Tilburg University

PREDICTING JOB SATISFACTION USING LOGISTIC REGRESSION, SUPPORT VECTOR MACHINE AND RANDOM FOREST MODELS

ELSBETH NIEUWKAMP

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

u636084

COMMITTEE

dr. Murat Kirtay
dr. Michal Klincewicz

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

May 19th, 2023

WORD COUNT

7987

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor dr. Murat Kirtay for his feedback and assistance throughout writing this thesis. I would also like to thank dr. Michal Klincewicz for his feedback.

PREDICTING JOB SATISFACTION USING LOGISTIC REGRESSION, SUPPORT VECTOR MACHINE AND RANDOM FOREST MODELS

ELSBETH NIEUWKAMP

Abstract

It is crucial for organizations to proactively stimulate job satisfaction among their employees. This not only enhances organizational success, but also yields societal benefits, including improvements in health and productivity which ultimately boosts the economic performance of a country. Recently, researchers have started to focus on job satisfaction prediction using supervised machine learning models. However, empirical studies on this topic remain limited and there is a strong demand for more research that combines Human Resource Management topics and machine learning. To fill this gap, this study will compare three machine learning models, namely a logistic regression, support vector machine and random forest model to find out which best predicts job satisfaction. In addition, this study aims to get a better understanding of the importance of the individual predictors in the models by use of the feature ablation method. Relationship satisfaction, environment satisfaction, overtime, job involvement, work-life balance, age and gender have been previously identified in the literature as being related to job satisfaction. Therefore, these features are included as predictors in the models. Data is obtained from Kaggle and consists of employee data. Findings indicate that the logistic regression model performs the best in predicting job satisfaction in terms of accuracy score. Relationship satisfaction, work-life balance, age and gender have been shown to be the most important predictors among the three models. As the performance of the three machine learning models are only slightly higher than chance level, findings should be interpreted with caution. It is concluded that future research that replicates this study is needed to utilize the findings in practice. Moreover, this study serves as a foundation for future studies to build upon.

Keywords: *job satisfaction, machine learning, classification*

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

For this thesis, data has been used from the data scientists platform Kaggle. The data is owned by Sripad Karthik and is anonymous. Furthermore, no data has been obtained from animals or humans. The original owner of the data and code remains the owner of the data and code during and after this thesis is finished. All images (Appendix A (page 37), Appendix C (page 39) and Appendix D (page 40)) were created by me. The code created for this thesis can be viewed via Github. Parts of the code were inspired on, adapted and reused from Kaggle, stack exchange, stack overflow and scikit-learn. More specific sources are indicated in the code. The generative language model 'ChatGPT' and 'Deepl Translator' were, after finishing my thesis, used for checking spelling and grammar. They were not used for generating text or typesetting text.

2 INTRODUCTION

Human Resource Management is the process of managing committed and qualified employees to achieve organizational objectives (Hecklau et al., 2016). With the emergence of Human Resource analytics, which involves analyzing employee data to make data-driven decisions in Human Resource departments, Human Resource Management has changed (Jain et al., 2021; Jones, 2014; Marler & Boudreau, 2017). In recent years, a shift has been going on from relying on intuition to using data-driven decision making which is accompanied by the help of machine learning techniques (Choi & Choi, 2022; Garg et al., 2022). This transformation is essential for an organizations success, since it enables them to gain valuable Human Resource insights that would otherwise have been lost (Choi & Choi, 2022). One insight that could be obtained by Human Resource analytics is regarding job satisfaction (Tomar & Gaur, 2020). Recently, there has been a renewed interest in job satisfaction since scholars started focusing on its prediction by use of supervised machine learning methods (e.g. Choi & Choi, 2022; Jain et al., 2021; Moro et al., 2021). However, since research about Human Resource analytics is still in its infancy there is much more to explore (Marler & Boudreau, 2017).

From a theoretical perspective, there are only a few empirical studies that analyzed the job satisfaction of employees by use of machine learning models. In addition, Garg et al. (2022) state that there is insufficient research that integrates machine learning with Human Resource Management. To fill this gap, this study will focus on both machine learning and job satisfaction. Various studies, for example from Choi and Choi (2022) and Moro et al. (2021), utilize different machine learning models to predict job satisfaction, indicating that there is not yet a single optimal model that provides a satisfactory solution in the literature. This study contributes to the literature in two ways. First, this research will compare three machine learning models, namely a logistic regression model, support vector machine model and a random forest model to find out which of them has the highest performance in predicting job satisfaction. Therefore, this study will ensure a valuable contribution to the existing literature by using a unique set of predictors and a distinct dataset, distinguishing it from previous studies. This will expand knowledge in the Human Resource field. Second, this research aims to get a better understanding of the predictors of job satisfaction using prediction modeling.

From a societal perspective, job satisfaction has been shown to have a direct relationship with people's subjective well-being (Judge & Kammeyer-Mueller, 2011). This means that employees who are satisfied with their job experience positive affect (i.e. pleasurable feelings), happiness and satisfaction in their daily lives (Harmon-Jones & Harmon-Jones, 2021; Judge & Kammeyer-Mueller, 2011). In turn, these people are more likely to find their lives meaningful, help others in society and enjoy health benefits (Isen, 2001; Judge & Kammeyer-Mueller, 2011; King et al., 2006). Furthermore, research suggests that job satisfaction and subjective well-being result in higher productivity, thereby improving the economic performance of a country in terms of gross domestic product (DiMaria et al., 2020; Judge & Kammeyer-Mueller, 2011; Nanda & Browne, 1977).

From a practical stance, job satisfaction is a key factor for an organization's success, growth and competitive advantage (Jung & Suh, 2019; Prajogo & Cooper, 2010). As job satisfaction is related to employee performance, low absenteeism rates and low turnover rates, it is important for organizations to ensure that their workforce is satisfied with their job (Koh & Boo, 2001). To achieve a satisfied workforce, organizations must have an understanding of the model that can best predict job satisfaction using available predictors. When these predictors are identified, the Human Resource department can maintain or modify employee policies and take proactive measures to ensure job satisfaction among employees (Choi & Choi, 2021; Rustam et al., 2021).

From a technological perspective, the machine learning models used in this study can be built on relatively simple and cheap devices. Also, only open source libraries are used. This is an advantage, since it is easy and cheap for Human Resource departments to implement these models, which can help them predict the job satisfaction of their employees.

Since comparable studies that predict job satisfaction showed that logistic regression, support vector machine and random forest models performed well (Jain et al., 2021; Rustam et al., 2021), these are utilized in this study. Jain et al. (2021) and Rustam et al. (2021) already compared several machine learning models in their studies. However, they used different predictors and a different dataset compared to those used in this study. Building on this idea, the main research question has been formulated:

> *Which of the following machine learning models, a support vector machine-, logistic regression- or random forest model, best classifies job satisfaction in terms of accuracy score?*

As relationship satisfaction, environment satisfaction, overtime, job involvement, work-life balance, age and gender have been shown to be positively related to job satisfaction, they are included as predictors in this study (e.g. Agbozo et al., 2017; Gopinath & Kalpana, 2020). Investigating which feature contributes the most to predicting job satisfaction is, as earlier mentioned, of interest to organizations and will therefore be examined in this study as well. All findings will be compared to existing literature. Therefore, the sub-questions are as follows:

Sub-RQ1 *Which predictors are most important in predicting job satisfaction?*

Sub-RQ2 *What can be concluded about the accuracy scores and contribution of the findings (in relation to existing research and literature)?*

Findings indicate that the logistic regression model performs the best in predicting job satisfaction compared to the other two models. Relationship satisfaction, work-life balance, age and gender have been shown to be important predictors. The findings should be interpreted with caution, as the machine learning models' performance is only slightly better than chance level.

3 RELATED WORK

Since the emergence of job satisfaction in the literature, it has been investigated extensively (Fida et al., 2019). Consequently, job satisfaction has been defined in several ways over these past years. For example, Aziri (2011) describes job satisfaction as the positive attitude of an employee about the job and Vroom (1964) as an employees' positive orientation towards all aspects of the job. Locke (1976) defines job satisfaction as 'the positive and pleasant affective state, which an individual hold about his or her job' (Locke, 1976 as cited in Zhu, 2013, p. 294). Locke (1976) his definition is often referenced in literature and widely accepted (Esfandiari & Kamali, 2016).

There are many antecedents of job satisfaction found in the literature. Some examples will be given, but it must be noted that the antecedents mentioned here are far from being exhaustive. For example, a meta-analysis from Saber (2014) shows that among others, age, autonomy, empowerment, organizational commitment and wages are common predictors of job satisfaction. A meta-analysis from Judge and Bono (2001) shows that the predictors self-esteem, self-efficacy, internal locus of control and emotional stability are positively related to job satisfaction. Since Jain et al. (2021) already compared machine learning models with internal marketing strategies (i.e. salary, recognition, appraisal, reward and promotion) as predictors of job satisfaction, these are not included in the present study.

3.1 *Predictors of job satisfaction*

The antecedents of job satisfaction used in this study are relationship satisfaction, environment satisfaction, overtime, job involvement, work-life balance and the demographic characteristics age and gender. In the upcoming section, the reason why these features are chosen will become clear.

3.1.1 *Relationship satisfaction*

The relationships that people have define both their identity (e.g. being a mother and an employee) and personality (e.g. being introverted and kind) (Mellor et al., 2008). Relationships are therefore an important part of an individual and can be found in different areas of life, such as their personal life and their professional life. Focusing on the latter, many different kinds of relationships can be built by employees within the workplace (Borzaga & Depedri, 2005). For example, relationships with colleagues and their supervisor(s). Relationships in the professional (i.e. work) domain can be either sources for joy, inspiration and learning or may be destructive and lead to frustration (Gersick et al., 2000). In organizational research, relationships at work offer two major benefits, which are emotional support and instrumental support. Emotional support encompasses sympathy, emotional advice, friendship and role-modeling, while instrumental support focuses more on the content of work such as assistance with a task, advising, recommendation and coaching (Alfes et al., 2016; Gersick et al., 2000; Pohl & Galletta, 2017).

