USING MACHINE LEARNING TECHNIQUES TO DETERMINE THE MOST INFLUENTIAL FACTOR IN WORKING HOURS OF NATIONAL MINIMUM WAGE WORKERS IN IRELAND

Usukhbayar Tsendgombo

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

September 2023

Supervisor: David McQuaid

**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Capstone Project |
| **Assessment Title:** | Using machine learning techniques to determine the most influential factor in working hours of national minimum wage workers in Ireland |
| **Supervisor Name:** | David McQuaid |
| **Student Full Name:** | Usukhbayar Tsendgombo |
| **Student Number:** | 2022418 |
| **Assessment Due Date:** | 22nd of September 2023 |
| **Date of Submission:** | 21st of September 2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Summary**

[1. Research Question 1](#_Toc146217755)

[2. Relevance 1](#_Toc146217756)

[3. Contribution 2](#_Toc146217757)

[4. Objectives 2](#_Toc146217758)

[5. Literature review 3](#_Toc146217759)

[5.1. Statutory minimum wage 3](#_Toc146217760)

[5.2. Working hours 5](#_Toc146217761)

[5.3. Machine learning analytic techniques 10](#_Toc146217762)

[6. Validity 14](#_Toc146217763)

[7. Sampling Strategy 15](#_Toc146217764)

[8. Primary Research Methodology 15](#_Toc146217765)

[8.1. Dataset and Collection Method 16](#_Toc146217766)

[8.2. Data preparation and Feature engineering 17](#_Toc146217767)

[8.3. Machine learning techniques 20](#_Toc146217768)

[8.4. Evaluate Metrics 30](#_Toc146217769)

[9. Results and Discussion 32](#_Toc146217770)

[10. Conclusions and Recommendation 45](#_Toc146217772)

[References 48](#_Toc146217773)

[Appendix A 54](#_Toc146217775)

**List of Figures**

[Figure 1: Plot of the Correlation heatmap 20](#_Toc145772946)

[Figure 2: Random Forest Classifier 22](file:////Users/usukhuu/Documents/1.%20CCT-2022/Semester%203_Thesis/0.%20Thesis/2022418_Project.docx#_Toc145772947)

[Figure 3: Decision tree Classification 24](#_Toc145772948)

[Figure 4: Sigmoid Function 25](#_Toc145772949)

[Figure 5: Employees by NACE Economic Sector (Source: Personal Collection) 30](#_Toc145772950)

[Figure 6: Component Bar Chart of Percentage of Employees by Education Level (Source: Personal Collection) 31](#_Toc145772951)

[Figure 7: Nationalities of NMW (Source: Personal Collection) 31](#_Toc145772952)

[Figure 8: A weekly usual working hours of employees (Source: Personal Collection) 32](#_Toc145772953)

[Figure 9: Confusion matrix of the RFC (100 estimator) – based on test and prediction (Source: Personal Collection) 33](#_Toc145772954)

[Figure 10: Feature importance (Source: Personal Collection) 34](#_Toc145772955)

[Figure 11: Train and Test accuracy (Source: Personal Collection) 35](#_Toc145772956)

[Figure 12: Confusion matrix of the DTC (4 max\_depth) – based on test and prediction (Source: Personal Collection) 36](#_Toc145772957)

[Figure 13: Confusion matrix of the DTC (4 max\_depth) – based on test and prediction (Source: Personal Collection) 37](#_Toc145772958)

[Figure 14: RBF SVM with Different Gamma, C and Degree (Source: Personal Collection) 37](file:////Users/usukhuu/Documents/1.%20CCT-2022/Semester%203_Thesis/0.%20Thesis/2022418_Project.docx#_Toc145772959)

[Figure 15: Linear SVM with Different Gamma, C, and Degree (Source: Personal Collection) 38](file:////Users/usukhuu/Documents/1.%20CCT-2022/Semester%203_Thesis/0.%20Thesis/2022418_Project.docx#_Toc145772960)

[Figure 16:Poly SVM with Different Gamma, C and Degree (Source: Personal Collection) 38](file:////Users/usukhuu/Documents/1.%20CCT-2022/Semester%203_Thesis/0.%20Thesis/2022418_Project.docx#_Toc145772961)

[Figure 17: RFECV - Feature selection by Recursive Feature Elimination with Cross-Validation (Source: Personal Collection) 39](#_Toc145772962)

[Figure 18: LR Accuracy with different C hyperparameter values (Source: Personal Collection) 39](#_Toc145772963)

[Figure 19: LR Accuracy with different C hyperparameter values (Source: Personal Collection) 40](#_Toc145772964)

[Figure 20: Training and Testing Accuracy (Source: Personal Collection) 41](#_Toc145772965)

[Figure 21: ANN- Training and Testing Loss (Source: Personal Collection) 41](#_Toc145772966)

[Figure 22: Feature importance by different tree-based models (Source: Personal Collection) 42](#_Toc145772967)

[Figure 23: Permutation importance by different tree-based models (Source: Personal Collection) 43](#_Toc145772968)

[Figure 24: Cross-Validation and Accuracy scores by ML algorithms (Source: Personal collection) 44](#_Toc145772969)

**List of Tables**

[Table 1: Data Type and Description 18](#_Toc145773538)

[Table 2: Types of Machine Learning 21](#_Toc145773539)

[Table 3: Table of activation functions 27](#_Toc145773540)

[Table 4: Table of the Feature importance 34](#_Toc145773541)

[Table 5: Cross-Validation and Accuracy of the selected Machine Learning Models 44](#_Toc145773542)

## 1. Research Question

The impact of a minimum wage change on employment outcomes has been studied extensively over the last several decades. Despite the accumulated evidence, it remains difficult to definitively state whether a minimum wage increase negatively affects employment outcomes (Paul and Seamus, 2023). In some cases, Potential labour demand shifts caused by the Minimum wage introduction may also manifest themselves as a change in employees’ hours of work (Mark and Joanna, 2006). Furthermore, The low Pay Commision found that A frequently reported response to the minimum wage was a unilateral reduction of workers’ hours by the employer. In some instances, the worker was required to maintain the same output in fewer hours. The Fawcett Society evidence cited an example of ‘someone who cleans an office [who] is now expected to cover the same space in less time to get the same total wages as before’ (Low Pay Comission, 2000, p. 105).

By following those statements, this research aims to investigate the effects of minimum wage changes on working hours by using machine learning techniques. Furthermore, determine the most influential factor in working hours of national minimum wage workers in Ireland. This research seeks to address the following questions:

* Does changes in the minimum wage have an impact on working hours?
* What is the most influential factor in working hours of national minimum wage workers?
* Who is really mostly work as minimum wage?
* What solutions could be proposed on a data analytics perspective to assist people who work minimum wage?

## 2. Relevance

As the economy recovered, a Low Pay Commission was established in Ireland in 2015. Their role is to make yearly recommendations to the Irish government on a minimum wage that is ‘fair and sustainable’ and will ‘assist as many low-paid workers as possible without harming overall employment and competitiveness’. Following recommendations from the Low Pay Commission, the minimum wage was increased in January 2016 from €8.65 to €9.15 per hour. Based on recommendations from the Low Pay Commission, further increases to the minimum wage were implemented in 2017 to €9.25 per hour, 2018 to €9.55 per hour, 2019 to €9.80 per hour, 2020 to €10.10 per hour and 2021 €10.20 per hour. Finally, the national minimum hourly rate became €11.30 on 1 January 2023. The category of employee: Aged twenty and above €11.30, Aged nineteen €11.30, Aged eighteen €11.30, Aged under eighteen €11.30 (Paul and Seamus, 2022).

In line with this approach, it is required to investigate how changes in the minimum wage affect the working hours of minimum wage workers. This requirement has led many researchers to focus on find the answer. Many research studies have yelded varying results and employed different methods, such as by working sector (Paul and Seamus, 2021, 2022; Mark and Joanna, 2006; Duncan, Andrew and Seamus, 2017), on teenagers who is ages 16 to 19 (Couch and Wittenburg, 2001), gender differences on worked hours (Richard, Rebecca and David (2009), and difference-in-difference (DID) methods (Mark, Andrea and Mark, 2012). However, through a literature review, it was found that these studies did not used machine learning techniques in their research.

In this research, aim to by using machine learning techniques to determine the most influential factor in working hours of national minimum wage workers in Ireland. As a result, if a machine learning techniques can identify most important feature feature (Age, working sector, education level, nationality, etc.), they can provide valuable information to policymakers. It can underscore the importance of addressing that development of that specific feature to potentially improve the working conditions and income prospects of minimum wage workers. Hence, understanding the factors that affect working hours among minimum wage workers can contribute to a better understanding of employment trends and disparities.

## 3. Contribution

The novelty of this work lies in the use of multiple machine learning techniques to identify the most important feature within the highest accuracy predictive model for the working hours of national minimum wage workers. The main contributions of this work are firstly choice of the best Machine Learning models and secondly determine the most important feature from multiple features, defining the most appropriate method for the proposed context.

## 4. Objectives

4.1 The objective of this project is to extract the most important feature from highly accurate and generalised machine learning models. The chosen model should be that must effectively predict working hours of national minimum wage workers.

4.1. To explore and compare the different hyperparameters of machine learning models to highly accurate and generalized machine learning model. Due to time on this project, supervised different machine learning models should be trained.

4.3. Collect comprehensive data from the administrative data sources to train machine learning models for enhance training and testing results. The data must reflect the needs of the project, which capture a range of employee characteristics including education, nation, marital status and sector of work. Importantly, it also captures the usual hours worked of the employee.

4.3. Identify features that have the most effect on the usual hours worked of the employee by Tree-based models, Permutation Importance, and compare the results of methods.

## 5. Literature review

### **5.1. Statutory minimum wage**

Low Pay Commission (2019) produced the research pieces, from 2015-2018. Which are A review of International Literature on Minimum Wages, A study of Sub-Minimum Wage rates for Young People, Minimum Wage Employment, the Effect of an Increase in the Minimum Wage on Hours Worked and Employment, An Examination of the Labour Market Transitions of Minimum Wage Workers, The Impact of a change in the National Minimum Wage on the distribution of hourly wages in Ireland.

A study of minimum wage employment in Ireland: the role of worker, household and job characteristics has been researched by Bertrand, Seamus, and Paul (2017). They used Irish data from 2013 and 2014 that includes a rich set of personal, family and job-related characteristics. They find that just under 5 per cent of workers were in receipt of the NMW in 2014, a figure below the comparable UK rate of 7.7 per cent. The proportion of female employees earning the NMW was 6.9 per cent, which compares to an incidence of 2.7 per cent among male employees. Because female.

NMW were more concentrated in certain sectors and occupations, and had a higher propensity to have lower occupational tenure, to be employed in small firms and to work part-time.

In addition, at 9 per cent, the incidence of minimum wage pay among non-Irish nationals was over twice that of Irish employees. With regard to age, young people aged 18–29 years had the highest exposure to NMW employment, at 13.9 per cent. Workers with lower levels of schooling were more likely to fall into the minimum wage category relative to graduates. The youth disadvantage became statistically insignificant within the multivariate framework when factors related to job type were included (such as part-time work, temporary contract, working in a small firm). As was the case for gender, these results suggest that much of the non-Irish national and low educational disadvantages are explained by a combination of job type variables and a higher relative concentration in low-paid occupations.

Finally from a policy perspective, Bertrand, Seamus, and Paul (2017) the research raises a number of important questions. While females’ risk of earning the NMW is low, it is clear that, for many females on the NMW, their low income relates to their part-time status and higher concentration in sectors such as ‘wholesale and retail’, ‘accommodation and food’ and ‘other’. Further research is required to understand the extent to which females who choose to work part-time can do so within their chosen occupations or are forced to switch to lower-paying sectors and occupations that are typically associated with part-time employment. Given that sectoral effects also appear to play a role in explaining the higher relative risk experienced by females, non-Irish nationals and young people, further investigation is required into the reasons underlying low pay in sectors where NMW employees are heavily concentrated.

Also, The effects of minimum wages on youth employment and income has been examined by Charlene (2016). The research shows that the bulk of the empirical evidence supports the prediction of the standard economic model that minimum wages reduce employment and create unemployment among youths. It also shows that reducing or eliminating the minimum wage for young unskilled workers reduces these negative effects. Thus, minimum wages may not be the best way to improve the labor market situation of unskilled youth. While some working youth will benefit from increased current earnings, others will suffer from reduced opportunities and lower lifetime earnings. Delays in labor market entry and work experience will reduce lifetime incomes for youths who are unable to find employment because of the minimum wage.

There is a substantial body of empirical evidence on the effects of a minimum wage on youth employment. Most of the studies have found negative effects on youth employment. A 2014 study of youth employment in the US showed a decline of 1.5% for teenagers Neumark, Salas, Wascher (2014). Also, David, N. and William. W. (1992) research mentioned a 1992 study of youth employment in the US found that a 10% increase in the minimum wage led to a 1–2% decline in the employment of teenagers and a 1.5–2% decline in the employment of young adults.

Paul, Seamus (2022) examined the effect of a minimum wage increase on weekly hours worked of minimum wage workers in Ireland. They used a dataset called the Earning Analysis using Administrative Data Sources (EAADS), which links earnings from administrative sources to Labour Force Survey data. The data are administered by the Central Statistics Office (CSO) in Ireland. The administrative earnings data are then linked to data from the Irish Labour Force Survey (LFS), which capture a range of employee characteristics including age, gender, education, marital status and sector of work. Importantly, for our analysis, it also captures the usual hours worked of the employee. They used seven years of repeated cross-sectional data, from 2012 to 2018.

MaCurdy (2015) research investigated the antipoverty efficacy of minimum wage policies. The summary of findings are advocates of higher minimum wages often cite helping poor families as the primary motive for raising its value. They argue that families primarily supported by low-wage earnings will receive a substantial portion of the benefits and, moreover, that increasing minimum wages imposes very little public or social cost. Supporters contend that employment impacts experienced by low-wage workers are small, if any at all, and the pass-through of labor costs to prices induces negligible changes.

### **5.2. Working hours**

The study allows for the assessment of the cumulative impact of three minimum wage increases that might be missed in frameworks that examine the policy effect for one particular year. In addition to examining the effects on minimum wage workers generally, we examine heterogeneous effects for minimum wage workers in different sectors, regions, with different contract types, and of different nationality. In assessing impacts across all minimum workers, while no immediate impact was observed following the 2016 rate rise, there was a cumulative impact, resulting in a decline of 0.9 hours per week among minimum wage employees following the three rate rises from 2016 to 2018. With respect to the other heterogeneous impacts, they found evidence of declines in weekly hours worked over the 2016 to 2018 period among employees in both the industry and accommodation and food sectors. Minimum wage workers in 2018 in the industry sector were working three hours less per week, while those in the accommodation and food sector were working 2.5 hours less per week. They also found that by 2018, non-Irish minimum wage workers were working three hours less per week than their higher paid, non-Irish counterparts.

As a result, Paul and Seamus (2022) report shows the fact many non-Irish minimum wage workers are located in the sectors that were most affected by the rate rises, namely industry and accommodation and food.

Mark and Joanna (2006) research shows impact of the introduction of the UK minimum wage on the working hours of low-wage employees using difference-in-differences estimators. The estimates using the employer-based New Earnings Surveys indicate that the introduction of the minimum wage reduced the basic hours of low-wage workers by between 1 and 2 hours per week. The effects on total paid hours are similar (indicating negligible effects on paid overtime) and lagged effects dominate the smaller and less significant initial effects within this.

In the standard textbook analysis of the impact of a minimum wage, it is seen as raising the wage above its market clearing level, leading to a reduction in the demand for labour. This is usually interpreted as a reduction in employment. However, the adjustment, as well as possibly taking place at the extensive margin, i.e. a reduction in the number of workers, can also take place at the intensive margin, i.e. a reduction in 4 the number of (paid) hours per worker. In the long run a firm’s choice of workers-hours mix depends on the extent of fixed costs of employment, the technology and productivity-hours schedule, the labour supply schedule faced, the presence and effectiveness of a union, etc. However in the short run, as Hamermesh (1993) observes, “employers are quicker to alter hours in response to shocks than they are to change levels of employment”.

