Predicting Early Retirement Intentions of Mature Workers using Machine Learning and Deep Learning Algorithms

FOTEINI RAFAELA KANTARA

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

Word count: 8635 words

STUDENT NUMBER

970967
f.r.kantara@tilburguniversity.edu

COMMITTEE

dr. E.O.J. Vanmassenhove
E.O.J.Vanmassenhove@tilburguniversity.edu

dr. G. Spigler
g.spigler@tilburguniversity.edu

LOCATION

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

DATE

December 3, 2021

ACKNOWLEDGMENTS

I would like to recognize the invaluable assistance of E.O.J.Vanmassenhove, who supported me and advised me throughout this difficult project. Her feedback was very insightful and critical, which made me get the best out of myself. Moreover, I would like to thank dr. G. Spigler for devoting time in reading and reviewing this thesis. Lastly, I would like to express my honest gratitude for W. Buchholz, who believed in me and supported me during this process.

PREDICTING EARLY RETIREMENT INTENTIONS OF MATURE WORKERS USING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

FOTEINI RAFAELA KANTARA

Contents

Abstract

Predicting whether a particular individual may retire early is important for both the organizations the governments, as, due to the ageing of the general population, there are concerns about the workforce supply in Europe. The present study investigates the retirement intentions of individuals and aims to explore whether supervised learning algorithms can offer new insights in the area of retirement intentions. For this purpose, a publicly available dataset from European Social Survey is utilized. Contrarily to the rest of the studies in the retirement intentions field, the present paper adopts an approach in predicting retirement intentions with two stages: a pre-processing feature selection stage, using a Support Vector Machine-Recursive Feature Elimination (SVM-RFE) model and a processing stage, using a Logistic Regression, a Support vector Machine (SVM), a Random Forest (RF) and a Multilayer Perceptron (MLP). Moreover, this study also examines whether deep learning algorithms can offer new prospects in predicting retirement intentions. The findings suggest an improvement in SVM's and RF's performance using a SVM-RFE model at the pre-processing stage. Additionally, the balanced accuracy scores for assessing models' performance indicate that machine learning models are sufficient for predicting early retirement intentions and MLP did not perform better than the LogReg and SVM, which achieved the optimal accuracy.

1 Data Source and Ethics

The present study did not involve collecting data from humans or animals, but used a secondary dataset. Specifically, it utilized the publicly available dataset [1] from European Social Survey European Research Infrastructure Consortium (ESS-RIC) (European Social Survey, n.d.), which investigates the attitudes, behaviours and beliefs of individuals. The data in this study were collected in 2010 through questionnaires. The participants of this study consented to the data collection, their participation was voluntarily and their anonymity and confidentiality were ensured. There are no information included in the dataset that could possibly used to identify a specific living human being. The complete information about the privacy rules followed by the ESS-RIC during the data collection is publicly available [2].

Since the dataset and part of the code were obtained from external sources, the author acknowledges no legal claims to this dataset or to these parts of the code. The code that was used for the purposes of this study is publicly available [3].

2 Introduction

The aim of this study is to explore the potentials of machine learning (ML) and deep learning (DL) algorithms in predicting intentions to retire early. The approach that was adopted to answer the study's research questions consists of two stages: the pre-processing and the processing. At the pre-processing part, the algorithm utilized for the feature selection was Support Vector Machine - Recursive Feature Elimination (SVM-RFE) in order to distinguish the factors that significantly influence the prediction of retirement intentions. For the processing stage, Logistic Regression (LogReg), Support Vector Machine (SVM), Random Forest (RF) and Multilayer Perceptron (MLP) were the algorithms used for estimating the intentions to retire early. Contrarily to the previous research on this field, this study also exploits the potentials of feature selection methods in determining the most significant features and DL techniques in predicting whether an individual is planning to retire early.

Retirement intentions is definitely a crucial topic from the societal point of view. The ageing of the general population, caused by the increase in longevity and the decrease in births (StaLine, 2021), in combination with an increase in the ratio of retired individuals to labor force (Ilmarinen,

[1] https://www.europeansocialsurvey.org/download.html?file=ESS5e03_4&y=2010
[2] https://www.europeansocialsurvey.org/about/privacy.html
[3] https://github.com/kantarafr/Thesis

2001) poses new challenges for the social security systems and workforce supply in Europe. Although the average retirement age has increased in the last few years, there is still a considerable number of people that decide to retire earlier than the general retirement age of 65 years (, n.d.; Reeuwijk et al., 2013). To increase the number of mature workers that stay in labor force, we need to understand the factors behind retirement (Browne, Carr, Fleischmann, Xue, & Stansfeld, 2019) and how these influence the decision to retire of an individual. If no action is taken, a labor shortfall is inevitable, with significant consequences for both societies and organizations.

From a practical standpoint, comprehending the reasons behind intentions to retire early and predicting whether someone would retire before the age of 65 is relevant for designing organizational policies or societal interventions to promote and increase employability of mature employees (Van Solinge & Henkens, 2014). Moreover, predicting how many people are going to retire prematurely can help the governments plan their expenditures and reform their pensions schemes.

The topic and the approach adopted by this study has also a scientific interest. Although there is an extensive literature focused on comprehending the reasons behind intentions of early retirement, such as metanalyses (e.g. Topa, Depolo, & Alcover, 2018), literature reviews (e.g. Browne, Carr, Fleischmann, Xue, & Stansfeld, 2019), empirical studies (e.g. Oakman & Wells, 2013; Reeuwijk et al., 2013), only a few studies employ ML and DL algorithms in combination with feature selection. In fact, to the author's knowledge, no one of the previous studies has used either feature selection or DL methods to examine the determinants behind intentions to retire. Thus, utilizing the advances in big data technology, this research aims to investigate how various supervised learning algorithms can provide new insights into the area of early retirement intentions.

2.1 Research Questions

The present research aims to explore the following research question:

> *To what extent can early retirement intentions be predicted through ML and DL algorithms built from a dataset containing information about attitudes, beliefs and behaviour patterns of individuals in Europe?*

To achieve this objective, the sub-sequent questions will be addressed:

RQ1 *To what extent does feature selection using the SVM-RFE model improves algorithm's predictive performance in classifying retirement intentions?*

RQ1 *Which features significantly contribute in predicting early retirement intentions?*

RQ2 *How do deep models, like MLP, compare to shallow models like LogReg, RF and SVM?*

2.2 Findings

According to the results of this paper, the feature selection resulted in increased performance for SVM and RF, while for the MLP and LogReg lead to a slight decrease in their predicting ability. Regarding the modelling approach of this study, it is suggested that some ML models, like SVM and LogReg, might be sufficient in predicting individuals that intend to retire early. For this classification task, the optimal performance was achieved by SVM.

The present study is structured as following: Section 3 discusses the related studies on the field of intentions to retire early and supervised learning; Section 4 elaborates on the method adopted by this study and Section 5 explains the experiments conducted in order to answer the research questions; Section 6 presents the results and Section 7 explains and evaluates the findings. Finally, this paper concludes with Section 8 where the key-points of this study are synthesized.

3 Related Work

This chapter explores the relevant work in the field of intentions to retire. First, the research area and retirement intentions are defined. Second, the relevant literature is presented, starting with the determinants of intentions to retire early and continuing with the work of previous studies using ML algorithms. Finally, this section concludes with aim and the contribution of this study.

3.1 Research Area

The present study intends to provide helpful insights in the research area of intentions to retire and early retirement. According to (K. Henkens & Leenders, 2010, p. 306), "retirement is no longer considered an abrupt event of switching from working life to non-working life". It has rather been conceptualized as a complex process (Beehr, Bennett, Shultz, & Adams, 2007; Beehr & Bowling, 2013) and as a progressive transition that develops over several years before the actual retirement (Shultz & Wang, 2011). Retirement as such might "constitute the end of a gradual process of mental

withdrawal from the labor force" which is usually characterized by lower motivation, commitment and productivity (K. Henkens & Leenders, 2010, p. 307). A few researchers have compared this "mental retirement" to other forms of disengagement from work, such as psychological withdrawal, burnout and absenteeism (C. J. I. M. Henkens & Solinge, 2003). Here, some approaches highlight that it is crucial to focus on characteristics of this transitioning period rather than the time of the actual retirement (Wang & Shi, 2014). Furthermore, they underline that retirements intentions is a robust indicator of the actual retirement (Beehr, 1986; Beehr et al., 2007).

Predicting when individuals intend to exit the labour force have important implications for both organizations and governments. In regards to the organizational importance, mature workers constitute usually a significant asset of an organization, as they possess significant experience and knowledge of the processes, the procedures and the products of the organization. Loosing this knowledge prematurely might affect organizationals resources (Auer & Fortuny, 2000). Hence, designing interventions and policies to ensure that an organization will retain its mature workers is of noteworthy attention for the organizations (Van Solinge & Henkens, 2014). With respect to the societal value of predicting the number of people that intend to retire earlier than the official retirement age, it can help governments in reforming policies, introducing benefits for staying longer in the labour force, creating interventions in influencing the factors that push towards early retirement. These can lead to reducing the government's expenditures by reducing the number of early retirees (Kuhn, Grabka, & Suter, 2021).

3.2 Literature Review

3.2.1 Determinants of retirement intentions

There is an extensive literature focused on understanding retirement intentions, such as empirical studies (e.g. Oakman & Wells, 2013; Reeuwijk et al., 2013), metanalyses (e.g. Topa, Depolo, & Alcover, 2018) and literature reviews (e.g. Browne, Carr, Fleischmann, Xue, & Stansfeld, 2019). It has been proposed that there are three main domains that influence retirement intentions: personal attributes, environmental factors and retirement determinants (Szinovacz, 2003). For example, there are multiple studies that investigate intentions to retire and how it is influenced by an individual's wellbeing (ten Have, van Dorsselaer, & de Graaf, 2015) household (K. Henkens & Van Solinge, 2002), social class (Mina-Riera & Voyer, 2020) and work characteristics (Wahrendorf, Dragano, & Siegrist, 2013). Thus,

retirement intentions are a multifaceted process that requires analysis of multiple variables in order to be comprehended.

3.2.2 Supervised learning models in intentions to retire early and related fields

Supervised learning techniques used for prediction have recently dominated the internal decision-making processes within companies due to their capabilities to produce reliable results (Talwar & Kumar, 2013). Within the field of Human Resource Management (HRM), ML and DL techniques have been used for several binary classification problems. For instance, Patel et al. (2020) used k-Nearest Neighbours (k-NNs), Decision Trees, RF and SVM to predict the employee attrition. Their findings revealed that RF was the best performing model for this binary classification task. In a similar study, Frierson and Si (2018) tried to predict the employees intending to leave the company using Decision Trees, Neural Networks (NNs), SVM, k-NNs, Naïve Bayes and LogReg. Their study concluded that LogReg had the best performance in this attrition classification task, with an accuracy of 87%. In another binary classification study in the field of HRM, Lima, Vieira, and de Barros Costa (2020) used SVM, MLP, Recurrent Neural Networks and a Long Short-term Memory algorithm for the task of estimating whether an employee would be a long term absentee. They concluded that MLP was the model with the optimal performance, achieving an accuracy of 78%.

