2021

Predicting Employee Performance: A Meta-Analysis and Systematic Review on Data Mining Methods

Turku Erengin
University of Central Florida

PREDICTING EMPLOYEE PERFORMANCE: A META-ANALYSIS AND SYSTEMATIC REVIEW ON DATA MINING METHODS

by

TÜRKÜ ERENGİN
B.A. Özyeğin University, 2018

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science in Industrial/Organizational Psychology
in the Department of Psychology
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Spring Term
2021

ABSTRACT

Data mining methods have been used to study a variety of topics in industrial and organizational psychology, including predicting employee performance. With the increased interest in predictive analytics in human resources, the present study aimed to review and explore the application of two commonly used data mining methods, decision trees (DTs) and artificial neural networks (ANNs), for predicting employee performance in organizational settings. Out of 103 studies reviewed, eight studies were retained and used for the meta-analyses. The number of employee performance classifications meta-analyzed was 2430 in total. The results suggested that both data mining methods showed good performance in employee performance prediction, although the difference between the overall effect sizes was not statistically significant. The theoretical and practical implications and the potential limitations were discussed, and recommendations were provided for future research directions. The current study was a first attempt to qualitatively and quantitatively evaluate the effectiveness of the data mining methods in predicting employee performance.

ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Shiyang Su, for her continuous support, patience, and guidance through each stage of my thesis. Your knowledge and encouragement pushed me to pursue my interest in an emerging topic and broaden my research skills.

I would like to acknowledge Dr. Victoria Pace and Dr. Steve Jex for serving as my committee members and for their kind help and valuable insights. In addition, I am grateful to have Dr. Jex as my advisor for a semester and for his immense knowledge that guides my thesis progress.

I would like to extend my thanks to Dani Draper, as you always provide unwavering help throughout my time at the program.

Lastly, I would like to thank my parents for their encouragement and patience. You are always there to support me. Finally, I would like to express my gratitude to my friends, the rest of my family and all the memories we collected during my study breaks. I would like to particularly thank Göktuğ Eren for his constant support and all the cherished time spent together. I am forever grateful to have them in my life.

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

Big data: Big data is a special type of data characterized by the large volume, high-speed generation, and various sources, types, and structures. It usually exceeds the capabilities of traditional methods and requires unique methods for utilization. The associated advantages involve improved decision-making, identification of the unexplored and novel patterns, and increased efficiency with the automatization of the processes.

Classification: Classification is a subtype of data mining tasks. It is used for determining the correct class of each case. It involves two steps as the learning (providing the correct classes for each case to guide the development of the model) and classification (assigning the unknown cases to their respective classes based on the developed model). Example data mining methods used for classification include Decision Trees and Artificial Neural Networks.

Data mining: Data mining appears to be a promising tool used to address the challenges of big data and used for searching the data to reveal any relationships and derive new information. It consists of different practices from diverse disciplines such as statistics, computational science, and information technology.

Human Resource Analytics: Human Resource Analytics involves the use of data to guide human resources practices. It adopts a data-driven approach to improve organizational processes and inform related decision-making. The three sublevels can be listed as descriptive, predictive, and prescriptive analytics.

Human Resource Predictive Analytics: Human Resource Predictive Analytics is the second sublevel of Human Resource Analytics. The goal is to use the past and current data to predict future occurren

CHAPTER ONE: INTRODUCTION

Human resource management (HRM) has been in the heart of companies throughout the ages. The changing nature of the workforce, economy, and the effects of globalization make organizations more prone to instability and external threats, putting human resources in a more critical position than it has ever been. Having a team of employees that perform well is a competitive advantage to the company. Additionally, the best practices of talent management include the utilization of organizational big data and nowadays, the implementation of advanced technology promotes the use of big data in decision-making by introducing new and more complex ways to collect, store, and analyze data. For these reasons, organizations are becoming more and more interested in using existing employee big data to understand future scenarios and improve decision-making.

A thorough understanding of the employees requires a combination of information from multiple data sources, including demographic data, performance data, compensation data, behavioral data, and social interaction data (Ryan & Herleman, 2016). With the new ways of collecting and storing data in organizations, the utilization of employee data for providing insights on the workforce has become more viable than ever, specifically through the use of big data approaches. Although there are multiple conceptualizations for big data across the literature, three main characteristics of such data are volume, velocity, and variety (King et al., 2016; Laney, 2001). Specifically, the volume is associated with the size of the data and can vary with the sample size or the number of variables measured; velocity refers to the pace of data being created for use; lastly, as the name itself implies, variety indicates various data types (King et al., 2016). The growing interest in rich information available to improve decision-making has created different research avenues for the studies of industrial and organizational (I/O) psychology and HRM (Oswald et al., 2020). For example, King et al. (2016) listed the wide range of big data applications in multiple human resources areas, such as selection, performance evaluation, occupational health, and even diversity. The big data movement has enabled organizational researchers and practitioners to take advantage of new opportunities, including the continuous use of data for more frequent performance evaluation, the prediction of occupational health related incidents for prevention, and the utilization of vast biodata for recruitment and selection of the candidates (King et al., 2016). These developments have also brought some unique challenges and called for new methods to handle employee data. Particularly, King et al. (2016) listed the most common challenges of utilizing big data as analysis, integration, and interpretation. These challenges make it imperative to develop and adopt new methods to overcome problems and pave the way for the use of big data. More specifically, machine learning and data mining methods have become popular due to their promising ways of analysis and interpretation of big data.

The recent upsurge toward using organizational data is shaped under the roof of Human Resource Analytics, an evidence-based approach for improving the decision-making processes (Fitz-enz & Mattox, 2014). According to Mishra et al. (2016), Human Resource Analytics is interested mainly in current situations such as the cost of a selection system and turnover/retention rates. However, Human Resource Predictive Analytics goes beyond analyzing the current situation and incorporates statistical methods, machine learning methods, and data mining methods to foresee the possible scenarios using existing employee data (Mishra et al., 2016). In the context of I/O psychology, employee turnover, severance pay acceptance, and employee performance are found to be the most commonly studied areas under Human Resource Predictive Analytics (Ekawati, 2019).

This increased interest in using organizational big data necessitates the introduction and application of novel methods in various organizational settings. Accordingly, this study aimed to evaluate the use of different data mining methods in employee performance prediction qualitatively and quantitatively for a comprehensive review of the emerging field. In the next sections, the importance of assessing and predicting employee performance was discussed, and Human Resource Analytics was introduced with a focus on predictive analytics. Lastly, the data mining methods were presented with their unique capabilities that exceed traditional methods in handling big data. To evaluate the effectiveness of different methods, meta-analyses were conducted to combine studies from the extant literature. The current study evaluated the various methods used in employee performance prediction and contributed to future research on the application of these methods in I/O psychology.

.

CHAPTER TWO: LITERATURE REVIEW

Employee Performance

Organizations are interested in certain outcomes to assess their strengths and weaknesses in the current competitive environment. These outcomes include selection, employee turnover, and employee performance (Strohmeier & Piazza, 2013). Among the outcomes that organizations are interested in, employee performance is one of the most critical constructs involved in many different HRM processes (Campbell & Wiernik, 2015). The variety of usage of performance appraisals reflects the organizations' great interest in the performance evaluation of the employees. Correspondingly, Campbell and Wiernik (2015) listed the most common purposes of performance appraisals as follows: research, legal defensibility, promotion, compensation, and other high-stakes appraisals, performance feedback, and performance improvement. For example, the appraisal of employee performance can be used to inform performance feedback and defend the legal liability of the personnel decisions (Campbell & Wiernik, 2015). For these reasons listed above, organizations have a great interest in more practical ways of evaluating employee performance levels.

Performance Appraisal

Frequency of Performance Appraisal

Different organizations vary in their performance appraisal practices, including the frequencies ranging from annually to quarterly, preferably semi-annually or quarterly (Aguinis, 2013). Although informal evaluation and communication of the performance occur all year long, the scheduled formal meetings are necessary parts of established performance appraisal systems. This nature of the current appraisal systems challenges the view of dual aspects of employee performance, characterized by the stable and dynamic aspects, by failing to focus on the dynamic aspect of the performance.

Sources of Performance Data

According to Sonnentag and Frese (2012), the characteristics of the workplace require a more comprehensive perspective to capture the employee performance fully. Traditional employee performance predictions involve building on past employee characteristics (e.g., attributes, behaviors, and performance) to make predictions related to future performance. However, this approach does not address the importance of the workplace context and its effects on employee performance. Correspondingly, Ouellette and Wood (1998) also highlighted the importance of the context in their meta-analysis by concluding “past behavior was a significant predictor within stable but not unstable contexts” (p. 68). Additionally, the traditional sources of performance data fail to reflect the dynamic aspect due to their limitations. More specifically, Aguinis (2013) stated the most common performance data sources as supervisor ratings assuming they have the best knowledge about employees. Also, ratings from peers, subordinates, and customers are frequently used as complementary sources of information. Yet, since the context in most organizations can be dynamic, the ratings from these multiple sources are likely to be altered within the short period before the appraisal and are not suitable for showing the real-time changes. These characteristics of the organizational contexts make it imperative to capture a more comprehensive view of the employees’ performance. This requirement, in turn, creates a need for the development of new methods for analyzing the data from discrete sources and for further evaluations of the employees.

Focus of Performance Appraisal

Another practice that varies among different organizations is the focus of performance measurement which can take three approaches: trait appraisals, behavior appraisals, and results appraisals (Aguinis, 2013). A trait appraisal approach focuses on measuring employees' relatively stable traits, including cognitive abilities and personality characteristics. Even though it is one of the three approaches of the performance appraisal and might appear as a favorable approach in certain circumstances, this approach is rarely adopted in organizations due to disadvantages associated with it (i.e., rater biases and errors, harder to change characteristics and directing the attention more towards the person rather than the performance). As the name itself implies, the behavior approach focuses on observable indicators of relevant competencies in the form of displayed behaviors. This approach enables employees to be evaluated through their observable behaviors and get specific feedback on areas that need to be improved. Lastly, the results approach incorporates quantitative metrics that are objective and thought to be indicators of performance. However, employees do not have as much control as their behaviors on some outcomes (i.e., number of sales) and might perceive such measures insufficient in reflecting their overall performance.