In the summary, Mark and Joanna (2006) presented broadly indicate a negative effect on hours, although the evidence is not unanimous and there is some variation in terms of both the magnitude and the significance of the estimated effects. The majority of the effect on total paid hours is found to be through the effect on basic hours. Typically, the effects on these two are very similar and the effect on paid overtime hours is minimal and insignificant. Lagged effects are found to dominate the initial effects. On the basis of the NES, the lagged effect on basic hours is estimated to be a reduction of between 1 and 2 hours per week. The lagged effect on total paid hours is very similar. The initial effect on basic hours is smaller (and in most cases insignificant). The LFS results are typically weaker than the corresponding NES ones in terms of significance. The estimates are generally found to be negative for both basic and total hours, and for both men and women. The NES total effect estimates indicate a reduction of between 1 and 2 hours per week in basic hours for both men and women, and similar for total paid hours.

Couch and Wittenburg (2001) examined the effect of minimum wage increases on the hours of work of teenagers (ages 16 to 19) using monthly data from the Current Population Survey. The primary reason for examining hours is that changes in aggregate employment might obscure an increase or decrease in labor demand as measured by hours of work. Consequently, the elasticities of labor demand estimated from employment data might provide a biased depiction of the overall responsiveness of labor utilization to changes in the minimum wage. They find that raising the minimum wage reduces hours worked by teens. Our results also indicate that estimates of the elasticity of teen labor demand with respect to the minimum wage based on employment data consistently understate the effect of minimum wage increases on labor utilization by 10 to 30 percent relative to those based on hours of work. The understatement of the impact of minimum wages on labor demand which occurs when aggregate employment rather than hours is examined is the result of employers choosing to decrease hours of teen workers who retain their jobs

Duncan, Andrew and Seamus (2017) examined the employment and hours impacts of the 1999 introduction of the UK National Minimum Wage (NMW) and the 2016 introduction of the UK National Living Wage (NLW) in Northern Ireland (NI) using Labour Force Survey data. This report provides neither the introduction of the NMW nor NLW impacted on average hours worked in Northern Ireland. On other hand, they conclude no additional reasons to expect that a further modest increase in the NLW would reduce average hours in Northern Ireland. Also, because the evidence of overall employment impacts of the NMW and NLW in NI is more mixed – no impact of the introduction of the NLW but a possible negative impact of the introduction of the NMW – this report provides no unambiguous message on whether a further modest increase in the NLW would reduce employment in NI.

Richard, Rebecca and David (2009) report investigates the impact of the 2001 to 2006 NMW upratings, a period where the NMW has risen substantially in excess of average earnings. Analysis of individual Labour Force Survey (LFS) and Annual Survey of Hours and Earnings (ASHE) data are presented along with local area analysis. The report shows hours worked most of the LFS models yield results that are not statistically significant, but in some cases they find the NMW is associated with a reduction in hours worked. There is some evidence to suggest that the larger upratings in 2001 and 2003 reduced basic hours worked amongst adult males. Overall, there is no evidence of a consistent impact on either basic hours or total hours. Similarly, the local area analysis finds no evidence of NMW impacts on hours worked.

Redmond, P., B. Maître, B. Seamus, M. and Konstantina, M. (2021) research found that the incidence of minimum wage employment varies considerably across EU countries. In Ireland, 9.6 per cent of employees are on the minimum wage. Countries with a relatively high incidence of minimum wage employment are Portugal (15.6 per cent), Germany (15.1 per cent), Poland (14.8 per cent), Hungary (14.2 per cent), Germany (14.0 per cent), Spain (14.0 per cent), UK (13.6%), Luxembourg (13.0 per cent) and Estonia (10.7 per cent). The incidence is low in Belgium (1.7 per cent), Netherlands (2.6 per cent) and Greece (4.5 per cent). The minimum wage rate in Ireland, in nominal terms, is the second highest of the 22 countries, after Luxembourg. However, in purchasing-power standard terms, the Irish minimum wage is just the seventh highest, behind Luxembourg, Germany, the Netherlands, Belgium, the UK and France. Also they found that Ireland and the Netherlands are the only two countries where there is no statistically significant difference in the incidence of minimum wage employment associated with gender. In all countries except Latvia, age is a strong predictor of minimum wage status. For example, in Ireland, employees aged above 29 years are five to eight percentage points less likely to be on the minimum wage relative to those under 29 years. In most countries, non-nationals are more likely to be on the 60 | Comparative Assessment of Minimum Wage Employment in Europe minimum wage than nationals; in Ireland, non-nationals are three percentage points more likely to be minimum wage employees than Irish nationals. Education level is also a significant factor in all countries. Tertiary-educated workers are less likely to be on the minimum wage compared to lower-educated workers. In all countries, working in accommodation and food or wholesale and retail increases the likelihood of earning the minimum wage.

The impact of the 2016 minimum wage increase on average labour costs, hours worked and employment in Irish firms has been researched by Paul and Seamus (2021). The Earnings Hours and Employment Costs Survey (EHECS) data accurately measure the proportion of employees in receipt of the national minimum wage across firms in Ireland. They use these data to: (a) carry out a detailed profile of minimum wage employment in Ireland, by examining the level and intensity of minimum wage employment across various sectors of the economy; (b) examine whether there was a greater increase in average labour costs in firms employing minimum wage employees following the 2016 rate rise, compared to firms with no minimum wage employees; (c) examine whether any changes occurred to hours worked, or the number of employees, in high intensity minimum wage firms following the 2016 minimum wage increase. They find that almost three-quarters of firms employ no minimum wage workers. Approximately 12 per cent of firms have less than 10 per cent of their workforce on the minimum wage. Just over three per cent of firms pay more than 50 per cent of their employees the minimum wage rate. Also, following the 2016 minimum wage increase, average weekly labour costs increased by 5.4 per cent more in firms with 100 per cent of employees on the minimum wage relative to firms with no minimum wage workers. However, the evidence suggests that these higher labour costs were confined to very high intensity firms, with more than 50 per cent of employees on the minimum wage. These firms account for just 3 per cent of all firms. For firms with between 10 and 50 per cent of employees on the minimum wage, they detected no statistically significant impact on average labour costs.

As a result, Paul, Seamus and Bertrand (2018) report shows an examination of the labour market transitions of minimum wage workers in Ireland. This study uses a new measure of minimum wage employment in Ireland, taken from the Quarterly National Household Survey (QNHS), to assess the degree to which individuals in receipt of the national minimum wage (NMW) transition in and out of NMW employment over a period of three quarters in 2016 and 2017. They do so using longitudinal data on 1,514 employees who were in receipt of the NMW in at least one of three consecutive quarters.

Consistent with much of the international evidence, they find that minimum wage employment often acts as a stepping stone to higher paid work. Of the 1,514 employees, 18 per cent remained on the minimum wage for all three quarters compared to approximately 30 per cent who transitioned from minimum wage to higher paid employment. The results also show that exits from NMW status to higher waged employment are achieved primarily through within-employer wage progression rather than between employer job change. Over 90 per cent of employees who transition to higher paid employment do not change occupation or employer. Approximately 13 per cent of the sample transitioned from higher pay to minimum wage employment, while 11 per cent transitioned to minimum wage employment from unemployment or inactivity.

Their multivariate analysis shows that Irish nationals, older workers, those with higher levels of education, full-time employees and those on permanent contracts are more likely to exit minimum wage employment to higher paid employment compared to non-nationals, younger persons, those with lower educational attainment, part-time workers and those on temporary contracts.

In UK, Mark, Andrea and Mark, T. (2012) research focus on three outcomes: employment retention, changes in working hours among employees, and the job finding probability of the unemployed. And the analysis uses difference-in-difference (DID) methods similar to previous studies of the impacts of the NMW, applied to data from the Labour Force Survey (LFS) . The research find some evidence that the NWM uprating had an impact on youth group. However, find no systematic effect on adults. But, the data size was not enough for this research, so their findings are only provisional.

### **5.3. Machine learning analytic techniques**

Machine learning is a field which focuses on building algorithms that make = predictions based on dataset. A machine learning task is designed to identify (learn) a function f: X/Y which maps 13 the input domain X (of data) to the output domain Y (of possible predictions) (Bekkerman and Bilenko 2011) Functions f are selected from different function groups, depending on the type of learning algorithm used. Mitchell (1997) describes "learning" as follows: "With regard to some class of tasks T and performance measure P, a computer program is said to learn from experience E if its performance at tasks in T, as calculated by P, increases with experience E" (Mitchell 1997). Quantitatively, the output metric P informs us how well a certain machine learning algorithm is doing. The precision of the system is typically chosen as the performance measure for a classification process, where accuracy is specified as the proportion for which the output is correctly generated by the system. Knowledge E that undergoes machine learning algorithms is data sets. Such datasets contain a collection of examples used to train and evaluate such algorithms.

Machine learning is the intersection between theoretically sound computer science and practically noisy data. Essentially, it’s about machines making sense out of data in much the same way that humans do Matthew (2017). Also, Andreas and Sarah (2017) define the most successful kinds of machine learning algorithms are those that automate decision-making processes by generalizing from known examples. In this setting, which is known as supervised learning, the user provide the algorithm with pairs of inputs and desired outputs, and the algorithm find a way to produce the desired output it given an input.

Sebastian and Vahid (2019) describes the three types of machine learning: supervised learning, unsupervised learning and reinforcement learning. The main goal in supervised learning is to learn a model from labelled training data that allows us to make predictions about unseen or future data. Here, the term “supervised” refers to a set of training examples (data inputs) where the desired output signals (labels) are ready to know. Khongdet, Alok, Mukesh and Tanya (2022) the most common supervised learning algorithm is the linear regression model. This algorithm use a linear relationship between the input data and the desired output variable.

**Linear regression**

Linear regression is an input variable against one or more than input variables in order to find the best fit line described Khongdet, Alok, Mukesh and Tanya (2022). Giuseppe (2017) have considered Linear Regression or ordinary least squares (OLS) is statistical model used to predict the relationship between independent and dependent variables. Linear models (LMs) are a class of models that are widely used in practice and LMs make a prediction using a linear function of the input features. Linear models are the simplest parametric methods and always deserve the right attention, because many problems, even intrinsically non-linear ones, can be easily solved with these models. A regression is a prediction where the target is continuous and its applications are several, so it's important to understand how a linear model can fit the data, what its strengths and weaknesses are, and when it's preferable to pick an alternative.

**Decision trees**

Decision treesare widelyused models for classification and regression tasks. Essentially, they learn a hierarchy of if/else question, leading to a decision Pang-Ning, Michael and Vipin (2006).Also,John, Brian and Aoife (2015) define the Decision tree models can be used for datasets that contain both categorical and continuous descriptive features. When a decision tree is used for classification tasks, it is most commonly referred to as a classification, when it is used for regression tasks, it is called a regression tree Lior and Oded (2010).

In addition, Lior and Oded (2010) have considered classification trees are frequently used in applied fields such as finance, marketing, engineering and medicine. The classification tree is useful as an exploratory technique. However, it does not attempt to replace existing traditional statistical methods and there are many other techniques that can be used to classify or predict the affiliation of instances with a predefined set of classes, such as artificial neural networks or support vector machines.

**Random forests**

Random forests (RF) is a type of machine learning algorithm that uses a combination of decision trees and bootstrap sampling to make predictions. A machine learning technique that uses a combination of Decision Trees and Random Forests to improve the performance of a model Khongdet, Alok, Mukesh and Tanya (2022). The combination of bagging, subspace sampling and decision tree is known as a random forest model John, Brian and Aoife (2015).

Leo (2002) defines that, Random forests are an effective tool in prediction. Because of the Law of Large Numbers they do not overfit. Injecting the right kind of randomness makes them accurate classifiers and regressors. Furthermore, the framework in terms of strength of the individual predictors and their correlations gives insight into the ability of the random forest to predict. Using out-of-bag estimation makes concrete the otherwise theoretical values of strength and correlation. For a while, the conventional thinking was that forests could not compete with arcing type algorithms in terms of accuracy.

His results dispel this belief, but lead to interesting questions. Boosting and arcing algorithms have the ability to reduce bias as well as variance (Schapire et al., 1998). The adaptive bagging algorithm in regression (Breiman, 1999) was designed to reduce bias and operates effectively in classification as well as in regression. But, like arcing, it also changes the training set as it progresses. Random inputs and random features produce good results in classification—less so in regression. The only types of randomness used in this study is bagging and random features. It may well be that other types of injected randomness give better results. For instance, one of the referees has suggested use of random Boolean combinations of features.

**Neural Networks**

Khongdet, Alok, Mukesh and Tanya (2022) mentioned the unsupervised learning is used to learn from data without any labels and the goal is to build a model that can automatically improve over time by making predictions or associations about new data. The most common unsupervised algorithm is the neural network. This algorithm uses a network of interconnected neurons to learn patterns in data. Neural networks are often used to learn how to classify data, or to predict the outcome of future event.

Doruk, Arindrajit, Attila and David (2021) have considered to assess the effect of the minimum wage on the labor market by applying modern machine learning tools to predict who is affected by the policy. They apply three main tree-based learning tools in the training data: decision trees, random forests and gradient boosting tree and they also explore the elastic net regularization of the logistic regression. A key advantage of the ML tools over the Card and Krueger (1995) approach is that they do not require the researcher to pre-specify the functional form of the prediction model, which is instead determined in a data-driven way. Then we compare the performance of various prediction models in the test data.

The best performing prediction comes from the gradient boosting tree model. At the same time, it is worth pointing out that the original linear prediction model proposed by Card and Krueger (1995) (with a judiciously chosen set of interactions) also performs relatively well, although a little worse than the state-of-the-art machine learning tools. When compared to a commonly used demographic groups in the literature (such as teens, or high school or less under the age of 30), the boosting approach can form groups with a similar number of (correctly classified) minimum wage workers while substantially reducing the number of (mis-classified) non-minimum wage workers. The gains in precision (i.e. the correctly classified share) for a given level of recall (i.e. share of minimum wage workers included in the group) is sizable when we limit attention to non-teen workers—a group that is of particular interest to policymakers.

Also, they implement an event study using 172 prominent minimum wage increases between 1979 and 2019. As a result, they find a clear increase in wages of affected workers and no change in employment. Furthermore, minimum wage increases have no effect on the unemployment rate, labor force participation, or labor market transitions. Overall, these findings provide little evidence of changing search effort in response to a minimum wage increase.

To capture the impact of the policy on a broad group of affected workers, they utilize modern ML techniques to estimate the likelihood that someone is a minimum wage worker. While the best performing prediction model does better than the linear prediction model of Card and Krueger (1995), the gap is not large. One implication of these findings is that minimum wage researchers who are not interested in investing in a Machine learning approach may do fairly well by simply applying the Card and Krueger linear probability prediction model.

Eventually, from the research some following researches and results found impact on the working hours of the national statutory minimum wage by gender, age, education and nationality. Doruk, Arindrajit, Attila and David (2021) have considered to assess the effect of the minimum wage on the U.S labor market by applying modern machine learning tools to predict who is affected by minimum wage increase. But they build a prediction model to explain the relationship between being a minimum wage worker and the demographic and educational variables.

It follows that the novelty of this topic area is capable to be an application of the machine learning functions and different specific variables. Investigating the core studies related to the minimum wage and working hours, this study proposes to using machine learning techniques to determine the most influential factor in working hours of national minimum wage workers which is unpresented area. By applying all the given concepts and theories to this project, ML models with method of choice, selection of hyperparameters (Li and Abdallah, 2020), feature importance method (Terence, Kerem, Christopher and Jeremy, 2018). All these elements were applied with a methodical data acquisition detailed in the Primary Research, Methodology and Ethics chapter and methods to avoid overfitting and underfitting, create highly accurate classification prediction of working hours of national minimum wage workers for all possible circumstances.

**6. Validity**

Machine Learning models used to determine the most influential factor in working hours of national minimum wage workers in Ireland by the use of dataset of employee working hours and wages, collected from reputable source. The data were chosen based on their suitability for the research question and trained and tested using accurate methodologies. The validity of the final result was given by the choice and accuracy of the machine learning algorithms and validation methods used to assess the performance of the models, sum of true predictions divided by the sum of all predictions.

Once the model was validated in terms of having no overfitting or underfitting, then the most appropriate and generalized models were selected for the analyzing procedure and the actual accurate of the models were counted and finally compared amongst themselves so the highest accuracy Machine learning model could be chosen.