When colleagues provide both emotional and instrumental support to each other, a favorable and enjoyable work atmosphere is created. As a result, positive feelings regarding the job are developed (Alfes et al., 2016; Gaan, 2008). When supervisors provide an employee with emotional and instrumental support, the employee feels that the supervisor supports and values him or her. This shows the willingness from the supervisor to invest in and to help the employee which will lead to feelings of job satisfaction (Gok et al., 2015).

From an empirical point of view, Lacy and Sheehan (1997) and Sousa-Poza and Sousa-Poza (2000), for example, show based on statistical analyses that a good relationship with colleagues is associated to job satisfaction and Sousa-Poza and Sousa-Poza (2000) show that a good relationship with the supervisor is essential for experiencing job satisfaction. In summary, an employee can be satisfied and positive about the relationships at work which can contribute to the overall job satisfaction of him or her (Ducharme & Martin, 2000). Based on the above, it is expected that employees who are satisfied with their relationships at work are likely to be satisfied with their job. This is why relationship satisfaction is expected to function as a predictor of job satisfaction.

3.1.2 *Environment satisfaction*

The work environment is anything that surrounds an employee at work and that can impact the tasks that the employee has to complete (Agbozo et al., 2017). Taking the view of Agbozo et al. (2017) it can be divided into three types. First, the physical work environment which concerns the physical aspects of the location where the job is completed, such as the temperature and machinery. Second, the physiological work environment which includes the elements in the work environment that have an impact on how an employee feels. For example, stress and prerequisites at work. Last, the social work environment which encompasses the relationships and forms of communication within the organization (Agbozo et al., 2017).

When there is a fit between the work environment and the employee, and hence the work environment suits the mental and physical abilities of the employee, the job can be performed well. Employees are then in an optimal state in which they can learn and work safely. In turn, employees will be satisfied with their job (Abou Elnaga, 2013). This is in line with empirical studies that also found, based on statistical analyses, that the work environment is linked to job satisfaction (e.g. Agbozo et al., 2017; Raziq & Maulabakhsh, 2015). Therefore, it is expected that an employee who is satisfied with the work environment, will also be satisfied with the job.

3.1.3 *Overtime*

Overtime can be described as putting in extra time and effort to make sure that the job gets finished (Mahmood et al., 2019). Research about overtime and job satisfaction is mixed (Zeytinoglu et al., 2013). For example, a study from Beckers et al. (2008) found, among others, that involuntary overtime is negatively related to job satisfaction while Mohr and Zoghi (2008) did not find any effect of overtime on job satisfaction. On the contrary, Green (2006) (as cited in Zeytinoglu et al., 2013) found that working long hours (i.e. overtime) is positively related to job satisfaction. In line with Green (2006), Mahmood et al. (2019) explain that overtime can be a good thing, depending on the amount of time an employee needs to work extra. Research from Shin et al. (2020), that is focused on the health sector, explains that employees who voluntarily work more hours experience a higher job satisfaction. Since research about this topic is mixed, it is interesting to include overtime as a predictor in the model and later on evaluate the effect using the feature ablation method.

3.1.4 *Job involvement*

Job involvement can be defined as being engaged in and actively participating in the job (Paullay et al., 1994; Saleh & Hosek, 1976). It is the degree to which an employee psychologically identifies him- or herself with the job. Employees who are highly involved in their job put their jobs at the centre of their lives (DeCarufel & Schaan, 1990).

When the job becomes an integral part of who an employee is as a person, the job becomes more meaningful and feelings of self-worth, empowerment and purpose at work are developed. In turn, feelings of job satisfaction will be strengthened (Bahjat Abdallah et al., 2017). This is consistent with other studies, which also found a correlation between job involvement and job satisfaction (e.g. Gopinath & Kalpana, 2020; Van Wyk et al., 2003). Therefore, it is expected that job involvement will function as a predictor of job satisfaction.

3.1.5 *Work-life balance*

Work-life balance can be described as the feeling of an employee that his or her work life and family- and personal life are balanced (Haar et al., 2014; Kasbuntoro et al., 2020). When employers encourage the balance between work and family life by practices (i.e. policies), employees will experience feelings of support and help which results in a higher level of job satisfaction (Agha, 2017; Farooqi & Arif, 2014). Empirical studies also found a link between these constructs (e.g. Haar et al., 2014; Hasan & Teng, 2017). Based on this idea, it is expected that work-life balance can act as a predictor of job satisfaction.

3.1.6 *Age*

The effect of age on job satisfaction has been researched extensively. Most researchers found a positive relationship between the two constructs, indicating that an older employee is more satisfied with their job than a younger employee (Dobrow et al., 2018). According to Clark et al. (1996), there are several explanations for this finding. One of these reasons is that older people have frequently changed jobs and have landed in more attractive jobs with higher status. Another possible reason is that older people may have more realistic job expectations due to several years of work experience. Ng and Feldman (2010) used the socioemotional selectivity theory to explain the finding. They describe that older employees have changed their view of life, from more negative to positive, due to people's changing beliefs of how long they still have to live (Dobrow et al., 2018).

Empirical studies that are in line with these findings are, for example, Dobrow et al. (2018) and Fitzmaurice (2012). Contradictory research, for example Clark et al. (1996), found a U-shaped relation between age and job satisfaction, which indicates that both the younger and older employees were more satisfied than the middle aged employees. Taking all of this into consideration, it is expected that age will be a predictor of job satisfaction.

3.1.7 *Gender*

Research carried out in the Western society often found that women are more likely to experience a higher job satisfaction than men (Huang & Gamble, 2015). Even though women often earn less money and have less promotional opportunities compared to men, there are explanations. For example, women tend to have lower expectations with regard to their job which results in higher job satisfaction (Huang & Gamble, 2015).

Several empirical studies, based on statistical analyses, found that women experience more job satisfaction than men (e.g. Perugini & Vladisavljević, 2019; Zou, 2015). On the contrary, studies that found no significant difference do also exist (e.g. Andrade et al., 2019). As a result, it is expected that gender will act as a predictor of job satisfaction.

In summary, relationship satisfaction, environment satisfaction, overtime, job involvement, work-life balance, age and gender were shown to have a significant relation with job satisfaction in the literature and are therefore included as predictors of job satisfaction in this study.

3.2 *Empirical review*

Over the last few years, several scholars started to focus on predicting job satisfaction by use of one or several machine learning models. In this section, studies that are most closely related to this study are briefly explained. Firstly, research from Moro et al. (2021), who predict job satisfaction based on employee reviews (i.e. text and scores) of the top fifteen IT companies in the United States by use of a support vector machine model. Based on the reviews, 40 features were obtained, such as the advantages and disadvantages of the job which were used as input to the support vector machine model. Results show that the mean absolute error and the mean absolute percentage were good enough to extract knowledge from the model.

On the contrary, research from Choi and Choi (2022) focused on logistic and linear regression. They used a data set from IBM that was originally developed to predict the attrition of employees. Input features were for example, employee tenure, age and job involvement. Focusing on binomial classification, the results show an accuracy of 59.86%.

Rustam et al. (2021) evaluate written (i.e. text) reviews of employees' job satisfaction. These reviews come from organizations such as Amazon and Facebook. The machine learning methods which they compare are logistic regression, random forest, support vector machine and gradient boosting. Next to this, Rustam et al. (2021) developed a deep neural network (i.e. multi-layered perceptron) to classify job satisfaction. The model performance was evaluated by use of accuracy, precision, recall and f1-score. The multi-layered perceptron outperformed the other models based on all performance measures. Random forest and logistic regression appeared to be the second best with an accuracy of 78.00% and support vector machine and gradient boosting had the lowest accuracy, namely 77.00%.

Other research, from Jain et al. (2021) compared a logistic regression, support vector machine, decision tree and random forest model while classifying job satisfaction. The obtained data, from a logistics company in India, consisted of internal marketing strategies, such as payment, recognition, appraisal, rewards and promotion. The four models are compared based on accuracy and sensitivity. The random forest model has been shown to be the best performing model with an accuracy of 87.30%, followed by the decision tree with an accuracy of 81.10% and the support vector machine model that has an accuracy of 78.50%. Logistic regression appeared to be the lowest in terms of accuracy score, which is 77.00%.

In conclusion, the studies discussed above report varying results regarding the sequence of performance levels of the machine learning models.

3.3 *Feature importance/contribution*

As machine learning models act as a black box, it is interesting to find out how much each feature contributes to the classification problem (Brusa et al., 2023). A study from Jain et al. (2021) explores the individual contribution of all features by looking at the decrease in mean Gini of their random forest model. On the contrary, Molnar et al. (2023) performs a model agnostic method (i.e. a method that works for all machine learning models), to find out the contribution of each feature in all machine learning models. Since this study would like to find out the contribution of each feature for all three machine learning models, a model agnostic method is used. More details can be found in section 4.2.3.

4 METHOD

In this section, a detailed description of the methods used will be given. First, the machine learning models will be explained and after that the experimental set-up will be described in detail.

4.1 *Machine learning models*

As explained in Section 3.2, several studies focused on predicting job satisfaction by use of machine learning models. In this section, the reasons for choosing the three machine learning models (i.e. logistic regression, support vector machine and random forest) will become clear and the models will be explained separately.