In the area of retirement and retirement intentions, the potentials of ML and DL algorithms have not been fully explored. Up to author's best knowledge, so far, only two studies employed ML techniques to estimate the early retirement or early retirement intentions. Particularly, Salazar and Penas (2019) analysed a dataset ($1500x11$) from private pension plans using RF, SVM, and LogReg to predict early retirement. It was suggested that LogReg could be a reliable model for estimating early retirement, as it achieved the highest accuracy score of 83%. In another study by Yun, Lee, Ji, and Lim (2017), KNNs, a shallow model of NNs with one layer, SVM, and Decision Trees were used in a larger dataset with fewer features ($14999x9$). Their results indicated that the Decision Trees had the highest accuracy (97%). However, the dataset used in this study might not have been completely appropriate for predicting retirement, as the target category referred more to attrition than to retirement.

Several studies in the field of Human Resource Management have suggested that it is necessary to adopt a complete approach in solving binary classification problems, with pre-processing and processing stages, in order to provide a reliable model and approach that can be used by the organizations. However, the studies in this field provide contrasting results. For example,Alduayj and Rajpoot (2018) used a filter feature selection

method to reduce the dimensionality of their dataset and predict attrition of employees. Their results indicated that their feature selection approach did not decreased the models' performance, and thus the whole dataset was used for the processing stage. There, they used SVM, RF and KNN, with SVM achieving the highest accuracy (69%). In contrast Najafi-Zangeneh, Shams-Gharneh, Arjomandi-Nezhad, and Hashemkhani Zolfani (2021) used a wrapper method for feature selection to eliminate the redundant features and enhance the algorithms' performance. At the processing stage, they adopted LogReg to predict the attrition of employees. Their feature selection method increased the model's performance from 78% to 81%. In a similar study, Shankar, Rajanikanth, Sivaramaraju, and Murthy (2018) used another wrapper feature selection method, called SVM-RFE, to reduce the number of the features in their dataset, which helped increase their models' performance (KNN, SVM, LogReg, Decision Tree and Naive Bayes).

3.3 Current Study

Utilizing the potentials of ML algorithms in the field of intentions to retire early is considerably new in the literature, with only two studies adopting supervised learning algorithms to address this topic. However, none of the studies have examined whether a feature selection method can potentially provide more reliable results or whether more deep, complex model structures are more appropriate in predicting intentions to retire.

Dealing with the limitations of the previous studies in this field, the present paper presents a two-stage framework, with pre-processing and processing, for creating a precise and reliable model in predicting individual's intentions to retire. First, at the preprocessing stage, the wrapper method SVM-RFE is introduced in order to eliminate redundant features and reduce the complexity of the algorithms (Guyon & Elisseeff, 2003). Second, this study also employs four models in order to predict intentions to retire early: a) a LogReg, b) a SMV, c) a RF and d) a MLP. The latter is utilized in order to examine whether a more complex algorithm is needed in order to accurately predict intentions to retire early.

4 Method

This chapter discusses the modelling approach adopted by this study. First, the feature selection method the SVM - RFE is presented. Subsequently, the mathematical principles behind the four chosen computational algorithms are illustrated.

4.1 Recursive Feature Elimination with Support Vector Machine

Large datasets with several noisy features might have a negative impact on the machine learning algorithms, as they increase complexity and the computational time (Guyon & Elisseeff, 2003). Thus, applying a feature selection method at the pre-processing stage might lead to an increased model prediction ability. The RFE algorithm has been proposed as an approach to select the most relevant features for a classification task and eliminate the irrelevant attributes (Guyon, Weston, Barnhill, & Vapnik, 2002). RFE is a wrapper method for feature selection, which uses another ML model and assesses the performance of this model by recursively dropping attributes from the dataset. The RFE using SVM is a competent feature selection method and has shown promising results in similar classification tasks (Shankar et al., 2018). The steps followed in an SVM-RFE model are presented in Figure 1.

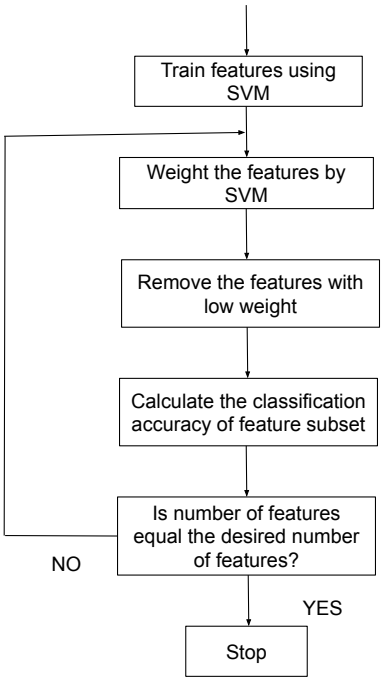


Figure 1: The process of an SVM-RFE model. Adjusted from Yousef et al. (2014)

4.2 Logistic Regression

LogReg was firstly proposed by Berkson (1944) and since then its use in has widely increased, with application in various fields, like medicine (Boateng & Abaye, 2019) and social sciences (Frierson & Si, 2018; LeBlanc & Fitzgerald, 2000). It is one of the most widely used parametric and classical methods for binary classification and it models the probability of an event, such as intentions to retire early/intentions to retire normally. In LogReg the independent features are combined linearly by using the coefficient values to predict the probability of an outcome, labelled in a binary way (0, 1). The main mathematical principle underpinning LogReg is the logit function - the natural logarithm of an odds that the outcome equals one of the two categories (Peng, Lee, & Ingersoll, 2002). The logit function is calculated as follows:

$$\log(P) = \ln(\frac{P}{1-P}) \tag{1}$$

where P the probability of an event. From the *logit* function, the logistic function is calculated, which is the inverse of logit, and it transforms the *logit* function into probabilities. It takes any real value between 0 and 1 (Peng et al., 2002). Figure 2 illustrates the relationship between the *logit* of an outcome occurring and the probabilities of an event. If the $logit(P)$ is a positive value, then the value of y will be 1 and if $\log(P)$ is a negative value, then the value of y will be 0. If P is more than 0.5, then y will be 1 and if it is less than 0.5, then y will be 0. The formula for calculating the logistic function is the following:

$$P = \frac{\exp^{\log(P)}}{1 + \exp^{\log(P)}} \tag{2}$$

LogReg was used as a baseline model, since it is characterized by simplicity, it has been used for similar tasks and achieved a high accuracy score (Frierson & Si, 2018; Salazar & Penas, 2019).

4.3 Support Vector Machine

SVM is a supervised learning algorithm (Cortes & Vapnik, 1995) that has been widely used for binary classification problems (Patel et al., 2020; Salazar & Penas, 2019). In SVM, all the data objects are depicted in an n-dimensional vector. SVM separates the data points into two classes by finding the hyperplane line which achieves the largest margin (see Figure3). Each margin is calculated as the sum of the smallest gap between the hyperplane line and the closest data point of each group. In that way, better generalization can be achieved (Yu & Kim, 2012). The mathematical

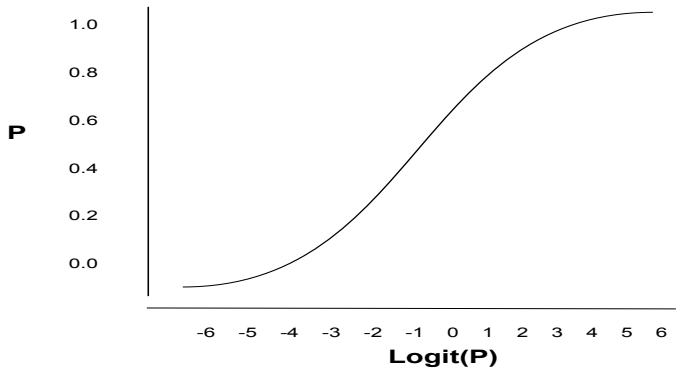


Figure 2: The logistic function. Adjusted from Cramer (2002).

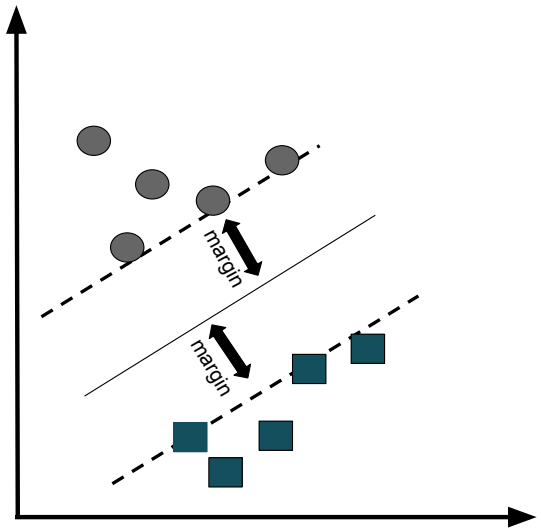


Figure 3: An SVM decision function illustrating the margin separating it from the data points of the two classes. Adjusted from Boyle (2011).

rationale behind SVM is the following:

$$mx + c = 0 \tag{3}$$

with m symbolizing the slope and c the intercept for a straight line. The hyperplane equation is written as:

$$H : w^T(x) + b = 0 \tag{4}$$

with b being the bias term. The distance of any line, $ax + by + c = 0$, for instance (x_0, y_0) is:

$$d = \frac{|ax_0 + by_0 + c|}{\sqrt{a^2 + b^2}} \tag{5}$$

and the distance of a hyperplane line with equation: $w^T\phi(x) + b = 0$ from a point $\phi(x_0)$ is the following:

$$d_H(\phi(x_0)) = \frac{|w^T(\phi(x_0) + b)|}{||w||_2} \tag{6}$$

and here the $||w||_2$ is the Euclidean norm for w (Yu & Kim, 2012).