**7. Sampling Strategy**

The sampling strategy of this research is purposive sampling. The population of this research is a specific group of workers (the Ireland statutory minimum wage workers). The choice of this non-probability and qualitative sampling method is related to the selecting a sample of employees from dissimilar industries (a sector of retail, food and accommodation, and health and social care), with varying levels of education and experience to accurately evaluate the impact of wage laws on the working hours of employees in Ireland.

Therefore, the source if the data must come from randomly selecting participants from a large amount of population. Purposeful sampling is a technique widely used in qualitative research for the identification and selection of information-rich cases for the most effective use of limited resources (Michael Patton 2002).

**8. Primary Research Methodology**

The reason why a machine learning technique was chosen as the key method was due to in accordance with the matters raised in the literature review: the machine learning methods used for analysis on how the minimum wage affects working hours using machine learning analytic techniques. A key advantage of the Machine learning tools does not require the researcher to pre-specify the functional form of the prediction model, which is instead determined in a data-driven way. Also, this can provide analyzing large-scale of datasets, such as administrative records or survey data to derive meaningful insights. Able to use statistical techniques and machine learning algorithms to identify patterns, correlations, and potential causal relationships.

Furthermore, by using quantitative analysis with machine learning algorithms, achievable to generate empirical evidence on the effects of the minimum wage on working hours, allowing for more precise and data-driven insights. Additionally, this approach could help uncover potential non-linear or interactive effects that traditional statistical methods may overlook.

### **8.1. Dataset and Collection Method**

Data collection method is Experimental research is secondary a quantitative method which is test a causal relationship. This method is perfectly suitable for the research objectives and questions. The most relevant, open-source and reliable dataset source is Earnings Analysis using Administrative Data Sources (EAADS), which links earnings from administrative sources to Labour Force Survey National Minimum Wage Estimates data. The data are administered by the Central Statistics Office (CSO) in Ireland.

When dealing with research, it is crucial to consider the ethical implications involved: the dataset generated cannot be shared out of the academic environment in which this research is inserted. The copyrights of Labour Force Survey National Minimum Wage Estimates to the Central Statistics Office (CSO) in Ireland. These records include all information collected directly for CSO statistical purposes in statutory or voluntary inquiries from persons, households, businesses and undertakings, or indirectly from the administrative records of public authorities held on completed questionnaires, worksheets or data bases. The CSO gathers information through both statutory (legally required) and voluntary inquiries. Statutory inquiries might be legally mandated data collection, while voluntary inquiries are those where participants provide data voluntarily.

The proposed research dataset and required independent variables were not available. Furthermore, after thorough data cleaning and statistical analysis, discovered the obtained datasets were not applicable in the chosen context. As a consequence of the erratic data collection, the most of time was concentrated on data preparation, find conducive of variables and feature engineering.

The dataset of Education level is the total number of Employees aged 15 years and over from the second quarter of 2016 to the third quarter of 2020. This dataset involves all level of the education level of the Total employees, Employees reporting earning National Minimum Wage or less and Employees reporting earning more than National Minimum Wage. Also, the type of the education level is “Primary or below”, “Lower secondary”, “post-secondary non-tertiary”, “Third level non-honours degree”, “Third level honours degree or higher” and “Higher secondary”.

The Nationalities of the total number which are Employees aged 15 years and over from the second quarter of 2016 to the fourth quarter of 2019. The dataset involves the Nationalities of the Total employees, Employees reporting earning National Minimum Wage or less and Employees reporting earning more than National Minimum Wage. Moreover, the dataset is six different categories of the Nationalities which are “Irish nationals”, “non-Irish nationals”, “United Kingdom”, “EU 15 excl. Irl. & UK”, “EU15 to EU28” and “Other nationals”.

NACE (Nomenclature of Economic Activities) is the European statistical classification of economic activities. NACE groups organizations according to their business activities. The dataset of those economic sectors was combined the total number of Employees aged 15 years and over from the second quarter of 2016 to the third quarter of 2020 with the national minimum wage earnings status.

The proposed dependent variable of worked hours datasets was generated by weekly usual worked hours of employees who has aged 15 years and over from the second quarter of 2016 to the third quarter of 2020 with the national minimum wage earnings status. Also, the dataset provides information of employees who is earning more than National Minimum Wage. From the all datasets, the key outcome variables are: Weekly usual worked hours, NACE economic sector, Education level and Nationality of the national minimum wage workers.

The data set consists of four Dataframe, the Education level 576 records, The Nationalities 420 records, NACE (Nomenclature of Economic Activities) 1296 records, the weekly usual worked hours 288 records with all 6 columns.

**8.2. Data preparation and Feature engineering**

Extensive exploratory data analysis and data preparation are a crucial first step in any data analysis process. Zahraa and Geoffrey (2017) defines as data preparation that before data can be analyzed they must be organized into an appropriate form. Data preparation is the process of manipulating and organizing data prior to analysis. Data preparation is typically an iterative process of manipulating raw data, which is often unstructured and messy, into a more structured and useful form that is ready for further analysis. The whole preparation process consists of a series of major activities (or tasks) including data profiling, cleansing, integration and transformation. Furthermore, Any model performance depends on data which is mean the process of creating representations of data that increase the effectiveness of a model (Jason, B. 2020).

Table 1: Data Type and Description

|  |  |
| --- | --- |
| Attributes | Data Type and Description |
| Statistic Label | Object, Age of NMW employees |
| Education Level | Object, Education level of NMW employees |
| NMW earnings status | Object, NMW Earnings status |
| Quarter | Object, Year and quarter |
| UNIT | Object, Thousand |
| VALUE | float64, Number of employees |
| Nationality | Object, Nationality of NMW employees |
| NACE Rev 2 Economic Sector | Object, Economic sector of NMW employees |
| Usual Hours Worked | Object, A weekly average working hours of NMW employees |

Data cleaning deals with the dataset preparation concentrates on getting the data prepared for further machine learning use. Data preparation is carried over by a set of rules associated with machine learning techniques. As the data is gathered from different source, its hold high potential to be unclean and inconsistent. This problem could lead to false conclusion of result. Hence, degree of clean data is proportional to the quality of result. (Omar, 2019). The data preparation process began with converting the date to a datetime format which was originally in an object type. This converting process was necessary because all dataset’s “Quarter” variables contain the unique time values. The correct datetime formats allows to process feature engineering, handling missing data and visualization. Furthermore, this step enhances the accuracy and effectiveness of the machine learning algorithms.

Next step of data preparation involved dropping irrelevant rows from each columns of the datasets. Because as mentioned earlier, the datasets contained a significant amount of irrelevant information. For instance, the counts of workers earning more than National Minimum Wage workers, “Not stated” entries, “Total employees”, All NACE economic sectors, among others. By removing these rows, essentially ensured that the remaining datasets are more relevant to the minimum wage workers information. Dropping these rows can improve the accuracy of machine learning models that perform on the cleaned dataset.

Real-world data often has missing values. Data can have missing values for a number of reasons such as observations that were not recorded and data corruption (Rodrick, Little and Rubin, 2002). Handling missing data is important as many machine learning algorithms do not support data with missing values. In the project’s datasets, missing values are checked and handled for machine learning algorithms.

Before training the machine learning models, the datasets were not suitable for a basis descriptive statistics and machine learning models because of indexes were all a string type. All the data of an observation should consistently belong to same datatypes. Therefore, the type variables of datasets were replaced index of string to integer. This step makes it suitable for machine learning models that require the numerical inputs instead of categorical strings.

In the final stage of feature engineering, a final dataset was created by selected valuable variables from four different datasets for use in machine learning methods. As mentioned in objective, those feature selection techniques can help with identify the most importance features from tree-based models.

The statistical relationship between two variables is called as correlation. Variables in a dataset can be related in various way i.e. when one variable could cause or depend on the values of other variable or when one variable could be slightly associated with another variable or when two variables could be dependent on the unknown third variable. A correlation could either be positive (when both variables change in same direction), neutral (when there is no relationship between the variables) and negative (when variables change in opposite directions). The performance of an algorithm degrades when two or more variables are tightly correlated with each other which is known as multicollinearity. When the variables can be related by having linear relationship or consistent is called as covariance. Correlation matrix is a covariance matrix where data points follow roughly straight-line trend to have an approximate linear relationship. Correlation describes the strength of the linear association i.e. it summarises the strength and direction of the linear (straight-line) association between two quantitative variables, denoted by r, where the values lies between -1 and +1. A positive value for r indicates a positive association, and a negative value for r indicates a negative association. The closer r is to 1 the closer the data points fall to a straight line; thus, the linear association is stronger. The closer r is to 0, making the linear association weaker. The highly negative and positive correlated data must be removed from the algorithm to get a better accuracy while predicting the dataset.

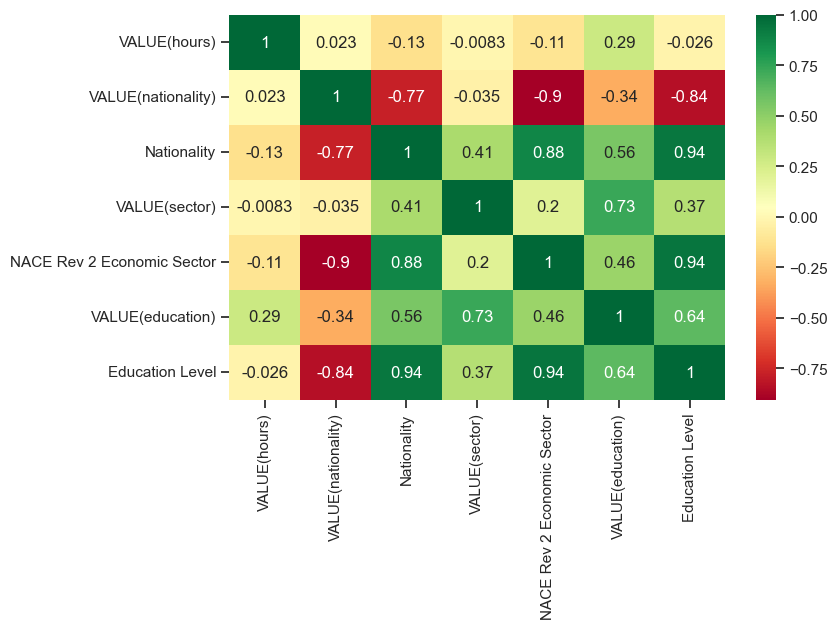


Figure 1: Plot of the Correlation heatmap

### **8.3. Machine learning techniques**

Based on the collected dataset, and the modelling techniques that better adopt to the problem are related to the supervised machine learning category. Machine learning algorithms that learn from input/output pairs are called supervised learning algorithms because a “teacher” provides supervision to the algorithms in the form of the desired outputs for each example that they learn from. While creating a dataset of inputs and outputs is often a laborious manual process, supervised learning algorithms are well understood and their performance is easy to measure (Andreas, Sarah. 2017). Classification is a subcategory of supervised learning in the goal is to predict the categorical class labels of new instances based on past observations. By the object of the project, aimed to predict the average weekly working hours of minimum wage workers. Furthermore, applied tree-based learning tools in the training data: Decision Trees, Random Forests, Support Vector Machine, Naive Bayes Classification and Deep learning technique in Neural Network**.** A key advantage of the Machine learning tools does not require the researcher to pre-specify the functional form of the prediction model, which is instead determined in a data-driven way (Doruk, Arindrajit, Attila, David. 2021).

There are different aspects to consider when explaining Machine learning. One of the main aspects is the ‘why’ of this tool. ML can solve three types of problems (Henke et al., 2016):

1. Classification: by identifying objects and recognizing text or audio. Classification also includes associations and recommendations through segmentation into clusters.

2. Prediction/Estimation: to predict and forecast possible outcomes

3. Generation: in this case for instance ML “can generate content from interpolating missing data to generating the next frame in a video sequence” (Henke et al., 2016).

This research project focuses on Classification and Prediction. ML is a technique that has evolved and has 3 main different types of learning depending on the data and purpose:

Table 2: Types of Machine Learning

|  |  |  |  |
| --- | --- | --- | --- |
| **Types of machine learning** | **Data type** | **Description** | **Form of analysis** |
| Supervised Learning | Labelled data | Through SL workflow the labelled training data is analysed through a ML algorithm to define a predictive model; thus, it can predict new unlabelled data inputs (Raschka, 2015). | Classification  Regression |
| Unsupervised Learning | Unlabelled data | Unsupervised Learning Algorithmsare based on the absence of any supervisor and therefore of absolute error measures: it’s useful when it’s necessary to learn how asset of elements can be grouped (clustered) according to their similarity (or distance measure)(Guiseppe, 2017). | Clustering Dimensionality reduction |
| Reinforcement Learning | System Agent | Reinforcement Learning (RL)is a popular and promising branch of AI that involves making smarter models and agents that can automatically determine ideal behavior based on changing requirements. | Reward System |

Different kinds of supervised machine learning algorithms were employed in this project. In supervised machine learning algorithms, the labeled training dataset is employed first of all to practice the fundamental algorithm.

After feature engineering process following procedures described, several models of machine learning were applied based on the successful techniques of minimum wage workers’ usual working hours and the literature review. Several Machine learning models were applied with 5 folds cross validation function, all of which were designed for categorical prediction. The based on the mean accuracy scores from cross-validation, the Random Forest classifier has the highest accuracy. The selected model achieved a mean accuracy of approximately 94.5%.

**Overfitting and Underfitting**

The avoidance of over-fitting and under-fitting help to improve the performance of machine learning and for avoiding those problems (Haider and Rafiqul, 2015).

Overfitting is the one of biggest problem in training neural networks is the over-fitting of training data. That means that the neural network at the certain time during the training period does not improve its ability to solve problem anymore. But just starts to learn some random regularity contained in the set of training patterns (Haider and Rafiqul, 2015). A statistical model is said to be overfitted when feed it a lot more data than necessary. To make it relatable, imagine trying to fit into oversized apparel. Over-fitting occurs when astatically model describes random error or nose instead of the underlying relationship (Piotrowski and Napiorkowski, 2013); (Chan, Kwong, Dillon, and Tsim, 2011); (Panchal, Ganatra, Shah, and Panchal, 2011); (Domingos, 2000).

Underfitting is the opposite of Over-fitting. This occurs when the model is incapable of capturing the variability of the data. For example suppose one is training a linear (y= ax+b not polynomial a and b are constant) classifier on a data set that is a parabola (Tom, 1995), (van der Aalst, Rubin, Verbeek, van Dongen, Kindler and Günther, 2010). The resultant classifier will have no predicative power nor will it able to properly map the training data (Lawrence, Giles, and Tsoi, 1997).

Some methods to avoid the problem of over-fitting and under-fitting in supervised machine learning: There are many methods (Kazushi, 2005); (Prechelt, 1999); (Schittenkopf, Deco and Brauer, 1997). :

1. Penalty methods:

• Map provides a penalty based on P(H).

• Minimum description length (MDL) principle.

• Structural risk minimization.

• Generalization cross-validation.

• Hold and cross-validation.

B. Early stopping for training

**K-Fold Cross-Validation**

The introduction to this note cast cross-validation as an alternative to the use of prior knowledge in selecting a classification method (Cullen, 1993). Cross-validation is similar to the repeated random subsampling method, but the sampling is done in such a way that no two test sets overlap. In k-fold cross-validation, the available learning set is partitioned into 70 k disjoint subsets of approximately equal size. Here, “fold” refers to the number of resulting subsets. This partitioning is performed by randomly sampling cases from the learning set without replacement. The model is trained using k − 1 subsets, which, together, represent the training set. Then, the model is applied to the remaining subset, which is denoted as the validation set, and the performance is measured. This procedure is repeated until each of the k subsets has served as validation set. The average of the k performance measurements on the k validation sets is the cross-validated performance (Daniel, 2019).

One of the most powerful features to avoid/prevent overfitting is cross-validation. The idea behind this is to use the initial training data to generate mini train-test-splits, and then use these splits to tune the model. In a standard k-fold validation, the data is partitioned into k-subsets also known as folds. After this, the algorithm is trained iteratively on k-1 folds while using the remaining folds as the test set, also known as holdout fold.

The cross-validation helps us to tune the hyperparameters with only the original training set. It basically keeps the test set separately as a true unseen data set for selecting the final model. Hence, avoiding overfitting altogether.

**Random Forest**

The Random Rorest is a tree-based ensemble learning technique.Random forests or random decision forests are a classification, regression, and other tasks machine learning algorithm for the ensemble. This works by creating a lot of decision trees at training time and outputting the class which is the class mode (classification) or mean prediction (regression) of the individual trees. Random forests offer a remedy for the problem of overfitting of the decision trees.

