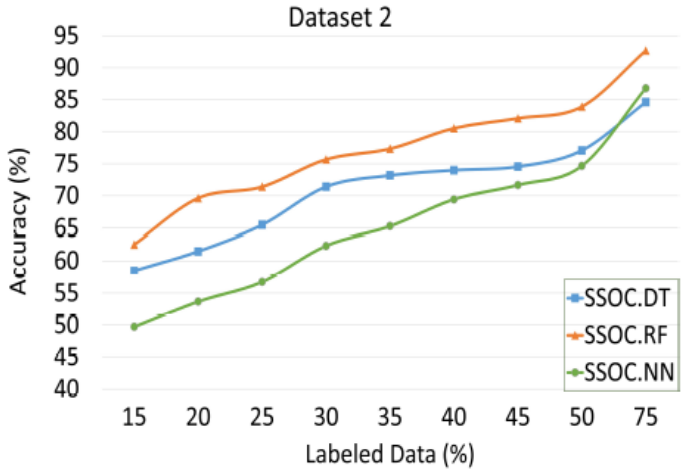


Figure 5.5 Comparison of methods in terms of accuracy on dataset 2 with different labeled data ratios

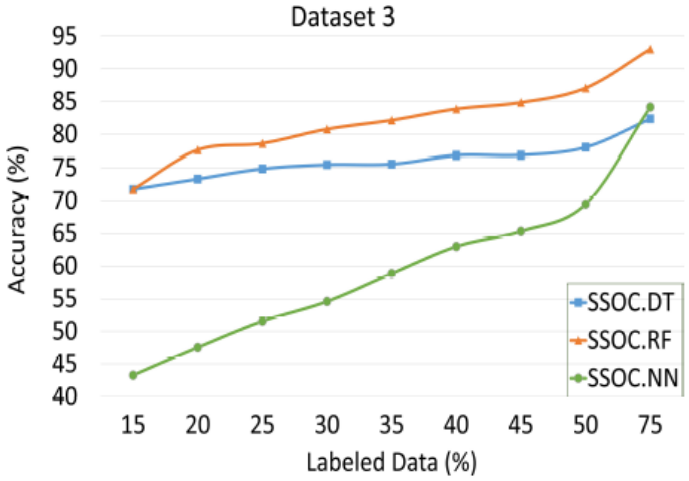


Figure 5.6 Comparison of methods in terms of accuracy on dataset 3 with different labeled data ratios

As has been empirically observed, SSOC with a single decision tree base learner followed SSOC.RF with high-value accuracy rates.

Table 5.7 shows the F-Score values obtained by the SSOCbased methods on the 25%, 50%, and 75% labeled data of three datasets. F-Score is ranged from 0 to 1, where 1 is the best value. Considering that being closer to 1 is better for F-Score, the

higher rate of labeled data resulted in an improvement in the F-Score value in all cases. For Dataset 1, an F-Score value of 0.93 was obtained by SSOC.RF when 75% labeled data. As has been observed for Dataset 2, the highest F-Score value (0.92) was obtained by SSOC.RF again. For Dataset 3, the highest performance (0.93) was achieved under the same conditions. As a result, SSOC.RF usually achieved better accuracy than the rest. This is probably because of the fact that Random Forest is an ensemble learning algorithm that builds many decision trees to improve performance.

Table 5.7 F-Score results obtained by the methods for 25%, 50%, and 75% labeled data

Dataset No	% Labeled	SSOC.DT	SSOC.RF	SSOC.NN
1	25	0.64	0.71	0.73
	50	0.73	0.85	0.85
	75	0.81	0,93	0.92
2	25	0.62	0.68	0.52
	50	0.75	0.82	0.71
	75	0.83	0.92	0.84
3	25	0.76	0.79	0.51
	50	0.78	0.87	0.70
	75	0.82	0.93	0.84

In this experiment, we compared the SSOC method with the traditional OC method (Frank & Hall, 2001) on 50% of labeled data in terms of accuracy. Figure 5.7 shows the results side-by-side for the same base learners (SSOC on the left and OC on the right), separately for each dataset. As can be observed, SSOC outperformed OC by obtaining higher accuracy values in all cases. For example, in Dataset 3, OC only reached the accuracy value of 73:85% with Random Forest base learner, while SSOC achieved an 87.06% accuracy value. For each dataset and with each base learner SSOC provided a significant increase in accuracy against OC.