4.4 Random Forest

RF algorithm is a classification method that was proposed by Breiman (2001). It is a supervised learning algorithm that has been extensively used in the field of social sciences for classification tasks (Frierson & Si, 2018; Patel et al., 2020; Salazar & Penas, 2019) and applied to a wide range of prediction problems. It is characterized by accuracy and its ability to handle high-dimensional feature spaces (Biau & Scornet, 2016). The algorithm constructs multiple decision trees and classifies the data points by the class chosen by the majority of the decision trees. In specific, an RF algorithm for classification creates a total of number of n subgroups from the training set, using bootstrap sampling and uses that to grow independent trees. For each random forest tree, the number of feature used for splitting nodes is randomly selected. The final random forest class prediction equal the majority vote of the individual random forest trees. By creating multiple decision trees, RF averts over-fitting over the training data (Hastie, Tibshirani, & Friedman, 2009).

4.5 Multilayer Perceptron

MLP is a deep, feedforward artificial NN (Haykin, 1994). It consists of at least three layers: an input layer, an output layer and one or more hidden layers. The number of neurons in the input layer are equal to the number of features in the dataset and the number of output layers are equal to the numbers of classes. When the algorithm receives the feature vector, it combines it with the bias terms and the initial weights in a weighted sum, which depends on the activation function. Each layer passes the result of the computation to the next one, till it passes the results to the output layer(Ramchoun, Idrissi, Ghanou, & Ettaouil, 2016). Moreover, the backpropagation technique helps the MLP to repetitively adjust the weight

values in the network. The structure of an MLP algorithm is presented in Figure 4. If we construct an MLP algorithm with an input layer of n_0 neurons $X = (x_0, x_1, ..., x_{n0})$ and utilized a sigmoid activation function:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{7}$$

The output of the first hidden layer would be

$$h_n^j = f(\sum_{k=1}^{n_{i-1}} w_{k,j}^0 x_k) \tag{8}$$

where $j = 1, ..., n_i$ symbolizes each neuron in the hidden layer $+1$, $w_{k,j}$ is the weight value between neuron k and layer i, n_i is the number of neurons in each hidden layer and h_i symbolizes each hidden layer. The output of the neurons in each hidden layer is computed as:

$$h_n^j = f(\sum_{k=1}^{n_{i-1}} w_{k,j}^{i-1} h_{i-1})^k \tag{9}$$

The final output of the MLP algorithm can be calculated by:

$$y_i = f(\sum_{k=1}^{K} w_{k,j}^N h_N^k) \tag{10}$$

$$Y = (y_1, .., y_{N+1}) = F(W, X) \tag{11}$$

where W is the matrix of weights, Y is the vector of output layer and F is the transformation function (Ramchoun et al., 2016). MLP has been used to predict similar task binary classification tasks (Lima et al., 2020) and showed good performance. MLP was employed in order to assess whether deep architectures are necessary for the accurate prediction of individuals who intend to retire early.

5 Experimental Setup

The aim of this chapter is to provide a detailed description of experimental set up of this study and consists of 6 subsections. Firstly, the characteristics of the dataset and the features used are described. Secondly, the preprocessing phase is explained in detail. Thirdly, the most significant insights from the exploratory data analysis are presented. Fourthly, the experimental procedure is outlined. Fifthly, the tools and packages utilized for the experimental setup are described. This chapter concludes with the evaluation metric used for assessing the performance of the algorithms. Moreover, the visualization of the overview of the experimental setup is depicted in subsection 5.7.

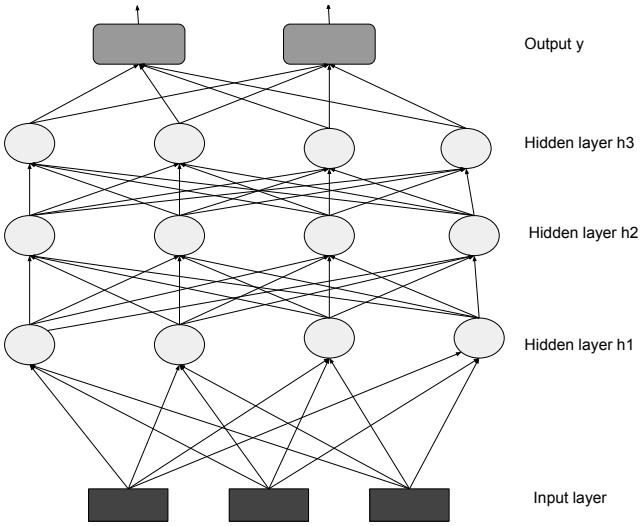


Figure 4: An MLP algorithm for a binary classification problem. Adjusted from Ramchoun et al. (2016).

5.1 Dataset

The experimental dataset in this study is obtained through a survey distributed by the National Center for Social Research in London, named the European Social Survey (ESS)(European Social Survey, n.d.). ESS is an academically driven survey, it has been conducted all over Europe since 2001 and it investigates beliefs, views, stereotypes and habits of individuals (European Social Survey, n.d.). For the purpose of this thesis, the data acquired in 2010 through round 5 were exploited, as this is the most recent survey that contains questions about an individual's work characteristics. As work environment or job characteristics has been found to play a crucial role in an individual's decision to retire (e.g. Wahrendorf, Dragano, & Siegrist, 2013), it was important to have features from this domain in the dataset.

The answers to the question "What age you would like to retired" served as the target variable. Although the data from this question were continuous, for the purposes of the present study, they were transformed into a binary variable. Specifically, all the instances that were less than the general retirement age of 65 years (Eurofound, 2021) were labelled as 1 and considered as early retirement intentions while the records with values more or equal to 65 were labelled as 0 and considered as intentions to retire at a normal retirement age.

The initial dataset included 52459 instances and 646 features. However, after the preprocessing phase (which is extensively elaborated in 5.2), due to list-wise deletion of multiple records and exclusion of features, the final dataset consisted of 12251 records and 132 features.

The dataset contained a wide variety of social and personal characteristics that might significantly influence an individual's intentions to retire early. Many of the attributes have been associated with intentions to retire early in previous studies(e.g. Blekesaune & Solem, 2005; Khan, Teoh, Islam, & Hassard, 2018; Schreurs, Van Emmerik, De Cuyper, Notelaers, & De Witte, 2011).

5.2 Pre-processing

Since preprocessing is crucial for a model's predictive behaviour (Li et al., 2019), several data cleaning and transformation techniques were applied to the raw dataset. At the initial stage, the records that participants had indicated "Not applicable", "Refused to answer", "No answer" or "I don't know" at the target variable were excluded, because it was meaningful to have the actual responses from the participants. Additionally, the participants that denoted that had already retired by the time that the data were collected were also dropped from the dataset. Moreover, the variables that were considered as redundant at first glance were excluded from the dataset. Furthermore, since this study concerns only mature workers, following the definition of a mature worker by Pitt-Catsouphes, Matz-Costa, and Brown (2011), list-wise deletion was applied to the records that indicated actual age or intentions to retire age below 40.

At the second phase, outliers and features having high percentage of missing values were handled. List-wise deletion was applied to the outliers on the continuous variables, like age and intentions to retire age. As outliers were considered all the records with age larger than 70 and intentions to retire age larger than 80. Subsequently, the continuous intentions to retire feature was binarized, with 1 being the category for early retirement intentions and 0 the category for normal retirement intentions (see Subsection 5.1). Furthermore, features with missing values of more than 20% were dropped (Enders, 2003). However, there were three variables that had from 20% to 40% of missing values and were kept in the dataset, namely: "How satisfied are you in your main job", "Satisfied with balance between time on job and time on other aspects" and " I would enjoy working in current job even if did not need money". Previous empirical studies have found that these features might be significant predictors of intentions to retire early (Pit & Hansen, 2014; Sibbald, Bojke, & Gravelle, 2003).

At the next stage, One-Hot Encoding was selected as the approach to encode the categorical data, since it is a relatively widespread approach for dealing with categorical data (Farahnakian & Heikkonen, 2018). This resulted in the final dataset, which consisted of 12251 observations and 132 attributes. After that, the dataset was split into train, validation and test set, with 70% of samples for training, 20% for validation and 10% for testing. Since the target variable was highly imbalanced (see Figure 8), the stratified sampling technique was used ensuring that the dataset was divided into homogeneous groups. Afterwards, the numeric features were normalized using MinMaxScaler. This scaler translates each value such that it lies in a range between 1 and 0.

At the final pre-processing stage, the missing values were replaced with missing value imputation methods, as it is a common solution in dealing with incomplete datasets (Lin & Tsai, 2020). In specific, Multivariate imputation and the function Iterative Imputer was used, which utilizes regression techniques and the information in the other features in order to estimate the missing values (Buck, 1960). In this technique, at each step a regressor takes one of the columns as the target variable, y, and the rest of the columns as predictors, X. The model fits on the X and the known y and tries to predict the unknown y. In this way, the missing values are imputed (Buck, 1960).

5.3 Exploratory Data Analysis

An exploratory examination of the data is important in any research analysis (Komorowski, Marshall, Salciccioli, & Crutain, 2016). The goals of any Exploratory Data Analysis (EDA) is to check the quality of the data, to calculate descriptives and explore the distributions of the features (Chatfield, 1986).

First, the distribution of the dependent variable (age of intentions to retire) before it was binarized was explored. Figure 5 shows the distribution of intentions to retire age by gender. As it can be inferred from the plot, there was a large number of people intending to retire at the age of 60. Moreover, the distribution of the actual age of the individuals is represented in Figure 6. As it can be seen, the age is slightly positively skewed to the right. In addition, the bivariate plot of age and intentions to retire age is depicted in Figure 7. It can observed that there is a large number of participants aged between 45 to 55 intended to retire at the age of 60.

Finally the intentions to retire age was plotted after it was transformed to binary variable. As it can be noticed from Figure 8, the sample was highly imbalanced, with 78% of the instances belonging to the positive class.