Figure 2: Random Forest Classifier

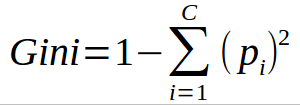
In the model apply, created three Random Forest models, each with different number of estimators. Each model has trained on the training data and made predictions using the individual models. Furthermore, performed ensemble averaging by taking the majority vote of the three models. This study involves combining the predictions from multiple random forest classifiers, and the ensemble's accuracy is evaluated to see if it outperforms the individual models. The model was applied with cross validation function of 5 folds, which achieved with 100 trees (estimators) and 15 predictors (features) tried at each split. The choice of 5 folds cross-validation suggests that the dataset was divided into 5 subsets (folds). The model was trained and evaluated five times, each time using a different fold as the test set and the remaining folds as the training set. This helps estimate the model's performance and generalization on unseen data. Overall, experiment aimed to find the best hyperparameters for the RF model while balancing model complexity and cross-validation to achieve the highest predictive performance.

After performance of the model, computed feature importance using the Random Forest Classifier and displayed them in a Pandas series (feature\_imp). Feature importance indicate the relative importance of each feature in making predictions with the RF model. These importances are typically used to understand which features have the most significant impact on the model's performance.

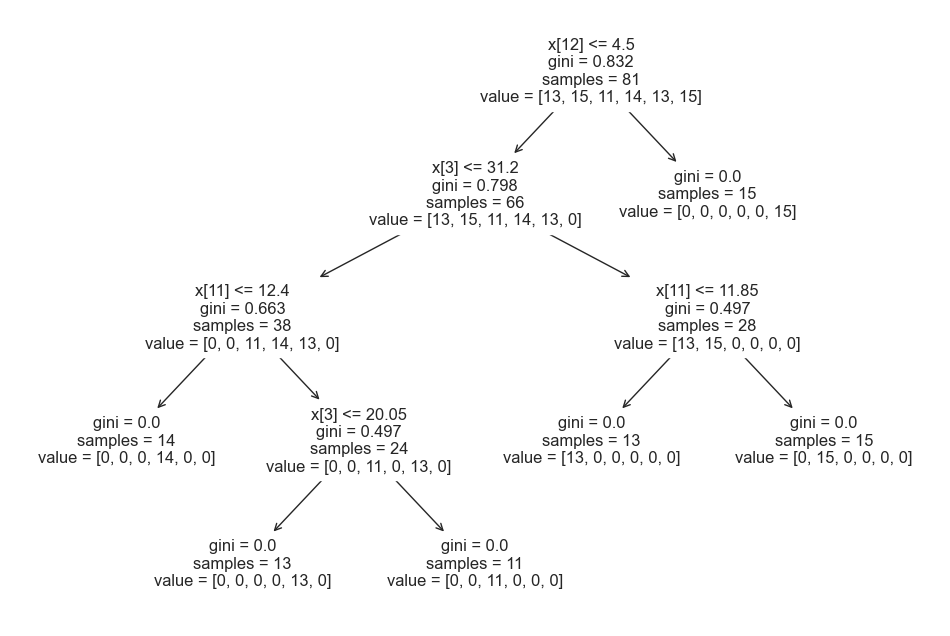
**Decision Tree**

The Decision Tree designs the logic of the decision in such a way that eval- uates and matches results for the classification of data items into a structure as like a tree. A decision tree recursively divides the feature (predictor) space into two in a way that reduces the prespecified loss function the most. More concretely, in the beginning the algorithm tries every possible split to divide the entire sample space into two and picks the one that diminishes the loss function the most. Subsequently, each subsample is treated as the new sample, and the first step is repeated. Once the splitting is over, it predicts the class of every observation according to the majority vote in the subspace (terminal node) to which the observation belongs.

Another attribute selection measure that CART (Categorical and Regression Trees) uses is the Gini index. It uses the Gini method to create split points.



where 𝑘 = 1, 2, 3, ..., 𝑚 represents the 𝑚 classes of the target variable. 𝑝𝑘 represents the proportion of samples that belong to class 𝑘. Gini index says, if randomly select two items from a population, they must be of the same class and probability for this is 1 if the population is pure. It works with the categorical target variable “Success” or “Failure”. It performs only binary splits. The higher the value of Gini, higher the homogeneity. CART (Classification and Regression Tree) uses the Gini method to create binary splits.



Root node

Decision node

Figure 3: Decision tree Classification

The data was split in 75% for training and 25% for testing and random state of 20. “X\_train” has a shape of (81, 16), which means 81 samples are in the training set, “X\_test” has a shape of (27, 15), indicating that 27 samples are in the testing set, “y\_train” has a shape of (81,), which corresponds to the labels for the training set, “y\_test” has a shape of (27,), which corresponds to the labels for the testing set. After the cross validation was carried out, the final training dataset set was stored in a csv file for further process. Before the applying the model, it was performed by varying the maximum depth “max\_depth” hyperparameter from 1 to 9. Furthermore, the highest testing accuracy model has chosen for further process. After all the process, applied the model that uses the Decision tree algorithm to predict the weekly usual working hours of minimum wage workers. The Decision tree model was compared using the accuracy score and confusion matrix.

**Naive Bayes Classification**

**Naive Bayes** is a simple, yet effective and commonly-used, machine learning classifier. It is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a Bayesian setting. It can also be represented using a very simple Bayesian network. Naive Bayes classifiers have been especially popular for text classification, and are a traditional solution for problems such as spam detection.

Gaussian Naïve Bayes (GaussianNB): This is a variant of the Naïve Bayes classifier, which is used with Gaussian distributions—i.e. normal distributions—and continuous variables. This model is fitted by finding the mean and standard deviation of each class.

**Logistic Regression**

Logistic regression is used for classification problems. The term regression is used in it because its underlying technique is the same as of Linear Regression. Decision boundary and cost function are the 2 important hyper-parameters for Logistic Regression. To predict in which category the class belongs to a threshold can be set. Based on this threshold, the obtained estimated probability is classified.

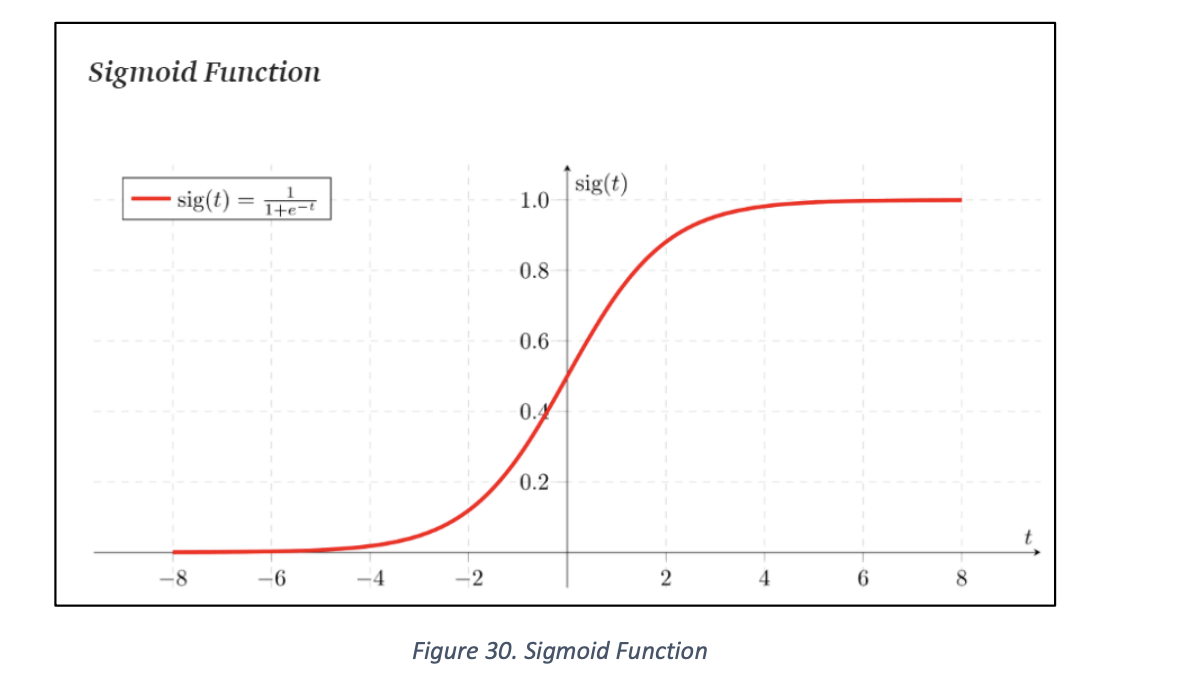


Figure 4: Sigmoid Function

If ‘Z’ goes to infinity, the target variable ‘Y’ will become 1 and if ‘Z’ goes to negative infinity, the target variable ‘Y’ will become 0. In this project, the target variable has three or more categories, so used Multinomial Logistic Regression.

In the project,The Logistic Regression start with Recursive Feature Elimination with Cross-Validation (RFECV) to select an optimal number of features for a Logistic Regression model. As (Puneet and Arun 2020) described, it is a wrapper method of feature selection. It removes the redundant and weak feature whose deletion least effects the training error and keeps the independent and strong feature to improve the generalization performance of the model. It uses the iterative procedure for feature ranking which is an instance of backward feature elimination. This technique first builds the model on the entire set of features and ranked the feature according to its importance. After that, it removes the least important feature and rebuilds the model again and recalculate the feature importance.

Then, feature selection method helped to identify the most important features for the model while considering cross-validation to ensure robustness. The method found 7 optimal features, which 'VALUE(hours)', 'VALUE(nationality)', 'Nationality', 'VALUE(sector)', 'NACE Rev 2 Economic Sector', 'VALUE(education)', 'Education Level'. After this, the training and testing dataset split “X\_train” has a shape of (81, 7), which means 81 samples are in the training set, “X\_test” has a shape of (27, 7), indicating that 27 samples are in the testing set, “y\_train” has a shape of (81,), which corresponds to the labels for the training set, “y\_test” has a shape of (27,), which corresponds to the labels for the testing set. For LogisticRegression and LinearSVC the trade-off parameter that determines the strength of the regularization is called C, and higher values of C correspond to less regularization (Andreas and Sarah, 2017). By the experiment conducted by (Andreas and Sarah, 2017), logistic regression models were trained different C hyperparameter values, and obtained the corresponding accuracy values for each C.

**Deep learning technique in Neural Network**

The neural networks proposed for this research are multilayer feedforward neural networks operating under supervised learning; they consist of three layers including one input layer, one hidden layer and one output layer. The input layer, the hidden layer, and the output layer are fully interconnected. The weights of the connections and the biases are initialized randomly by the system, and adjusted through the learning process.

Artificial Neural Network (ANN) is capable of learning any nonlinear function. Hence, these networks are popularly known as Universal Function Approximators. ANNs have the capacity to learn weights that map any input to the output. One of the main reasons behind universal approximation is the activation function. Activation functions introduce nonlinear properties to the network. This helps the network learn any complex relationship between input and output. Table 1 shows the compares the properties of several activation functions.

Table 3: Table of activation functions

|  |  |  |
| --- | --- | --- |
| **Name** | **Plot** | **Range** |
| identity |  | (-∞,∞) |
| Binary step |  | {0,1} |
| Logistic, sigmoid, soft step |  | (0,1) |
| Hyperbolic tangent tanh |  | (-1,1) |
| Rectified linear unit ReLu |  | [0, ∞) |
| Softplus |  | (0, ∞) |

Neural Network links a set of input nodes 𝑥𝑖 = (𝑥1, 𝑥2, ... , 𝑥𝑛) existing in the input layer with a set of one or more output nodes 𝑦𝑗 = (𝑦1, 𝑦2, ... , 𝑦𝑚) existing in the output layer through an intermediate hidden layer.

Nodes in each layer are activated once they reach the layer threshold value 𝜃𝑖. This matching is realized by finding an unknown function h (Rohaifa, Abdellatif, Raddouane and Rdouan).

yj = h(x1,x2,...,xn)

𝑖∈[1,𝑛]and𝑗∈[1,𝑚]

The research intends to use Keras which is an open source of a high-level Neural Network library, written in Python and runs on Theano and TensorFlow. First, data should be collected and preprocessed into input dataset and desired output dataset. Second, should build and design his network by choosing the type of learning. As well as by fixing the network parameters, for example net input, activation function, number of Epoch, Batch, number of neurons in each layer, etc.

As the project objective, to predict the minimum wage workers’ average weekly working hours using ANN model based on 15 input variables. The creation of the model begins from setting up the environment, installing necessary libraries like Scipy and Keras. The learning model to map rows of input variables (X) to an output (y). Then, split X, y variables into Training and Testing parts and each of variables’ test size is configuration 0.20. The data stored in a 2D array where the first dimension is rows and the second dimension is columns, e.g. After this, specified the number of neurons or nodes in the layer as the first argument, and specified the activation function the activation argument. In the model, used the rectified linear unit activation function referred to as “relu” on the first two layers and the Sigmoid function in the output layer. Because, the model’s output neuron produces a probability between 0 and 1, and the neuron with highest probability choose as the predicted class. Training occurs over epochs and each epoch is split into batches. Fit the model with 100 epochs and batch size was 32. Finally, the input layer, the hidden layer, and the output layer are fully interconnected.

**8.4. Evaluate Metrics**

To evaluate the model performance different evaluation metrics have been used like Accuracy, F1-Score. A confusion matrix is often used for classification model evaluation. In a confusion matrix, the number of correct and incorrect predictions are summarized, and count values are broken down into the classes.

Confusion matrix consists of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). TP means observation is positive and prediction is also positive. TN means observation is negative and prediction is also negative. FP means observation is negative but predicted positive. FN means observation is positive but predicted negative.

Accuracy

Formula for accuracy is given by:

For classification problems accuracy may not be the right evaluation metric as it assumes equal costs for both kinds of errors. Ability of a classifier to select all cases that needs to be selected and reject all cases that needs to be rejected is called accuracy (Kotu and Deshpande, 2019)

Precision is calculated as:

Recall is calculated as:

High recall indicates a class has been classified correctly. It can be defined as proportion of relevant cases that were found among all the relevant cases (Kotu and Deshpande, 2019).

F1-Score:

It can be defined as harmonic mean of both precision and recall.

K-FOLD CROSS VALIDATION - For robust results, the research further implements 10-fold cross validation to validate the results obtained.

**Feature importance**

Feature importance is the most useful interpretation tool, and data scientists regularly examine model parameters (such as the coefficients of linear models), to identify important features. Feature importance is available for more than just linear models. Most random Forest (RF) implementations also provide measures of feature importance. The mean decrease in impurity importance of a feature is computed by measuring how effective the feature is at reducing uncertainty (classifiers) or variance (regressors) when creating decision trees within RFs (Terence, Kerem, Christopher and Jeremy, 2018).

Also (Terence, Kerem, Christopher and Jeremy, 2018) described Permutation importance is a common, reasonably efficient, and very reliable technique. It directly measures variable importance by observing the effect on model accuracy of randomly shuffling each predictor variable. This technique is broadly-applicable because it doesn't rely on internal model parameters, such as linear regression coefficients (which are really just poor proxies for feature importance). Breiman and Cutler also described *permutation importance*, which measures the importance of a feature as follows. Record a baseline accuracy (classifier) or R2 score (regressor) by passing a validation set or the out-of-bag (OOB) samples through the Random Forest. Permute the column values of a single predictor feature and then pass all test samples back through the Random Forest and recompute the accuracy or R2. The importance of that feature is the difference between the baseline and the drop in overall accuracy or R2 caused by permuting the column. The permutation mechanism is much more computationally expensive than the mean decrease in impurity mechanism, but the results are more reliable. The permutation importance strategy does not require retraining the model after permuting each column; we just have to re-run the perturbed test samples through the already-trained model.

**9. Results and Discussion**

After the methodology was applied to gather the dataset, the final document contained combinations of NACE Economic sector, Education level, Nationality and Weekly usual worked hours datasets of numpy arrays. As mentioned in the previous sections, the research aims to analyse the relationship between minimum wage workers and their average weekly working hours.