Due to the different advantages and disadvantages offered by these performance appraisal approaches; organizations are most likely to select and combine approaches that are most suitable to the organizational context. Therefore, the inclusion of big data in congruence with the chosen approach would be a valuable source of information by introducing more objective and reliable data. Through the incorporation of sensory data (e.g., activity and interaction monitoring), text data (e.g., emails, documents, interviews with the supervisors), social media data (e.g., social media activity), and other types of big data (King et al., 2016), organizations can improve their practices regardless of the preferred approach. For example, sensory data can be used to track the employees' interactions to address any issues related to group work or the employees' workplace activity to make sure specific procedures are being followed (Macey & Fink, 2020). Additionally, the utilization of big data would greatly benefit from combining different sources for making decisions based on the most objective way possible. Since most human resource practices are expected to be legally defensible, the methods' objectivity in predicting employee performance has topmost importance in the development and application of these methods. Thus, the researchers try to develop the most accurate and easily interpretable methods for prediction. The study by Nedelcu et al. (2020) exemplified how a data mining method can perform as good as a manager's evaluation of an employee. Accordingly, the data mining methods were perceived negatively by the managers due to using inadequate parameters for the employees' performance prediction (Nedelcu et al. (2020).

Criteria of Performance Appraisal

The fourth point that might differ across organizations and impact performance-related decisions is the evaluation criteria used for the employee performance assessment. Particularly, most performance criteria are measured as continuous variables whereas treated as dichotomous variables. Usually, the dichotomy includes whether or not an employee meets a predefined standard. Congruently, Behrman and Perreault (1982) suggested employee performance is generally measured as a continuous variable with both the collection of objective and subjective data. However, for the sake of convenience, the results of the data are generally categorized under groups. For instance, employee performance results tend to be classed as either “Pass” or “Fail,” depending on the predefined standard. Other common categories in classifying employee performance include categories such as “Acceptable/Unacceptable,” “High performers/Low performers,” and “Behaving in a certain way/Not.” Farrington & Loeber (2000) stressed the advantage of such dichotomization that “it greatly simplifies the presentation of results and produces meaningful findings that are easily understandable to a wide audience.” (p.102). Thus, such classifications can leverage employee-related decision-making for human resources professionals. These classifications are easily made by different data mining methods which report the results in various categories. For example, some studies preferred to predict the employee performance for two different levels, whereas others used more classes by adding a class for average performers (Vijayalakshmi et al., 2020). In this way, the human resources professionals can easily interpret the employees’ performance level and take the necessary actions accordingly.

Prediction of Performance

The best interest of organizations lies in applying the best practices in performance management to ensure retaining the best performers in the organization and addressing any issues related to employee performance in the fastest way. Data mining methods allow organizations to use multiple sources of data to accurately predict future employee performance while considering different aspects of employee performance. More specifically, data mining methods aim to predict employee performance by identifying the most relevant attributes and guiding the decision-making accordingly. As mentioned previously, this will result in a comprehensive evaluation of the employee, which can provide a legal basis and serve as an indication of an objective process. On the contrary, more traditional performance assessment methods tend to rely on a more particular aspect of the employees' performance (e.g., task performance) at a specific point in time. This big data movement was grouped under the technology-enhanced assessment among newly introduced methods (Campbell & Wiernik, 2015). Furthermore, Illingworth et al. (2016) summarized the anticipated advantages of the incorporation of such big data methods as enabling employees to "monitor their behavior, evaluate progress toward performance goals, identify opportunities for improvement, and receive recommendations regarding additional training that is available to target specific performance deficiencies" (p.268). For these reasons, the recent technological advancement as a means to leverage employee performance data would also help the organizations make better decisions.

All in all, human resource operations require a great deal of effort in handling data from multiple sources and for different purposes. Using various data sources to improve the accuracy of predicting and classifying future employee performance levels would benefit all the stakeholders in the situation, including the organizations' decision-makers and the employees. As an example, organizations can make more sound decisions about their resources, such as investing in an employee or not. The accuracy of such decisions may impact the return of investment and the competitive advantage across many well-managed organizations. On the other hand, using these methods can enhance the employee perceptions of the fairness of the performance evaluation. Since such methods accommodate the human bias, the results appear more favorable to the employees. Therefore, it can be concluded that the introduction of the new data mining methods in organizations provides a promising way for achieving the best practices in the use of employee performance data.

Human Resource Analytics

Human resource processes can leverage the use of data from a variety of sources to improve the quality of decision-making with the introduction of recent technological advancements (van den Heuvel & Bondarouk, 2017). This progress has emerged under the concept of Human Resource Analytics that aims to "find the best path through a mass of data to uncover hidden value" (Fitz-enz & Mattox, 2014, p.4). According to Fitz-Enz & Mattox (2014), analytics can be further divided into three categories as descriptive, predictive, and prescriptive analytics. With its easiest form, descriptive analytics are used to collect organizational-related data to explore the areas of improvement and cost reduction. It can provide highly valuable information and direct attention to areas of the organization where interventions as solutions may be needed. It involves the usage of past data and their relationship with the current variables. For example, turnover and attrition rates, as well as the recruitment and selection budget, are usually investigated by descriptive analytics. The second form, predictive analytics, as the name itself implies, aims to make predictions. The past and current data are this time used congruently to assess the likelihood of future events. Finally, prescriptive analysis can be considered as the highest level of Human Resource Analytics that involves the demonstrations of possible future scenarios. It can be seen as an extra step beyond the predictive analysis by offering multiple decisions and demonstrating the possible future scenarios.

Performance evaluation was used as an example to illustrate the use of data mining methods for different types of analytics. For the first level of Human Resource Analytics, descriptive analytics, human resources professionals want to keep track of the employees' performance by collecting relevant information from the appraisals. For example, the average scores of the supervisor evaluations, the ratings received from the peer-evaluations, and the increased cost with the promotion decision can be recorded for the evaluations. Such information is used for defining the current situation by focusing on the levels of performance. One step further, predictive analytics aims to explain what will presumably happen by using predictive data mining methods. Sesil (2014) highlighted that the potential biases inherent in decision-making appear as challenges to traditional methods in performance evaluation. Correspondingly, the data mining methods draw researchers' attention because they allow for objective and comprehensive evaluation through learning the existent data and grounding the employee performance to the relevant attributes. Lastly, the prescriptive analysis focuses on the illustration of the possible future scenarios. In other words, by leveraging the use of advanced technological methods, including data mining, the goal is to determine the best course of action (Sesil, 2014) regarding the performance-related decisions as depicted in this case.

For this study, predictive analytics of employee performance in the context of I/O psychology was the primary focus. It is important to expand on the changing organizational practices with the introduction of big data and innovative methods to utilize such data. As stated by Han et al. (2012), "predictive mining tasks perform induction on the current data in order to make predictions" (p.15) which in turn appears as a promising way to improve organizational decision-making. On the contrary, traditional methods of predictive analytics in human resources generally include regression and linear methods, whereas current advancements in the field of computer science introduced more advanced methods, including nonlinear methods and machine learning/data mining methods (Fitz-enz & Mattox, 2014). Various researchers have implemented data mining and classification methods to predict and classify employees' performance (Strohmeier & Piazza, 2013). For example, Kirimi and Moturi (2016) built a classification model to predict the performances of public management development employees by using decision tree algorithms, whereas Asanbe et al. (2016) incorporated another data mining method, neural network, to predict the performance of teachers in a higher education institution. The results of those studies reported the accuracy of the methods as the indicator of the model's classification performance. However, the findings of the past studies showed mixed results in terms of the precision of the classification for each method and which method overperformed the others, specifically in the prediction and classification of the employee performance.

Data Mining

Predictive analytics in the context of I/O psychology adopt a variety of methods such as statistics, modeling, and data mining (Fitz-enz & Mattox, 2014). These terms are frequently used interchangeably without fully reflecting the true meanings, making it crucial to address the differences for this study. To start with, statistics can be defined as a collection of mathematical equations used to identify the outcome of the object's behaviors as grouping them under a target class by considering the probability distributions (Han et al., 2012). Moreover, predictive modeling utilizes statistical equations and data mining methods to forecast new or future scenarios (Shmueli, 2010). It should be noted that data mining incorporates different concepts, tools, and methods from a diverse body of domains, including computational areas like machine learning and database systems as well as statistics (Hand, 1999). Thus, data mining goes above and beyond the scope of traditional statistics, primary analysis (i.e., collection of data for a predetermined question (Hand, 1998)) by building on top of the existing methods through bringing more advanced computational methods into use and enable "secondary data analysis of large databases aimed at finding unsuspected relationships which are of interest" (p.112). Consequently, I/O psychology starts to incorporate the aforementioned technological advancements in computer science and statistics to study HRM issues and, in turn, identify the unexplored best practices towards studying human behavior in the workplace (Strohmeier & Piazza, 2013). Correspondingly, data mining methods have been recently used in psychology studies due to their promising solutions to the prediction and classification problems.

Data mining is the "process of automatically discovering useful information in large data repositories" (Tan et al., 2014, p.2). With the tools that it offers, the data mining methods aim to find useful patterns and predict a future outcome. For example, these methods are used to determine the most relevant attributes for good performers and predict the different levels of performance. By this means, the underlying patterns that impact the employee performance levels were explored, and a categorical class for the employee performance was provided. It is important to acknowledge that this study uses a specific type of predictive task as classification. Although there are other types of predictive tasks (i.e., regression), this study only focused on the meta-analysis of employee performance classification models. The details of the concept of classification and several types of classification models are presented below.

Classification

Classification tasks refer to assigning outcomes to predefined categories based on the predictors provided in the dataset. Correspondingly, the classification methods can be defined as “a systematic approach to building classification models from an input data set” (Tan et al., 2014, p.148). Generally, the success of these methods depends on the accuracy levels in which they can make predictions for that particular classification task. According to Tan et al. (2014), the accuracy of each classification can be calculated as follows:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (1)$$

Besides, there are other widely used performance metrics for assessing the models’ capabilities. These metrics can be listed as sensitivity, specificity, positive predictive value, negative predictive value, and confusion matrices. By taking these performance metrics of particular data mining methods, this study aims to evaluate the accuracy of different classification methods across studies and conclude which data mining classification method outperforms others in predicting employee performance in the organizational context. The literature review showed that artificial neural networks (ANNs), decision tree (DT), Naive Bayes, and support vector machines are the most commonly used classification methods in the studies on employee performance prediction/classification (Jantan et al., 2009). Also, some methods appear to be more favorable by their unique characteristics. By taking the commonality of use and the specific advantages associated into account, two data mining methods for classification were selected as the focal interest. These two methods were DTs and ANNs. Although they have different characteristics and offer unique advantages, these two methods apply to similar problems while producing comparable accuracy levels in their prediction capabilities (Mitchell, 1997). In consideration of this, these two methods were further compared in terms of effectiveness and drawbacks. The following sections explain and demonstrate each of the methods in detail.

Decision Tree

The DT is “a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf-node (or terminal node) holds a class label.” (Han et al., 2012, p.330).

As shown in Figure 1, the use of the DT method in classification involves testing a tuple starting from the root node and creating a path to a leaf node with the class label. In other words, a case begins to be tested from the very beginning of the flowchart and continues passing on predictors based on the cut-off points until it reaches the right end of the tree, its class (Oswald & Putka, 2016). The splits were used to separate the nodes into more precise sub-nodes to guide the path to the correct classification. On the other hand, the branches represented the possible options for the next decision. The ease of application and interpretation, the ability to handle multidimensional data, and the simplicity and fastness of the classification are the main reasons for its popularity (Han et al., 2012).