Since Moro et al. (2021) found that the support vector machine model performed well in predicting job satisfaction, the model is used to find out how well it performs compared to other machine learning models. Next to Moro et al. (2021), Jain et al. (2021) and Rustam et al. (2021) also found promising results of the support vector machine model. To test the rather low performance of the logistic regression model of Choi and Choi (2022) and because logistic regression performs well in research from Jain et al. (2021) and Rustam et al. (2021) this model is utilized as well. Leaving the deep neural network aside since it can be a form of unsupervised learning (Mathew et al., 2021) which is not the focus of this study, Jain et al. (2021) found random forest to be the best performing model and Rustam et al. (2021) found the random forest model to perform equally good as their logistic regression model. A random forest yields classifications that are more accurate than a decision tree (Lan et al., 2020). This is the reason why a decision tree is not included in this study. As research from Rustam et al. (2021) reveals that gradient boosting had the lowest accuracy score, it has been decided to exclude it from this study as well.

In summary, the present study will focus on comparing a logistic regression-, support vector machine- and a random forest model.

4.1.1 *Logistic regression*

Logistic regression is an algorithm that predicts a binary dependent variable using the logistic function (Jain et al., 2021; Rustam et al., 2021). A binary variable has two possible outcomes such as 'yes' and 'no' or 'satisfied' and 'not satisfied'. The logistic regression model classifies the binary variable into 0 and 1, where 1 represents the class the model is trying to predict. The independent variables of the model can be both binary and continuous (Jain et al., 2021). The logistic regression model uses maximum likelihood estimation to estimate probabilities. Maximum likelihood estimation is a method that tries to find the best collection of parameters for which the data has the highest probability (Czepiel, 2002).

4.1.2 *Support vector machine*

Support vector machine is another algorithm that can be used for binary classification and it is capable of handling linear as well as non-linear classification problems (Cortes & Vapnik, 1995; Rustam et al., 2021). Its aim is to find the line (or hyperplane), that consists of n-dimensions, to correctly classify data points (Jain et al., 2021). The number of dimensions in the hyperplane is determined by the number of features in the model. For instance, if there are seven features in the model, there will be n-1 dimensions, which equates to six dimensions (Amarappa & Sathyanarayana, 2014). In addition to a linear kernel, support vector machine has other kernels that can be used to classify data points that cannot be separated linearly (Ben-Hur et al., 2008).

4.1.3 *Random forest*

Random forest, as initiated by Breiman et al. (1984), involves building an ensemble (i.e. pool) of decision trees. This ensemble is created by randomly selecting decision trees and replacing features either with or without replacement. This process is called bagging. In this process, the most successful feature for splitting the node is selected. Measures that are often used for selecting the most successful feature are Gini index, entropy and information gain. Once a sufficient number of decision trees have been built, their output is aggregated and the final result is obtained by averaging them. Due to the random selection within this model, it is less likely to overfit the data. Random forest can be applied to both regression- and classification problems (Jain et al., 2021; Nápoles, 2022).

4.2 *Experimental set-up*

In this section, the dataset and pre-processing steps will be explained. In addition, a detailed description of the experimental procedure, the software and packages and evaluation metrics used will be elaborated on. A visualisation of the methodology and machine learning pipeline can be found in Appendix A (page 37).

4.2.1 *Dataset description*

The dataset used in this study is the 'Employee Performance Prediction' dataset that can be accessed via Kaggle (Karthik, 2022). This dataset, that is provided in an Excel format, is owned by Sripad Karthik and consists out of 29 columns (i.e. features) and 5044 rows (i.e. employees). No information is given about the type of company or the participants. Therefore, the assumption is made that it is simulated data. The data is chosen to predict the job satisfaction of employees, as far as is known, it has not been previously used for this purpose. Since the dataset was designed to predict employee performance, not all features appear to be relevant to predict job satisfaction. To test and confirm what is found in theory, features were selected from the dataset based on a literature review (see Section 3). As mentioned before, the seven input features that are included in the model are relationship satisfaction, environment satisfaction, overtime, job involvement, work-life balance, age and gender. The target variable is job satisfaction. The features are binary, categorical and continuous in nature[1]. For example, job satisfaction is measured on a 5-point likert scale and is labeled as 'very dissatisfied', 'dissatisfied', 'neutral', 'satisfied' and 'very satisfied'.

4.2.2 *Pre-processing*

After loading the data from Microsoft Excel into Python, several exploratory data analysis tasks were carried out. First, the columns (i.e. variables) that were not of interest to this study were deleted. This left the dataset with a total of 8 columns. For example, the variable 'marital status' is excluded since research found that being married or not did not effect the job satisfaction of employees (Azim et al., 2013). Table 1 (see Appendix B (page 38)) gives an overview of the 8 features that are left in the dataset and their characteristics. After that, missing values were detected. In total, 34 rows contained missing values in the gender, relationship satisfaction and work-life balance columns. These rows were deleted, which left the dataset with 5010 rows. To get a feeling of how the data is distributed, histograms were created (see Appendix C (page 39)). It appears that all features, except for age, have roughly the same number of employees for each answer category. Additionally, the descriptive statistics such as the mean of age and percentages 'females', 'males' and 'not-specified' in the dataset were obtained. To spot outliers, a boxplot was created

[1] It must be noted that the variables relationship satisfaction, environment satisfaction, job involvement, work-life balance and job satisfaction which are all measured on a 5-point likert scale, are treated as continuous variables (i.e. interval scales) as is generally done in research in the social sciences (Treiblmaier & Filzmoser, 2011; Wu & Leung, 2017)

for the continuous variable age. No outliers were found (see Appendix D (page 40)). Outliers were not checked for the binary, categorical and continuous variables measured on the 5-point likert scales, since there are only a limited amount of response categories on these types of scales making it impossible to find outliers (Treiblmaier & Filzmoser, 2011).

Two new features were created within the dataset. First, the 'gender' feature. 'Gender' was originally displayed as 'female', 'male' and 'not specified'. Since most machine learning methods need input features that are numerical, 'gender' was recoded into 0 (= female), 1 (= male) and 2 (=not specified). This procedure is called label encoding. The binary feature 'overtime' was already displayed as 0 (= no overtime) and 1 (= overtime). The other features did not have to be recoded. Second, the 'job satisfaction' feature. Several scholars with similar research transformed their continuous target variable into a binary variable (Choi & Choi, 2021; Naburi et al., 2017, e.g.). Jeong and Lee (2016) dichotomized a continuous variable with a 5-point likert scale by transforming 1 to 3 from the original scale to 'disagree' (0) and 4 and 5 to 'agree' (1). They concluded that the scales that were transformed performed well. Additionally, Naburi et al. (2017) transformed job satisfaction with a 5-point likert scale by transforming 'very dissatisfied', 'dissatisfied' and 'neutral' into 'dissatisfied' (0) and 'satisfied' and 'very satisfied' into 'satisfied' (1). Their scale turned out to be reliable. Therefore, the same will be done in this study with the target variable job satisfaction that has a 5-point likert scale as well. Answer categories 1 to 3 will become 'not satisfied' and 4 and 5 will become 'satisfied'. The rationale behind this procedure is that only employees that specifically indicate that they are satisfied are classified as satisfied employees (i.e. 4 and 5). Due to this procedure, there is an imbalance in the data. The minority class (i.e. satisfied class) now consists of 1997 employees, while the majority class (i.e. not satisfied class) now consists of 3013 employees. Therefore, the under sampling technique was carried out, which means that the size of the majority class was randomly decreased to the size of the minority class (Dal Pozzolo et al., 2015). While it removes data from the dataset, it is seen as a simple and fast method to create a balanced dataset, which is why this method has been chosen (Dal Pozzolo et al., 2015)

Additionally, data has been standardized. Standardizing the data means that data is scaled in a way that the mean is zero and the standard deviation is one. Standardization is useful for machine learning methods, such as logistic regression, which is normally distributed and has a Gaussian distribution (Ali et al., 2014). Since the kernel of the support vector machine model is based on distance, it is important that it is scaled before the model implementation as well. The support vector machine is looking for a large distance between the support vectors and the hyperplane. If one features' value is large, it will overrule the other features' values which will affect the calculation of the distance. Rescaling the features prevents one feature to be dominating and ensures all features to have an equal impact on the calculation of the distance (Jrieke, 2016).

One oddity was found when going through the pre-processing steps. As can be seen in Appendix C (page 39), all answer categories of each variables' scale got more or less the same amount of answers. For example, job involvement got 968 answers on answer category 1, 988 answers on answer category 2, 1035 answers on category 3, 1001 answers on category 4 and 1018 answers on category 5. Since all data is distributed in a similar way, it might be the case that data is distributed randomly and that therefore the model cannot make a correct prediction.

4.2.3 *Description of experimental procedure*

To answer the main research question, which focuses on the comparison of three machine learning models, the accuracy scores needed to be generated. Before the machine learning models were run to obtain the accuracy scores, data was split in such a way that 70% of the data was assigned to the training set and 30% to the test set. This is in line with research from Choi and Choi (2022) and Holgado-Apaza et al. (2023). The data split was stratified, meaning that the data in the train- and test set consist of the same proportion of each target class as the whole dataset (Sklearn, 2023a). In addition, a random seed was set to make sure the models produce the same output each time the program is runned (Sklearn, 2023e). The three models were trained on the train set and tested on the test set. K-fold cross validation was carried out, where k was set to 10 since this is common in research (Nematzadeh et al., 2015). Within this cross-validation procedure data is split into k partitions (also called folds). Within one iteration, one partition is set aside for testing and the other partitions are used for training. In total, there are k iterations. In the end, the accuracy scores obtained are averaged to get the final model accuracy score (Yadav & Shukla, 2016). Also, hyperparameters were tuned. Hyperparameters are values that a user can set manually to achieve the best predictive performance of a model (Probst et al., 2019). This is important, because default hyperparameters do not necessarily generate the best results (Schratz et al., 2019). Gridsearch, which carries out an exhaustive search to find the best set of hyperparameters, has been performed (Sklearn, 2023d). An overview of the hyperparameters that were tuned including a short description of these hyperparameters can be found in Table 2. After executing these procedures, the accuracy scores were compared among the three machine learning models[2].