The figure 6 shows that the distribution of NMW workers across different sectors of the economy in the given data set varies significantly. The 44% of NMW employees are concentrated in the "Services (G to U)" sector, representing a significant proportion of total employees. It also shows the importance of NMW workers in the "Wholesale and retail trade, repair of motor vehicles and motorcycles (G)" sector. On the other hand, the sectors "Education (P)" and "Professional, scientific and technical activities (M)" have a significantly lower number of NMW workers.

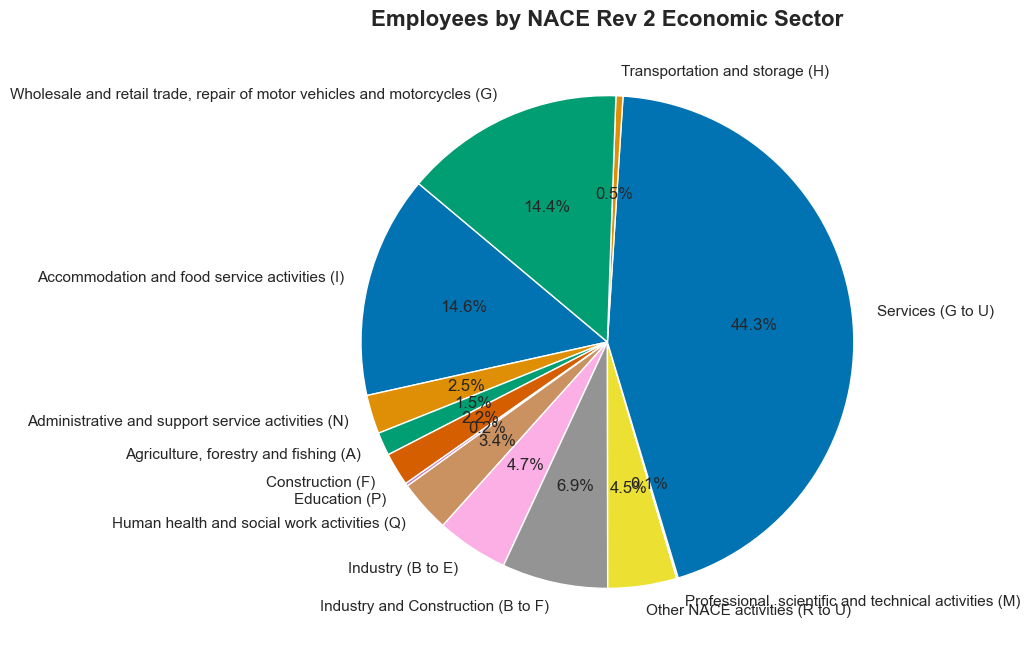


Figure 5: Employees by NACE Economic Sector (Source: Personal Collection)

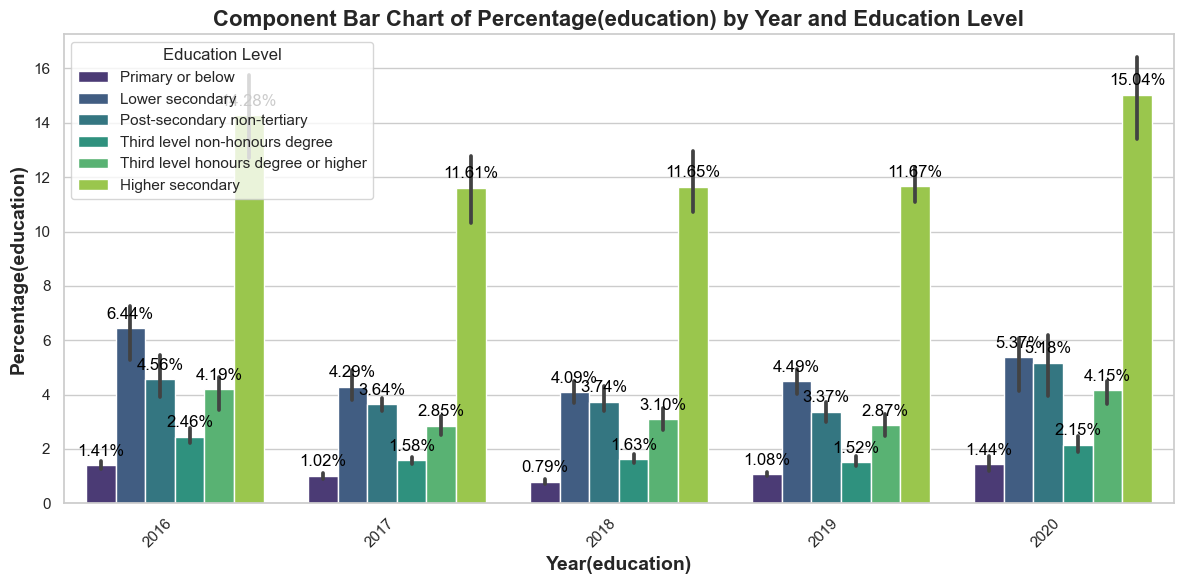


Figure 6: Component Bar Chart of Percentage of Employees by Education Level (Source: Personal Collection)

Figure 7 shows the number of NMW employees by nationality. The results were different than the proposed hypotheses, which was mentioned that most of the minimum wage workers would be non-Irish. Furthermore, the highlight finding of the result was that the number of Irish NMW workers increased by 64.9% to 72.8% each year, while the number of non-Irish nationals decreased by 2.4%.

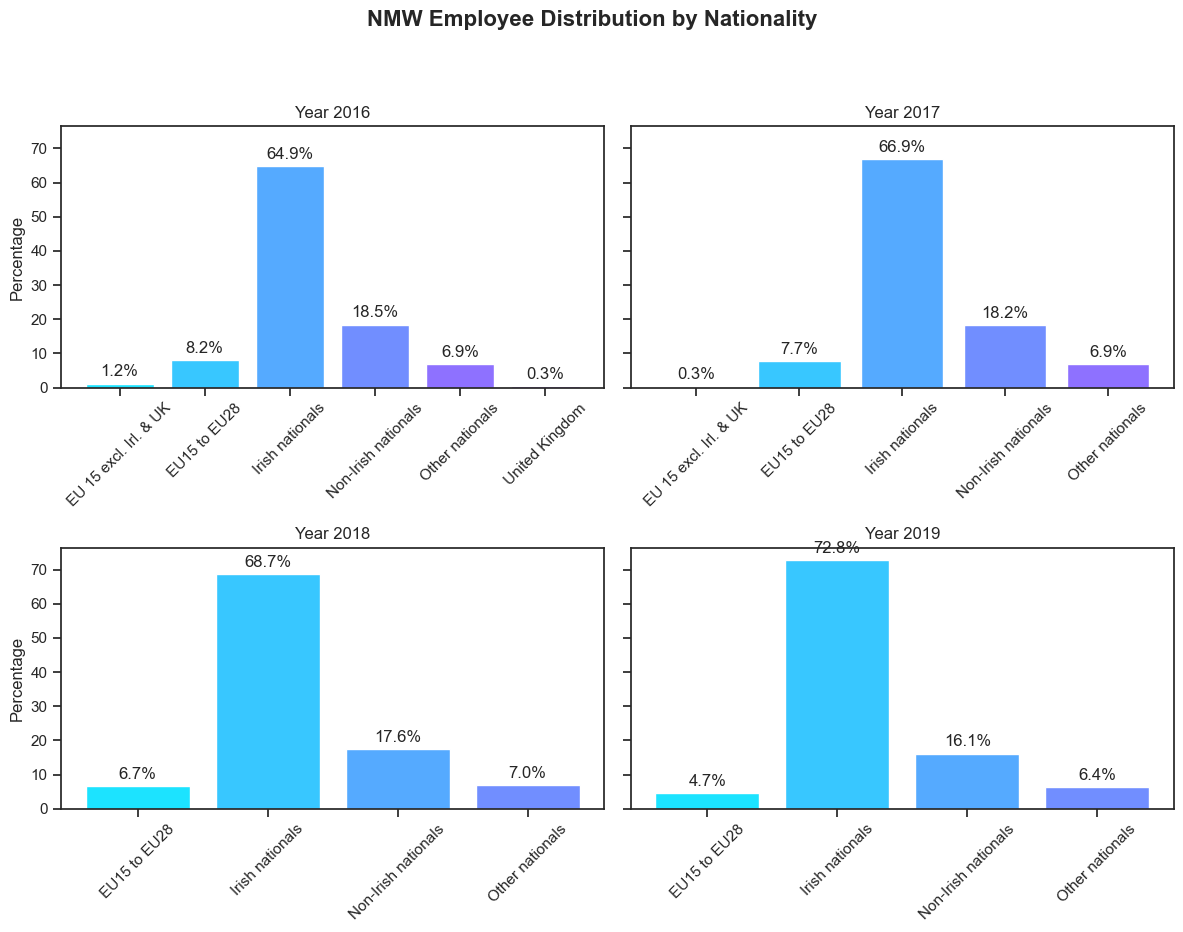


Figure 7: Nationalities of NMW (Source: Personal Collection)

Based on the statistical information, it was observed that from 2016 to 2020 the majority of minimum wage workers worked between 10-29 hours per week. Furthermore, highlight of the study found that the of number of minimum wage workers working these usual weekly hours decreased towards the end of 2020.

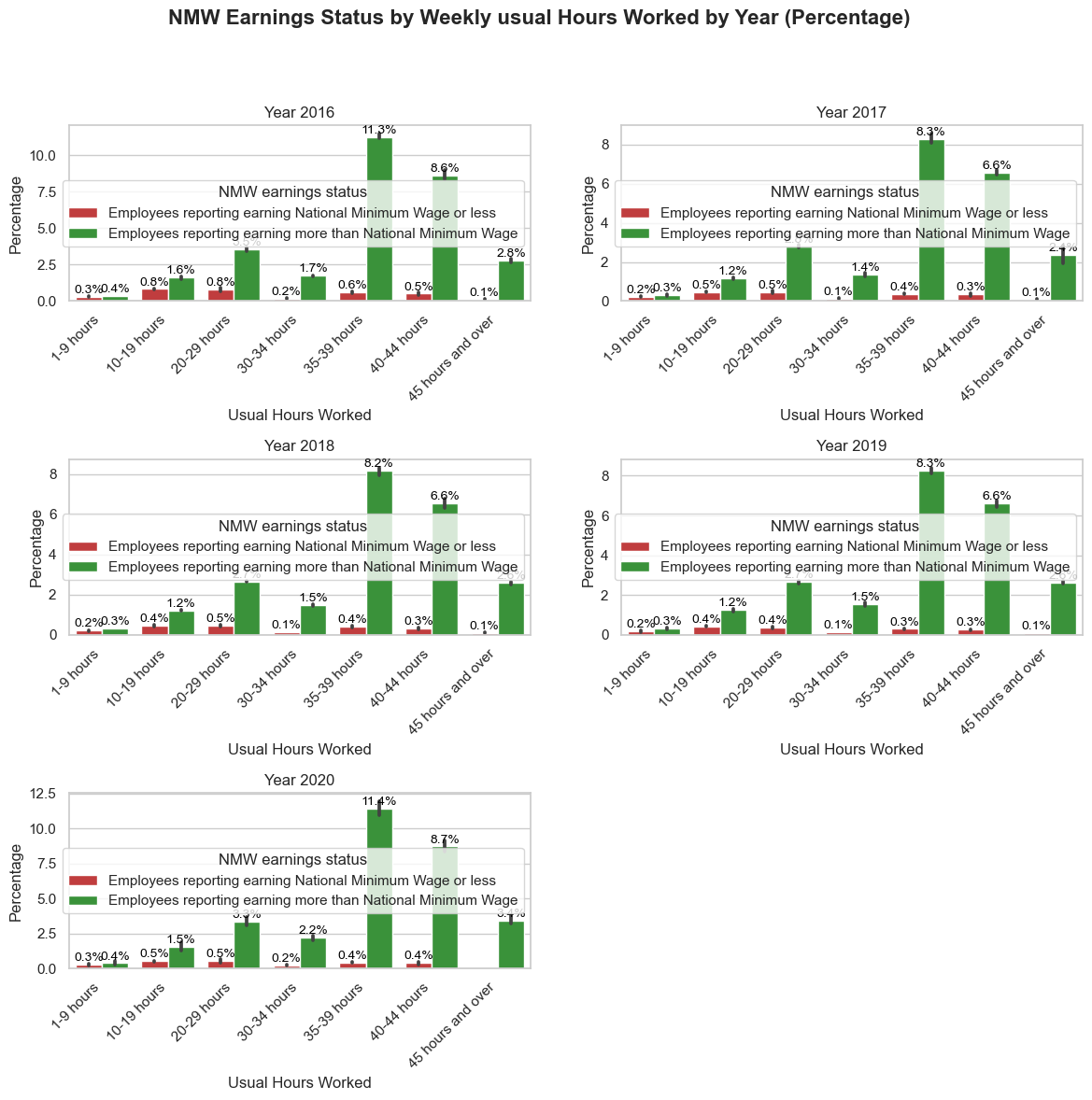


Figure 8: A weekly usual working hours of employees (Source: Personal Collection)

**Data Analysis**

The study performed machine learning tests and deep learning tests in the prediction of the NMW workers’ weekly working hours. Machine learning techniques such as the Support Vector Classification (SVC), Naive Bayes (GaussianNB), Logistic Regression (LR), Random Forest Classifier (RFC), and Decision Tree Classifier (DTC) were used in the study. Also, the deep learning technique used in this study was the Artificial Neural Network (ANN).

Random Forest Classifier

The first machine learning model Random Forest Classifier (RFC) applied to three different number trees 100, 150, 200 (estimator). The different estimators performed same result, which is the model accuracy ended with 0.89, which means it correctly classifies 89% of the instances in the dataset. The macro average and weighted average F1 scores are both 0.89, indicating good overall model performance.

precision recall f1-score support

1-9 hours 1.00 1.00 1.00 5

10-19 hours 0.60 1.00 0.75 3

20-29 hours 0.83 0.71 0.77 7

30-34 hours 1.00 0.75 0.86 4

35-39 hours 1.00 1.00 1.00 5

40-44 hours 1.00 1.00 1.00 3

accuracy 0.89 27

macro avg 0.91 0.91 0.90 27

weighted avg 0.91 0.89 0.89 27

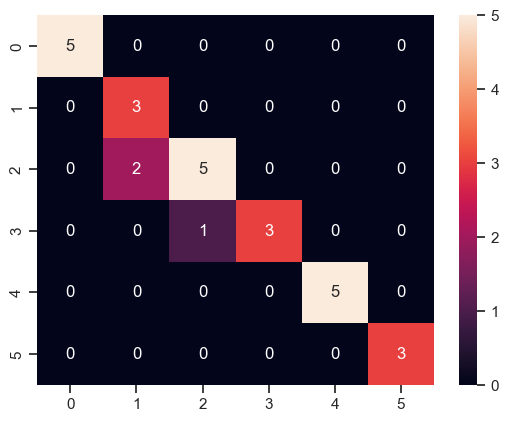


Figure 9: Confusion matrix of the RFC (100 estimator) – based on test and prediction (Source: Personal Collection)

Based on the model, the Scikit-learn Feature importance method was applied with features of the training data (15 features). By the Feature importance method, the Education level was the highest feature importance score (0.23), the NACE economic sector values were the second (0.12), the Nationality was third feature importance score (0.10).