It should also be noted that there are different types of decision tree algorithms. The most well-known and widely used ones are iterative dichotomiser 3 (ID3) (Quinlan, 1986), C4.5 (Quinlan, 1993), and classification and regression tree (CART) (Breiman et al., 1984), respectively. Especially, ID3 and C4.5 algorithms are frequently used in the literature of Human Resource Predictive Analytics for human talent prediction (Jantan et al., 2010), employee selection (Rabcan et al., 2017), and employee performance prediction (Nasr et al., 2019). Moreover, Magesh et al. (2013) used the DT method to make promotion-decisions in an educational setting, whereas Kirimi and Moturi (2016) applied it to predict employee performance in the context of public service. In addition to these, with its promising classification performance, the decision tree method has been implemented for predicting target behaviors in many diverse areas of science, including information technology, education, biology, medicine, and human resources (Al-Radaideh & Al Nagi, 2012).

Artificial Neural Networks (ANNs)

According to Scarborough and Somers (2006), ANNs are “practical pattern recognition tools” (p.6) that are being implemented in various areas, including organizational behavior. A neural network is “a set of connected input/output units in which each connection has a weight associated with it.” (Han et al., 2012, p.398). Specifically, neurons create layers of nodes between the input (predictor) and output (criterion). The number of the nodes depends mainly on the designer as well as their connections and other characteristics such as the number of layers in between. The associated weight mentioned above is used to determine the threshold of a particular neuron, which indicates whether it becomes activated and transfers that activation to other nodes (Oswald & Putka, 2016). An illustration of a simple ANN is provided in Figure 2. The circles represent the neurons or nodes involved in the model whereas, the arrows indicate the information flow in between. Specifically, the input neurons receive the information about the attributes to transfer them to the next layer. With the activation of certain neurons, the processing of the input in the correct class requires the network to adapt the weight following the learning (Han et al., 2012). By doing that, ANNs can differ in terms of their training algorithm or learning rule. For example, if each input’s correct class is provided during the training to adapt the weight accordingly, such a network is called a supervised network.

Moreover, supervised networks can be further divided into different types such as backpropagation, Bayesian probabilistic, and Radial Basis Function (Scarborough & Somers, 2006). Among other supervised networks, backpropagation is found to be one of the most commonly used types which date back to the 1980s. Rumelhart et al. (1986) introduced backpropagation as a new learning approach that can overcome the challenges presented by earlier and simpler methods such as the perceptron-convergence procedure. Specifically, the interest in backpropagation was shaped due to the limitations introduced by Rosenblatt's (1958) study on this procedure as a simple neural network. The advancements of technology and recurring interest in ANNs have enabled overcoming the associated challenges and, eventually, developing more advanced methods that are capable of learning the linearly inseparable concepts (Han et al., 2012).

Secondly, if no such correct class for input is provided, it falls under an unsupervised network that primarily operates by trial and error during the learning phase (Scarborough & Somers, 2006). According to Han et al. (2012), one of the biggest advantages of ANNs is their suitability for inputs and outputs that are continuous. In addition to that, a comparison of the prediction and classification performances with other traditional statistical methods yields promising results for incorporating this method in organizational science (Scarborough & Somers, 2006). Collins & Clark (1993) emphasized the promising applications of ANNs in the workplace by reporting higher accuracy levels than regression models in the first study of ANNs in I/O psychology. Since then, it has been getting more and more popular with its promising solutions for predicting and classifying organizational behavior. To exemplify, Minbashian et al. (2010) compared ANNs with multiple regression in the context of personality and work performance. Jantan et al. (2010) also implemented ANNs and other data mining methods for public sector employees' performance prediction based on several attributes such as work outcome, knowledge and skill, and activities and contribution.

Based on the different characteristics that each method has, it is critical to choose a suitable method depending on the problem. Mitchell (1997) specified the most appropriate problems for each method by taking the general capabilities of the methods and characteristics of the context into account. Correspondingly, DTs are found to be most appropriate when the "instances are represented by attribute-value pairs," "the target function has discrete output values," and "the training data contain errors and/or missing attribute values" (Mitchell, 1997, p.54). These characteristics reflect the prevalence of DTs in organizational studies (Oswald et al., 2020), given the characteristics of predictors and criteria in workplace data. Correspondingly, the study by Kirimi & Moturi (2006) can be used as an example to demonstrate an instance with the assessment of specific attributes (i.e., qualification) and their respective values (i.e., Ph.D., Master's, Undergraduate). Also, the performance decision classes (i.e., meet expectation, exceed expectation, need improvement, outstanding, and do not meet minimum standards) show how the output is represented by discrete categories.

On the other hand, ANNs can handle real-valued inputs and outputs as well as other forms of outputs such as discrete values and vectors of these. While having distinctive characteristics compared to DTs, ANNs are also found to be suitable for similar problems described above for DTs. Accordingly, Mitchell (1997) concluded the accuracies of ANNs, and DTs are frequently similar. Therefore, this study chose to compare these two methods to provide insight into their classification performances.

Data Mining in Organizations

In recent years, the applications of different data mining methods have been widespread among many organizations for a variety of purposes. With the increasing popularity of data mining methods, it is important to understand these methods' potential benefits and limitations. According to Gonzalez et al. (2017), potential benefits include the ability to handle big data, increased efficiency in the costs, time, and workload whereas, the potential challenges consist of stakeholder reactions (i.e., employees, job applicants, general public), legal and ethical basis for the utilization of data, and the dependency of prediction success on the quality of data and the applied method. Moreover, changes in the collecting, storing, and analyzing data have favorable outcomes, including reducing the associated costs and saving employees' time and energy. Both scientific research and more applications are needed to maximize the effectiveness of the use of these methods. On the other hand, the potential limitations should be minimized and appropriately addressed to extend the applications. Correspondingly, Johnson & Verdicchio (2017) stated that many people are currently not familiar with such methods, which creates unrealistic perceptions of the applications, including unfairness and inaccuracy. Additionally, the legal and ethical concerns might impact both individuals and organizations. To overcome these concerns, organizations are expected to get consent from the employees to collect, use, and analyze their data. Correspondingly, an increasing number of organizations require compliance with specific regulations for ensuring the appropriate use of data mining methods in human resources processes (Gonzalez et al., 2017).

Overview and Goals of Current Study

The literature review revealed that there had been an increased number of attempts of leveraging data mining methods in organizational science on various topics, including the prediction of employee performance. Despite the existing and growing number of studies on the data mining methods used in employee performance prediction, there has been a lack of a systematic review or meta-analysis that compares these methods' performances in predicting employee performance. Due to this reason, this study aims to thoroughly review and summarize the applications of different data mining methods in the prediction and classification of employee performance in organizational settings. Studies on two frequently used classification methods, DTs and ANNs, were compared and presented to yield further clarification for the performances of different data mining methods in predicting employees' future performance.

.

CHAPTER THREE: METHODOLOGY

Study Selection

Search Strategy

Electronic databases including Web of Science, EBSCOHost, ProQuest, IEEE Xplore, and ScienceDirect were systematically searched to identify potential studies. Additionally, efforts were put into locating other possible studies that were not covered in the electronic databases. For this purpose, book chapters, theses, and dissertations were checked to identify additional studies. Specifically, ProQuest Dissertations & Theses Global was used to identify the relevant theses and dissertations, whereas Google Scholar was searched to identify any other sources that might not appear in the previous searches. Combinations and variations of the following keywords were used to guide the literature search: *data mining, machine learning, neural network, decision tree, prediction, classification, employee performance, job performance, performance appraisal, performance evaluation, personnel performance*. Moreover, the key researchers were contacted to obtain information of ongoing or unpublished studies in the field. However, no response was received. Lastly, additional studies were identified through backward and forward searches. The references of the previously found studies were checked to yield studies that have been missed from the previous literature search. On the other hand, more recent studies were identified through a forward search by checking the related studies' citing works.

Inclusion and Exclusion Criteria

The titles and abstracts of the identified studies were screened for the initial selection based on the following inclusion and exclusion criteria. Studies that used DTs and ANNs for employee data, reported contingency tables and accuracy measures (e.g., sensitivity, specificity, positive and negative predictive values) of the methods were selected for inclusion. Publications that use other data mining methods for classification and that are not available in the full-text form or English were excluded from the final selection of the studies. Studies with non-employee samples were also excluded since this study focuses on employee performance prediction. Lastly, the studies were excluded if they used hierarchical models, ensemble methods (e.g., Random Forest), or hybrid algorithms (e.g., Bayesian Neural Network) for prediction and classification. The purpose was to capture the actual effect size truly.

Study Quality Assessment

After determining the studies to be included in the meta-analysis, the study quality assessment score was used as a final criterion to make inclusion and exclusion decisions. To ensure the data provided by the studies are useful for the meta-analysis, each study was assessed based on the following questions developed by the author and the primary supervisor:

- Are the predictors of employee performance used in the classifier model clearly defined?
- Is the data mining method used for prediction and classification clearly stated?
- Are the performance levels predicted by the classifier explicitly provided?

- Is the effect size provided? If not, can it be calculated with the available accuracy measures?
- Are the findings generalizable to similar populations?

Each question was answered as one of the three options (i.e., “Yes,” “No,” and “Somewhat Yes”) based on the judgment of the author. Specifically, the scores for the associated options were determined as “Yes” =1, “No” =0, and “Somewhat Yes” =0.5 and were used to create an overall study quality score by adding them together. Correspondingly, the studies with a total score of 5 are considered “high quality,” whereas a total score smaller than 2 represents a “low quality” study. Any score between 2 to 5 is regarded as “medium quality” and found to be eligible for inclusion in the final analysis along with the “high quality” studies.

Coding procedure

Along with the inclusion criteria above, a coding scheme has been created to gather the necessary information from each selected study. This coding scheme includes the title of the study, name of the authors, publication year, source, data mining method(s) used for classification, the specific type of the occupation, sample size, attributes used to predict employee performance, number of attributes used in the prediction of performance, the classification type used (e.g., binary vs. multinomial) and, and the type of accuracy measures of employee performance prediction models (e.g., accuracy, specificity, sensitivity). The coding was done by the author under the supervision of the primary advisor.

Data Extraction

Data Collection

With the study selection, 103 studies were identified and retrieved for more detailed evaluation. For the meta-analytic investigation, eight articles were selected for the final analysis based on inclusion and exclusion criteria and quality assessment. Studies that used either DTs or ANNs or both methods were included in the analysis. Moreover, studies using one method on multiple samples were also collected.