Table 2
Hyperparameter tuning

Hyperparameters	Model	Description
Solver	LOR	The algorithm used (lbfgs, newton-cg, liblinear)
Penalty	LOR	None, elasticnet, L1, l2
C	LOR, SVM	Regularization parameter
kernel	SVM	The type of kernel used (rbf, poly, sigmoid, linear)
Gamma	SVM	Coefficient of the kernel
N_estimators	RF	The amount of trees in the random forest
Max_features	RF	The amount of features needed for the best split
Max_depth	RF	The depth of the trees in the random forest set to a maximum
Min_samples split	RF	The number of samples needed to split a node set to a minimum
Min_samples leaf	RF	The number of samples needed to go to a leaf node set to a minimum
Bootstrap	RF	If bootstrap samples are used in the tree-building process (True or False)

Note. LOR is logistic regression, SVM is support vector machine and RF is random forest. Sources: (Sklearn, 2023b, 2023c, 2023f)

To address the first sub research question, which examines the contribution of each feature to the model, the feature ablation method has been carried out. This method involved training the machine learning model with all features to see its performance. Subsequently, one input feature was removed iteratively and the performance of the model was evaluated. The changes in accuracy score resulting from the removal of each feature provided insight into their importance and relevance (Covert et al., 2020).

4.2.4 *Packages and software*

The software that is used in this study is Spyder IDE using Python version 3.9. The library of sci-kit learn was mostly used. For example, for data splitting and carrying out the logistic regression, support vector machine and grid search. Furthermore, pandas, nympy and matplotlib were utilized.

4.2.5 *Evaluation metrics*

The evaluation metric used in this study is the accuracy score, which is one of the most frequently used in binary classification problems (Chicco & Jurman, 2020). The accuracy score is the proportion of correct predictions out of the total number of predictions and works well with a balanced target variable (Brodersen et al., 2010). It can be easily understood by looking at a confusion matrix. A confusion matrix is used to evaluate the performance of a classification model and compares the actual values with the predicted values (Görtler et al., 2022). Table 3 shows a confusion matrix.

[2] The experimental procedure described above was also tested on imbalanced data using the f1-score evaluation metric. Although the results were not necessarily worse, they were less easily interpretable due to the nature of f1-score. Since the accuracy and f1-scores for the three models were around chance level, this study proceeded with the balanced dataset.

Table 3
Confusion matrix

	Predicted values		
Actual values		Not satisfied	Satisfied
	Not satisfied	True negative	False positive
	Satisfied	False negative	True positive

5 RESULTS

This section presents the results of the three machine learning models and discusses the contribution of each feature.

5.1 *Performance of the three machine learning models*

In Table 4, an overview of the accuracy scores for the three machine learning models is given. It can be observed that logistic regression achieves the highest accuracy score (52.04%), while support vector machine achieved the lowest accuracy score (50.54%). The accuracy score of random forest falls in between the two (51.13%). What stands out is that all three accuracy scores are rather low and around chance level. Another noteworthy point is that the training score of the random forest model is considerably higher than that of the test set [3].

Table 4
Accuracy scores logistic regression, support vector machine and random forest

Model	Train set	Test set
	Accuracy (%)	Accuracy (%)
Logistic regression	52.49	52.04
Support vector machine	53.24	50.54
Random Forest	58.82	51.13

Note. Train N = 2795, test N = 1199.

More details about the statistics can be found in the Appendices. Tables 5, 6 and 7 in Appendix E (page 41) present the classification reports for the logistic regression, support vector machine and random forest models, respectively. The classification reports include information about the precision, recall, f1-scores, accuracy and support of each model. The confusion matrices can be found in Appendix F (page 43).

5.2 *Feature importance/contribution*

As explained, feature ablation has been performed to identify the feature that contributes the most to the prediction of job satisfaction. Tables 8, 9, and 10 present an overview of the accuracy scores and the change in the

[3] The train accuracy is approximately seven percent higher than the test accuracy, which suggests potential overfitting. As adviced by Ellis (2021), the depth of the tree and the minimum number of samples to split a node were kept low and decided during hyperparameter tuning. This resulted in the train and test accuracy reported here. As steps were taken and as Breiman and Cutler (n.d.) explains that random forest does not overfit no further changes were made.

original accuracy score resulting from the removal of individual features for the logistic regression, support vector machine, and random forest models, respectively. In the upcoming sections, the most notable scores will be highlighted.

5.2.1 *Logistic regression*

Table 8 shows that removing age from the model results in a 2.83 percentage point (pp) decrease in accuracy score and removing gender leads to a 1.50 pp decrease in accuracy score. These findings indicate that both age and gender play an important role in predicting job satisfaction. The table also shows that removing environment satisfaction or overtime from the model leads to an increase in accuracy score of 0.25 pp and 0.09 pp, respectively. This means that the model predicts better without environment satisfaction or overtime in the model.

Table 8
Feature importance logistic regression

	Accuracy (%)	Decrease/increase (pp)
Relationship satisfaction	51.13	-0.91
Environment satisfaction	52.29	0.25
Overtime	52.13	0.09
Job involvement	50.71	-1.33
Work-life balance	51.54	-0.50
Age	49.21	-2.83
Gender	50.54	-1.50

Note. The accuracy score represents the accuracy score obtained by the model without the corresponding feature. The change in accuracy score reflects the decrease or increase in accuracy compared to the original score of 52.04. pp is percentage point.

5.2.2 *Support vector machine*

What can be seen in Table 9 is that relationship satisfaction and work-life balance are the most important factors in predicting job satisfaction, as their removal from the model results in a decrease in accuracy of 1.33 pp and 0.33 pp, respectively. It is also worth mentioning that removing overtime from the model results in a decrease in accuracy of 0.33 pp. What stands out is that removing job involvement from the model leads to an increase in accuracy of 1.17 pp, indicating that its exclusion from the model leads to a higher accuracy score.

Table 9
Feature importance support vector machine

	Accuracy (%)	Decrease/increase (pp)
Relationship satisfaction	49.21	-1.33
Environment satisfaction	50.21	-0.33
Overtime	50.21	-0.33
Job involvement	51.71	1.17
Work-life balance	49.87	-0.67
Age	49.96	-0.58
Gender	50.21	-0.33

Note. The accuracy score represents the accuracy score obtained by the model without the corresponding feature. The change in accuracy score reflects the decrease or increase in accuracy compared to the original score of 50.54. pp is percentage point.

5.2.3 *Random forest*

Table 10 displays several features with highly negative changes in accuracy scores. Gender, work-life balance and age show a decrease in accuracy scores of -3.34 pp, -2.34 pp and -2.26 pp respectively, when removed from the model. These findings highlight the crucial role that all three features play in predicting job satisfaction. It is noteworthy that overtime shows a decrease in accuracy score of 0.50 pp when being removed from the model. Conversely, job involvement shows an increase in accuracy score of 0.50 pp, indicating that the model performs better when job involvement is excluded from the model.

Table 10
Feature importance logistic regression

	Accuracy (%)	Decrease/increase (pp)
Relationship satisfaction	49.12	-2.01
Environment satisfaction	49.96	-1.17
Overtime	50.63	-0.50
Job involvement	51.63	0.50
Work-life balance	48.79	-2.34
Age	48.87	-2.26
Gender	47.79	-3.34

Note. The accuracy score represents the accuracy score obtained by the model without the corresponding feature. The change in accuracy score reflects the decrease or increase in accuracy compared to the original score of 51.13. pp is percentage point.

5.2.4 *Comparison*

When comparing the three tables above, several similarities and differences between the logistic regression, support vector machine and random forest model come to light. To facilitate the discussion of these similarities and differences, Table 11 provides a clear overview of the signs of the change in accuracy scores for each feature in the three models. Looking at the similarities, relationship satisfaction, work-life balance, age and gender are all contributing to the prediction of job satisfaction individually, as the accuracy scores decrease when any of these features are removed from the model. In contrast, differences can be observed in how environment satisfaction, overtime, and job involvement affect the accuracy score. When these three features are individually excluded from the model, the accuracy score will either increase or decrease depending on which of the three models is being looked at. For example, removing overtime from the logistic regression model will lead to a better predicting model, while removing overtime from the support vector machine and random forest model will lead to a worse predicting model.

Table 11
Change in accuracy score for LOR, SVM and RF

	LOR	SVM	RF
Relationship satisfaction	-	-	-
Environment satisfaction	+	-	-
Overtime	+	-	-
Job involvement	-	+	+
Work-life balance	-	-	-
Age	-	-	-
Gender	-	-	-

Note. LOR is logistic regression, SVM is support vector machine and RF is random forest. A minus (-) stands for decrease and a plus (+) stands for increase in original accuracy score when removing the corresponding variable.

6 DISCUSSION

In this study, three machine learning models - logistic regression, support vector machine and random forest - were compared to determine which one most accurately predicts job satisfaction. Additionally, this study examined the contribution of each individual feature in predicting job satisfaction.

The results show that logistic regression was best in predicting job satisfaction, followed by the random forest model as the second best model. The support vector machine model appeared to be the poorest in predicting job satisfaction. All three machine learning models performed only slightly better than chance level and their differences were small. With regard to each feature's contribution, it can be seen that age and gender are important predictors in the logistic regression model, relationship satisfaction and work-life balance in the support vector machine model and age, gender and work-life balance in the random forest model. The least important predictors are environment satisfaction and overtime in the logistic regression model and job involvement in the support vector machine- and random forest model. Overall, the results show that relationship satisfaction, work-life balance, age and gender are contributing to the prediction of job satisfaction in all three models. The contribution of environment satisfaction, overtime and job involvement in predicting job satisfaction varied depending on the machine learning model.