Table 4: Table of the Feature importance

|  |  |
| --- | --- |
| Education Level | 0.231286 |
| VALUE(education) | 0.164616 |
| NACE Rev 2 Economic Sector | 0.129423 |
| VALUE(hours) | 0.118944 |
| VALUE(nationality) | 0.117775 |
| Nationality | 0.103103 |
| VALUE(sector) | 0.074766 |
| Year(sector) | 0.010993 |
| Month(nationality) | 0.009908 |
| Year(nationality) | 0.009391 |
| Year(hours) | 0.008806 |
| Year(education) | 0.007725 |
| Month(education) | 0.004744 |
| Month(hours) | 0.004551 |
| Month(sector) | 0.003971 |

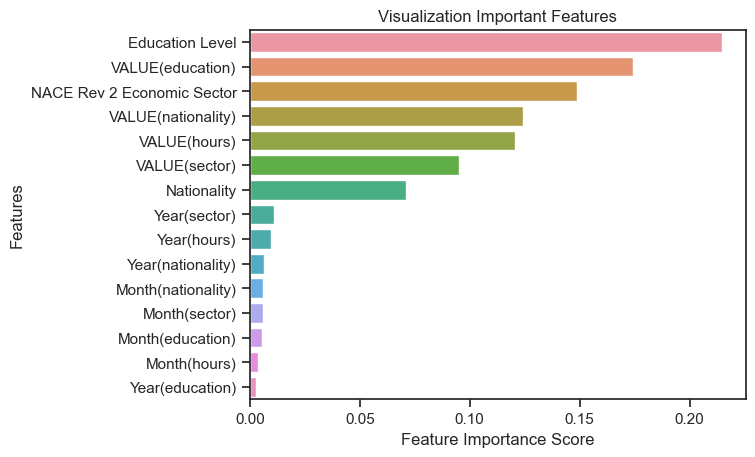


Figure 10: Feature importance (Source: Personal Collection)

Decision Tree model

After the Random Forest classifier models were analysed, the Decision Tree models were applied with the same training and testing data split. The performance of the first Decision tree classifier was analysed by varying the maximum depth (max\_depth) hyperparameter from 1 to 9.

max\_depth = 1: Training Accuracy = 34.48%, Testing Accuracy = 28.57%

max\_depth = 2: Training Accuracy = 51.72%, Testing Accuracy = 42.86%

max\_depth = 3: Training Accuracy = 67.82%, Testing Accuracy = 61.90%

max\_depth = 4: Training Accuracy = 100.00%, Testing Accuracy = 80.95%

max\_depth = 5: Training Accuracy = 100.00%, Testing Accuracy = 100.00%

max\_depth = 6: Training Accuracy = 100.00%, Testing Accuracy = 100.00%

max\_depth = 7: Training Accuracy = 100.00%, Testing Accuracy = 100.00%

max\_depth = 8: Training Accuracy = 100.00%, Testing Accuracy = 80.95%

max\_depth = 9: Training Accuracy = 100.00%, Testing Accuracy = 80.95%

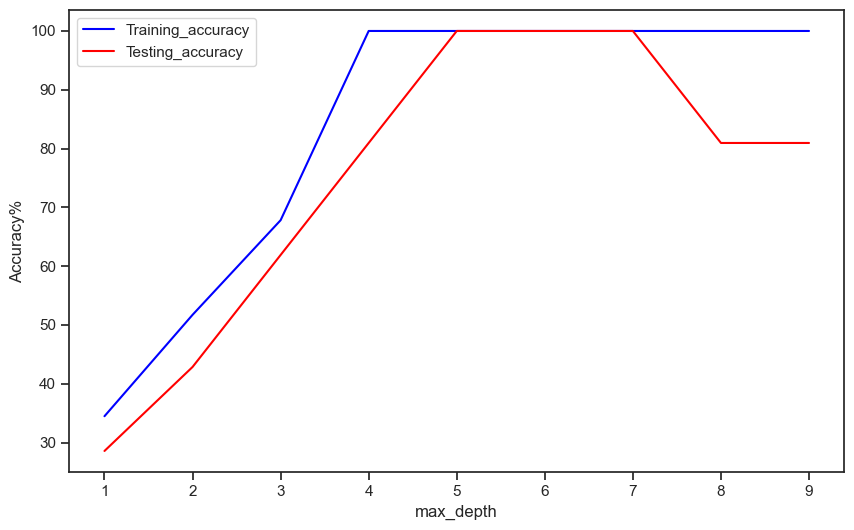


Figure 11: Train and Test accuracy (Source: Personal Collection)

These results highlight the importance of finding the right balance for ‘max\_depth’ to prevent overfitting and choose the best generalization. A “max\_depth” of 4 was reasonable choice in this dataset, as it provided the highest testing accuracy without overfitting. After experiment of hyperparameters, the model accuracy ended with 0.81, which means it correctly classifies 81% of the instances in the dataset. The F1 scores of macro average was 0.77 and weighted average was 0.75.

precision recall f1-score support

1-9 hours 1.00 1.00 1.00 3

10-19 hours 0.00 0.00 0.00 4

20-29 hours 0.43 1.00 0.60 3

30-34 hours 1.00 1.00 1.00 4

35-39 hours 1.00 1.00 1.00 4

40-44 hours 1.00 1.00 1.00 3

accuracy 0.81 21

macro avg 0.74 0.83 0.77 21

weighted avg 0.73 0.81 0.75 21

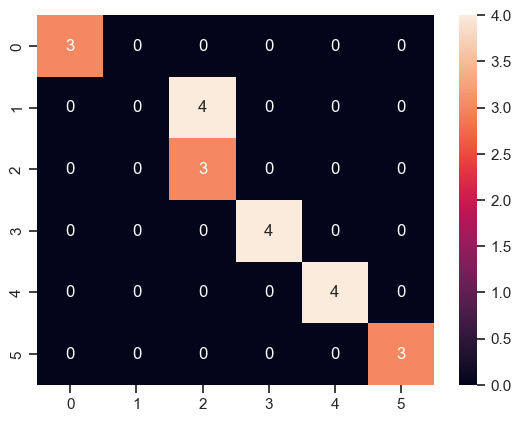


Figure 12: Confusion matrix of the DTC (4 max\_depth) – based on test and prediction (Source: Personal Collection)

Naive Bayes Classification model has achieved the accuracy of 1.0, which means it has made correct predictions for all instances in the dataset based on the confusion matrix. The accuracy of 1.0 indicates that the model has achieved perfect classification on the data. Additionally, for each individual class ('1-9 hours', '10-19 hours', etc.), have precision, recall, and F1-score values of 1.00, indicating perfect performance for each class. Overall, Naïve Bayes classification model performed the best fit for this dataset, achieved perfect classification result.

precision recall f1-score support

1-9 hours 1.00 1.00 1.00 3

10-19 hours 1.00 1.00 1.00 4

20-29 hours 1.00 1.00 1.00 3

30-34 hours 1.00 1.00 1.00 4

35-39 hours 1.00 1.00 1.00 4

40-44 hours 1.00 1.00 1.00 3

accuracy 1.00 21

macro avg 1.00 1.00 1.00 21

weighted avg 1.00 1.00 1.00 21

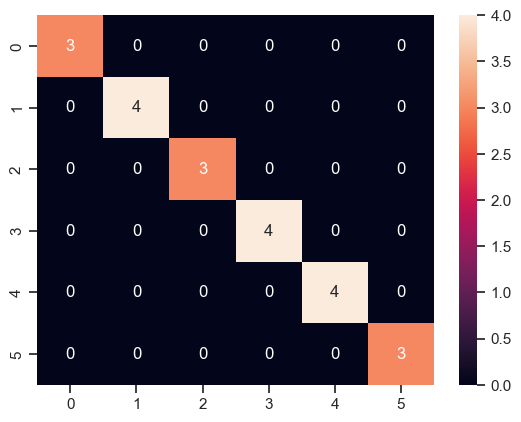


Figure 13: Confusion matrix of the DTC (4 max\_depth) – based on test and prediction (Source: Personal Collection)

### Support Vector Machine

Support Vector Machine (SVM) applied to the dataset with the default hyperparameter, the accuracy score was 0.428. From all the SVM hyperparameters applied in the cross validation of 10 folds.

Therefore, another cross validation was applied specifically for the SVM with linear kernel with 10 folds. The results of the Polynomial and RBF kernel showed clear underfitting, with results below 15% for training and testing. However, most of the linear kernel models tested across range of C values, higher accuracy was observed with only one-fold had slightly lower accuracy. And the mean accuracy 98.18% indicates that the SVM model with a linear kernel performs very well on the dataset. After kernel selection, following experiment was perform range of different regularization parameter C variable on the dataset. The following experiment achieved consistent accuracy.

Figure 14: RBF SVM with Different Gamma, C and Degree (Source: Personal Collection)

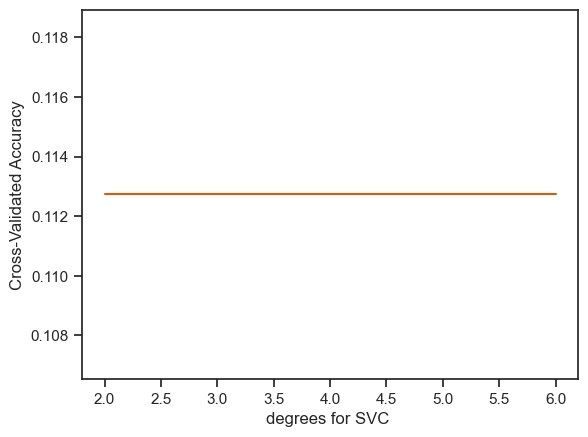
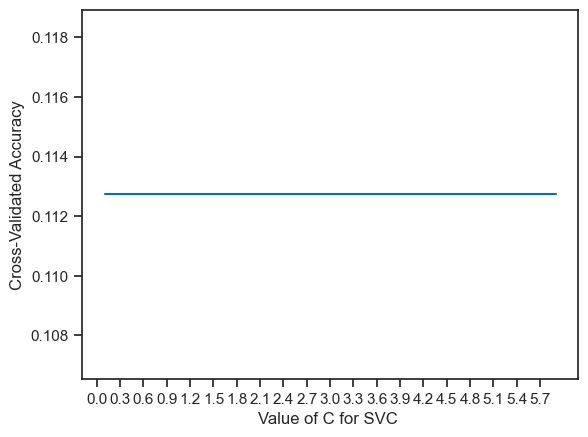
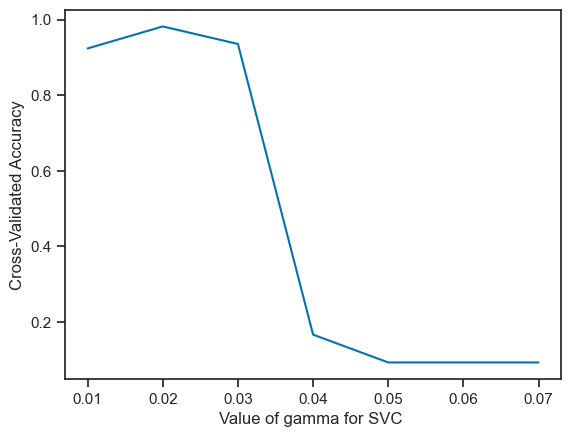


Figure 15: Linear SVM with Different Gamma, C, and Degree (Source: Personal Collection)

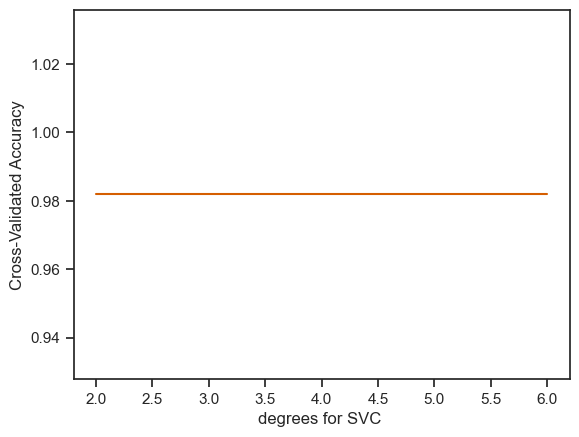
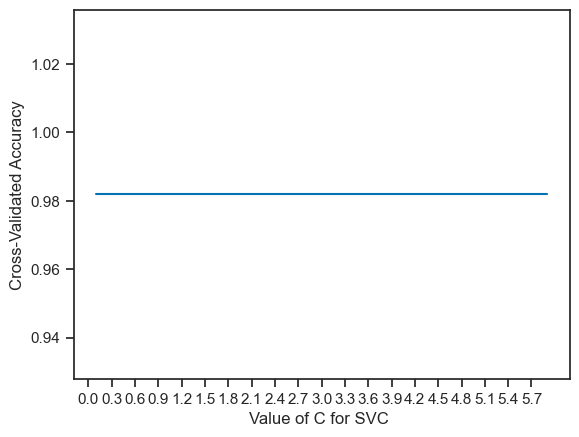
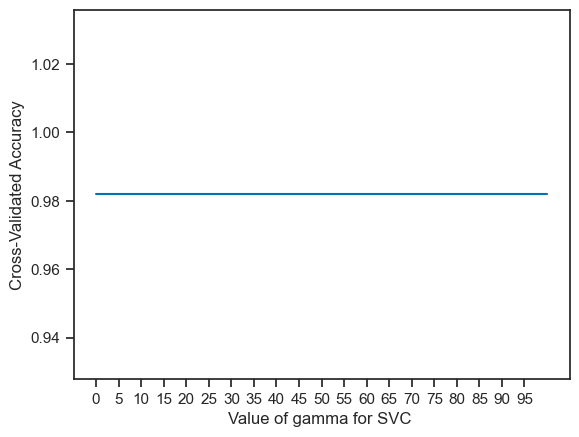
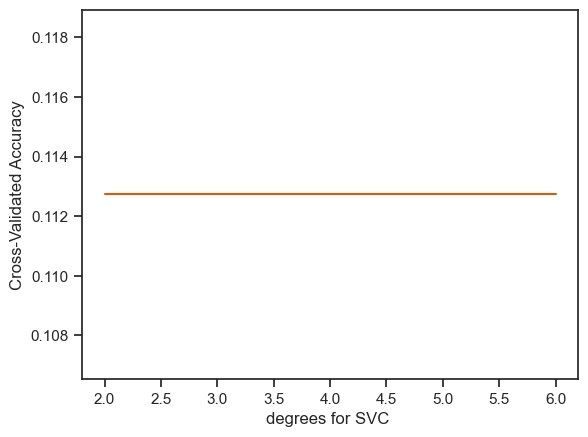
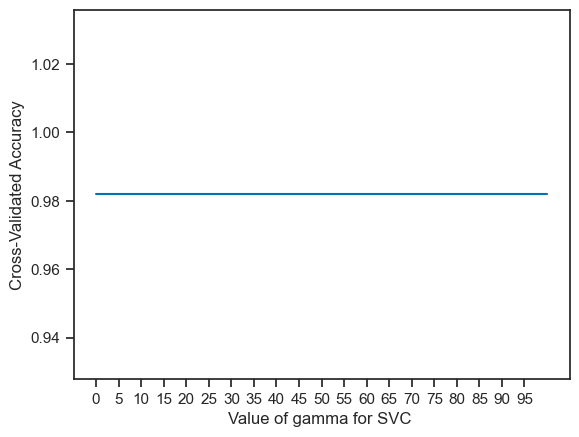
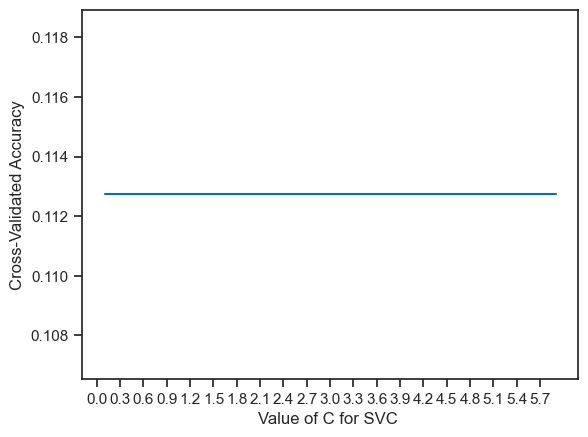


Figure 16:Poly SVM with Different Gamma, C and Degree (Source: Personal Collection)



Applied to the SVM linear model after evaluated the best performing hyperparameters of the model. The final testing accuracy was 0.95 and 10 fold Cross-validation score was 0.98.

Logistic Regression

The Logistic Regression model began with Recursive Feature Elimination with Cross-Validation (RFECV) to select an optimal number of features for a Logistic Regression model. The method found 7 optimal features, which 'VALUE(hours)', 'VALUE(nationality)', 'Nationality', 'VALUE(sector)', 'NACE Rev 2 Economic Sector', 'VALUE(education)', 'Education Level'. Following this feature selection method, the training and testing dataset split as follows: (108 samples, 7 features), (108 samples), (81 samples, 7 features), (27 samples, 7 features), (81 samples) and (27 samples).