Effect Sizes

A standardized form of effect size is required to measure the accuracy of various data mining methods across studies. After careful consideration of the different accuracy measures, odds ratios were found to be the most appropriate forms of effect sizes for dichotomous outcomes (Borenstein et al., 2009). Specifically, the odds ratio was selected as the effect size due to its usage as an indicator of the accuracy in classifications of binary outcomes. As suggested by Macaskill et al. (2010), the diagnostic odds ratios (DOR) can be calculated as follows:

$$DOR = \frac{True\ positives * True\ negatives}{False\ positives * False\ negatives} \tag{2}$$

Correspondingly, the odds ratios of each study were calculated based on reported contingency tables. The review of the literature pointed to differences in reporting the accuracy measures of different models. More specifically, the percentage rate was the most commonly used accuracy measure. Still, some studies provided other accuracy measures, including specificity, sensitivity, positive and negative predictive values, as well as the numbers of true positive, true negative, false positive, and false negative in contingency tables. Additionally, the confusion matrices of multiclass predictions (e.g., 3x3 matrices) were converted into 2x2 matrices to calculate the desired effect size. Due to the lack of necessary information for the calculation of effect sizes and the respective confidence intervals, some studies were further excluded. Specifically, studies that report confusion matrices were used in the final analysis. Also, a data extraction form was created to ensure the consistency of gathering the quantitative data. The author performed the data extraction with the supervision by the primary advisor.

Lastly, DORs range from 0 to infinity, with higher values indicating better prediction accuracy and values between 0 and 1 indicating poor prediction accuracy, as the odds for predicting the correct class of employee's performance are smaller than the odds for predicting the wrong class. The results are significant if the confidence intervals do not contain 1. In other words, the confidence interval and its overlap with the null value (i.e., no effect (DOR=1)) was used to indicate the statistical significance (Szumilas, 2010). DORs were converted into a logarithmic scale for analysis purpose and the results were converted back into the primary ratio forms for reporting purpose (Borenstein et al., 2009).

Statistical Methods

Meta-analytical Model

The statistical models that are most commonly used in the majority of the meta-analyses are the fixed-effect model and the random-effects model (Borenstein et al., 2009). As the name implies, the main underlying assumption behind the fixed-effect model is the existence of a true effect across studies that only varies by the sampling error. On the other hand, the random-effects model takes the variability of the true effect across studies into account, regardless of the sampling error. In other words, the effect sizes can vary depending on the other characteristics of the study, such as the attributes used for prediction purposes, different types of jobs, and additional factors that possibly differ across different studies. By taking the population of the studies included in the meta-analysis into consideration and following the recommendations suggested by Borenstein et al. (2009), a random-effects model was selected. Specifically, one of the most common methods in the random-effects model, the DerSimonian and Laird method (DerSimonian & Laird, 1986), was used for the analysis.

The variance of a random-effects model includes within-study and between-study variance (Borenstein et al., 2009). Also, the differences between studies require assigning different weights to cover the unique influences of each study without letting the overall effect to be highly influenced by the studies with large sample sizes. As Borenstein et al. (2009) stated, the appropriate weights for each study can be calculated through the inverse of its variance with the following formula:

$$w_i^* = \frac{1}{v_i^*} \tag{3}$$

Specifically, the variance (v_i^*) for a study involved in a random-effects model can be calculated as following:

$$v_i^* = v_i + \tau^2 \tag{4}$$

where v_i stands for the within-study variance whereas tau-squared (τ^2) represents the between-studies variance. The formula presented below is used to calculate the summary estimate by taking the associated weights of the studies into account and using the observed effect:

$$\frac{\Sigma w_i y_i}{\Sigma w_i^*} \tag{5}$$

Lastly, the Z-test for two independent samples was used to assess the statistical significance of the differences between overall DOR. More specifically, the difference between logs DORs was divided by the standard error of the difference.

In addition to the meta-analyses of DORs using a random-effects approach, the sensitivity, specificity, and Q* also provide valuable information in terms of a model's prediction performance. The investigation of these values is possible through a variety of ways such as the summary receiver operating characteristics (SROC) curve (Moses et al., 1993), the bivariate model (Reitsma et al., 2005), and the hierarchical summary receiver operating characteristics (HSROC) (Rutter & Gatsonis, 2001). More specifically, the bivariate model and the HSROC are considered as the hierarchical models as they offer mixed/multilevel analyses of the relationship between the sensitivity and specificity of the models. All of these three methods are characterized by different strengths and weaknesses. Specifically, the SROC is considered as a fixed-effects approach, whereas the bivariate model uses a random-effects approach that also takes the between study variation into account (Takwoingi et al., 2015). Correspondingly, the bivariate model was selected over the SROC for the studies using DTs, due to the characteristics of these studies such as the high heterogeneity and the number of studies included in the analysis (*k*= 7). It should be also noted that the bivariate and HSROC models have similar statistical properties in the absence of covariates (Harbord et al., 2007), creating a possibility to obtain the HSROC curve.

On the other hand, the SROC was preferred for the studies using ANNs due to the limited number of included studies and associated study characteristics (e.g., low heterogeneity). The hierarchical models (i.e., the bivariate model, HSROC) can be adversely impacted by the small number of studies (Dahabreh et al., 2012) and in turn can result in failure of parameter estimation (Takwoingi et al., 2015). Similarly, Cooper and Hedges (1994) recommended the use of fixed-effects models when the number of studies included do not exceed five. Accordingly, a random-effects meta-analysis was conducted for the DOR regardless of the low heterogeneity, whereas a fixed-effect selection was made, and presented SROC curve for studies using ANNs.

Test for Heterogeneity

Following the recommendations by Borenstein et al. (2009), the heterogeneity was assessed separately from the decision of a fixed or random-effects model. To estimate the heterogeneity of the effects across studies, the tau-squared (τ^2) and the percentage of variance among studies due to their heterogeneity (I^2) were reported. As suggested by Borenstein et al. (2009) τ^2 represents the between-study variance and indicates the distribution of the true effects. On the other hand, I^2 is used to demonstrate the variance among studies in relation to the total variance in effect sizes as a proportion (Card, 2012). Lastly, the Cochran's Q test (i.e., standard chi-squared test) is used to assess if the true effects sizes are identical across studies (Sutton et al., 2000). As such, τ^2 and I^2 were used instead of Cochran's Q test since it has low power of assessing heterogeneity (Gavaghan et al., 2000), especially for the meta-analyses with a small number of studies. To interpret the results, the heterogeneity was grouped into three levels based on values of I^2, with "High" for above 75%, "Moderate" for 50% - 75%, and "Low" for 25% - 50% respectively (Higgins et al., 2003).

Lastly, study subgroups were created to further investigate the changes in the overall effect size as a result of exclusion. Accordingly, the studies with certain characteristics (e.g., required zero-cell corrections, used promotion as the target class) were excluded for comparing the absence and presence of the associated characteristics. The goal was to detect any meaningful variances in the overall effect sizes that might be associated with the characteristics of these studies.

Test for Publication Bias

Previous studies used the funnel plots to assess the publication bias such that the symmetry of the plots around the mean effect size indicates an absence of publication bias (Borenstein et al., 2009). However, as suggested by Lau et al. (2006), the minimum number of studies required for using the funnel plot asymmetry as a test of publication bias is ten. Since the number of studies retained in the analyses was less than ten, the funnel plot asymmetry was not evaluated. Alternatively, Begg's Test (Beggs & Mazumdar, 1994) was used for the evaluation of studies using DTs. The publication bias of the studies using ANNs could not be evaluated due to the small number of studies retained in the analysis.

Analytical Software

The meta-analysis software developed specifically for the accuracy data, Meta-DiSc (Zamora et al., 2006), was used to conduct the quantitative analysis in combination with the MedCalc (MedCalc Software, Ostend, Belgium) and MetaDTA (Freeman et al., 2019).

CHAPTER FOUR: RESULTS

Search Results

Based on the inclusion criteria and the scores of the study quality assessment, eight studies with a total of 2430 employees were used in the meta-analysis. More specifically, these studies yielded seven pooled effect sizes for the studies of DT methods whereas and three pooled effect sizes for the studies of ANN methods. The studies included in the meta-analysis were marked with an asterisk in the references.

Meta-Analytical Results

The overall effect sizes for the random-effect models were calculated separately for the two data mining methods of interest, along with other summary statistics including sensitivity and specificity. Figures 3 – 5 are the forest plots for the DOR, sensitivity, and specificity analyses of the DTs. The figures show there are seven studies using DTs for employee performance prediction and the DOR is between 0.20 and 418.71, the sensitivity is between 0.75 and 1, and the specificity is between 0 and 0.97. The results of the DT methods for employee performance prediction showed an overall DOR of 82.20 (k=7, 95% CI [28.98, 233.15]). Since upper and lower ends of the confidence interval exceed 1, the odds of successfully predicting an employee’s performance level was significantly greater than the odds of unsuccessful prediction. In other words, the DTs have a significantly higher successful prediction rate of the employee performance compared to unsuccessful prediction rate. Additionally, the pooled sensitivity of the DT was 0.89 (95% CI [0.87, 0.91]), whereas the pooled specificity was 0.93 (95% CI [0.91, 0.95]).

Similarly, Figure 6 – 8 are the forest plots created for the DOR, sensitivity, and specificity analyses of the ANNs. The figures show there are two studies and three effects sizes using ANNs for employee performance prediction and the DOR is between 76.25 and 112.93, the sensitivity is between 0.87 and 0.91, and the specificity is 0.92. The results of the studies using ANNs for employee performance prediction demonstrated an overall DOR of 106.04 (k=2, %95 CI [67.76, 165.96]). The results were also found to be significant, as the confidence interval of the DOR did not contain 1, suggesting the effectiveness of the ANN method in predicting employee performance. In other words, ANNs were associated with significantly higher odds of successful prediction of the employee performance. The pooled sensitivity of the ANN was 0.90 (95% CI [0.87, 0.92]) and the pooled specificity was 0.92 (95% CI [0.89, 0.95]). Lastly, the result of the Z-test indicated the difference between the DTs and ANNs in terms of DORs were non-significant (z= 0.04, p= 0.9681, two-tailed). In the same vein, the confidence intervals overlap, supporting the statistical non-significance of the difference.

The results of the SROC curve and the HSROC curve also supported DTs and ANNs in predicting employee performance. Specifically, as shown in Figure 9, the HSROC curve for DTs produced by the bivariate model appeared to be close to the top left corner of the plot, representing effective prediction performances. Additionally, the summary estimate of the sensitivity and specificity, as shown with the blue square, is found within the 95% confidence interval that is shown in the area with dashed line. On the other hand, the 95% prediction interval is presented with the dotted line, indicating the range that new studies will fall into.