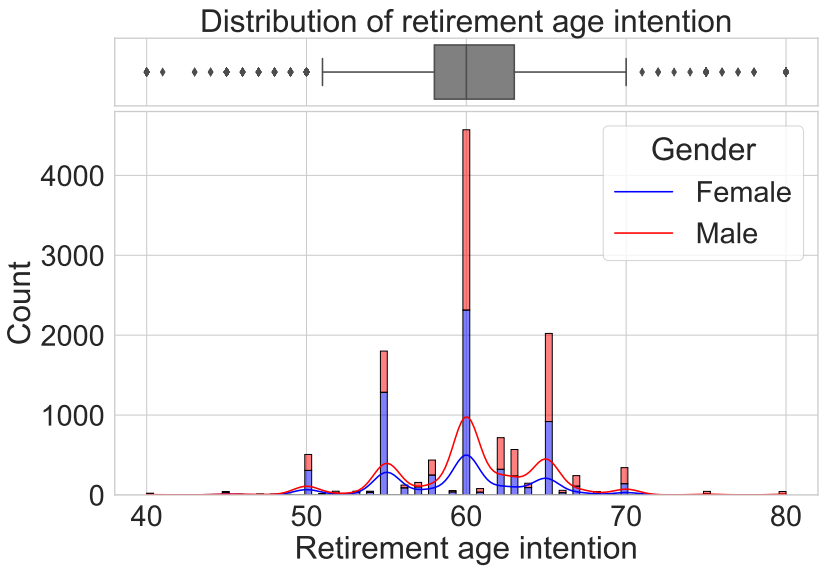


Figure 5: Distribution of Intentions to retire age by gender

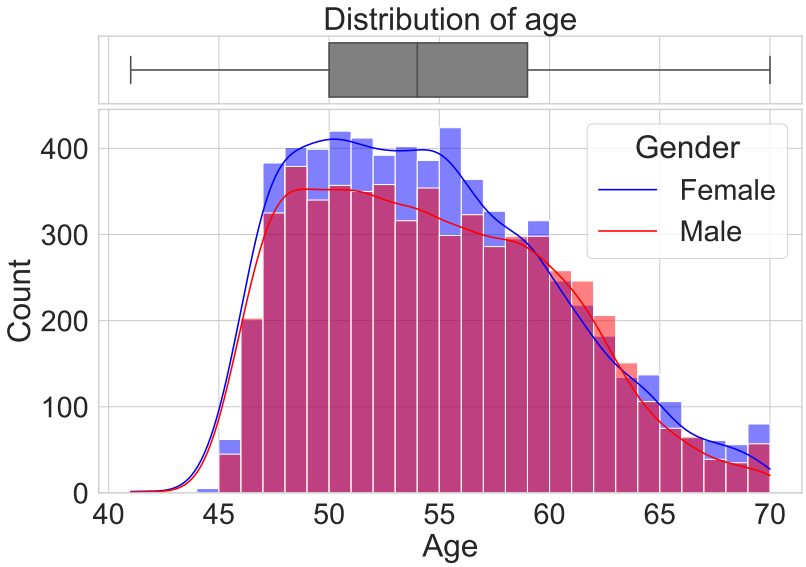


Figure 6: Distribution of age by gender

For ensuring that the data were appropriate for conducting LogReg, the assumptions of multicollinearity and the linearity of log odds and continuous features were examined. For the multicollinearity assumption of LogReg, Pearson's correlations coefficient were used in order to asses how strongly two variables were associated. According to Senaviratna and

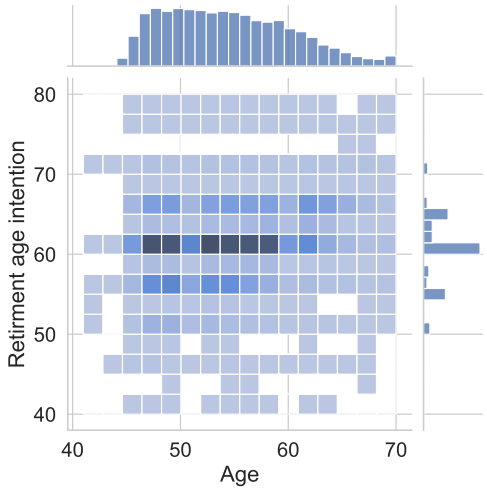


Figure 7: Distribution of age by intentions to retire age

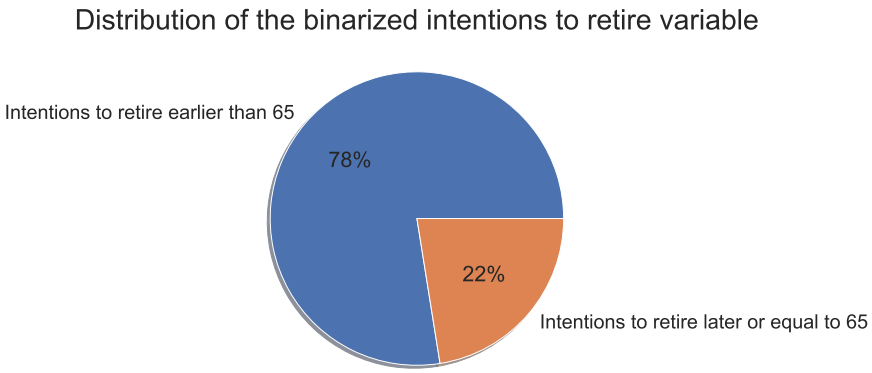


Figure 8: Age of intentions to retire binary

Cooray (2019), as a rule of the thumb, correlations greater than 0.8 or 0.9 indicate a serious problem of multicollinearity. There were no correlations 0.8 or above were observed. Moreover, to examine the linearity with the logodds for continuous features, the graphs in Appendix A (page 38) were plotted. It can be observed at the independent variables follow a linear distribution with the log odds. Thus, it could be inferred that the data were appropriate for performing LogReg. Since Random Forest (Smith, Ganesh, & Liu, 2013), SVM (Mahadevan, Shah, Marrie, & Slupsky, 2008) and MLP (Smith et al., 2013) are non parametric, no further tests were conducted for those algorithms.

5.4 Experimental Procedure

This section describes the experimental procedure utilized to achieve the research aims of this study. Firstly, the feature elimination method is presented. Subsequently, the ML and DL algorithms and the architectural configurations are laid out.

5.4.1 Recursive Feature Elimination with SVM

In order to select the most relevant features for the analysis, the SVM-RFE technique was used. For this purpose, the function *RFE* from the package Scikit-learn (Pedregosa, 2011) was utilized. The SVM-RFE was trained using the training set. The parameters of an RFE model that can be adjusted, according to the documentation of Scikit-learn[4], are the following: a) the estimators (which in this case was the SVC model, with the parameter 'balanced' for adjusting for class imbalanced data and kernel linear), b) the number of features to select and c) the number of features to remove at each iteration. The number of feature to remove each time was set to 10 and the number of features to select iterated over various values from 10 to 90 in order to select the optimal number of estimators. After each SVM-RFE model with the different number of attributes was trained, it was used to predict the validation set and the balanced accuracy was measured. Moreover, in order to compare whether the feature elimination technique resulted in higher performance for the models and reduced noise in the dataset, the default configurations of all the models described below were trained on the initial dataset and on the dataset after the feature selection method and their performances were compared.

5.4.2 Logistic Regression

For the LogReg algorithm, the function *LogisticRegression* was used from Scikit-learn (Pedregosa, 2011). According to the documentation for LogReg from Scikit-learn[5], there are many parameters that can be optimized for LogReg. Due to the class imbalance in this study, the class weight parameter was set to 'balance', adjusting the weights inversely proportional to the sizes of the classes. The LogReg model was tuned for the values of regularization of strength in order to decrease the magnitude of the parameters and reduce over-fitting.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

5.4.3 Support Vector Machine

The function *SVC* from Skikit-learn (Pedregosa, 2011) was used for SVM. The hyper parameters that SVM was tuned for were a) the C values, which account for the penalty, and b) the kernel. Moreover, the class weight was also set to 'balanced' and the rest of the hyper-parameters were kept with their default values[6].

5.4.4 Random Forest

The function *RandomForestClassifier* from Scikit-learn (Pedregosa, 2011) was used for RF. The class weight was adjusted to 'balanced', the hyper-parameters "number of trees in the forest" and "the maximum number of features to take into account when looking for the best split" were optimized. The rest of the parameters were kept with their default settings[7].

5.4.5 Multilayer Perceptron

For the MLP algorithm, the function *MLPClassifier* was used from Scikit-learn (Pedregosa, 2011). The number of neurons in each hidden layer was set to 50 and 4 hidden layers were constructed. The hyper parameters that were optimized were a) the batch sizes and b) the activation function, while the maximum number of iterations were set to 2000. The other hyperparameters of the MLP were kept with their default values[8].

5.4.6 Tuning Algorithms

To find the optimal parameters for the algorithms, the package RandomizedSearchCV was used from Scikit-learn(Pedregosa, 2011) [9], with number of iterations set to 20. Random Search has been found to be an effective and efficient technique for optimization, with some studies finding that it can perform as good as or even better than GridSearch (Bergstra & Bengio, 2012). Random Search selects a grid of hyper parameter values and combines then randomly to train the model. Moreover, 5-fold cross validation was applied, with number of repetitions set to 3. After the optimal architectural configurations were found for each model, they were used to predict the test data and evaluate the balanced accuracy of the models.

[6] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[7] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[8] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

[9] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

5.5 Algorithms and Packages

For the purposes of the analyses, Python 3.8.3 programming language (Van Rossum & Drake Jr, 1995) was utilized. Moreover, the libraries Numpy 1.19.5 (Harris et al., 2020) and Pandas 1.1.1 (McKinney et al., 2010) were mainly used for data manipulation. For the data transformation, missing values imputation, the tuning, the training and the prediction stages Scikit-learn 0.24.2 (Pedregosa, 2011) and Statsmodels 0.13.1 (Seabold & Perktold, 2010) were used. Lastly, the package matplotlib 3.3.1 (Hunter, 2007) and seaborn 0.11.2 (Waskom et al., 2017) were used for visualization purposes.

5.6 Evaluation Criteria

To evaluate the quality of each algorithm and of each combination of hyper parameters, five cross validation experiments were performed for each architecture configuration. This model validation technique was conducted using as a sample both the training and validation sets. Out of the five cross validation experiments for each architecture configuration, the average score was kept to represent the configuration quality. Finally, the best set of hyper parameters for each algorithm was selected by keeping the architecture configuration with the highest score. This architecture configuration was trained again and used to predict the test data, for which the final algorithm accuracy was kept to represent the quality of each model.

As choosing a meaningful metric of generalizability is important in order to determine how well the algorithm has performed on a given dataset (Brodersen, Ong, Stephan, & Buhmann, 2010), the metric that was used to assess all the algorithms' performances was balanced accuracy. Due to the target category being highly imbalanced (Figure8), it was necessary to adopt a metric that is not sensitive to imbalanced data, contrarily to accuracy and precision (Tharwat, 2020). Balanced accuracy is a metric which has been proved to an accurate estimation of model's performance over datasets that are characterized by high imbalance (Bekkar, Djemaa, & Alitouche, 2013; Korkmaz, 2020). The balanced accuracy is a measure that combines specificity and sensitivity metrics and is calculated as follows:

$$BalancedAccuracy = \frac{1}{2}(Sensitivity + Specificity) = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+TP}) \tag{12}$$

The balanced accuracy scores of the best performing configuration for SVM, RF and MLP were compared with that of LogReg, which was used as the baseline model.

5.7 Overview

The process of the experimental setup is depicted in Figure 5.7.