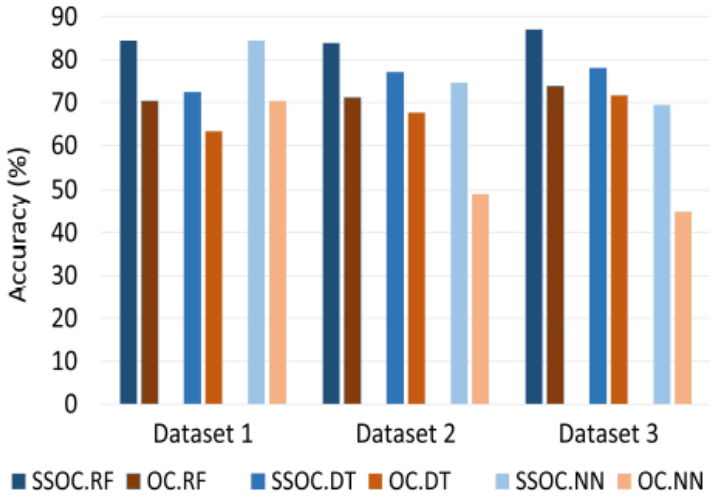


Figure 5.7 SSOC vs OC

CHAPTER SIX
DISCUSSION

6.1 Main Findings

The results obtained in this study have given us an idea about the successes we have achieved with our method. In the experimental studies, 10 different datasets in the ordinal structure were used. In addition, the results were compared with different labeled percentages in these datasets. While comparing SSOC and YATSI, 25%, 50%, and 75% labeled data percentages were taken and five different base learner values were compared. When the results are reviewed, even as low as 25% labeled percentage; Good results were obtained such as 95.30% in the Car Evaluation dataset, 99.58% in the Nursery dataset, 95.53% in the User knowledge modeling (UKM) dataset, and 97.11% in the Volcanoes on venus (B6) dataset. With these results, the SSOC algorithm is ahead of the compared YATSI algorithm with a significant difference in success. In comparison with the YATSI algorithm, according to Wilcoxon test results, our algorithm with SVM base learner was stronger than YATSI and very strong with all other base learner algorithms.

We also compared our algorithm with the OC algorithm to test the semi-supervised effects that increase our success. In this comparison, the same datasets with a 50% labeled ratio and the same base learner algorithms were used. In this comparison, remarkable successful results were obtained. Using RF base learner, 87.80% with Automobile dataset, 96.92% with UKM dataset, and 98.17% with Volcanoes dataset; By using NN base learner, 99.98% accuracy rates with the Nursery dataset were obtained. When the results are evaluated, it has been revealed that the SSOC algorithm is more powerful than the OC algorithm.

The main findings can be concluded as follows:

- First, The proposed SSOC method resulted in a significant improvement over the existing YATSI method.

- The accuracy of prediction slightly increased as the number of labeled data samples increased.
- When the SSOC method was tested in combination with popular classification algorithms (DT, SVM, KNN, RF, and NN), the SSOC-RF algorithm achieved significantly better accuracy (93.57 %) than the rest.
- Semi-supervised ordinal classification methods (SSOC-DT, SSOC-SVM, SSOC-RF, and SSOC-NN) generally exceeded their supervised ordinal counterparts (OC-DT, OC-SVM, OC-RF, and OC-NN) in terms of accuracy when labeling at most half of the ordinal instances.

In our previous studies, we evaluated good accuracy values with the SSOC algorithm. In addition, it has been determined that the algorithm is suitable for education, especially considering that the data for measuring student achievement in education is in an ordered structure and labeled data is less. To prove the compatibility of the algorithm with the data in the field of education, educational datasets were studied and comparisons were made. For this purpose, 3 different datasets were studied and values were obtained with different labeled ratios between 15% and 75%. Here are the remarkable results; In the third dataset, 71.74% accuracy was achieved with only 15% labeled percentage and 93% accuracy with 75% labeled percentage. Table 5.7 shows that 3 different base learner algorithms get good F-Score values for each dataset. Success was achieved with all three datasets, as shown in Figure 5.7, in the comparison with the OC algorithm. These results show that the algorithm is also suitable for the education field.