The SROC curve (Figure 10) demonstrated good prediction capabilities of ANNs for employee performance. Specifically, the SROC curve showed effectiveness of performance prediction with AUC= 0.9647 (**SE=** 0.0060) and index Q*=0.9115 (**SE=** 0.0092). The AUC is offered as a measure of model's performance and expected to have a value in between 0 and 1. Although the interpretation of the value depends on the research context, the rule-of-thumb evaluates a value closer to 1 as "excellent" prediction (Bradley, 1997). Besides, the Q* statistic is used to display the point in which the sensitivity and specificity values are equal and corresponds to the intersection of the line of identity and the curve. In addition to the AUC, higher values of Q*, representing higher sensitivity and specificity, can be interpreted as a better performance of the model (Jones & Athanasiou, 2005).

Test for Heterogeneity

The analysis of the studies using DTs yielded high heterogeneity, as the I^2 showed 85.46% of variation across studies due to heterogeneity. Such a value corresponded to "High heterogeneity" with the percentage being greater than 75% (Higgins et al., 2003) and required further investigation of the possible sources of heterogeneity. On the contrary, the analysis of the studies using ANNs showed "Low heterogeneity" (Higgins et al., 2003), as the I^2 was 0%. Additionally, the associated τ^2 values that reflect the true variance between studies (Borenstein et al., 2009) were also used to inform the evaluation of heterogeneity. Accordingly, the random-effects model analysis of the DTs resulted in τ^2= 1.3355. On the other hand, τ^2 was 0 in the analysis of ANNs that supported the homogeneity of the data.

Additionally, sensitivity analyses were conducted to understand the effects of certain studies on the results. The analyses were repeated after removing studies that required zero-cell corrections (e.g., Nedelcu et al., 2020; Rajesh, 2017) or used the promotion decision based on the employee performance as a target class (Nedelcu et al., 2020) from the DT dataset. Removing these studies had a moderate impact on the results such that the pooled DORs vary discernibly. Specifically, the pooled DOR became 120.06 (k=5, 95% CI [48.323, 298.32]) after removing the two studies required zero-cell corrections, showing a moderate increase. The difference can be explained by the poor performance of the DT method used in Rajesh (2017) study, having no true negative predictions. Similarly, Dahabreh et al. (2012) suggested that the zero-cell corrections can underestimate the accuracy of the model. On the other hand, the pooled DOR after removing the study by Nedelcu et al. (2020) was smaller compared to the previous analysis, with the value of 74.21 (k=6, 95% CI [24.810, 221.98]). This analysis resulted in a smaller difference with the original DORs, supporting the similarity of Nedelcu et al. (2020)'s effect size with other studies.

Test for Publication Bias

Following the recommendations of Lau et al. (2006), the publication bias of the DT studies was evaluated based on the Begg's Test using MedCalc (MedCalc Software, Ostend, Belgium). Accordingly, Kendall's-Tau was -0.2381 (p=0.4527), suggesting there was not significant publication bias.

CHAPTER FIVE: DISCUSSION

The purpose of this study is to evaluate the performance of the most commonly used data mining methods in predicting employee performance. Given the increased importance of performance management and the growing need for incorporating Human Resource Predictive Analytics in organizational decision making this study aims to elucidate the utility of data mining applications in one of the most crucial human resources processes. The rest of the chapter discusses the findings of the study, theoretical and practical implications, and limitations of the study, along with the recommendations for future research.

Findings

The results of the quantitative analysis revealed promising outcomes, with respect to the performance of DTs and ANNs in employee performance prediction. The pooled DORs, sensitivity and specificity values, and the results of the were in congruence with the previous studies of data mining methods. More specifically, the results suggested there were no statistically significant differences between the overall DORs of DTs and ANNs.

The heterogeneity statistics for the two data mining methods were in the two extremes, necessitating further explanations. As indicated by the I^2 value of 85.46, a high variation was apparent between the studies using DT methods. One potential explanation of such a high heterogeneity might be lying under the fact that different studies used different algorithms of DT. More specifically, the DT algorithms that were used included CART (Agaoglu, 2016; Mellisa, 2019), C4.5 (Kirimi & Moturi, 2016; Nedelcu et al., 2020; Rajesh, 2017; Wagh, 2010), and C5.0 (Agaoglu, 2016; Vijayalakshmi et al., 2020). The discrepancy on DT algorithms contributes to the variation across studies. However, due to the insufficient number of studies for the subgroup analysis, such an effect could not be covered by the present study. Additionally, taking the diverse set of employee samples working at different organizations into account, the thresholds for employees' performances are expected to vary. This implies that the heterogeneity is inherent among the studies, as a result of the diverse sample characteristics.

On the contrary, the heterogeneity for the studies using ANN was extremely small, as suggested by the $I^2 = 0$. This can be explained by the nature of data used for the analysis. More specifically, the two effect sizes out of the three were coming from the same study applying the same ANN method to two different samples. Accordingly, the heterogeneity was found to be lower because of the smaller between-study variances for two effect sizes derived from the same study.

Theoretical and Practical Implications

To the author's best knowledge, this is the first study of comparison of performances of different data mining algorithms in employee performance prediction. Thus, this study contributes to the extant literature on Human Resource Analytics. Besides organizational researchers, researchers from information systems, education, computer science, and engineering have also shown interest in the application of data mining methods in the organizational context. On the other hand, Gonzalez et al. (2017) highlighted the discrepancy between the frequency of data mining applications and implementations in organizations and the scholarly research. Such discrepancy indicates an emerging need to conduct more scientific studies to study the applications of data mining methods. With a more comprehensive understanding of the best practices for the application of DTs and ANNs for employee performance prediction, models with better performance can be developed to address organizations' needs.

Besides the theoretical implications above, the current study offers an overview of the current scientific developments of data mining methods to organizational practitioners. Since organizational practitioners are expected to use innovative methods including data mining methods to address practical problems of big data, it is very important to advance their understanding of such methods (Gonzalez et al., 2017). Having a deeper understanding of the underlying mechanisms used for the prediction would benefit practitioners in making associated decisions. For example, Oswald et al. (2020) highlighted the possibility of using a large number of variables in decision-making as an advantage, introduced by the big data movement in organizations. As can be seen from the current study, the data mining methods utilized many different variables to predict the employee performance, resulting in a more comprehensive evaluation. Additionally, new sources of data (e.g., sensory data and text data) can be incorporated into the analyses to unravel relationships that might be overlooked previously (Oswald et al., 2020). Correspondingly, this study could provide guidelines for organizational practitioners by summarizing one of the data mining application avenues.

Limitations and Future Research

There are certain limitations of the present study. First, the number of studies and effect sizes used in this meta-analysis is relatively small. There is not a specific required number of studies to conduct a meta-analysis (Borenstein et al., 2009). The nature of the topic of interest plays a role in this situation, since it involves a narrow, emerging field in the intersection of various disciplines. Furthermore, these disciplines differ in terms of their practices for presenting the results, creating missing data for the current analysis.

Secondly, the lack of information provided in the identified primary studies resulted in certain adjustments being made for the purposes of confidence interval calculations. More specifically, the confusion matrices for multiclass classification models were treated as binary classifiers by combining the values of different classes with using the associated weights. This adjustment was needed to obtain the true positive, true negative, false positive, and false negative values even though other performance measures for effect size calculation (i.e., sensitivity, specificity) are available in the study. Although the calculations took into the associated weights into account, grouping the different classes together was a potential drawback of this study due to the creation of confusion matrices differently for the studies with binary and multiclass outcomes.

Other than the small number of studies and effect sizes, just as all, this meta-analysis study was affected by the quality of the primary studies included in the analysis. Although the study quality assessment has been done prior to the inclusion of studies, it should be acknowledged that different studies might have different ways of applying data mining methods and conceptualizing the model that in turn introduce heterogeneity among studies. Additionally, the potential publication bias was assessed and appeared as a potential limitation. Correspondingly, there are some debates about the power of Begg's Test in detecting the biases. Sterne et al. (2000) stated that the Begg's Test does not provide adequate power in the assessment of publication bias in meta-analysis with few studies. Similarly, it is important to acknowledge that in the case of large between-study variability, the performance of Begg's Test suffers which is apparent in the analysis of DTs ($I^2 = 85.46$). It should be also noted that the alternative test proposed by Harbord et al. (2006) has similar power in highly heterogeneous cases.

Following the discussed search strategy, efforts were put towards identifying unpublished studies in the grey literature to mitigate the effects of publication bias on the study's results. As such, unpublished studies including theses and dissertations were considered. After the application of exclusion/inclusion criteria, one unpublished study was included in the analysis. For these reasons, it is expected and recommended that the future research would be able to identify more studies in the topic and minimize the effects of publication bias to provide more comprehensive conclusions.

The selection of the effect sizes for the studies using multiple DT algorithms should be also noted as another limitation of this study. There were instances of multiple effect sizes from the same study that required a systematic approach to selecting the appropriate effect sizes for the analysis. In such cases, the algorithm with the highest accuracy rate was selected for inclusion as an effect size. A single effect size was then chosen from the studies that report multiple effect sizes based on the same sample. On the other hand, if two effect sizes were based on different samples, both were used in the analysis. For example, in the analysis of studies using ANNs, two of the three effect sizes belong to the same algorithm of the same study, but on a different sample. Correspondingly, only taking the best performing algorithms' effect sizes might result in different conclusions, suggesting a need for combining the dependent effect sizes in the future research. After the combination of multiple effect sizes, the average effect size can be determined and used to evaluate the performances of different algorithms as potential moderators.

Additionally, the selection of the comparison methods should be noted as a limitation of this study. The meta-analyses of DORs were conducted using random-effects models and were not affected by the results of the heterogeneity tests. On the other side, the meta-analysis of studies using ANNs used the fixed-effects approach due to the limited number of studies. Specifically, the SROC developed by Moses et al. (1993) was used to quantify the performance of the ANNs. The fixed-effects approach overlooked the heterogeneity between studies (Takwoingi et al., 2015), creating a potential drawback in the interpretation of the results. With the availability of more studies in future, a hierarchical approach could be used to overcome the issues inherent in the SROC curves (Takwoingi et al., 2015).

Lastly, the use of a bivariate model as a hierarchical approach yielded valuable and more robust information regarding the performance of studies using DTs. As the sensitivity and specificity were accounted separately in the analysis, the results were expected to reflect the within- and between-study variation. The findings were provided as the summary estimate and corresponding confidence and prediction areas, together with the HSROC curve. However, due to the limitations of available software, the summary estimate score and other associated values (e.g., AUC) for the curve could not be calculated. It is recommended that future research use more advanced statistical softwares to derive the respective values from the parameters provided. Additionally, the AUCs of HSROC curves can be compared with the inclusion of more studies with ANNs, that results in using a hierarchical model.