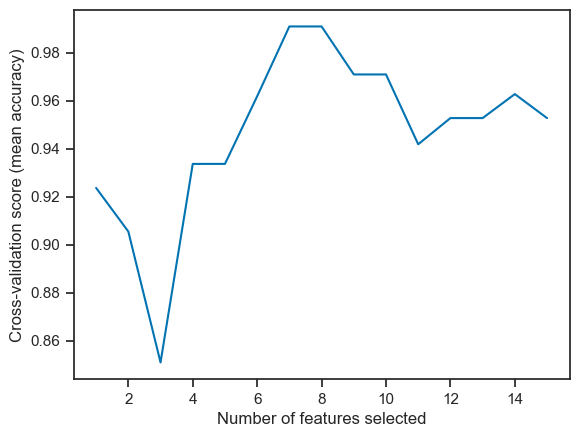


Figure 17: RFECV - Feature selection by Recursive Feature Elimination with Cross-Validation (Source: Personal Collection)

The model has evaluated range of C values for testing. Following experiment helped to find the highest accuracy C hyperparameter, which the most reasonable selection was 1.0, as it provided high accuracy without overfitting the data. When C equal 0.1 and 1 to 96.30%, the latter values remain constant for C equal 100, which is mean might overfitting, and the model may not generalize well to unseen data.

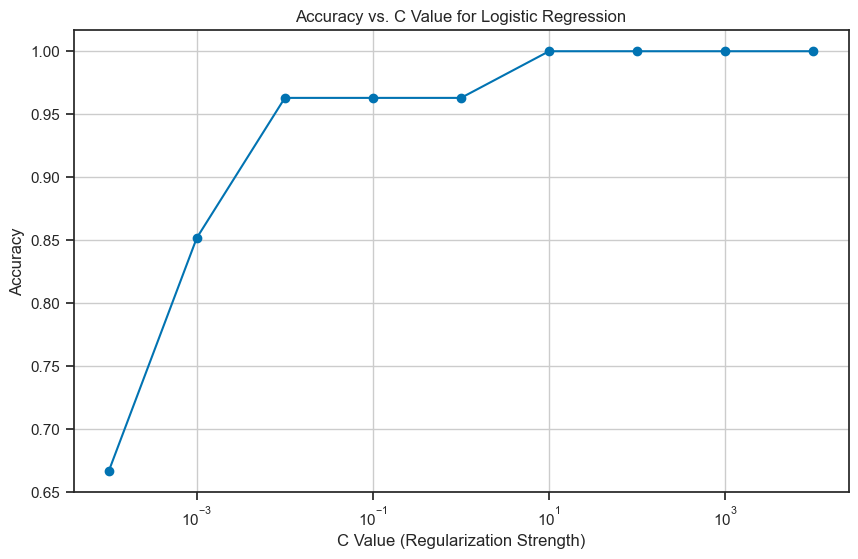


Figure 18: LR Accuracy with different C hyperparameter values (Source: Personal Collection)

Thereby, the final hyperparameter selected C equal 1, solver=”liblear”, multi\_class=’ovr’. The testing accuracy was 96.29% and F1 scores of macro average and weighted average were both 0.96. The results are better than another promising model, the LR model with C equal 1.

lassification Report:

precision recall f1-score support

1-9 hours 1.00 1.00 1.00 7

10-19 hours 0.88 1.00 0.93 7

20-29 hours 1.00 0.67 0.80 3

30-34 hours 1.00 1.00 1.00 4

35-39 hours 1.00 1.00 1.00 1

40-44 hours 1.00 1.00 1.00 5

accuracy 0.96 27

macro avg 0.98 0.94 0.96 27

weighted avg 0.97 0.96 0.96 27

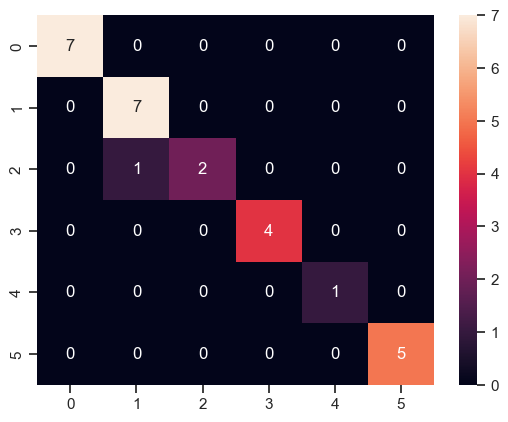


Figure 19: LR Accuracy with different C hyperparameter values (Source: Personal Collection)

Artificial neural networks

The neural networks proposed for this research are multilayer feedforward neural networks operating under supervised learning; they consist of three layers including one input layer, one hidden layer and one output layer.

The input layer, the hidden layer, and the output layer are fully interconnected. The weights of the connections and the biases are initialized randomly by the system, and adjusted through the learning process. The creation of the model starts from install Scipy (including numpy), Keras and backend (Tensorflow). There are seven input variables and one output variable which is weekly usual worked hours. The learning model to map rows of input variables (X) to an output (y). Then, split X, y variables into Training and Testing parts and each of variables’ test size is configuration 0.25. The data stored in a 2D array where the first dimension is rows and the second dimension is columns, e.g. After this, specify the number of neurons or nodes in the layer as the first argument, and specify the activation function the activation argument. In this model, used the rectified linear unit activation function referred to as “relu” on the first two layers and the Sigmoid function in the output layer. Because, the model’s output neuron produces a probability between 0 and 1, and the neuron with highest probability choose as the predicted class.

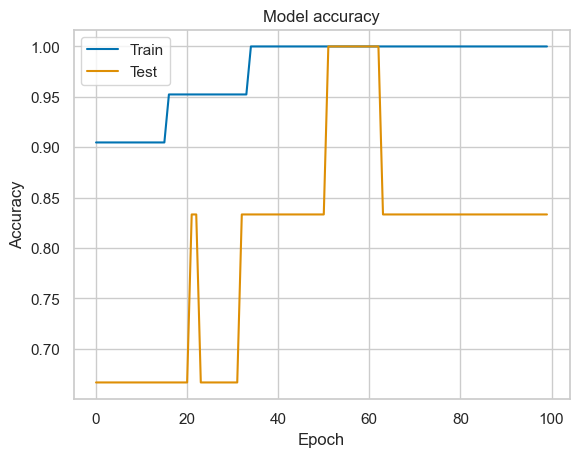


Figure 20: Training and Testing Accuracy (Source: Personal Collection)

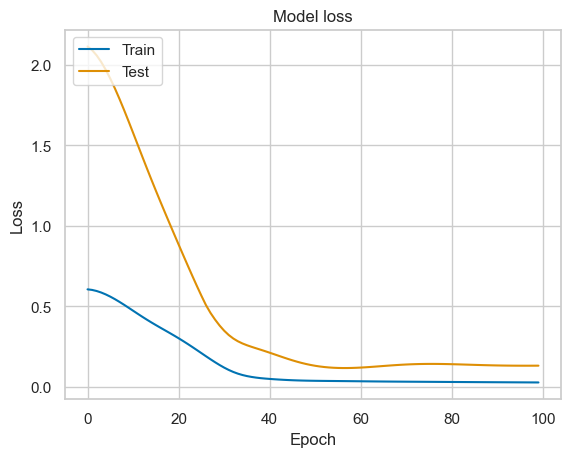


Figure 21: ANN- Training and Testing Loss (Source: Personal Collection)

Training occurs over epochs and each epoch is split into batches. Fit the model with 100 epochs and batch size is 10. The loss parameter specifies the loss function that the model will use during training. In this case, it's set to 'binary\_crossentropy', which used for binary classification problems. Finally, the training accuracy reached 98.7% and testing accuracy 96.30%, whereas the training loss was 0.0028 an the testing loss was 0.0505.

**Feature importance**

Feature importance rankings among different tree-based models, the tunned Random Forest and Decision Tree model. The consistency of results across different feature importance techniques, including Permutation Importance. Each subplot shows the features sorted by their importance for the respective model. Based on the importance bar charts for the Random Forest and Decision Tree models, “Education level” has the highest importance value from both models. The other feature, such as "NACE Rev 2 Economic Sector” and "VALUE(nationality)” have higher importance in the Random Forest model. The study found that "Education Level" appears to be the most influential feature for the models, followed by "VALUE(education)."

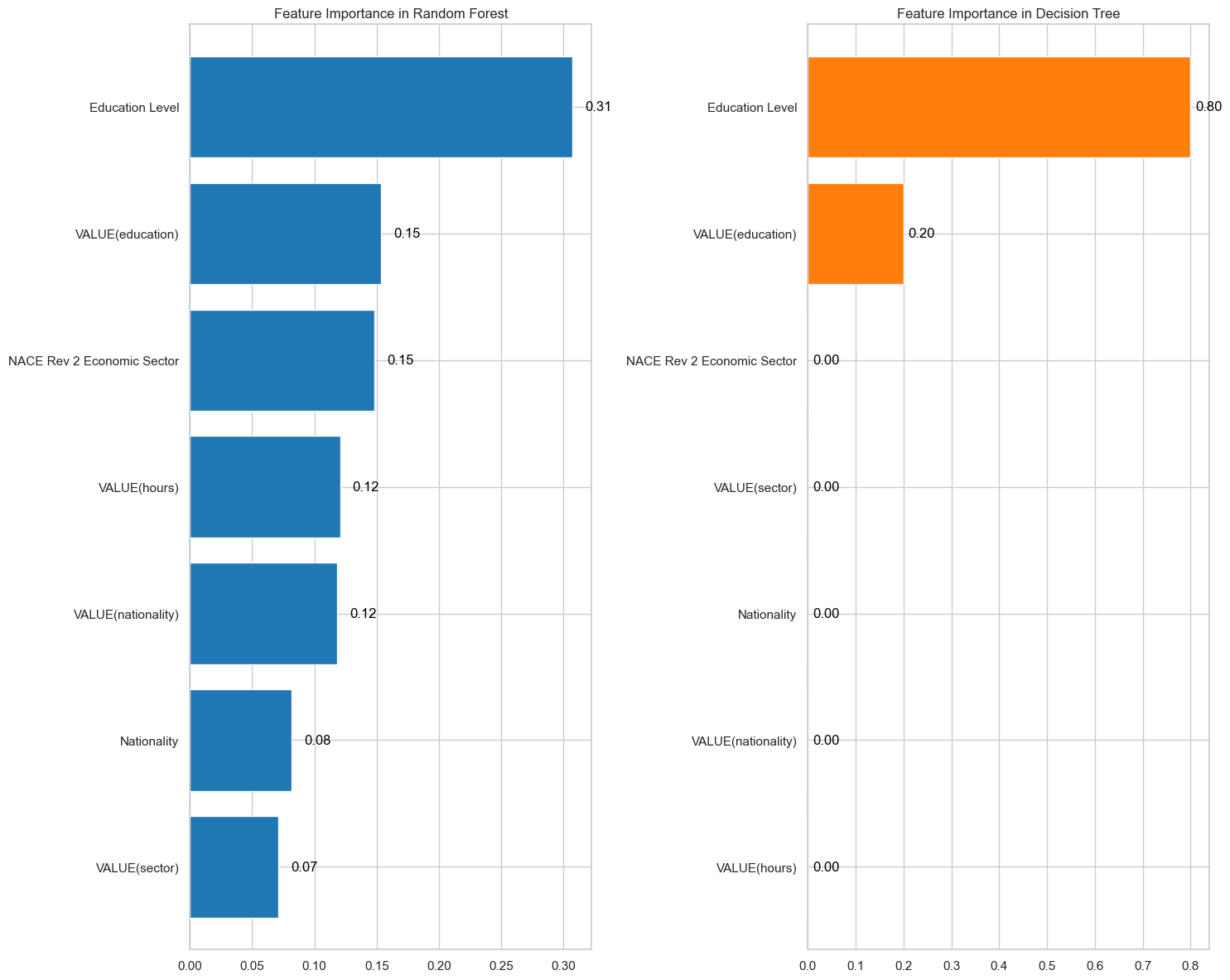


Figure 22: Feature importance by different tree-based models (Source: Personal Collection)

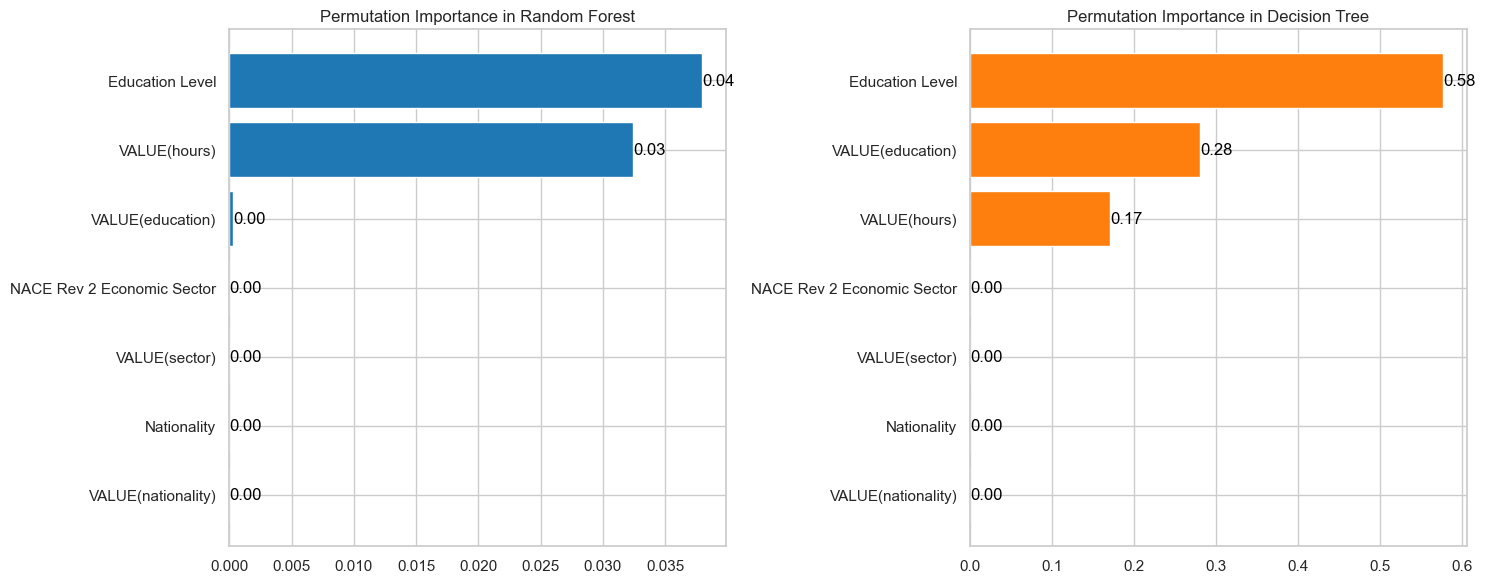


Figure 23: Permutation importance by different tree-based models (Source: Personal Collection)

**10. Conclusions and Recommendation**

This research aimed to using machine learning techniques to determine the most influential factor in working hours of national minimum wage workers in Ireland. This study used information of the Ireland statutory minimum wage workers between 2016 and 2020, which is four different datasets from one source, and made use of some data preprocessing techniques to ‘clean’ the datasets. The datasets underwent preprocessing to handle missing values, remove duplicates, adjust datatypes and perform feature engineering. The dataset was found to have a considerable amount of irrevelant information. And the process of data preparation involved dropping irrelevant rows from each columns of the datasets. In the final stage of feature engineering, the final dataset was created by selected valuable variables from four different datasets for use in machine learning methods. After completing this process, dataset has been prepared for train machine learning model and captured a range of MW employee characteristics including education, nationality, sector of work and the number of weekly usual work hours.