In the upcoming sections, these findings will be evaluated in relation to existing research and literature as described in Section 3. The study's contribution to literature and society will also be discussed, followed by a discussion of limitations and directions for future research.

6.1 *Performance of the three machine learning models*

Research from Rustam et al. (2021) shows that the logistic regression model performs the best, which is in line with this study. Although they found random forest to be equally good, this study finds it to be the second best model. The present study found the support vector machine to be the lowest performing model, which was consistent with findings from Rustam et al. (2021). Findings from Jain et al. (2021) do not align with this study, as they found the random forest model to be the best and the logistic regression to be the poorest performing model. Choi and Choi (2022) found a rather low performance level of the logistic regression model, which is consistent with this studies findings. In terms of sequences of the models, there are both similarities and differences between the findings of earlier research and this study. This could be because earlier research has not yet found a consistent resolution and because the input features used in each study differ from one another.

To investigate why the accuracy scores of the three models are around chance level, two possible reasons will be given. One reason is the small dataset size of 3994 samples. Only 70% of the data is available for training the models, which may not be enough for the models to recognize patterns. Training the models on a larger dataset would result in recognizing more patterns in the data, which might lead to a better performance of the models (Barman et al., 2019). For example, Osisanwo et al. (2017) state that the support vector machine model needs a large dataset in order to achieve maximum performance in terms of accuracy score. A second reason for the performance around chance level could concern patterns in the data found during exploratory data analysis. As mentioned in Section 4.2.2 and as can be seen in Appendix C (page 39), samples (i.e. employees) are almost equally divided among each answer category. This may contribute to the difficulty in finding patterns in the data, as patterns may not exist or may be too small to detect. Therefore, the models might report accuracy scores around chance level.

6.2 *Feature importance/contribution*

As previous research on job satisfaction prediction use different sets of input features compared to this study, it is not feasible to compare the most contributing features of this study's models to those of existing research. However, it is possible to compare findings across the three models with what was expected in literature. As described in Section 3.1, it was expected that all features would play a role in predicting job satisfaction. Overtime was considered a special case, since previous research has had mixed results. The findings of this study indicate that relationship satisfaction, work-life balance, age and gender are significant predictors of job satisfaction, which is consistent with the literature review. Moreover, findings show that the contribution of overtime to job satisfaction prediction varies depending on which model is examined. This finding also aligns with previous literature.

Contrary to expectations, environment satisfaction and job involvement showed different results among the three models. A possible reason for these varying results could be the presence of mediating effects that were not considered in this study as stated by Lee and Brand (2005). This means that there is no direct relation between environment satisfaction and job satisfaction and job involvement and job satisfaction, but that there is a third variable that provides an explanation for their association (MacKinnon et al., 2007). For instance, Newsham et al. (2009) found that satisfaction with management mediates the relationship between work environment and job satisfaction and Bayraktar et al. (2017) showed that reward acts as a mediating variable between job involvement and job satisfaction. This shows an opposing view that argues that the satisfaction with the work environment and job involvement are not directly related to job satisfaction. Another reason for the varying results between environment satisfaction and job satisfaction may be attributed to a non-linear relationship between the two variables as suggested by Warr (1990). According to his explanation, the work environment and a person's reactions, such as job satisfaction, are not linearly related. Instead, there is an optimal level of work conditions in the environment that leads to job satisfaction, while too little or too many work conditions can have negative effects on job satisfaction (Noblet et al., 2009). It could be the case that the support vector machine and random forest model were able to capture this non-linear relationship, while the logistic regression model was not able to. Lastly, Taheri et al. (2020) argues that the work environment does not contribute to job satisfaction, but that it can only prevent an employee from experiencing dissatisfaction with the job. This suggests another different perspective on the relationship between work environment and job satisfaction.

6.3 *Scientific and societal impact*

This study makes several contributions to the scientific literature. First, it contributes to the literature on machine learning and Human Resource Management by identifying the best machine learning model (i.e. logistic regression) for predicting job satisfaction. Second, it confirms a large part of the individual feature's contribution as expected by literature. However, it is worth noting that that the results are only slightly better than chance level, so caution is advised when interpreting these findings. Nonetheless, this study provides valuable suggestions for future research and can be used by future studies to build upon.

This study provides valuable insights not only for scientific purposes, but also for society and practice. Job satisfaction has a significant impact on working individuals in society. When people are satisfied with their job, they experience satisfaction with their life, and experience advantages in their physical and mental health. Moreover, they are expected to live longer (Kaplan et al., 1991). Therefore, organizations need to understand which features contribute to the prediction of job satisfaction and which machine learning model best can be used. The findings of this study suggest that the logistic regression model is the most effective method for organizations to predict job satisfaction, and that, among others, relationship satisfaction and work-life balance are important predictors. Organizations can promote a work-life balance by offering flexible work opportunities and empowering individuals to determine when they need to take actions to improve their balance (Kelliher et al., 2019). In addition, managers can create a safe atmosphere in which mutual trust is stimulated, which fosters relationship building within an organization (Shin et al., 2012). By promoting these and other features from the model, organizations can increase job satisfaction, leading to above mentioned and earlier stated societal benefits (see Section 2). However, it is important to note that further research is needed to confirm these findings before they can be recommended to organizations for practical use and societal benefits can be reaped.

6.4 *Limitations and future research*

Several limitations and recommendations for future research will be given. First, a relatively small dataset was used in this study. As it is harder for machine learning models to find patterns in a small dataset, it is recommended to use a larger dataset in future studies (Barman et al., 2019; Choi & Choi, 2022). Choi and Choi (2022) suggest to use over a million data points in a future dataset to predict job satisfaction. In line with this view, future studies could utilize the oversampling technique to balance data. As a result, more data will be maintained and the dataset will become larger.

Second, the performance levels of the three models are only slightly higher than chance level. Therefore, future research is necessary that replicates this study. According to Zhao et al. (2019), it is best to try the different machine learning models in different settings, meaning that it is advisable to test them on different datasets of varying sizes. When this study is replicated in different settings and contexts, confidence in the results will be strengthened and generalizations to theory and practice can be safely made (Flint et al., 2013).

In addition to the feature ablation method, future studies could employ other model agnostic or model specific methods to assess the contribution of individual features. Permutation feature importance is one such method, but it requires a significant amount of computational power (Hapfelmeier et al., 2022). Because this study aimed to keep the analyses technologically accessible for simple and affordable devices, as stated in the introduction (see Section 2), this method was not used here. However, if computationally feasible, future studies could explore other methods to evaluate the contribution of individual features. Exploring different feature importance methods can provide more certainty about the importance of the features in predicting job satisfaction. This can help organizations to better understand how to improve job satisfaction among their employees.

7 CONCLUSION

Job satisfaction is a crucial factor for an organization's success while also yielding a range of societal benefits. Therefore, this study aims to predict the job satisfaction of employees by comparing several machine learning models to answer the following research question *"Which of the following machine learning models, a support vector machine-, logistic regression- or random forest model, best classifies job satisfaction in terms of accuracy score?"*. The logistic regression model emerges as the best performing model in terms of accuracy score. The accuracy score is only slightly higher than chance level. Therefore, this model cannot be utilized in practice and future research is needed that replicates this study.

Two sub-questions are formulated. The first sub-question focuses on finding out which predictors are most important in predicting job satisfaction. Among the three machine learning models, relationship satisfaction, work-life balance, age and gender emerge as the most important predictors. This is evident from the relatively large decrease in accuracy scores compared to other features. As the machine learning models need to perform well in order to assess the contribution of individual features, future research is needed to confirm the most important predictors. Therefore, findings cannot be utilized in practice.

The second sub-question focuses on comparing the accuracy scores and the findings in relation to existing research and literature. The accuracy scores of the three machine learning models are very low compared to previous studies. Logistic regression has been found to be the best performing model in one study, contrary to another study. During the feature ablation analysis, relationship satisfaction, work-life balance, age, gender and overtime align with literature. However, work environment and job involvement do not align with literature as they behave inconsistently between the three models.

Overall, additional research is necessary to confirm the findings before making generalizations and recommendations to organizations. Only then, accompanying societal benefits could be realized. Nevertheless, this study contributes to the existing literature by providing a foundation for future studies to build upon.

REFERENCES

Abou Elnaga, A. (2013). Exploring the link between job motivation, work environment and job satisfaction. *Journal Of Business and Management*, *41*.

Agbozo, G. K., Owusu, I. S., Hoedoafia, M. A., & Atakorah, Y. B. (2017). The effect of work environment on job satisfaction: Evidence from the banking sector in ghana. *Journal of human resource management*, *5*(1), 12–18.

Agha, K. (2017). Work-life balance and job satisfaction: An empirical study focusing on higher education teachers in oman. *International Journal of Social Science and Humanity*, *7*(3), 164–171.

Alfes, K., Shantz, A., & van Baalen, S. (2016). Reducing perceptions of overqualification and its impact on job satisfaction: The dual roles of interpersonal relationships at work. *Human Resource Management Journal*, *26*(1), 84–101.

Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H. (2014). Data normalization and standardization: A technical report. *Mach Learn Tech Rep*, *1*(1), 1–6.

Amarappa, S., & Sathyanarayana, S. (2014). Data classification using support vector machine (svm), a simplified approach. *Int. J. Electron. Comput. Sci. Eng*, *3*, 435–445.