Conclusion

Overall, this study aimed to serve as a first step in integrating the extant literature on data mining methods in employee performance prediction. The results provided evidence for the promising predictive capabilities of two most commonly used data mining methods in organizational settings, DT and ANN, with capturing the diverse sources of employee data to predict future performance levels. The two most commonly applied data mining methods for employee performance prediction were compared, and the results suggested no statistical difference between the overall DORs of DTs and ANNs. Nevertheless, both of the methods showed good performance capabilities, as can be seen from the associated DORs, sensitivity and specificity values, and SROC or HSROC curves. As with all meta-analyses, there were certain limitations of the study which necessitates the importance of interpreting the results with precautions. Directions of future research that advance the applications of various data mining methods to improve the human resource decision making were also discussed.

APPENDIX: SUMMARY OF THE STATISTICAL ANALYSES

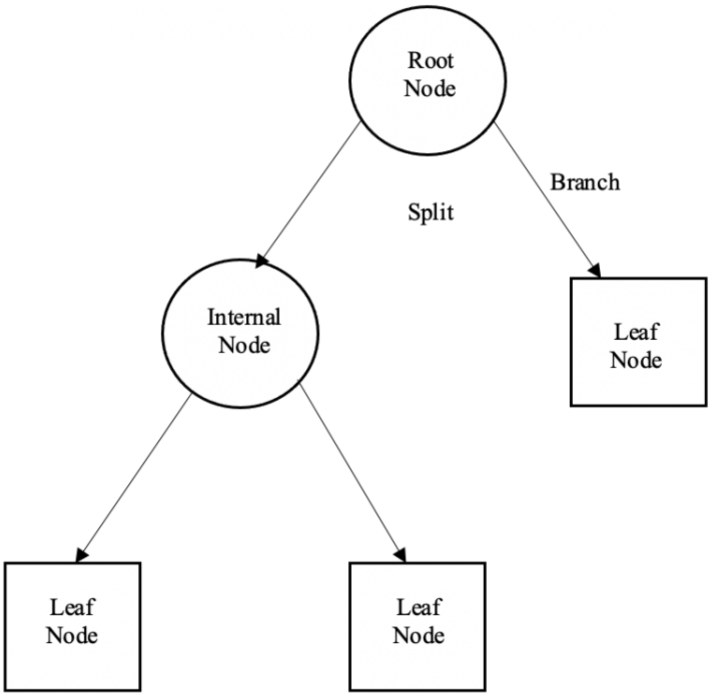

Figure 1: An Example of Simple Decision Tree

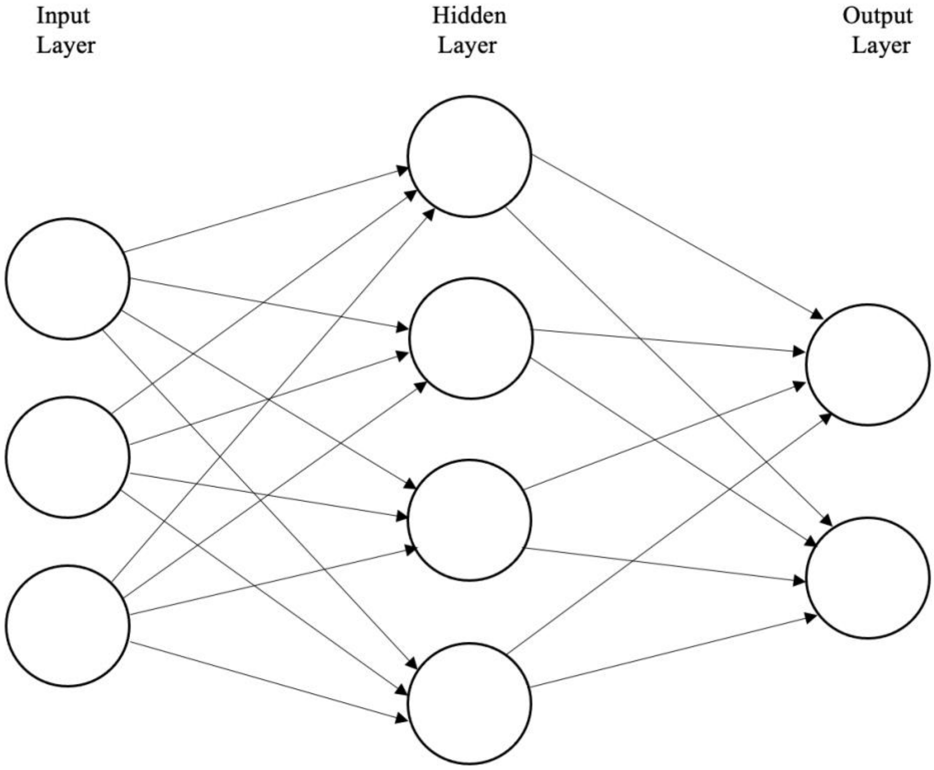

Figure 2: An Example of Simple Artificial Neural Network

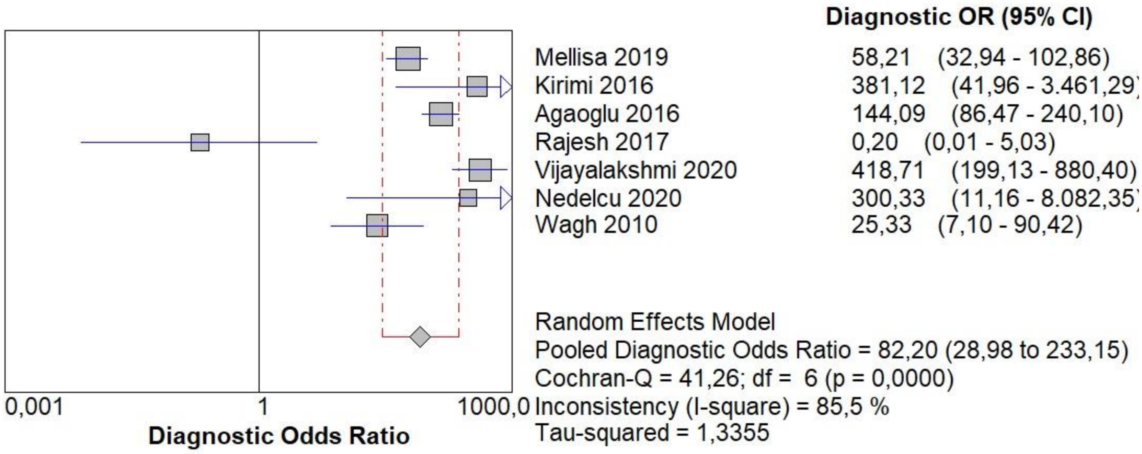

Figure 3: Forest Plot for Diagnostic Odds Ratio of Decision Tree Studies

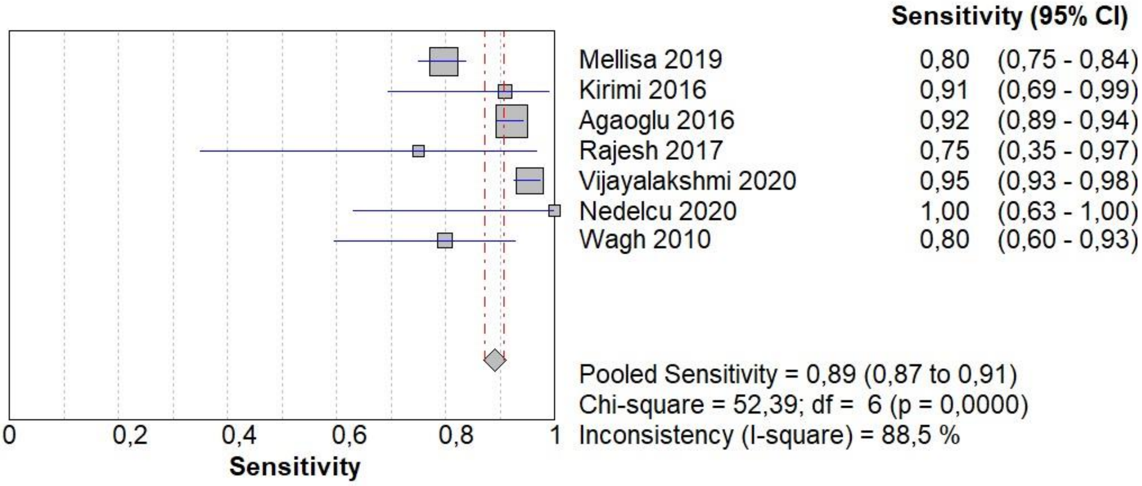

Figure 4: Forest Plot for Sensitivity of Decision Tree Studies

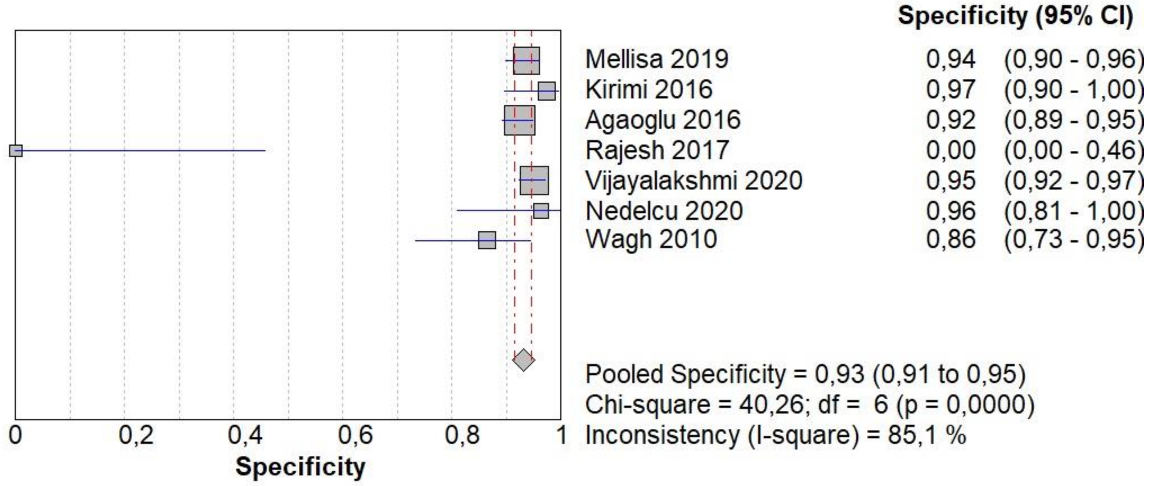

Figure 5: Forest Plot for Specificity of Decision Tree Studies

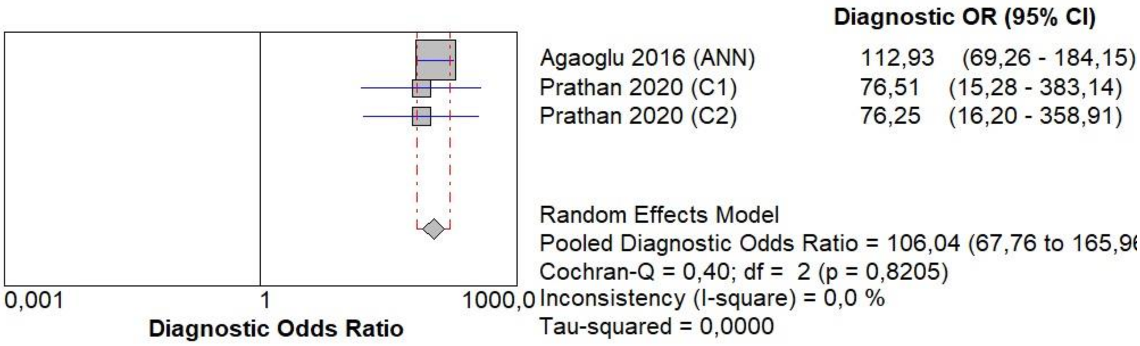

Figure 6: Forest Plot for Diagnostic Odds Ratio of Artificial Neural Network Studies