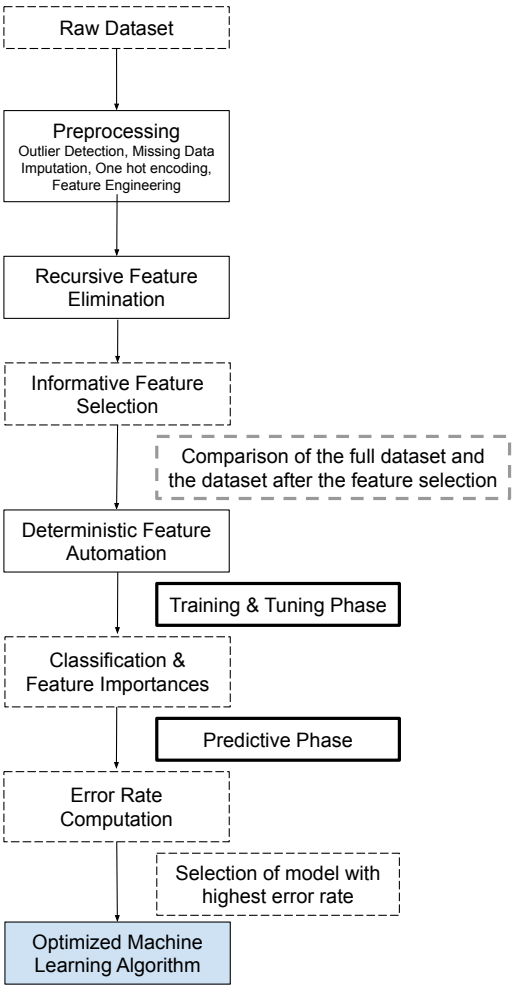


Figure 9: Overview of the Experimental Setup of study. Flowchart of development of a machine learning algorithm adapted from Bisele et al. (2017)

6 Results

This chapter is dedicated to present the results of the experiments. It is organized in two parts: a) the results of the SVM-RFE method and the features weights and b) the results of tuning the models and the performance of the final algorithms on the test set.

6.1 Feature Selection

The results from the SVM-RFE model are presented in Figure 10. As it can be seen, the model with 20 features performed better than the rest, producing a balanced accuracy of 0.70 on the validation set. The Table 1 shows the results. As it can be noted, for SVM and RF the feature selection technique resulted in higher balanced accuracy score. Contrarily, for LogReg and MLP, the balanced accuracy score with using all the feature in the dataset was higher. However, since the performance of at least two out of four algorithms was improved and feature selection results in reduced complexity and run time (Guyon & Elisseeff, 2003), the dataset after the feature selection was used for the processing stage.

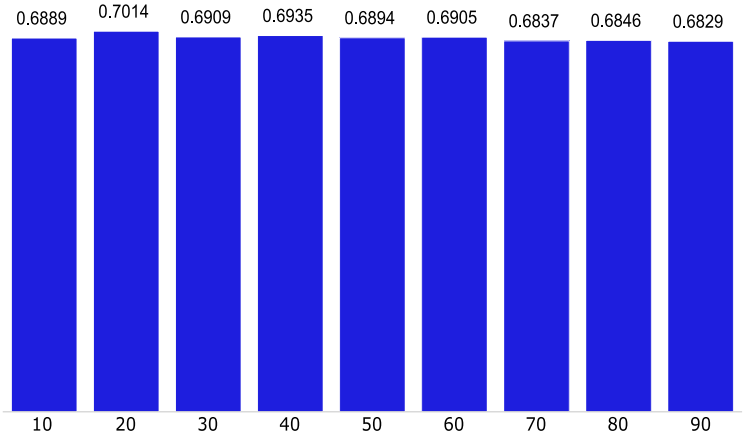


Figure 10: Accuracies on the validation set by the number of features

The 20 features presented in Figure 11 were considered as the most important attributes for an individual's intentions to retire early. According to Furey et al. (2000), the absolute sizes of the feature weights could be used as a ranking criterion. Thus, the "age" of the participant, "the years of working" and whether someone "enjoys his work" were considered as the three most important predictors of intentions to retire. It can be inferred

Table 1: *BalancedAccuracy* score for classifying intentions to retire.

Models	*BalancedAccuracy*(131 features)	*BalancedAccuracy* (20 features)
LogReg	0.71	0.69
SVM	0.692	0.694
RF	0.52	0.56
MLP	0.64	0.62

that the features belong to one of the following categories: family-related, individual-related, work-related or community-related (see Appendix B, page 41, for the table of features and the corresponding domain).

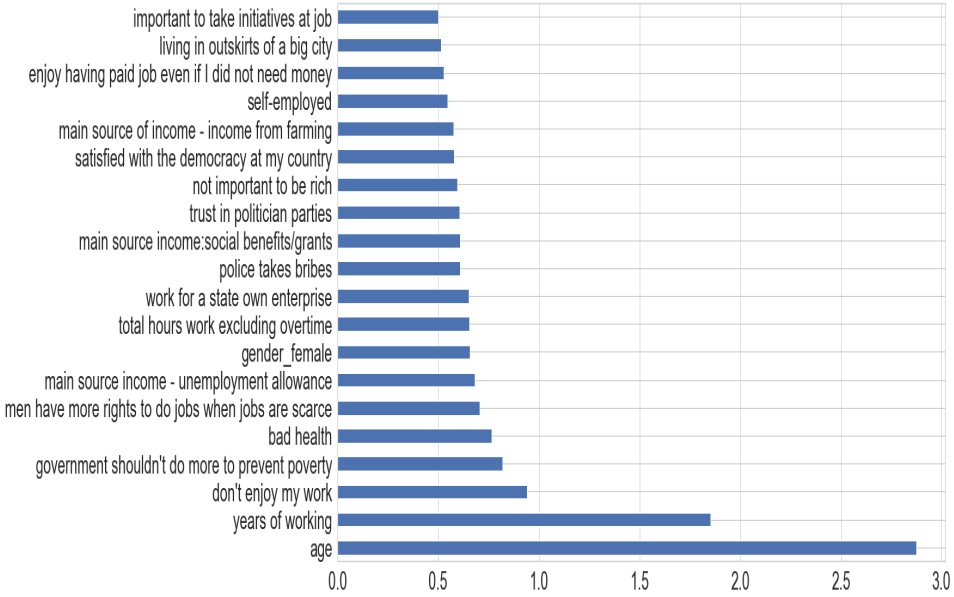


Figure 11: Absolute Feature Weight Coefficients from the SVM-RFE algorithm.

6.2 Performance of ML and DL algorithms

To explore which ML or DL algorithm is more reliable in predicting the intentions to retire early, the dataset with the 20 most important features was used. The following subsections explain in detail the specific architectural configuration used for each model as well as each model's final performance on the test set.

6.2.1 LogReg

The predictive performance of LogReg was proved to be a fairly reliable for the task of predicting intentions to retire early. The architectural configuration that yielded the highest balanced accuracy during the cross-validation sampling technique was the one with C value of 0.1. This configuration was used to predict the test test and produced an balanced accuracy of 69%. The summary of the prediction results is shown in Figure 12. It can be observed that LogReg performed fairly sufficiently in predicting both the negative (191 True Negative), which is the class with the less instances, and positive class (665 True Positive).

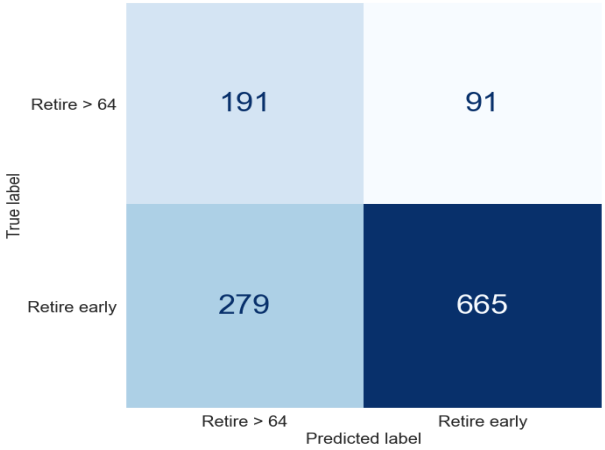


Figure 12: Confusion Matrix for LogReg

6.2.2 SVM

The SVM model yield the best predicting performance amongst all the algorithms. The optimal hyperparameters for performing this classification task with highly imbalance data were a C of 0.05 in combination with a radial basis function (RBF) kernel. For this classification task, the RBF allowed for non linear decision region. This particular configuration generated a balanced accuracy of 70% on the test set. The predictive performance of the SVM is depicted in Figure 13. As it can be seen, SVM predicted comparatively well in the positive class (623 True positive) and had the best performance in predicting the negative class (207 True Negative).

6.2.3 RF

The RF model showed a fairly poor performance in estimating the instances that express intentions to retire early. After tuning the model, the optimal hyperparameters were observed at number of trees in the forest 10 and

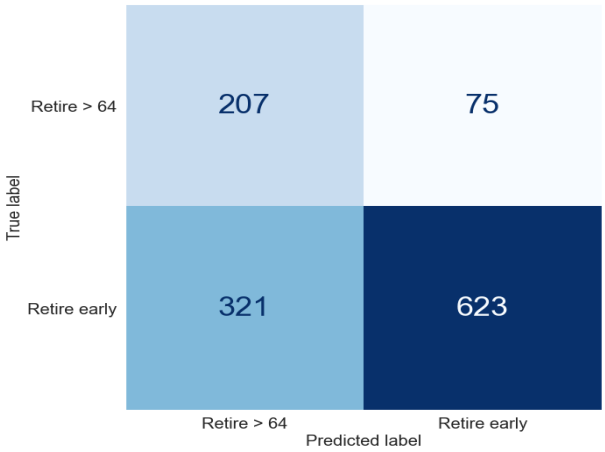


Figure 13: Confusion Matrix for SVM

number of features to consider when looking for best split 20. This algorithmic structure resulted in a balanced accuracy score of 58% on the test dataset. An overview of the model's performance is presented in Figure 14. It can be noticed that RF performed poorly in recognizing the negative class (74 True Negative) while it performed well for predicting the positive class (851 True Positive).