Andrade, M. S., Westover, J. H., & Peterson, J. (2019). Job satisfaction and gender. *Journal of Business Diversity*, *19*(3).

Azim, M. T., Haque, M. M., & Chowdhury, R. A. (2013). Gender, marital status and job satisfaction an empirical study. *International Review of Management and Business Research*, *2*(2), 488.

Aziri, B. (2011). Job satisfaction: A literature review. *Management Research and Practice*.

Bahjat Abdallah, A., Yousef Obeidat, B., Osama Aqqad, N., Al Janini, K., Na'el, M., & Dahiyat, S. E. (2017). An integrated model of job involvement, job satisfaction and organizational commitment: A structural analysis in jordan's banking sector. *Communications and Network*, *9*(01), 28–53.

Barman, R., Deshpande, S., Agarwal, S., Inamdar, U., Devare, M., & Patil, A. (2019). Transfer learning for small dataset. *Proceedings of the National Conference on Machine Learning*, *26*.

Bayraktar, C. A., Araci, O., Karacay, G., & Calisir, F. (2017). The mediating effect of rewarding on the relationship between employee involvement and job satisfaction. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *27*(1), 45–52.

Beckers, D. G., Van der Linden, D., Smulders, P. G., Kompier, M. A., Taris, T. W., & Geurts, S. A. (2008). Voluntary or involuntary? control over overtime and rewards for overtime in relation to fatigue and work satisfaction. *Work & Stress*, *22*(1), 33–50.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., & Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS computational biology*, *4*(10).

Borzaga, C., & Depedri, S. (2005). Interpersonal relations and job satisfaction: Some empirical results in social and community care services. *Economics and social interaction: Accounting for interpersonal relations*, 132–153.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. *Belrnont: WadsWorth*.

Breiman, L., & Cutler, A. (n.d.). Random forests [Accessed on May 11, 2023]. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. *2010 20th international conference on pattern recognition*, 3121–3124.

Brusa, E., Cibrario, L., Delprete, C., & Di Maggio, L. G. (2023). Explainable ai for machine fault diagnosis: Understanding features' contribution in machine learning models for industrial condition monitoring. *Applied Sciences*, *13*(4), 2038.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*, 1–13.

Choi, Y., & Choi, J. (2022). Job satisfaction prediction and machine learning technique.

Choi, Y., & Choi, J. W. (2021). A study of job involvement prediction using machine learning technique. *International Journal of Organizational Analysis*, *29*(3), 788–800.

Clark, A., Oswald, A., & Warr, P. (1996). Is job satisfaction u-shaped in age? *Journal of occupational and organizational psychology*, *69*(1), 57–81.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273–297.

Covert, I., Lundberg, S. M., & Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, *33*, 17212–17223.

Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: Theory and implementation. *Available at czep. net/stat/mlelr. pdf*, *83*.

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. *2015 IEEE symposium series on computational intelligence*, 159–166.

DeCarufel, A., & Schaan, J.-L. (1990). The impact of compressed work weeks on police job involvement. *Canadian Police College Journal*.

DiMaria, C. H., Peroni, C., & Sarracino, F. (2020). Happiness matters: Productivity gains from subjective well-being. *Journal of Happiness Studies*, *21*(1), 139–160.

Dobrow, S. R., Ganzach, Y., & Liu, Y. (2018). Time and job satisfaction: A longitudinal study of the differential roles of age and tenure. *Journal of management*, *44*(7), 2558–2579.

Ducharme, L. J., & Martin, J. K. (2000). Unrewarding work, coworker support, and job satisfaction: A test of the buffering hypothesis. *Work and occupations*, *27*(2), 223–243.

Ellis, C. (2021). Random forests overfitting [Accessed on May 11, 2023]. https://crunchingthedata.com/random-forest-overfitting/

Esfandiari, R., & Kamali, M. (2016). On the relationship between job satisfaction, teacher burnout, and teacher autonomy. *Iranian Journal of Applied Language Studies*, *8*(2), 73–98.

Farooqi, Y., & Arif, B. (2014). Impact of work life balance on job satisfaction and organizational commitment among university teachers: A case study of university of gujrat, pakistan. *International Journal of Multidisciplinary Sciences and Engineering*.

Fida, M. K., Khan, M. Z., & Safdar, A. (2019). Job satisfaction in banks: Significance of emotional intelligence and workplace environment. *Scholars Bulletin*, *5*(9), 504–512.

Fitzmaurice, C. (2012). Job satisfaction in ireland: An investigation into the influence of self-esteem, generalised self-efficacy and affect.

Flint, D., Haley, L. M., & McNally, J. J. (2013). Individual and organizational determinants of turnover intent. *Personnel Review*.

Gaan, N. (2008). Stress, social support, job attitudes and job outcome across gender. *ICFAI Journal of Organizational Behavior*, *7*(4), 34–44.

Garg, S., Sinha, S., Kar, A. K., & Mani, M. (2022). A review of machine learning applications in human resource management. *International Journal of Productivity and Performance Management*, *71*(5), 1590–1610.

Gersick, C. J., Dutton, J. E., & Bartunek, J. M. (2000). Learning from academia: The importance of relationships in professional life. *Academy of Management Journal*, *43*(6), 1026–1044.

Gok, S., Karatuna, I., & Karaca, P. O. (2015). The role of perceived supervisor support and organizational identification in job satisfaction. *Procedia-Social and Behavioral Sciences*, *177*, 38–42.

Gopinath, R., & Kalpana, R. (2020). Relationship of job involvement with job satisfaction. *Adalya journal*, *9*(7), 306–315.

Görtler, J., Hohman, F., Moritz, D., Wongsuphasawat, K., Ren, D., Nair, R., Kirchner, M., & Patel, K. (2022). Neo: Generalizing confusion matrix visualization to hierarchical and multi-output labels. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–13.

Haar, J. M., Russo, M., Suñe, A., & Ollier-Malaterre, A. (2014). Outcomes of work–life balance on job satisfaction, life satisfaction and mental health: A study across seven cultures. *Journal of vocational behavior*, *85*(3), 361–373.

Hapfelmeier, A., Hornung, R., & Haller, B. (2022). Sequential permutation testing of random forest variable importance measures. *arXiv preprint arXiv:2206.01284*.

Harmon-Jones, E., & Harmon-Jones, C. (2021). On defining positive affect (pa): Considering attitudes toward emotions, measures of pa, and approach motivation. *Current Opinion in Behavioral Sciences*, *39*, 46–51.

Hasan, N., & Teng, L. S. (2017). Work-life balance and job satisfaction among working adults in malaysia: The role of gender and race as moderators. *Journal of Economics, Business and Management*, *5*(1), 18–24.

Hecklau, F., Galeitzke, M., Flachs, S., & Kohl, H. (2016). Holistic approach for human resource management in industry 4.0. *Procedia Cirp*, *54*, 1–6.

Holgado-Apaza, L. A., Carpio-Vargas, E. E., Calderon-Vilca, H. D., Maquera-Ramirez, J., Ulloa-Gallardo, N. J., Acosta-Navarrete, M. S., Barrón-Adame, J. M., Quispe-Layme, M., Hidalgo-Pozzi, R., & Valles-Coral, M. (2023). Modeling job satisfaction of peruvian basic education teachers using machine learning techniques. *Applied Sciences*, *13*(6), 3945.

Huang, Q., & Gamble, J. (2015). Social expectations, gender and job satisfaction: Front-line employees in c hina's retail sector. *Human resource management journal*, *25*(3), 331–347.

Isen, A. M. (2001). An influence of positive affect on decision making in complex situations: Theoretical issues with practical implications. *Journal of consumer psychology*, *11*(2), 75–85.

Jain, D., Makkar, S., Jindal, L., & Gupta, M. (2021). Uncovering employee job satisfaction using machine learning: A case study of om logistics ltd. *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2020, Volume 2*, 365–376.

Jeong, H. J., & Lee, W. (2016). The level of collapse we are allowed: Comparison of different response scales in safety attitudes questionnaire. *Biom Biostat Int J*, *4*(4), 00100.

Jones, K. (2014). Conquering hr analytics: Do you need a rocket scientist or a crystal ball. *Workforce Solutions Review*, *5*(1), 43–44.

Jrieke. (2016). Why scaling is important for the linear svm classification? https://stats.stackexchange.com/q/224201

Judge, T. A., & Bono, J. E. (2001). Relationship of core self-evaluations traits—self-esteem, generalized self-efficacy, locus of control, and emotional stability—with job satisfaction and job performance: A meta-analysis. *Journal of applied Psychology*, *86*(1), 80.

Judge, T. A., & Kammeyer-Mueller, J. D. (2011). Happiness as a societal value. *Academy of management perspectives*, *25*(1), 30–41.

Jung, Y., & Suh, Y. (2019). Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews. *Decision Support Systems*, *123*, 113074.

Kaplan, R., Boshoff, A., & Kellerman, A. (1991). Job involvement and job satisfaction of south african nurses compared with other professions. *Curationis*, *14*(1), 3–7.

Karthik, S. (2022). *Employee performance prediction*. Retrieved February 13, 2023, from https://www.kaggle.com/datasets/sripadkarthik/employee-performance-prediction

Kasbuntoro, D. I., Maemunah, S., Mahfud, I., Fahlevi, M., & Parashakti, R. D. (2020). Work-life balance and job satisfaction: A case study of employees on banking companies in jakarta. *International Journal of Control and Automation*, *13*(4), 439–451.

Kelliher, C., Richardson, J., & Boiarintseva, G. (2019). All of work? all of life? reconceptualising work-life balance for the 21st century. *Human resource management journal*, *29*(2), 97–112.

King, L. A., Hicks, J. A., Krull, J. L., & Del Gaiso, A. K. (2006). Positive affect and the experience of meaning in life. *Journal of personality and social psychology*, *90*(1), 179.