All these evaluations showed us that we could achieve good results even in low labeled percentages of our algorithms in this field. But the same results will not be valid for all kinds of datasets, the algorithm has proven its power in datasets with ordinal structures.

CHAPTER SEVEN
APPLICATION

7.1 Web Application for SSOC

This thesis proposes a new method, named SSOC, in which "ordinal classification" and "semi-supervised learning" concepts are combined for the first time. In other words, it introduces a new concept of "Semi-supervised Ordinal Classification". To make the proposed SSOC method accessible via the web interface, we developed a web application. The web interface consists of dataset selection, labeled data percentage selection, and a base learner algorithm selection screen. The parameters taken from the screen are evaluated by the classification algorithm running on the server, and the results are shared on the screen. Thanks to this interface, our method has been made available to the whole world with no need to develop an environment.

In this study, we developed a web application. A web application is software that runs on web server and is accessed by the user through a web browser. A client-server model was designed. We used a Tomcat server as a web server.

The application was developed on the J2EE platform with the Eclipse framework and Java programming language. Weka library was used to run data mining algorithms.

In the web application, three selections can be made: dataset selection, labeled data split (%) selection, and base learner selection. The following can be chosen as a base learner: Random Forest, K-Nearest Neighbors, Decision Trees, Neural Networks, and Support Vector Machines.

Figure 7.1 shows the general architecture of our web application. All evaluation process has been made on the server-side. The application takes all parameters from users, then the server-side runs the algorithm, and finally, the results have been

shown to the user by an interface. The application enables the use of the proposed SSOC method via a web interface to evaluate new test data for prediction.

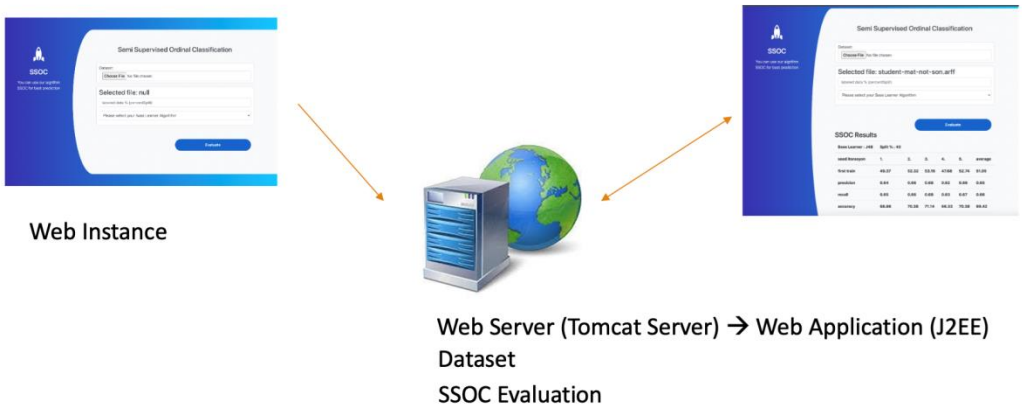


Figure 7.1 SSOC web application architecture

Figure 7.2 shows a screenshot of the web application. Here, the user can make parameter selections. On the interface, there is a file upload field that enables a dataset selection. In addition, a field allows the user to select the percentage of labeled data, and finally, another field makes it possible to select the base learner algorithm.