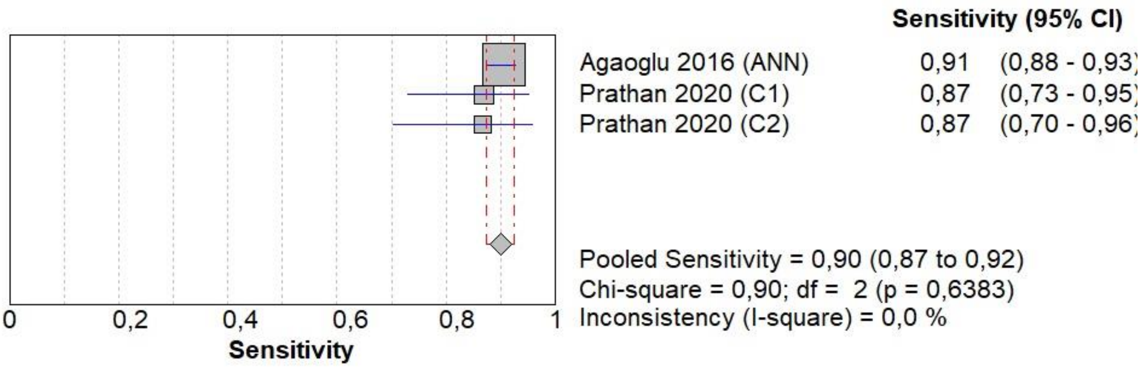

Figure 7: Forest Plot for Sensitivity of Artificial Neural Network Studies

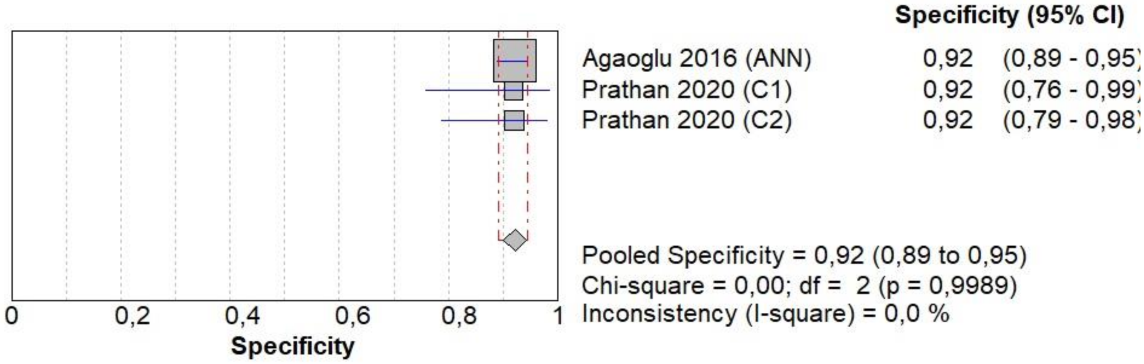

Figure 8: Forest Plot for Specificity of Artificial Neural Network Studies

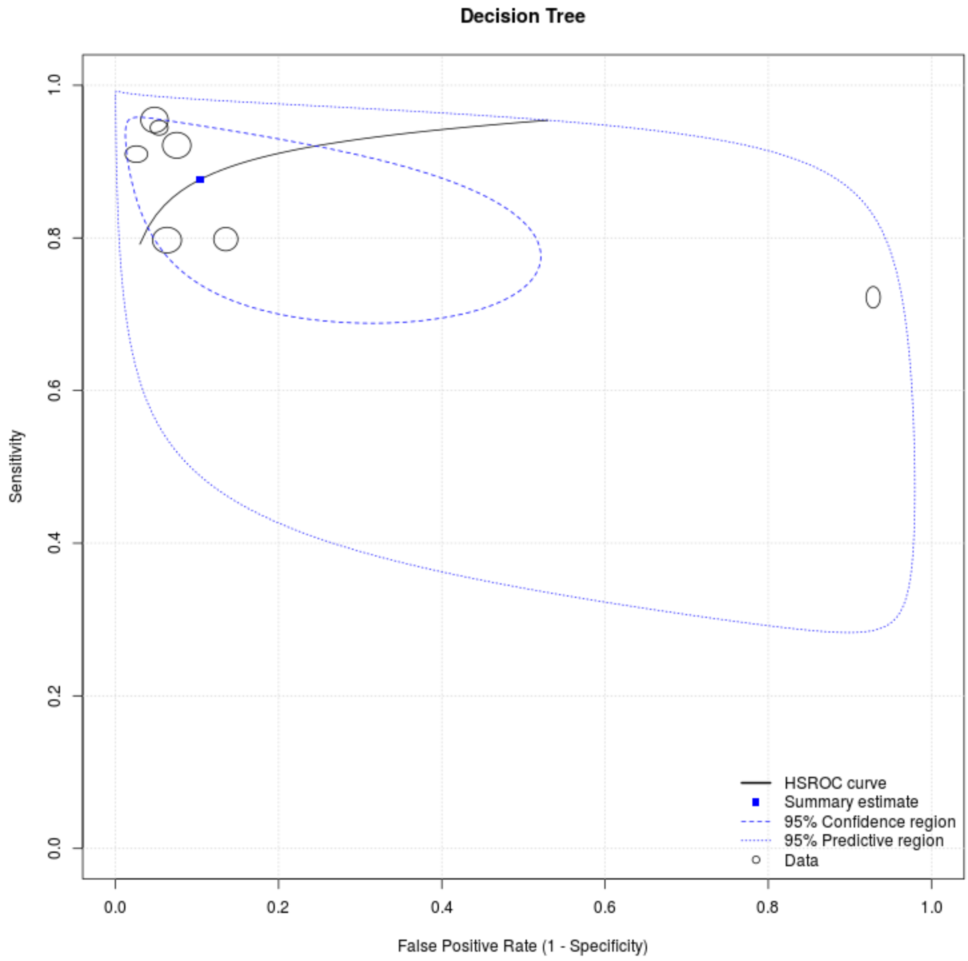

Figure 9: HSROC Plot for Decision Tree Studies

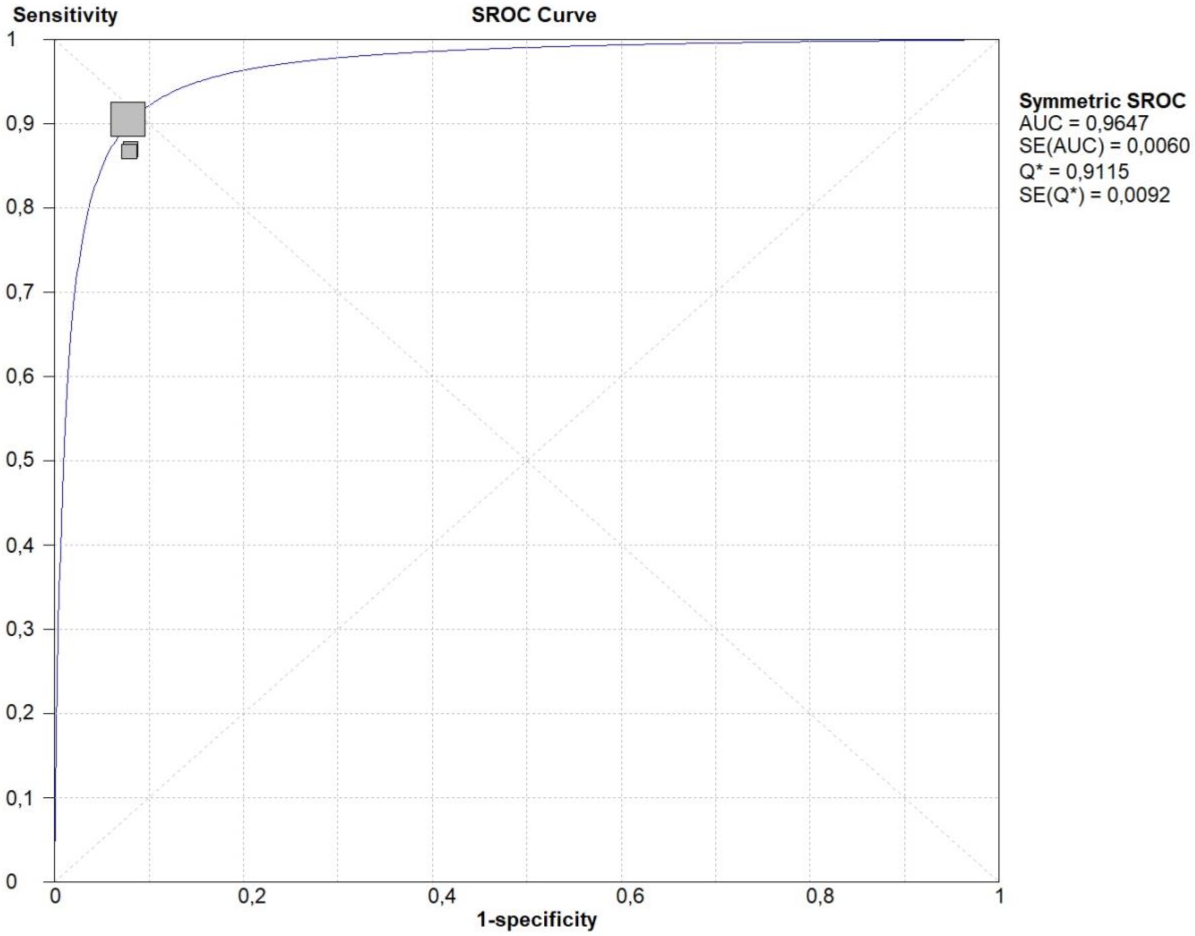

Figure 10: SROC Plot for Artificial Neural Network Studies

REFERENCES

Aguinis, H. (2013). *Performance Management* (3rd ed.). Pearson.