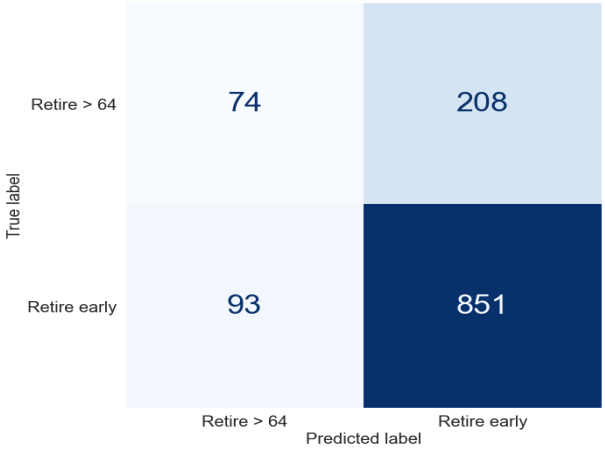


Figure 14: Confusion Matrix for RF

6.2.4 MLP

The results from the RandomSearch indicated that the optimal hyperparameters for the MLP were a batch size of 500 and an activation function of *tahn*. This configuration structure lead to a balanced accuracy of a 64% on the test set for this binary classification task. The confusion matrix

in Figure 15 shows the output of the model's predictive performance. It can be noted that MLP adequately predicted the negative class (124 True Negative) while it also predicts fairly well the positive class (785 True Positive). The Table 2 presents an overview of all algorithms' performances.

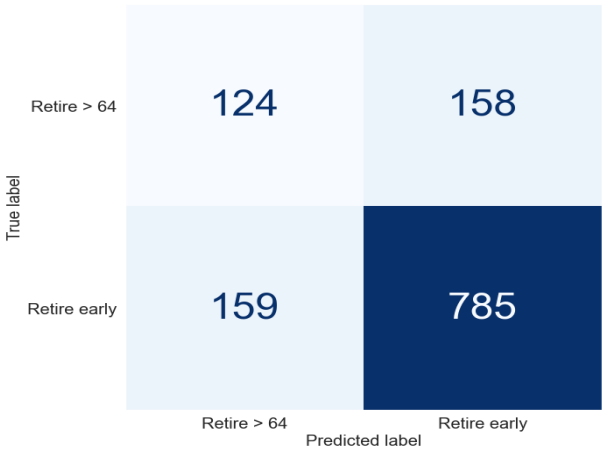


Figure 15: Confusion Matrix for MLP

Table 2: *BalancedAccuracy* score for classifying intentions to retire.

Models	*BalancedAccuracy* score
LogReg ($C = 0.1$)	0.69
SVM ($C = 0.01, kernel = rbf$)	**0.70**
RF ($number_estimators = 10, max_features = 20$	0.58
MLP ($activation = tahn, batch_size = 500$)	0.64

7 Discussion

This sections discusses the outcomes of the study. Initially, the purpose of the study is explained. Right after, the interpretation and the description of the study findings are provided. Finally, the limitations of this research and the suggestions for future research are presented. This section concludes with the contributions of this study.

7.1 Purpose of the study

The purpose of this study was to explore whether supervised learning techniques can offer new prospects in the field of intentions to retire.

For this purpose, first it was examined whether the use of the SVM-RFE model can effectively eliminate the noise from a dataset and improve the predictive ability of the algorithms. Moreover, the features selected as more prominent to the prediction of intentions to retire early were inspected. Finally, various ML models (LogReg, SVM, RF) and one DL model (MLP) were utilized to predict intentions to retire early. The aim of the latter experiment was to explore whether deep and complex algorithms are of added value to predict intentions to retire early.

7.2 Findings

One of the questions that this study explored was whether an SVM-RFE feature selection model would improve the models' performance. The literature suggests that using a feature selection model might be beneficial for a predictive algorithm, especially when the dimensions of the dataset is large, as it can decrease the noise and the complexity in the dataset and reduce the training time of the algorithm (Guyon & Elisseeff, 2003). Moreover, several studies that used wrapper methods and the SVM-RFE algorithm on tasks in HRM have proved that the predictive performance of the algorithms can improve, if the right subset is chosen (Najafi-Zangeneh et al., 2021). In partial agreement with those studies, the findings of the present paper suggest that the performance of some algorithms, like SVM and RF, can improve by using an SVM-RFE algorithm. However, some other algorithms might not benefit from the application of it, like LogReg and MLP, which is in accordance with Alduayj and Rajpoot (2018). The reason behind this might have been that either these methods require a larger training size or that the fact that this study used only a data-driven approach in feature selection, without extensively making use of prior knowledge (Chu et al., 2012). This might have resulted in eliminating the less informative features and maintaining only the most-informative features (Chu et al., 2012). However, due to the fact that there was only a slight decrease in these models' performance, the number of features chosen by the SVM-RFE were used for the processing stage, as the advantages of using it outweigh the disadvantages.

Secondly, this study explored the features that significantly contribute in predicting intentions to retire early. The 20 attributes that were the most influential for predicting intentions to retire are presented in Figure 11. It can be noted that these features belong to one of the following domains: family-related, community-related, individual-related, or work-related. These domains are also mentioned in the scientific literature as crucial for predicting intentions to retire (Elder & Johnson, 2018; Urick, Hollensbe, Masterson, & Lyons, 2017).

Finally, the final aim of this study was to investigate whether, due to multifaceted problem being investigated, DL models, like MLP, perform better than ML models, like LogReg, SVM and RF. Previous research in related field have shown that, in some tasks, LogReg, SVM and RF are sufficient for binary classifications in the field of HRM (Alduayj & Rajpoot, 2018; Salazar & Penas, 2019). According to the study results, algorithms like SVM and LogReg can accurately predict intentions to retire early, while MLP and RF had lower predictive performances. Additionally, SVM achieved the highest balanced accuracy score of 70%. The lower performance of the MLP might have been a result of not sufficient training data (Tweedie, Singh, & Holmes, 1996).

7.3 Limitations and Future Research

Despite having made several efforts to ensure that the results are reliable and valid, this study is constrained by a few limitations, which should be considered when interpreting the findings of this study.

Firstly, the data used for this study were collected in 2010. Thus, more research is needed in order to confirm whether the features selected as more important by the SVM-RFE algorithm could be still considered as important factors for intentions to retire early. Future studies with more recently collected data could confirm whether the 20 predictors of intentions to retire are still the most significant predictors of early retirement intentions. In addition, future research could explore whether countries specific factors, like regulations regarding retirement, might also affect the intentions to retire early (Salazar & Penas, 2019).

Secondly, since the feature selection method used was an RFE using an SVM algorithm and the best performing model in terms of balanced accuracy was an SVM model, there could be a possible bias. The final dataset might have included more favourable features for the performance of SVM. Therefore, future research could use different selection techniques and assess whether models' predictions changes and whether SVM continues to be the model with the highest performing ability.

Moreover, no one-sided conclusions can be drawn as to whether DL algorithms are not of an added value to predicting intentions to retire early, as this might differ depending on the dataset. In addition, the lower performance of the MLP algorithm could have been subject to not sufficient dataset. NNs usually require large amount of data for learning the categories correctly (Tweedie et al., 1996). Therefore, future studies could use larger datasets in order to confirm whether DL could provide more insight in the area of intentions to retire.

7.4 Contribution

The present study contributes to the research field of intentions to retire and early retirement by exploring whether a feature selection model can be used to choose the most prominent features and investigating whether DL models are of added value in predicting intentions to retire early. Firstly, following the approach used by Najafi-Zangeneh et al. (2021) in the related field of employee attrition, this study shows that adopting an SVM-RFE algorithm for eliminating the redundant features can beneficial for the predictive ability of some algorithms. Considering how multifaceted issue retirement intentions is, a feature selection algorithm might be of a great significant in order to choose the most important factors. Moreover, this study shows that ML algorithms can accurately predict early retirement intentions while more research is needed to assess whether DL models can offer more insights into retirement intentions field.

In respect to the societal contribution, this study offers a useful approach for both organizations and governments. In specific, the study's approach can be used by the governments that own the necessary data to track and control individuals that intend to retire early but also plan their expenditures according that. These findings can also be beneficial for the organizations, as they can use the approach suggested to identify mature employees that intend to retire early. In that way they can plan specific interventions and provide specific benefits that aim to prolong the stay of those mature workers in the organizations

8 Conclusion

The present study explores whether supervised learning algorithms can offer new prospects in the field of retirement intentions. This is achieved through three sub questions.

The first sub question is *"To what extent does feature selection using the SVM-RFE model improves algorithm's predictive performance in classifying retirement intentions?"*. The findings of this research indicate that using an SVM-RFE at the preprocessing stage for eliminating the noise in the dataset can be advantageous for SVM's and RF's predictive performance.

The second sub question is *"Which features significantly contribute to predicting early retirement intentions?"*. According to the findings, there were 20 features selected in total as more influential for predicting intentions to retire early.

The final sub question is *"How do deep models, like MLP, compare to shallow models like LogReg, RF and SVM?"*. The experiments of this study revealed

that LogReg and SVM performed better than MLP, while RF had the lowest performance. Moreover, SVM achieved the highest performance.

References

F. C. f. P. (n.d.). *Retirement ages.* Retrieved from https://www.etk.fi/en/work-and-pensions-abroad/international-comparisons/retirement-ages/

Alduayj, S. S., & Rajpoot, K. (2018). Predicting employee attrition using machine learning. In *2018 international conference on innovations in information technology (iit)* (pp. 93–98).

Auer, P., & Fortuny, M. (2000). *Ageing of the labour force in oecd countries: Economic and social consequences*. International Labour Office Geneva.

Beehr, T. A. (1986). The process of retirement: A review and recommendations for future investigation. *Personnel psychology*, *39*(1), 31–55.

Beehr, T. A., Bennett, M. M., Shultz, K., & Adams, G. (2007). Examining retirement from a multi-level perspective. *Aging and work in the 21st century*, 277–302.

Beehr, T. A., & Bowling, N. A. (2013). Variations on a retirement theme: Conceptual and operational definitions of retirement. In *The oxford handbook of retirement*.

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, *3*(10).

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American statistical association*, *39*(227), 357–365.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227.

Bisele, M., Bencsik, M., Lewis, M. G., & Barnett, C. T. (2017). Optimisation of a machine learning algorithm in human locomotion using principal component and discriminant function analyses. *PloS one*, *12*(9), e0183990.

Blekesaune, M., & Solem, P. E. (2005). Working conditions and early retirement: a prospective study of retirement behavior. *Research on Aging*, *27*(1), 3–30.

Boateng, E. Y., & Abaye, D. A. (2019). A review of the logistic regression model with emphasis on medical research. *Journal of Data Analysis and Information Processing*, *7*(4), 190–207.

Boyle, B. H. (2011). *Support vector machines: data analysis, machine learning and applications*. Nova Science Publlishers, Incorporated.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th*

international conference on pattern recognition (pp. 3121–3124).

Browne, P., Carr, E., Fleischmann, M., Xue, B., & Stansfeld, S. A. (2019). The relationship between workplace psychosocial environment and retirement intentions and actual retirement: a systematic review. *European journal of ageing*, *16*(1), 73–82.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society: Series B (Methodological)*, *22*(2), 302–306.

Chatfield, C. (1986). Exploratory data analysis. *European journal of operational research*, *23*(1), 5–13.

Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., Lin, C., Initiative, A. D. N., et al. (2012). Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage*, *60*(1), 59–70.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Cramer, J. S. (2002). The origins of logistic regression.

Elder, G. H., & Johnson, M. K. (2018). The life course and aging: Challenges, lessons, and new directions. In *Invitation to the life course: Toward new understandings of later life* (pp. 49–81). Routledge.

Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological methods*, *8*(3), 322.

Eurofound. (2021, Oct). *Retirement.* Retrieved from https://www.eurofound.europa.eu/topic/retirement

European Social Survey. (n.d.). *About the european social survey european research infrastructure esseric.* Retrieved 5.10.2021, from https://www.europeansocialsurvey.org/

Farahnakian, F., & Heikkonen, J. (2018). A deep auto-encoder based approach for intrusion detection system. In *2018 20th international conference on advanced communication technology (icact)* (pp. 178–183).

Frierson, J., & Si, D. (2018). Who's next: Evaluating attrition with machine learning algorithms and survival analysis. In *International conference on big data* (pp. 251–259).

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, *16*(10), 906–914.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*,

46(1), 389–422.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020, September). Array programming with NumPy. *Nature*, *585*(7825), 357–362. Retrieved from https://doi.org/10.1038/s41586-020-2649-2 doi: 10.1038/s41586-020-2649-2

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random forests. In *The elements of statistical learning* (pp. 587–604). Springer.

Haykin, S. (1994). Neural networks: A comprehensive foundation, prentice hall ptr. *Upper Saddle River, NJ, USA*.

Henkens, C. J. I. M., & Solinge, H. (2003). *Het eindspel: Werknemers, hun partners en leidinggevenden over uittreden uit het arbeidsproces*. Koninklijke Van Gorcum.

Henkens, K., & Leenders, M. (2010). Burnout and older workers' intentions to retire. *International Journal of Manpower*.

Henkens, K., & Van Solinge, H. (2002). Spousal influences on the decision to retire. *International Journal of Sociology*, *32*(2), 55–74.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi: 10.1109/MCSE.2007.55

Ilmarinen, J. E. (2001). Aging workers. *Occupational and environmental medicine*, *58*(8), 546–546.

Khan, A., Teoh, K. R., Islam, S., & Hassard, J. (2018). Psychosocial work characteristics, burnout, psychological morbidity symptoms and early retirement intentions: a cross-sectional study of nhs consultants in the uk. *BMJ open*, *8*(7), e018720.

Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory data analysis. *Secondary analysis of electronic health records*, 185–203.

Korkmaz, S. (2020). Deep learning-based imbalanced data classification for drug discovery. *Journal of chemical information and modeling*, *60*(9), 4180–4190.

Kuhn, U., Grabka, M. M., & Suter, C. (2021). Early retirement as a privilege for the rich? a comparative analysis of germany and switzerland. *Advances in Life Course Research*, *47*, 100392.

LeBlanc, M., & Fitzgerald, S. (2000). Logistic regression for school psychologists. *School Psychology Quarterly*, *15*(3), 344.

Li, Z., Xie, Y., Shangguan, L., Zelaya, R. I., Gummeson, J., Hu, W., & Jamieson, K. (2019). Towards programming the radio environment with large arrays of inexpensive antennas. In *16th {USENIX} symposium on networked systems design and implementation ({NSDI} 19)* (pp. 285–300).

Lima, E., Vieira, T., & de Barros Costa, E. (2020). Evaluating deep models

for absenteeism prediction of public security agents. *Applied Soft Computing*, *91*, 106236.

Lin, W.-C., & Tsai, C.-F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, *53*(2), 1487–1509.

Mahadevan, S., Shah, S. L., Marrie, T. J., & Slupsky, C. M. (2008). Analysis of metabolomic data using support vector machines. *Analytical chemistry*, *80*(19), 7562–7570.

McKinney, W., et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference* (Vol. 445, pp. 51–56).

Mina-Riera, N., & Voyer, V. (2020). Early retirement, social class, and family relationships in cloutier's bonne retraite, jocelyne (2018). *The Gerontologist*, *60*(6), 1011–1019.

Najafi-Zangeneh, S., Shams-Gharneh, N., Arjomandi-Nezhad, A., & Hashemkhani Zolfani, S. (2021). An improved machine learning-based employees attrition prediction framework with emphasis on feature selection. *Mathematics*, *9*(11), 1226.

Oakman, J., & Wells, Y. (2013). Retirement intentions: what is the role of push factors in predicting retirement intentions? *Ageing & Society*, *33*(6), 988–1008.

Patel, A., Pardeshi, N., Patil, S., Sutar, S., Sadafule, R., & Bhat, S. (2020). Employee attrition predictive model using machine learning. *International Research Journal of Engineering and Technology (IRJET)*, *7*(05).

Pedregosa, F. (2011). G. varoquaux, a. *Gramfort, V. Michel, B. Thirion, o. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, *96*(1), 3–14.

Pit, S. W., & Hansen, V. (2014). Factors influencing early retirement intentions in australian rural general practitioners. *Occupational Medicine*, *64*(4), 297–304.

Pitt-Catsouphes, M., Matz-Costa, C., & Brown, M. (2011). The prism of age: managing age diversity in the twenty-first-century workplace. In *Managing an age-diverse workforce* (pp. 80–94). Springer.

Ramchoun, H., Idrissi, M. A. J., Ghanou, Y., & Ettaouil, M. (2016). Multilayer perceptron: Architecture optimization and training. *Int. J. Interact. Multim. Artif. Intell.*, *4*(1), 26–30.

Reeuwijk, K. G., De Wind, A., Westerman, M. J., Ybema, J. F., Van der Beek,

A. J., & Geuskens, G. A. (2013). 'all those things together made me retire': qualitative study on early retirement among dutch employees. *BMC public health*, *13*(1), 1–11.

Salazar, J. d. J. R., & Penas, M. d. C. B. (2019). Scoring and prediction of early retirement using machine learning techniques:: application to private pensions plans. In *Anales del instituto de actuarios españoles* (pp. 119–145).

Schreurs, B., Van Emmerik, H., De Cuyper, N., Notelaers, G., & De Witte, H. (2011). Job demands-resources and early retirement intention: Differences between blue-and white-collar workers. *Economic and Industrial Democracy*, *32*(1), 47–68.

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th python in science conference.*

Senaviratna, N., & Cooray, T. (2019). Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics*, 1–9.

Shankar, R. S., Rajanikanth, J., Sivaramaraju, V., & Murthy, K. (2018). Prediction of employee attrition using datamining. In *2018 ieee international conference on system, computation, automation and networking (icscan)* (pp. 1–8).

Shultz, K. S., & Wang, M. (2011). Psychological perspectives on the changing nature of retirement. *American Psychologist*, *66*(3), 170.

Sibbald, B., Bojke, C., & Gravelle, H. (2003). National survey of job satisfaction and retirement intentions among general practitioners in england. *Bmj*, *326*(7379), 22.

Smith, P. F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of neuroscience methods*, *220*(1), 85–91.

StaLine. (2021). *Bevolking; kerncijfers.* Open data Cbs. Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37296ned/table?ts=1616744934832

Szinovacz, M. E. (2003). Contexts and pathways: Retirement as institution, process, and experience. *Retirement: Reasons, processes, and results*, *6*, 52.

Talwar, A., & Kumar, Y. (2013). Machine learning: An artificial intelligence methodology. *International Journal of Engineering and Computer Science*, *2*(12), 3400–3404.

ten Have, M., van Dorsselaer, S., & de Graaf, R. (2015). Associations of work and health-related characteristics with intention to continue working after the age of 65 years. *The European Journal of Public Health*, *25*(1), 122–124.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing*

and Informatics.

Topa, G., Depolo, M., & Alcover, C.-M. (2018). Early retirement: a meta-analysis of its antecedent and subsequent correlates. *Frontiers in Psychology*, *8*, 2157.

Tweedie, F. J., Singh, S., & Holmes, D. I. (1996). Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, *30*(1), 1–10.

Urick, M. J., Hollensbe, E. C., Masterson, S. S., & Lyons, S. T. (2017). Understanding and managing intergenerational conflict: An examination of influences and strategies. *Work, Aging and Retirement*, *3*(2), 166–185.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial* (Vol. 620). Centrum voor Wiskunde en Informatica Amsterdam.

Van Solinge, H., & Henkens, K. (2014). Work-related factors as predictors in the retirement decision-making process of older workers in the netherlands. *Ageing & Society*, *34*(9), 1551–1574.

Wahrendorf, M., Dragano, N., & Siegrist, J. (2013). Social position, work stress, and retirement intentions: a study with older employees from 11 european countries. *European sociological review*, *29*(4), 792–802.

Wang, M., & Shi, J. (2014). Psychological research on retirement. *Annual review of psychology*, *65*, 209–233.

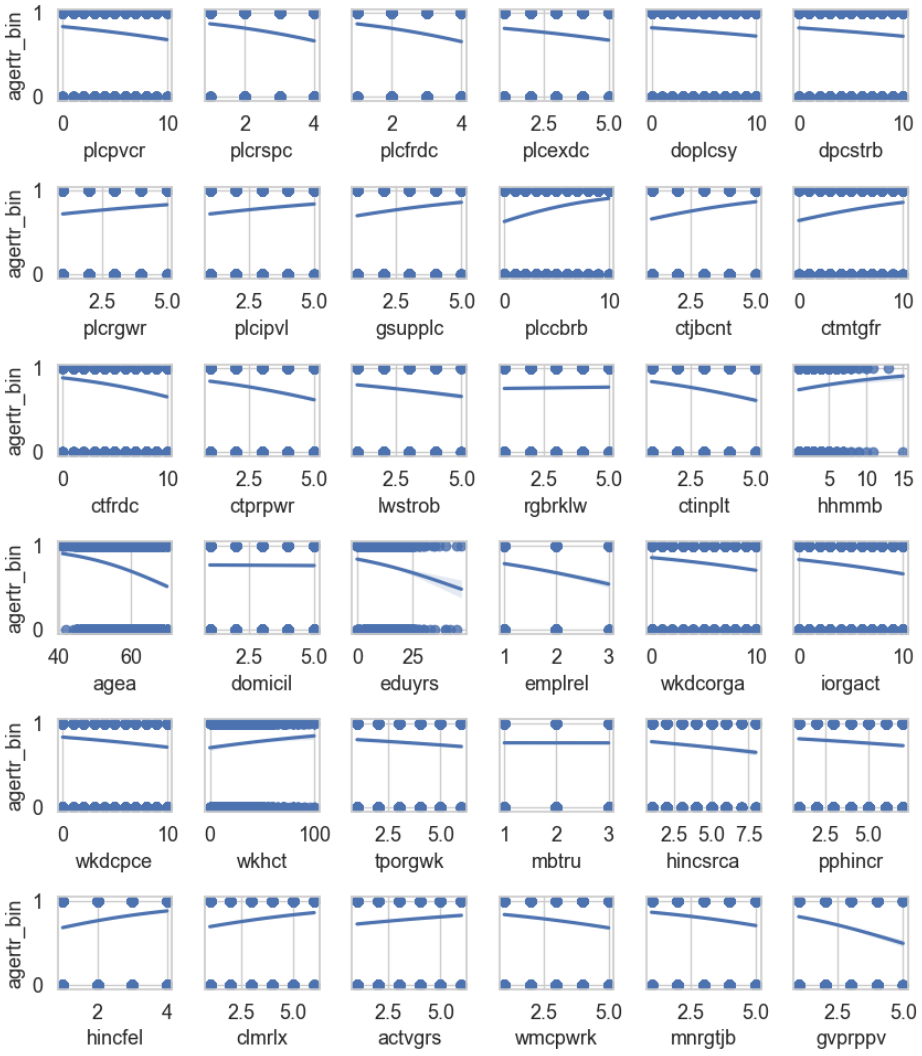
Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., ... Qalieh, A. (2017, September). mwaskom/seaborn: v0.8.1 (september 2017). Retrieved from https://doi.org/10.5281/zenodo.883859 doi: 10.5281/zenodo.883859