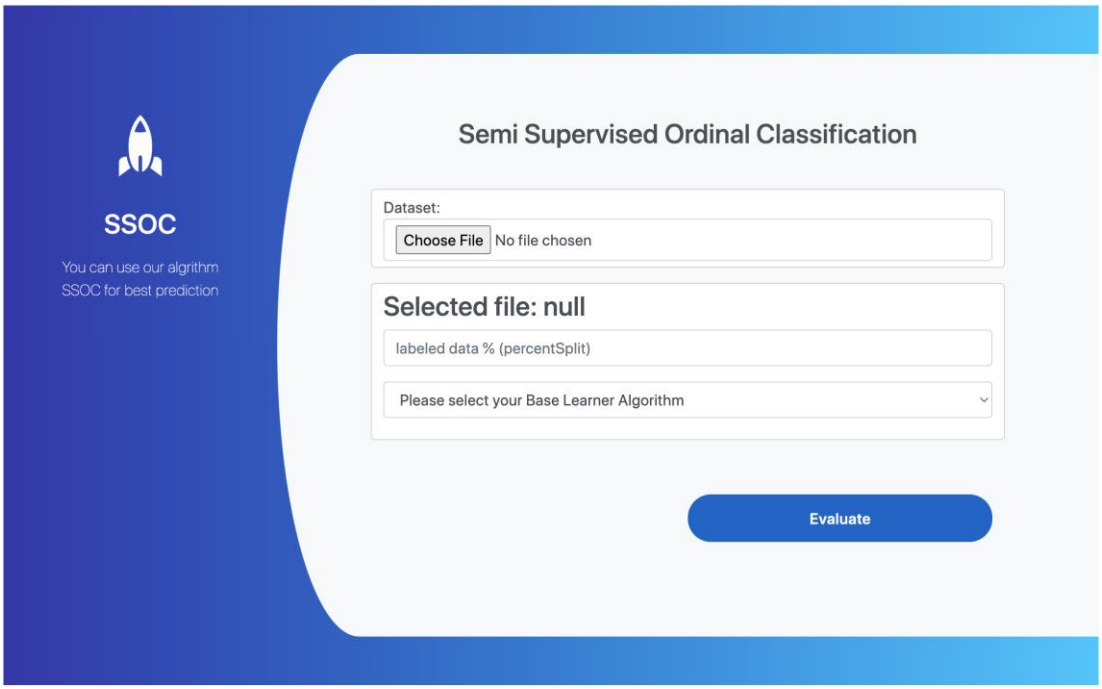


Figure 7.2 SSOC web application user interface

Figure 7.3 shows another screenshot displaying the results of the SSOC method working with the parameters we have chosen. As can be seen in the figure, a sample result table is shown after selecting a sample dataset, percentage, and base learner algorithm options.