Koh, H. C., & Boo, E. H. (2001). The link between organizational ethics and job satisfaction: A study of managers in singapore. *Journal of Business Ethics*, *29*, 309–324.

Lacy, F. J., & Sheehan, B. A. (1997). Job satisfaction among academic staff: An international perspective. *Higher education*, *34*(3), 305–322.

Lan, T., Hu, H., Jiang, C., Yang, G., & Zhao, Z. (2020). A comparative study of decision tree, random forest, and convolutional neural network for spread-f identification. *Advances in Space Research*, *65*(8), 2052–2061.

Lee, S. Y., & Brand, J. L. (2005). Effects of control over office workspace on perceptions of the work environment and work outcomes. *Journal of environmental psychology*, *25*(3), 323–333.

Locke, E. (1976). The nature and causes of job satisfaction. *The handbook of industrial and organizational psychology*, *31*.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annu. Rev. Psychol.*, *58*, 593–614.

Mahmood, Y. N., Raewf, M. B., & AL-Hamadany, Z. S. (2019). A study on the perceptual relationship between overtime and output. *Cihan University-Erbil Journal of Humanities and Social Sciences*, *3*(1), 27–31.

Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of hr analytics. *The International Journal of Human Resource Management*, *28*(1), 3–26.

Mathew, A., Amudha, P., & Sivakumari, S. (2021). Deep learning techniques: An overview. *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020*, 599–608.

Mellor, D., Stokes, M., Firth, L., Hayashi, Y., & Cummins, R. (2008). Need for belonging, relationship satisfaction, loneliness, and life satisfaction. *Personality and individual differences*, *45*(3), 213–218.

Mohr, R. D., & Zoghi, C. (2008). High-involvement work design and job satisfaction. *ILR Review*, *61*(3), 275–296.

Molnar, C., König, G., Bischl, B., & Casalicchio, G. (2023). Model-agnostic feature importance and effects with dependent features: A conditional subgroup approach. *Data Mining and Knowledge Discovery*, 1–39.

Moro, S., Ramos, R. F., & Rita, P. (2021). What drives job satisfaction in it companies? *International Journal of Productivity and Performance Management*, *70*(2), 391–407.

Naburi, H., Mujinja, P., Kilewo, C., Orsini, N., Bärnighausen, T., Manji, K., Biberfeld, G., Sando, D., Geldsetzer, P., Chalamila, G., et al. (2017). Job satisfaction and turnover intentions among health care staff providing services for prevention of mother-to-child transmission of hiv in dar es salaam, tanzania. *Human resources for health*, *15*, 1–12.

Nanda, R., & Browne, J. J. (1977). Hours of work, job satisfaction and productivity. *Public Productivity Review*, 46–56.

Nápoles, G. (2022). *Pattern classification* [Powerpoint slides], Tilburg University.

Nematzadeh, Z., Ibrahim, R., & Selamat, A. (2015). Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. *2015 10th Asian control conference (ASCC)*, 1–6.

Newsham, G., Brand, J., Donnelly, C., Veitch, J., Aries, M., & Charles, K. (2009). Linking indoor environment conditions to job satisfaction: A field study. *Building Research & Information*, *37*(2), 129–147.

Ng, T. W., & Feldman, D. C. (2010). The relationships of age with job attitudes: A meta-analysis. *Personnel psychology*, *63*(3), 677–718.

Noblet, A., Rodwell, J., & Allisey, A. (2009). Job stress in the law enforcement sector: Comparing the linear, non-linear and interaction effects of working conditions. *Stress and Health: Journal of the International Society for the Investigation of stress*, *25*(1), 111–120.

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., Akinjobi, J., et al. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128–138.

Paullay, I. M., Alliger, G. M., & Stone-Romero, E. F. (1994). Construct validation of two instruments designed to measure job involvement and work centrality. *Journal of applied psychology*, *79*(2), 224.

Perugini, C., & Vladisavljević, M. (2019). Gender inequality and the gender-job satisfaction paradox in europe. *Labour Economics*, *60*, 129–147.

Pohl, S., & Galletta, M. (2017). The role of supervisor emotional support on individual job satisfaction: A multilevel analysis. *Applied Nursing Research*, *33*, 61–66.

Prajogo, D. I., & Cooper, B. K. (2010). The effect of people-related tqm practices on job satisfaction: A hierarchical model. *Production Planning and Control*, *21*(1), 26–35.

Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, *20*(1), 1934–1965.

Raziq, A., & Maulabakhsh, R. (2015). Impact of working environment on job satisfaction. *Procedia Economics and Finance*, *23*, 717–725.

Rustam, F., Ashraf, I., Shafique, R., Mehmood, A., Ullah, S., & Sang Choi, G. (2021). Review prognosis system to predict employees job satisfaction using deep neural network. *Computational Intelligence*, *37*(2), 924–950.

Saber, D. A. (2014). Frontline registered nurse job satisfaction and predictors over three decades: A meta-analysis from 1980 to 2009. *Nursing Outlook*, *62*(6), 402–414.

Saleh, S. D., & Hosek, J. (1976). Job involvement: Concepts and measurements. *Academy of management journal*, *19*(2), 213–224.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109–120.

Shin, Oh, S. J., Kim, J., Lee, I., & Bae, S.-H. (2020). Impact of nurse staffing on intent to leave, job satisfaction, and occupational injuries in korean hospitals: A cross-sectional study. *Nursing & health sciences*, *22*(3), 658–666.

Shin, Taylor, M. S., & Seo, M.-G. (2012). Resources for change: The relationships of organizational inducements and psychological resilience to employees' attitudes and behaviors toward organizational change. *Academy of Management journal*, *55*(3), 727–748.

Sklearn. (2023a). Cross-validation: Evaluating estimator performance [Accessed on April 30, 2023]. https://scikit-learn.org/stable/modules/cross_validation.html#stratification

Sklearn. (2023b). Sklearn.ensemble.randomforestclassifier [Accessed on April 30, 2023]. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Sklearn. (2023c). Sklearn.linear$_{m}$*odel.logisticregression* [Accessed on April 30, 2023]. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Sklearn. (2023d). Sklearn.model$_{s}$*election.gridsearchcv* [Accessed on April 30, 2023]. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Sklearn. (2023e). Sklearn.model$_{s}$*election.train*$_{t}$*est*$_{s}$*plit* [Accessed on April 30, 2023]. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

Sklearn. (2023f). Sklearn.svm.svc [Accessed on April 30, 2023]. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Sousa-Poza, A., & Sousa-Poza, A. A. (2000). Well-being at work: A cross-national analysis of the levels and determinants of job satisfaction. *The journal of socio-economics*, *29*(6), 517–538.

Taheri, R. H., Miah, M. S., & Kamaruzzaman, M. (2020). Impact of working environment on job satisfaction. *European Journal of Business and Management Research*, *5*(6).

Tomar, S., & Gaur, M. (2020). Hr analytics in business: Role, opportunities, and challenges of using it. *Journal of Xi'an University of Architecture & Technology*, *12*(7), 1299–1306.

Treiblmaier, H., & Filzmoser, P. (2011). Benefits from using continuous rating scales in online survey research.

Van Wyk, R., Boshoff, A., & Cilliers, F. (2003). The prediction of job involvement for pharmacists and accountants. *SA Journal of Industrial Psychology*, *29*(3), 61–67.

Vroom, V. H. (1964). Work and motivation.

Warr, P. B. (1990). Decision latitude, job demands, and employee well-being. *Work & Stress*, *4*(4), 285–294.

Wu, H., & Leung, S.-O. (2017). Can likert scales be treated as interval scales?—a simulation study. *Journal of Social Service Research*, *43*(4), 527–532.

Yadav, S., & Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. *2016 IEEE 6th International conference on advanced computing (IACC)*, 78–83.

Zeytinoglu, I. U., Yılmaz, G., Keser, A., Inelmen, K., Uygur, D., & Özsoy, A. (2013). Job satisfaction, flexible employment and job security among turkish service sector workers. *Economic and Industrial Democracy*, *34*(1), 123–144.

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2019). Employee turnover prediction with machine learning: A reliable approach. *Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 2*, 737–758.

Zhu, Y. (2013). A review of job satisfaction. *Asian Social Science*, *9*(1), 293.

Zou, M. (2015). Gender, work orientations and job satisfaction. *Work, employment and society*, *29*(1), 3–22.