The machine learning techniques used to develop the models used to evaluate and predict the minimum wage workers working hours by education level, nationality and working sector. The Support Vector Classification (SVC), Naive Bayes (GaussianNB), Logistic Regression (LR), Random Forest Classifier (RFC), Decision Tree Classifier (DTC) and Deep Learning (ANN) are assessed in this research. This study found the Considering all the machine learning models assessed, the deep learning (ANN) model and Naive Bayes (GaussianNB) both showed higher cross-validation and training and testing accuracy than the other algorithms in the prediction of working hours of minimum wage workers. Furthermore, the study found that the Education Level of NMW workers appears to be the most influential factor in determining the working hours of NMW workers. The performance of the model was evaluated using the Mean Cross-validation accuracy score and Machine Learning Model Accuracy. The performance of the models is outlined below table 5.

|  |  |  |
| --- | --- | --- |
| **Algorithms** | **Cross-validation** | **Accuracy** |
| Support Vector Classification (SVC) + linear | 0.42 | 0.95 |
| Naive Bayes (GaussianNB) | 0.94 | 1.0 |
| Logistic Regression (LR) + “ovr” multi\_class + “liblinear” solver | 0.90 | 0.96 |
| Random forest Classifier (RFC) + 100 Estimator | 0.90 | 0.89 |
| Decision Tree Classifier (DTC)+4 Max\_depth | 0.92 | 0.81 |
| Deep Learning (ANN) | 0.96 | 0.96 |

Table 5: Cross-Validation and Accuracy of the selected Machine Learning Models

(Source: Personal collection)

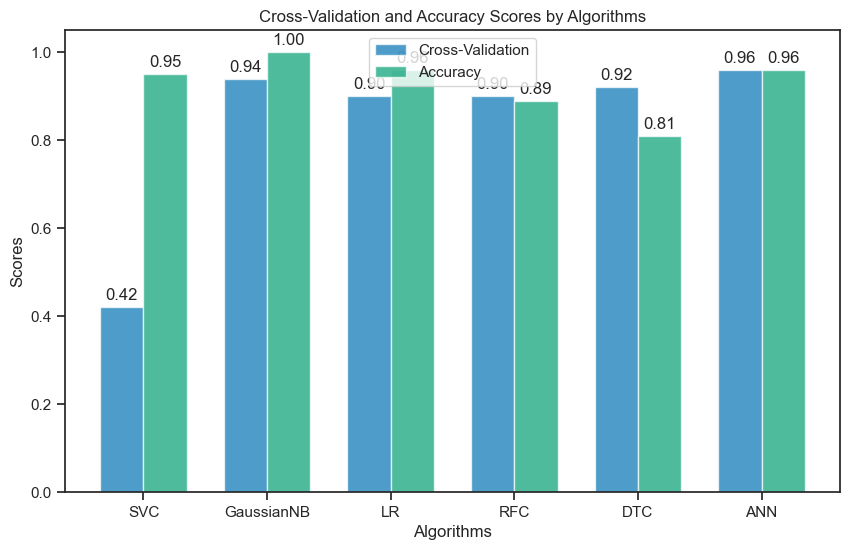
****

Figure 24: Cross-Validation and Accuracy scores by ML algorithms (Source: Personal collection)

In this study, I have successfully developed highly accurate predictive machine learning models for predict the weekly usual working hours of the NMW workers. Notably, the Education Level of NMW workers appears to be the. Conversely, nationality and the economic working sector in which NMW workers were found to have small impact on their working hours. Most importantly, my research demonstrated that there is no significant interaction between increases of the national minimum wage and the working hours of the minimum wage workers. These findings contribute valuable insights to my understanding of the dynamics between minimum wage policies and labor force behavior.

.

## References

Andrea, G., Herwig, I. and Daniele, P. The minimum wage and the tax and benefit system in Ireland. *A note for the Irish Low Pay Commission.*

Adams, C., Jonathan, M., and CarlyWill, S. (2018). The Minimum Wage and Search Effort, *NBER Working Paper No. 25128.*

Ahn., Tom., Peter, A., and Walter, W. (2011). The Distributional Impacts of Minimum Wage Increases When Both Labor Supply and Labor Demand Are Endogenous, *Journal of Business and Economic Statistics, 29(1): 12–23.*

Andreas, C. M. and Sarah, G. (2017). Introduction to Machine Learning with Python, Publishers.

Arindrajit, D., Lester, T.W. and Michael, R.(2010). Minimum Wage Effects across State Borders: Estimates Using Contiguous Counties, Review of Economics and Statistics.

Aurélien, G. and O’reilly, M. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition.

Barrett, A. and McGuinness, R. (2012). The Irish Labour Market and the Great Recession. *CESifo DICE Report 2/2012.*

Bertrand, M., Seamus, M. and Paul, R. (2017). A study of minimum wage employment in Ireland: the role of worker, household and job characteristics. *Economic and Social Research Institute (ESRI), The low Pay Commission.*

Belman, D., Wolfson, P. and Nawakitphaitoon, K. (2015). Who Is affected by the minimum wage? *Industrial Relations: A Journal of Economy and Society.*

Blanc, G., Lange, J., Mingda, Q. and Li-Yang, T. (2022). Properly Learning Decision Trees in almost Polynomial Time.

Breiman, L. (2001). Random forests. Machine Learning 45.

Bergin, A. & Kelly, E. (2020). The Labor Market in Ireland, 2000-2018. *IZA World of Labor.*

Brown, C., Gilroy, C. and Kohen, A. (1982). The effect of the minimum wage on employment and unemployment. *Journal of Economic Literature.*

Card, D. and Krueger, A. (1995). Myth and Measurement: The New Economics of the Minimum Wage, *New Jersey: Princeton University Press.*

Charlene. M. K. (2016). The effects of minimum wages on youth employment and income.

Charles, S. (2009). Phyton for Everybody, Exploring data using python 3.

Couch, K.A. and D.C. Wittenburg (2001). The response of hours of work to increases in the minimum wage’, Southern Economic Journal.

[Giuseppe, B. (2017). Machine Learning Algorithms: A reference guide to popular algorithms for data science and machine learning](https://www.amazon.co.uk/Machine-Learning-Algorithms-reference-algorithms-ebook/dp/B072QBG11J/ref=sr_1_1?keywords=9781785884511&linkCode=qs&qid=1694082931&s=books&sr=1-1)**.**

Daniel, L. and Chantal, L. (2014). Discovering Knowledge In Data: An Introduction To Data Exploration, Second Edition,

Daniel, B. (2019). Cross-validation. *Data science laboratory, Tokyo institiute of technology*

David, N., M. I. Salas, and William. W. (2014) Revisiting the minimum wage- employment debate: Throwing out the baby with the bathwater? *Industrial and Labor Relations Review 67:3,* pp. 608–648.

David, N. and William. W. (1992) Employment effects of minimum and subminimum wages: Panel data on state minimum wage laws. *Industrial and Labor Relations Review 46:1*, pp. 55–81.

David, C. (1992). Do minimum wages reduce employment? A case study of California, 1987–89, *ILR Review, 46(1): 38–54.*

Dickens, R., R. Riley and D. Wilkinson (2015). A Re-examination of the Impact of the UK National Minimum Wage on Employment.

Doruk, C., Arindrajit, D., Attila, L. and Ben, Z. (2019). The effect of minimum wages on low-wage jobs. Quarterly Journal of Economics 134.

Doruk, C., Arindrajit, D., Attila, L. and David, Z. (2021). Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes. *Seeing beyond the Trees.*

Duncan, M., Andrew, P. and Seamus, M. (2017). Employment and Hours Impacts of the National Minimum Wage and National Living Wage in Northern Ireland.

Elish, K., Seamus, M. (2017). A study of sub-minimum wage rates for young people. *Final Report to the Low Pay Commission.*

Gerard, B. (2016). A Random forest guided tour.

Giuseppe, B. (2017). Machine Learning Algorithms.

Giuseppe, B. (2017). Machine Learning Algorithms : Build Strong Foundation for Entering the World of Machine Learning and Data Science with the Help of This Comprehensive Guide, *APA 7th Edition (American Psychological Assoc.), MLA 9th Edition (Modern Language Assoc.)*

Guangfei, W., Jie, Z., Yanli, F., Aixing, H. and Jun, Y. (2020) A novel hybrid feature selection method based on dynamic feature importance.

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., and Guru, S. (2016). The Age Of Analytics: Competing In A Data-Driven World.

Jason, B. (2016). Feature Importance and Feature Selection With XGBoost in Python.

Jason, B. (2020). Dara Preparation for Machine Learning, *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python.*

Jiawei, H., Micheline, K. and Jian, P. (2012). Data mining, 3rd Edition.

John, P.M. and Luka, M. (2021) Machine learning for Dummies. 2nd Edition.

John, D.K., Brian, M.N. and Aoife, D. (2015). Fundamentals of machine learning for predictive data analytics. *Algorithms, Worked Examples and Case studies.*

John, A.R. (2007). Mathematical Statistics and Data Analysis, 3rd Edition, University of California, Berkeley.

Joseph, O.C. (2013). An Architecture for Big Data Analytics.

Judith, D., Paul, R. and Seamus, M. (2019). The prevalence and effect on hours worked of the minimum wage in Ireland: a sectoral and regional analysis. *ESRI Research Series 93. Low Pay Commission.*

Khongdet, P., Alok, K.S., Mukesh, K.S. and Tanya, S. (2022). Fundamental of Machine Learning.

Lior, R. and Oded, M. (2010). Data mining with Decision trees, Theory and Application, 2nd edition.

Leo, B. (2001). Random forests, *Machine learning, 45(1): 5–32.*

Low Pay Commission. (2019). Three Year Report (2015-2018). *Low Pay Commission.*

Low Pay Commission. (2018). A Review of Current Practices in relation to Tips & Gratuities.

Low Pay Commission (2016). Recommendations for the National Minimum Wage. *Low Pay Commission.*

MaCurdy, T. (2015). How effective is the minimum wage at supporting the poor?, *Journal of Political Economy, Vol. 123, No. 2, pp. 497-545.*

Machin, S., Manning, A. and Rahman, L. (2003). Where the minimum wage bites hard: introduction of minimum wages to a low wage sector.

Mark, B.S., Joanna, K.S. (2006). The Other Margin: Do Minimum Wages Cause Working Hours Adjustments for Low-Wage Worker? *University of Warwick,* *University of York.*

Mark, B., Andrea, S. and Mark, T. (2012). The Impact of the National Minimum Wage on Earnings, Employment and Hours through the Recession. *A report to the Low Pay Commission.*

Matt, H. (2019). Machine learning Rocket Reference. Working with Structured Data in Python.

Matthew, K. (2017). Thoughtful Machine Learning with Python, A Test-Driven Approach.

McFlynn, P. (2015). Public Sector Employment in Northern Ireland. NERI Research InBrief (no 20).

McGuinness, S. & Redmond, P. (forthcoming). Estimating the Effect of an Increase in the Minimum Wage on Hours Worked and Employment in Ireland. *Low Pay Commission (Ireland).*

Meer, J. and West, J. (2016). Effects of the minimum wage on employment dynamics. Journal of Human Resources.

Mark, S. (2009). Programming in Python 3*, A Complete Introduction to the Python Language Second Edition.*

Mark, S. and Joanna, S. (2006). The other Margin: Do minimum wages cause working hours adjustment for Low-Wage workers?

Michael, Q.P. (2002). Qualitative Research & Evaluation Methods. *Patton.- 3rd edition.*

Mohanad, A., Stefano, F., Angelo, G., Vincenzo, P. and Fabio, S. (2021). A Survey of Unsupervised Generative Models for Exploratory Data Analysis and Representation Learning.

Neumark, D. (2017). The employment effects of minimum wages: some questions we need to answer. *National Bureau of Economic Research*.

Neumark, D. and Wascher, W. (2006) Minimum Wages and Employment: A Review of Evidence from the New Minimum Wage Research. *National Bureau of Economic Research*

Nolan, B., O'Neil, D., and Williams, J. (2002). The Impact of the Minimum Wage on Irish Firms. ESRI Policy Research Series No. 44.

Ormerod, C. and Ritchie, F. (2007). Issues in the measurement of low pay. *Economic and Labour Market Review.*

Pang-Ning, T., Michael, S. and Vipin, K. (2006). Introduction to Data Mining, 1st Edition, Chapter 4.

Paul, R. and Seamus, M. (2022). Heterogeneous effects of minimum wage increase on hours worked. *ESRI Research Series 132. Low Pay Commission.*

Paul, R. and Seamus, M. (2021). The impact of the 2016 minimum wage increase on average labour costs, hours worked and employment in Irish firms. *ESRI Research Series Number 118.*

Paul, R., Karina, D. and Seamus ,M. (2019). The impact of a change in the national minimum wage on the distribution of hourly wages and household income in Ireland. *ESRI Research Series 86. Low Pay Commission.*

Paul, R., Seamus, M. and Bertrand, M. (2018). An examination of the labour market transitions of minimum wage workers in Ireland. *ESRI Research Series Number 75.*

Rahil, S. (2018). Feature Selection Techniques in Machine Learning with Python.

Ramzan, B., Bajwa, I. S., Jamil, N., Amin, R. U., Ramzan, S., Mirza, F., and Sarwar, N. (2019). An Intelligent Data Analysis for Recommendation Systems Using Machine Learning. *Scientific Programming*.

Raschka, S. (2015). Python machine learning. *Packt publishing open source.*

Redmond, P., B. Maître, S. McGuinness and K. Maragkou (2021). A Comparative Assessment of Minimum Wage Employment in Europe, *ESRI Research Series Number 123.*

Richard, D., Rebecca, R. and David, W. (2009). The employment and hours of work effects of the changing national minimum wage. *Report prepared for the Low Pay Commission.*

Richard, D. (2015). How are minimum wages set? *Author's main message (pp.1).*

Roderick, L. and Donald, R. (2002). Statistical Analysis with Missing Data

Lan, Y. and Jason, A.C. Use Case and Performance Analyses for Missing Data Imputation Methods in Big Data Analytics, *Computer Science Department California State Polytechnic University Pomona.*

Seamus, M. and Paul, R. (2018). Estimating the effect if an increase in the Minimum wage on hours worked and employment. *IZA DP No. 11623*

Sebastian, R. and Vahid, M. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2, 3rd Edition.

Sebastian, R. (2015). Python Machine Learning.

Sebastian, R. Liu, YH. And Mirjalili, V. (2016). Python: deeper insights into machine learning.

Sebastian, R. Liu, YH. And Mirjalili, V. (2022). Machine Learning with PyTorch and Scikit-Learn.

Shagar, M. and Alexander, J.S. (2002). Advanced lectures on Machine learning.

Stewart, M.B. and Swaffield, J.K. (2008). The other margin: do minimum wages cause working hours adjustments for low-wage workers?

Shrestha, A. and Mahmood, A. (2019). Review of Deep Learning Algorithms and Architectures.

Thakur, A. (2020) Approaching (Almost) Any Machine Learning Problem.

Terence, P., Kerem, T., Christopher, C. and Jeremy, H. (2018). Beware Default Random Forest Importances

Puneet, M. and Arun, S.Y. (2020). Improving the Classification Accuracy using Recursive Feature Elimination with Cross-Validation

Wellington, K. and Absalom, E. (2023). Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives.

Wes, M. (2017). Python for Data Analysis, *Data Wrangling with Pandas, NumPy, and IPython, Second edition.*

Zahraa S.A. and Geoffrey. (2017). Data Preparation, *Research Gate.*

Zavodny, M. (2000). The effect of the minimum wage on employment and hours.

Zhenyun, D., Xiaoshu, Z., Debo, C., Ming, Z., and Shichao, Z. (2015). Efficient kNN classification algorithm for big data.

**Appendix A**

Administrative Data Sources (EAADS), which links earnings from administrative sources to Labour Force Survey National Minimum Wage Estimates data. The data are administered by the Central Statistics Office (CSO) in Ireland.

1. The dataset of Education level is the total number of Employees aged 15 years and over from the second quarter of 2016 to the third quarter of 2020.

Link: <https://data.cso.ie/table/MWA27>

1. he Nationalities of the total number which are Employees aged 15 years and over from the second quarter of 2016 to the fourth quarter of 2019. Link: <https://data.cso.ie/table/MWA25>
2. A weekly usual worked hours datasets was generated by worked hours of employees who has aged 15 years and over from the second quarter of 2016 to the third quarter of 2020 with the national minimum wage earnings status. Link: <https://data.cso.ie/table/MWA15>
3. NACE (Nomenclature of Economic Activities) is the European statistical classification of economic activities. NACE groups organizations according to their business activities. The dataset of those economic sectors was combined the total number of Employees aged 15 years and over from the second quarter of 2016 to the third quarter of 2020 with the national minimum wage earnings status. Link: <https://data.cso.ie/table/MWA07>

The copyrights of Labour Force Survey National Minimum Wage Estimates to the Central Statistics Office (CSO) in Ireland. These records include all information collected directly for CSO statistical purposes in statutory or voluntary inquiries from persons, households, businesses and undertakings, or indirectly from the administrative records of public authorities held on completed questionnaires, worksheets or data bases. The CSO gathers information through both statutory (legally required) and voluntary inquiries. Statutory inquiries might be legally mandated data collection, while voluntary inquiries are those where participants provide data voluntarily.