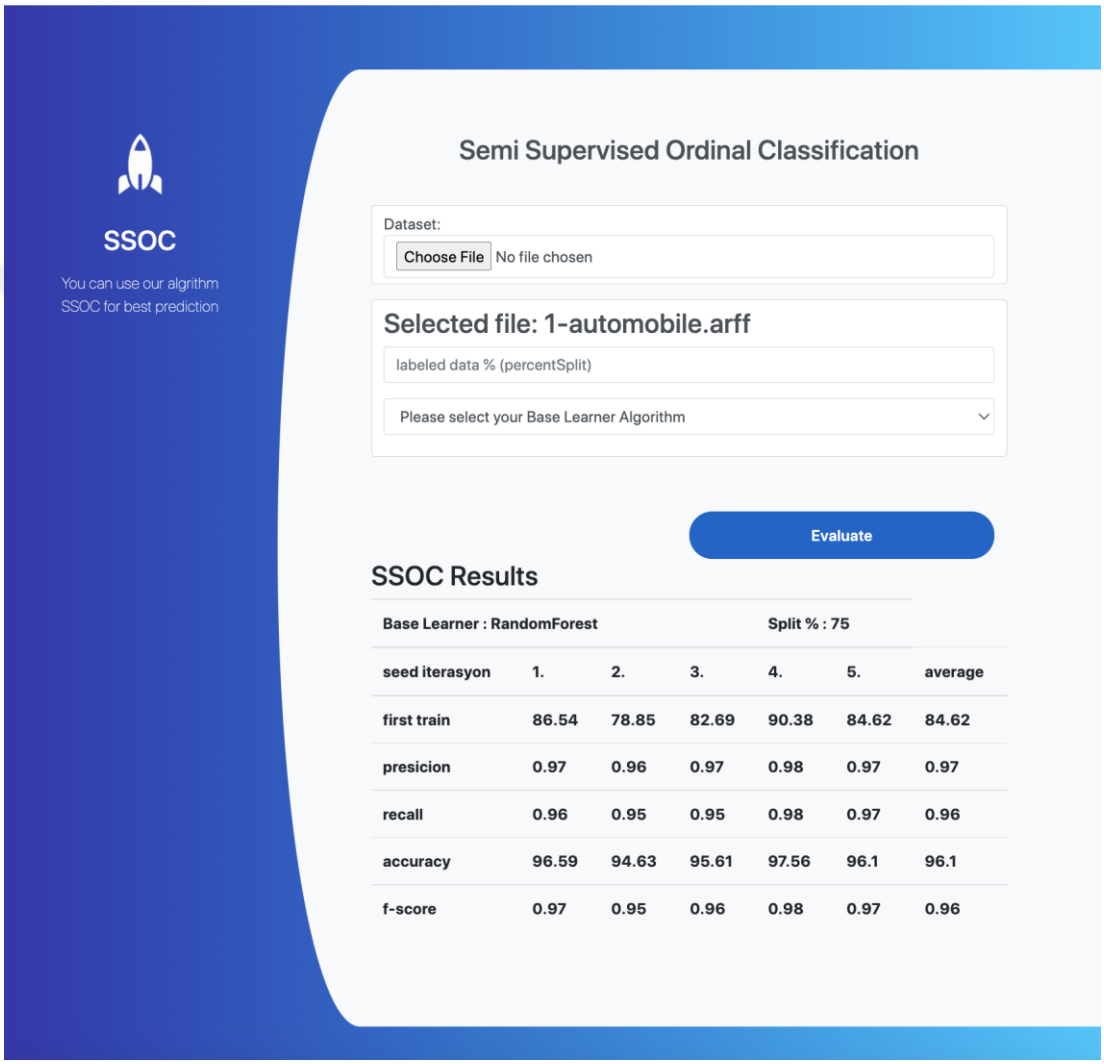


Figure 7.3 SSOC web application results page

This application enables us to make the SSOC method accessible to users. Hence, the SSOC method can be tested without downloading any library or providing an environment.

CHAPTER EIGHT
CONCLUSION AND FUTURE WORK

8.1 Conclusion

Ordinal classification differs from multiclass (nominal) classification in that there exists some ordering among the class labels such as low, medium, and high. The existing semi-supervised methods provide a nominal classification task. The key issue for our study, thus, is to design a new method that can effectively combine "ordinal learning" and "semi-supervised learning" paradigms for the categorical class labels for the first time. The purpose of our study is to effectively use a small number of labeled ordinal data and a large number of unlabeled ordinal data for classification model construction. From this perspective, our study is very important since unlabeled ordinal data is cheap and abundantly available, and labeling ordinal data is a time-consuming and costly process requiring expert knowledge. It aims to maximize the classification performance of the model through appending unlabeled samples to the labeled ordinal data while minimizing the human effort.

This paper introduces a new concept of "semi-supervised ordinal classification" for categorical class labels. It proposes a new algorithm, called SSOC, which takes into account the relationships between the class labels during semi-supervised learning. In the proposed SSOC method, a classifier is firstly trained on an initially small number of labeled ordinal samples to classify unlabeled instances. After that, the resulting classifier is re-trained with its estimations by enlarging its initial labeled set.

In the experimental studies, we evaluated the performance of the proposed algorithm (SSOC) on various ordinal datasets. The SSOC method resulted in a significant improvement over the existing YATSI method.

The proposed SSOC method was tested in the field of education. It was applied to three ordinal educational datasets. Since the general problem is the lack of labeled

data, the proposed method can increase the accuracy rate by using unlabeled data as well.

To make the proposed method accessible and available, we developed a web application. With the web interface, it is possible to work on any dataset with the selected base learner.

8.2 Future Work

In this study, the concepts of semi-supervised learning and ordinal classification were brought together. In the future, the concept of ensemble learning can be added to this structure. In other words, an ensemble semi-supervised ordinal classification approach can be implemented by using the proposed method as an ensemble member. Furthermore, the success of the ensemble-based approach can be tested in the field of education.

REFERENCES

Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, *24*(4), 567-598.

Agaoglu, M. (2016). Predicting instructor performance using data mining techniques in higher education. *IEEE Access*, *4*, 2379-2387. https://doi.org/10.1109/ACCESS.2016.2568756

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, *9*(8), 119-136. http://dx.doi.org/10.14257/ijdta.2016.9.8.13

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177-194. https://doi.org/10.1016/j.compedu.2017.05.007

Buenaño-Fernandez, D., Villegas-CH, W., & Luján-Mora, S. (2019). The use of tools of data mining to decision making in engineering education—A systematic mapping study. *Computer Applications in Engineering Education*, *27*(3), 744-758. https://doi.org/10.1002/cae.22100