APPENDIX A

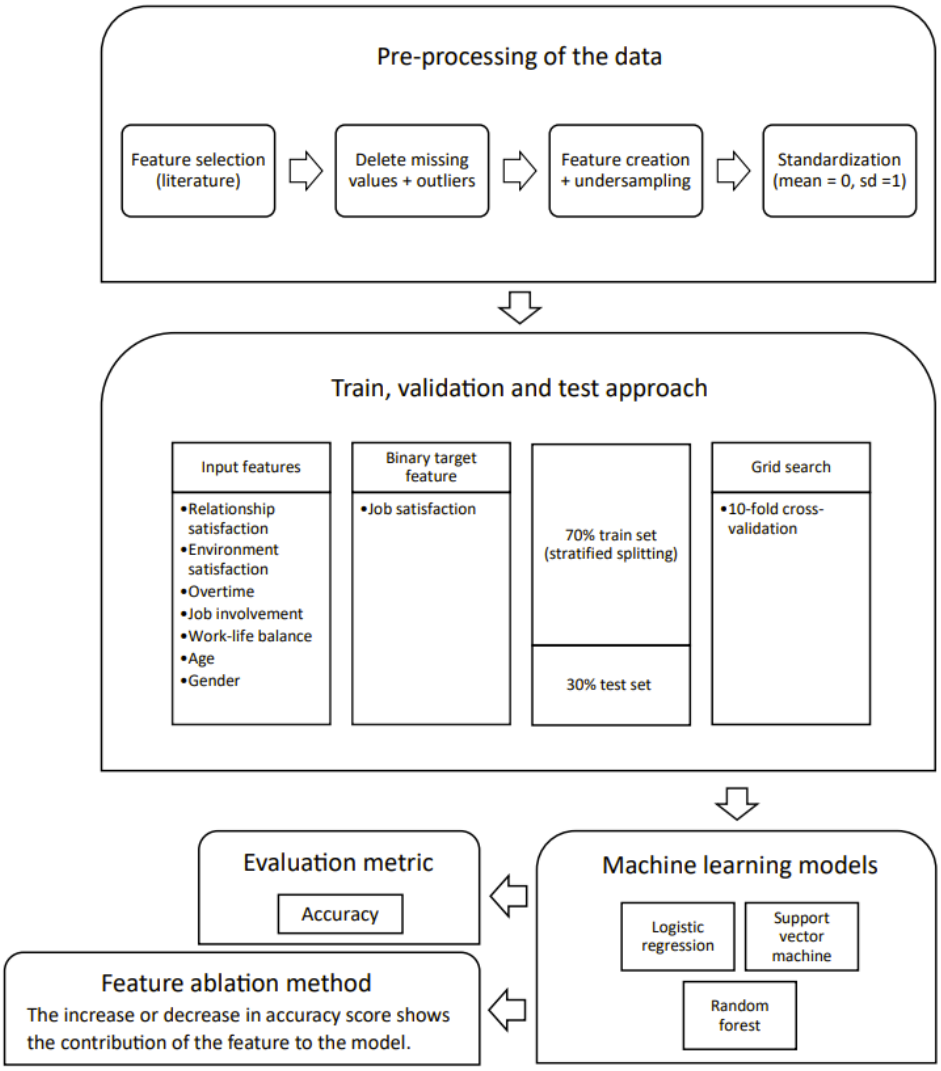
Pre-processing of the data
Feature selection (literature)
Delete missing values + outliers
Feature creation + undersampling
Standardization (mean = 0, sd =1)
Train, validation and test approach
Input features
•Relationship satisfaction
•Environment satisfaction
•Overtime
•Job involvement
•Work-life balance
•Age
•Gender
Binary target feature
•Job satisfaction
70% train set (stratified splitting)
30% test set
Grid search
•10-fold cross-validation
Machine learning models
Logistic regression
Support vector machine
Random forest
Evaluation metric
Accuracy
Feature ablation method
The increase or decrease in accuracy score shows the contribution of the feature to the model.

Figure 1: Visualisation of methodology and pipeline

APPENDIX B

Table 1
Characteristics dataset

Features	Type	Percentages	Min	Max	SD	Mean
Age	Numerical		22	50	8.37	35.86
Gender	Categorical (1-3)					
Male		32.00				
Female		33.99				
Not specified		34.01				
Relationship satisfaction	Continuous (1-5)					2.99
Environment satisfaction	Continuous (1-5)					2.98
Work-life balance	Continuous (1-5)					3.03
Overtime	Binary (0-1)					
No		49.50				
Yes		50.50				
Job involvement	Continuous (1-5)					3.02
Job satisfaction	Binary (0-1)					
Not satisfied		60.14				
Satisfied		39.86				

Note. Min is minimum, Max is maximum, SD is standard deviation.

APPENDIX C

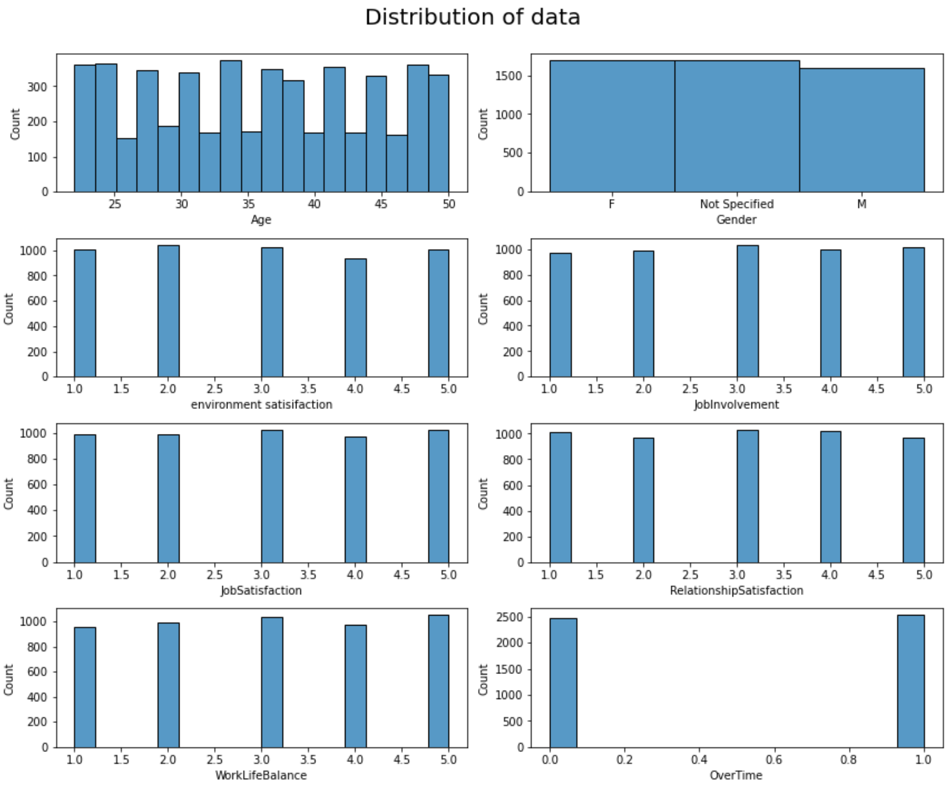

Figure 2: Histogram data distribution

APPENDIX D

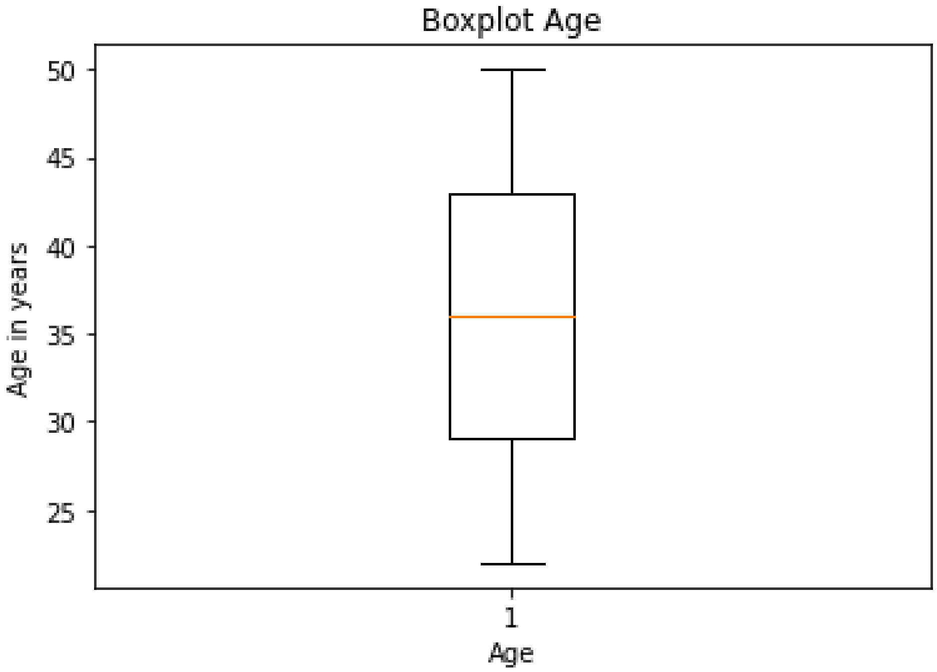

Figure 3: Boxplot age

APPENDIX E

Table 5
Classification report logistic regression (%)

	Precision	Recall	F1-score	Support
Not satisfied	52.13	51.00	51.56	600
Satisfied	51.96	53.09	52.52	599
Accuracy			52.04	1199
Macro avg	52.05	52.04	52.04	1199
Weighted avg	52.05	52.04	52.04	1199

Note. C=0.001, solver = newton-cg, penalty = none.

Table 6
Classification report support vector machine (%)

	Precision	Recall	F1-score	Support
Not satisfied	50.57	52.00	51.27	600
Satisfied	50.52	49.08	49.79	599
Accuracy			50.54	1199
Macro avg	50.54	50.54	50.53	1199
Weighted avg	0.50.54	50.54	50.53	1199

Note. C=1, gamma = 0.01, kernel = rbf.

Table 7
Classification report random forest (%)

	Precision	Recall	F1-score	Support
Not satisfied	50.96	62.17	56.01	600
Satisfied	51.39	40.07	45.03	599
Accuracy			51.13	1199
Macro avg	51.17	51.12	50.52	1199
Weighted avg	51.17	51.13	50.52	1199

Note. n_estimators = 100, min_samples_split = 2, min_samples_leaf = 2, max_features = 'sqrt', max_depth = 4, bootstrap = False.

APPENDIX F

Table 8
Confusion matrix logistic regression

	Predicted values		
Actual values		Not satisfied	satisfied
	Not satisfied	306	294
	Satisfied	281	318

Table 9
Confusion matrix support vector machine

	Predicted values		
Actual values		Not satisfied	satisfied
	Not satisfied	312	288
	Satisfied	305	294

Table 10
Confusion matrix random forest

	Predicted values		
Actual values		Not satisfied	satisfied
	Not satisfied	373	227
	Satisfied	359	240