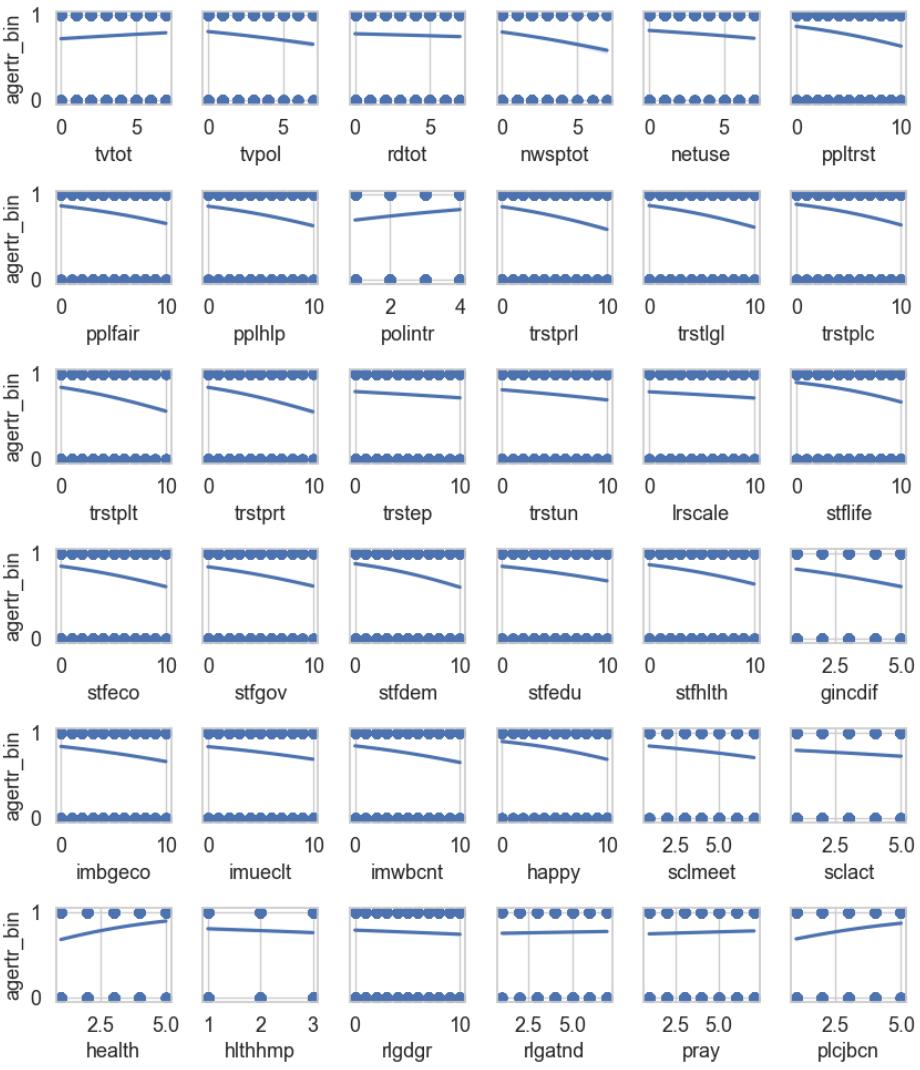
Yousef, M., Najami, N., Abedallah, L., Khalifa, W., et al. (2014). Computational approaches for biomarker discovery. *Journal of Intelligent Learning Systems and Applications*, *6*(04), 153.

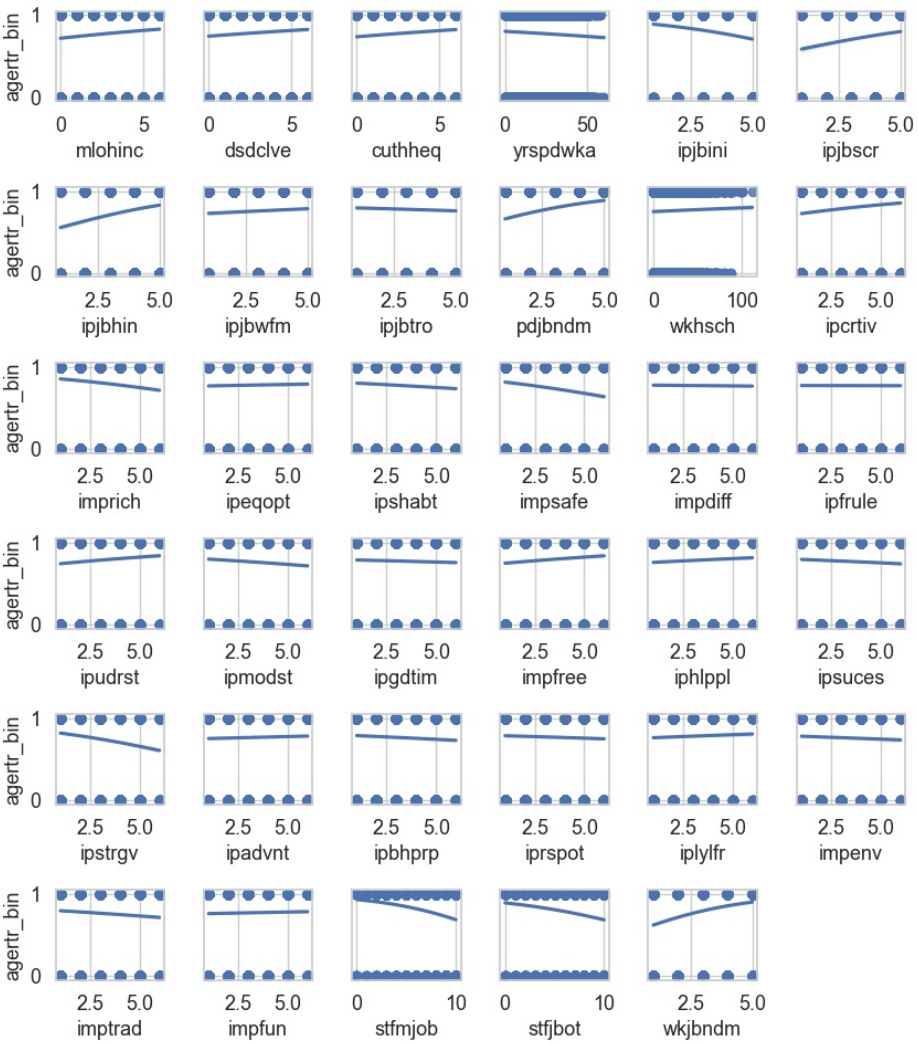
Yu, H., & Kim, S. (2012). Svm tutorial-classification, regression and ranking. *Handbook of Natural computing*, *1*, 479–506.

Yun, Y.-D., Lee, S.-H., Ji, H.-S., & Lim, H.-S. (2017). Development of retirement prediction model based on work life profile using machine learning method. *The Journal of Korean Association of Computer Education*, *20*(1), 87–97.

Appendix A: Linearity of independent variables and log odds




agertr_bin
1
0
tvtot
tvpol
rdtot
nwsptot
netuse
ppltrst
pplfair
pplhlp
polintr
trstprl
trstlgl
trstplc
trstplt
trstprt
trstep
trstun
lrscale
stflife
stfeco
stfgov
stfdem
stfedu
stfhlth
gincdif
imbgeco
imueclt
imwbcnt
happy
sclmeet
sclact
health
hlthhmp
rlgdgr
rlgatnd
pray
plcjbcn


agertr_bin
1
0
0 5
mlohinc
0 5
dsdclve
0 5
cuthheq
0 50
yrspdwka
2.5 5.0
ipjbini
2.5 5.0
ipjbscr
2.5 5.0
ipjbhin
2.5 5.0
ipjbwfm
2.5 5.0
ipjbtro
2.5 5.0
pdjbndm
0 100
wkhsch
2.5 5.0
ipcrtiv
2.5 5.0
imprich
2.5 5.0
ipeqopt
2.5 5.0
ipshabt
2.5 5.0
impsafe
2.5 5.0
impdiff
2.5 5.0
ipfrule
2.5 5.0
ipudrst
2.5 5.0
ipmodst
2.5 5.0
ipgdtim
2.5 5.0
impfree
2.5 5.0
iphlppl
2.5 5.0
ipsuces
2.5 5.0
ipstrgv
2.5 5.0
ipadvnt
2.5 5.0
ipbhprp
2.5 5.0
iprspot
2.5 5.0
iplylfr
2.5 5.0
impenv
2.5 5.0
imptrad
2.5 5.0
impfun
0 10
stfmjob
0 10
stfjbot
2.5 5.0
wkjbndm

Appendix B: Domains that affect intentions to retire early and features selected from SVM-RFE

Domain	Feature
Community/Social	statisfied with the democracy at my country
	trust in politician parties
	police takes bribes
	government shouldn't do more to prevent poverty
Family	living in outskirts of a big city
Work	important to take initiatives at job
	enjoy having paid job even if I did not need money
	self-employed
	years of working
	don't enjoy my work
	total hours work excluding overtime
	work for a state own enterprise
Individual	age
	bad health
	men have more rights to do jobs when jobs are scarcce
	main source income - unemployment allowance
	gender female
	main source income - social benefits/grants
	not important to be rich
	main source income - income from farming