Cardoso, J. S., & Domingues, I. (2011, December). Max-coupled learning: application to breast cancer. In *2011 10th International Conference on Machine Learning and Applications and Workshops* (Vol. 1, pp. 13-18). IEEE. https://doi.org/10.1109/ICMLA.2011.93

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357.
https://doi.org/10.1613/jair.953

Chen, K., & Wang, S. (2010). Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(1), 129-143.
https://doi.org/10.1109/TPAMI.2010.92

Cortez. P., & A. Silva. A., (2008). Using Data Mining to Predict Secondary School Student Performance. In *A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference* (pp. 5-12). Porto, Portugal.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247-256.
https://doi.org/10.1016/j.chb.2017.01.047

Deshmukh, V.B., Mangalwede, S., & Rao, D.H. (2018). Student Performance Evaluation Using Data Mining Techniques for Engineering Education. *Advances in Science, Technology and Engineering Systems Journal*, *3*, 259-264.
http://dx.doi.org/10.25046/aj030634

Dornaika, F., Dahbi, R., Bosaghzadeh, A., & Ruichek, Y. (2017). Efficient dynamic graph construction for inductive semi-supervised learning. *Neural Networks*, *94*, 192-203.
https://doi.org/10.1016/j.neunet.2017.07.006

Driessens, K., Reutemann, P., Pfahringer, B., & Leschi, C. (2006, April). Using weighted nearest neighbor to benefit from unlabeled data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 60-69). Springer, Berlin, Heidelberg.
https://doi.org/10.1007/11731139_10

Fan, Y., Liu, Y., Chen, H., & Ma, J. (2019). Data Mining-Based Design and Implementation of College Physical Education Performance Management and Analysis System. *International Journal of Emerging Technologies in Learning*, *14*(6).

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*, 335-343.
https://doi.org/10.1016/j.jbusres.2018.02.012

Frank, E., & Hall, M. (2001, September). A simple approach to ordinal classification. In *European Conference on Machine Learning* (pp. 145-156). Springer, Berlin, Heidelberg.
https://doi.org/10.1007/3-540-44795-4_13

Guruler, H., & Istanbullu, A. (2014). Modeling student performance in higher education using data mining. In *Educational Data Mining* (pp. 105-124). Springer, Cham.
http://dx.doi.org/10.1007/978-3-319-02738-8_4

Gutiérrez, P. A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., & Hervas-Martinez, C. (2015). Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, *28*(1), 127-146.

https://doi.org/10.1109/TKDE.2015.2457911

Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, *36*, 469-478. https://doi.org/10.1016/j.chb.2014.04.002

Jayakumar, S., Parameswari, R. & Akila, A. (2019). Enhancing the quality education using predictive and descriptive data mining model. *International Journal of Recent Technology and Engineering*, *8*(3), 2277-3878.

Jiang, D., Ma, P., Su, X., & Wang, T. (2014). Distance metric based divergent change bad smell detection and refactoring scheme analysis. *International Journal of Innovative Computing, Information and Control*, *10*(4), 1519-1531.

Lang, R., Lu, R., Zhao, C., Qin, H., & Liu, G. (2020). Graph-based semi-supervised one class support vector machine for detecting abnormal lung sounds. *Applied Mathematics and Computation*, *364*, 124487. https://doi.org/10.1016/j.amc.2019.06.001

Lang, S., Bravo-Marquez, F., Beckham, C., Hall, M., & Frank, E. (2019). Wekadeeplearning4j: A deep learning package for weka based on deeplearning4j. *Knowledge-Based Systems*, *178*, 48-50. https://doi.org/10.1016/j.knosys.2019.04.013

Li, W., Meng, W., & Au, M. H. (2020). Enhancing collaborative intrusion detection via disagreement-based semi-supervised learning in IoT environments. *Journal of Network and Computer Applications*, *161*, 102631. https://doi.org/10.1016/j.jnca.2020.102631

Livieris, I. E., Drakopoulou, K., Tampakas, V. T., Mikropoulos, T. A., & Pintelas, P. (2019). Predicting secondary school students' performance utilizing a semi-supervised learning approach. *Journal of Educational Computing Research*, *57*(2), 448-470. https://doi.org/10.1177/0735633117752614

Lu, M. (2017). Predicting college students English performance using education data mining. *Journal of Computational and Theoretical Nanoscience*, *14*(1), 225-229. https://doi.org/10.1166/jctn.2017.6152