*Agaoglu, M. (2016). Predicting instructor performance using data mining techniques in higher education. *IEEE Access, 4*, 2379-2387.

Al-Radaideh, Q.A., & Al-Nagi, E. (2012). Using data mining techniques to build a classification model for predicting employee's performance. *International Journal of Advanced Computer Science and Applications, 3*(2), 144 – 151.

Asanbe, M. O., Osofisan, A. O., & William, W. F. (2016). Teachers' performance evaluation in higher educational institution using data mining technique. *International Journal of Applied Information Systems*, *10*(7), 10-15.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101.

Behrman, D. N., & Perreault, W. D., Jr. (1982). Measuring the performance of industrial salespersons. *Journal of Business Research, 70*, 355-370.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*, 1145-59.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Routledge.

Campbell, J. P., & Wiernik, B. M. (2015). The modeling and assessment of work performance. *Annual Review of Organizational Psychology and Organizational Behavior, 2*, 47–74. http://doi.org/10.1146/annurev-orgpsych-032414-111427

Card, N. A. (2012). *Applied Meta-Analysis for Social Science Research.* The Guilford Press.

Collins, J. M., & Clark, M. R. (1993). An application of the theory of neural computation to the prediction of workplace behavior: An illustration and assessment of network analysis. *Personnel Psychology, 46*(3), 503–524. https://doi.org/10.1111/j.1744-6570.1993.tb00882.x

Cooper, H., & Hedges, L. (1994). *Handbook of Research Synthesis, The*. Russell Sage Foundation.

Dahabreh, I. J., Trikalinos, T. A., Lau, J., & Schmid, C. (2012). An empirical assessment of bivariate methods for meta-analysis of test accuracy. *Agency for Healthcare Research and Quality*, *12*(13).

DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*, 177–188.

Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences*, *22*(4), 387-397.

Farrington, D. P., & Loeber, R. (2000). Some benefits of dichotomization in psychiatric and criminological research. *Criminal Behaviour and Mental Health*, *10*(2), 100–122. https://doi.org/10.1002/cbm.349

Fitz-Enz, J., & Mattox, J., II. (2014). *Predictive analytics for human resources*. Wiley.

Freeman, S. C., Kerby, C. R., Patel, A., Cooper, N. J., Quinn, T., & Sutton, A. (2019). Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA. *BMC Medical Research Methodology, 19*. https://doi.org/10.1186/s12874-019-0724-x

Gavaghan, D. J., Moore, A. R., & McQuay, H. J. (2000). An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. *Pain, 85*(3), 415–424. https://doi.org/10.1016/S0304-3959(99)00302-4

Gonzalez, M. F., Capman, J. F., Oswald, F. L., Theys, E. R., & Tomczak, D. L. (2019). “Where’s the I-O?” Artificial intelligence and machine learning in talent management systems. *Personnel Assessment and Decisions, 5*(3), 33-44.

Han, J., Kamber, M. & Pei, J. (2012). *Data mining concepts and techniques* (3rd ed.). Morgan Kaufmann Publishers

Hand, D. J. (1998). Data mining: Statistics and more?. *The American Statistician, 52*(2), 112-118.

Hand, D. J. (1999). Statistics and data mining: Intersecting disciplines. *ACM SIGKDD Explorations, 1*(1), 16-19.

Harbord, R. M., Egger, M., Sterne, J. A. C. (2006). A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statist. Med., 25*, 3443–3457.

Harbord, R. M., Deeks, J. J., Egger, M., Whiting, P., & Sterne, J. A. C. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics, 8*(2), 239-251.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical research ed.), 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Illingworth, A. J., Lippstreu, M., & Deprez-Sims, A. (2016). Big data in talent selection and assessment. In S. Tonidandel, E. B. King & J. M. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology* (pp. 242-276). Routledge/Taylor & Francis Group.

Jantan, H., Hamdan, A. R., & Othman, Z. A. (2009). Potential data mining classification techniques for academic talent forecasting. In *Ninth International Conference on Intelligent Systems Design and Applications* (pp. 1173–1178). http://dx.doi.org/10.1109/ISDA.2009.64.

Jantan, H., Puteh, M., Hamdan, A. R., & Othman, Z. A. (2010). Applying data mining classification techniques for employee's performance prediction. In *Knowledge Management 5th International Conference* (pp. 645–652).

Johnson, D. G., & Verdicchio, M. (2017). Reframing AI discourse. *Minds & Machines, 27,* 575-590.

Jones, C. M., & Athanasiou, T. (2005). Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests. *The Annals of Thoracic Surgery*, *79*(1), 16–20. https://doi.org/10.1016/j.athoracsur.2004.09.040

King, E. B., Tonidandel, S., Cortina, J. M., & Fink, A. A. (2016). Building understanding of the data science revolution and I-O psychology. In S. Tonidandel, E. B. King & J. M. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology* (pp. 1-18). Routledge/Taylor & Francis Group.

*Kirimi, J. M., & Moturi, C. A. (2016). Application of data mining classification in employee performance prediction. *International Journal of Computer Applications*, *146*(7), 28-35.

Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. *META Group Research Note.*

Lau, J., Ioannidis, J. P. A, Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *BMJ (Clinical research ed.), 333*, 597-600. 10.1136/bmj.333.7568.597.

Macaskill, P., Gatsonis, C., Deeks, J. J, Harbord, R. M., & Takwoingi, Y. (2010). Analysing and presenting results. In J. J. Deeks, P. M. Bossuyt, & C. Gatsonis (Eds.), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. The Cochrane Collaboration.

Macey, W. H., & Fink, A. A. (2020). *Employee surveys and sensing.* Oxford University Press.

Magesh, N., Thangaraj, P., Sivagobika, S., Praba, S., Mohana Priya, R. (2013). Evaluating the performance of an employee using decision tree algorithm. *International Journal of Engineering Research & Technology, 2*(4), 2814-2830.

MedCalc Software. (2021). *MedCalc for Windows.*[Computer Software]. Retrieved from https://www.medcalc.org/

*Mellisa, I. (2019). Building data mining decision tree model for predicting employee performance. *Journal of Applied Information, Communication and Technology, 6*(2), 75-86.

Minbashian, A., Bright, J. E. H., & Bird, K. D. (2010). A comparison of artificial neural networks and multiple regression in the context of research on personality and work performance. *Organizational Research Methods, 13*(3), 540–561. https://doi.org/10.1177/1094428109335658

Mishra, S. N., Lama, D.R., & Pal, Y. (2016). Human resource predictive analytics (HRPA) for HR management in organizations. *International Journal of Scientific & Technology Research, 5*(5), 33-35.

Mitchell, T. M. (1997). *Machine learning.* McGraw-Hill.

Moses, L. E., Shapiro, D., & Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med, 12*(14):1293-1316.

Nasr, M. M., Shaaban, E., & Samir, A. (2019). A proposed model for predicting employees' performance using data mining techniques: Egyptian case study. *International Journal of Computer Science and Information Security, 17*(1), 31-40.

*Nedelcu, A., Nedelcu, B., Sgarciu, A. I., Sgarciu, V. (2020). Data mining techniques for employee evaluation. *2020 12th International Conference on Electronics, Computers and Artificial Intelligence, Romania.*

Oswald, F. L., & Putka, D. J. (2016). Statistical methods for big data: A scenic tour. In S. Tonidandel, E. B. King & J. M. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology*. Routledge/Taylor & Francis Group.

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in industrial-organizational psychology and human resource management: Forward progress for organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*, 505-533.

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin, 124*, 54-74.

*Prathan, S., & Ow, S. H. (2020). Determining the best-fit programmers using Bayes' theorem and artificial neural network. *IET Software*, *14*(4), 433-443.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81–106. https://doi.org/10.1007/BF00116251

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* Morgan Kaufmann Publishers

Rabcan, J., Vaclavkova, M., & Blasko, R. (2017). Selection of appropriate candidates for a type position using C4.5 decision tree. In *the International Conference on Information and Digital Technologies, Zilina* (pp. 332-338). doi: 10.1109/DT.2017.8024318.

*Rajesh, A. (2017). Data mining classification of teaching evaluation using WEKA. *International Journal of Latest Trends in Engineering and Technology, 9*(1), 82-86.

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, *58*(10), 982-990.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386–408. https://doi.org/10.1037/h0042519

Rumelhart, D., Hinton, G. & Williams, R. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536. https://doi.org/10.1038/323533a0

Rutter, C. M., Gatsonis, C. A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*, *20*(19), 2865-2884.

Ryan, J., & Herleman, H. (2016). A big data platform for workforce analytics. In S. Tonidandel, E. B. King & J. M. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology* (pp. 36-59). Routledge/Taylor & Francis Group.

Scarborough, D., & Somers, M. J. (2006). *Neural networks in organizational research: Applying pattern recognition to the analysis of organizational behavior.* APA Books.

Sesil, J. C. (2014). *Applying advanced analytics to HR management decisions: Methods for selection, developing incentives, and improving collaboration.* Pearson.

Shmueli, G. (2010). To explain or to predict?. *Statistical Science, 25*(3), 289-310.

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*(11), 1119–1129. https://doi.org/10.1016/s0895-4356(00)00242-0

Strohmeier, S., & Piazza, F. (2013). Domain-driven data mining in human resource management: A review of current research. *Expert Systems with Applications, 40*(7), 2410-2420.

Sonnentag, S., & Frese, M. (2012). Dynamic performance. In S. W. J. Kozlowski (Ed.), *Oxford library of psychology. The Oxford handbook of organizational psychology* (Vol.1). (pp. 548–575). Oxford University Press.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A, & Song, F. (2000). Methods for meta-analysis in medical research. *Wiley*.

Szumilas M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *19*(3), 227–229.

Takwoingi, Y., Guo, B., Riley, R. D., & Deeks, J. J. (2015). Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Statistical Methods in Medical Research*, *26*(4), 1896–1911. https://doi.org/10.1177/0962280215592269

Tan, P. N., Steinbach, M., & Kumar, V. (2014). *Introduction to data mining*. Pearson.

van den Heuvel, S., & Bondarouk, T. (2017). The rise (and fall?) of HR analytics: A study into the future application, value, structure, and system support. *Journal of Organizational Effectiveness: People and Performance, 4*, 157-178.

*Vijayalakshmi, V., Panimalar, K., & Janarthanan, S. (2020). Predicting the performance of instructors using machine learning algorithms. High Technology Letters, 26(12), 694-705.

*Wagh, R. S. (2010). Employee performance prediction model. [Unpublished master's thesis. Christ University.

Zamora, J., Abraira, V., Muriel, A., Khan, K., & Coomarasamy, A. (2006). Meta-Disc: A software for meta-analysis of test accuracy data. *BMC Medical Research Methodology, 6*(31).