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, *103*, 1-15. https://doi.org/10.1016/j.compedu.2016.09.005

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, *33*(1), 107-124. https://doi.org/10.1111/exsy.12135

Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, *115*, 36-51. https://doi.org/10.1016/j.dss.2018.09.001

Perez-Ortiź, M., Gutiérrez, P. A., & Hervás-Martínez, C. (2013). Projection-based ensemble learning for ordinal regression. *IEEE Transactions on Cybernetics*, *44*(5), 681-694. https://doi.org/10.1109/TCYB.2013.2266336

Pérez-Ortiz, M., Gutiérrez, P. A., Carbonero-Ruz, M., & Hervás-Martínez, C. (2016). Semi-supervised learning for ordinal Kernel Discriminant Analysis. *Neural Networks*, *84*, 57-66. https://doi.org/10.1016/j.neunet.2016.08.004

Rahangdale, A., & Raut, S. (2019). Clustering-based transductive semi-supervised learning for learning-to-rank. *International Journal of Pattern Recognition and Artificial Intelligence*, *33*(12), 1951007. https://doi.org/10.1142/S0218001419510078

Sati, N. U. (2020). A novel semisupervised classification method via membership and polyhedral conic functions. *Turkish Journal of Electrical Engineering & Computer Sciences*, *28*(1), 80-92. http://dx.doi.org/10.3906/elk-1905-45

Sánchez-Monedero, J., Gutiérrez, P. A., Tiňo, P., & Hervás-Martínez, C. (2013). Exploitation of pairwise class distances for ordinal classification. *Neural Computation*, *25*(9), 2450-2485. https://doi.org/10.1162/NECO_a_00478

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414-422. https://doi.org/10.1016/j.procs.2015.12.157

Srijith, P. K., Shevade, S., & Sundararajan, S. (2013, September). Semi-supervised gaussian process ordinal regression. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 144-159). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40994-3_10

Stanescu, A., & Caragea, D. (2015). An empirical study of ensemble-based semi-supervised learning approaches for imbalanced splice site datasets. *BMC Systems Biology*, *9*(5), 1-12. https://doi.org/10.1186/1752-0509-9-S5-S1

Tsuchiya, T., Charoenphakdee, N., Sato, I., & Sugiyama, M. (2021). Semisupervised Ordinal Regression Based on Empirical Risk Minimization. *Neural Computation*, *33*(12), 3361-3412. https://doi.org/10.1162/neco_a_01445

Ünal, F. (2020). Data mining for student performance prediction in education. *Data Mining-Methods, Applications and Systems*.

Unal, F., BIrant, D., & ŞEKER, Ö. (2021). A new approach: semisupervised ordinal classification. *Turkish Journal of Electrical Engineering & Computer Sciences*, *29*(3), 1797-1820. http://dx.doi.org/10.3906/elk-2008-148

Unal, F., & Birant, D. (2021, June). Educational Data Mining Using Semi-Supervised Ordinal Classification. In *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-5). IEEE. https://doi.org/10.1109/HORA52670.2021.9461278

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373-440. https://doi.org/10.1007/s10994-019-05855-6

Wang, X., Yang, I., & Ahn, S. H. (2019). Sample efficient home power anomaly detection in real time using semi-supervised learning. *IEEE Access*, *7*, 139712-139725.

https://doi.org/10.1109/ACCESS.2019.2943667

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & DATA, M. (2005). Practical machine learning tools and techniques. *In DATA MINING*, *2*, 4.

Xu, P., Lu, W., & Wang, B. (2019). A semi-supervised learning framework for gas chimney detection based on sparse autoencoder and TSVM. *Journal of Geophysics and Engineering*, *16*(1), 52-61.
https://doi.org/10.1093/jge/gxy004

Yang, Y., Liu, J., Xu, S., & Zhao, Y. (2016). An extended semi-supervised regression approach with co-training and geographical weighted regression: a case study of housing prices in Beijing. *ISPRS International Journal of Geo-Information*, *5*(1), 4.
https://doi.org/10.3390/ijgi5010004

Zhu, S. (2019). Research on data mining of education technical ability training for physical education students based on Apriori algorithm. *Cluster Computing*, *22*(6), 14811-14818.
https://doi.org/10.1007/s10586-018-2420-8