UNIVERSITY OF ICELAND

M.S. Thesis
in Applied Statistics

Predictive performance of imputation methods

Evaluating imputation methods in predicting labour force status

Friðrik Þór Bjarnason

February 2024

FACULTY OF PHYSICAL SCIENCES

Predictive performance of imputation methods

Friðrik Þór Bjarnason

30 ECTS thesis submitted in partial fulfillment of a
MAS degree in Applied Statistics

Advisors
Anna Helga Jónsdóttir
Stefanía Benónísdóttir
Ólafur Már Sigurðsson

Examiner
Violeta Calian

Faculty of Physical Sciences
School of Engineering and Natural Sciences
University of Iceland
Reykjavik, February 2024

Predictive performance of imputation methods
Evaluating imputation methods in predicting labour force status

30 ECTS thesis submitted in partial fulfilment of a MAS degree in Applied Statistics.

Faculty of Physical Sciences
School of Engineering and Natural Sciences
University of Iceland
Tæknigarður, Dunhaga 5
107, Reykjavik
Iceland

Telephone: 525 4000

Bibliographic information:
Friðrik Þór Bjarnason, 2024, *Predictive performance of imputation methods,* MAS thesis, Faculty of Physical Sciences, University of Iceland, 46 pp

Abstract

Key statistics about the status of Iceland's labour force market come from the labour force survey conducted by Statistics Iceland. Response rate has however been declining since 2011 and certain groups are less likely to respond (Sigurðsson, 2022). Therefore, the survey results may be at risk of response bias. Currently weights are utilized to adjust for the bias. However, utilizing additional techniques like imputation may be suitable in this scenario. Due to unit non-response being the primary form of non-response in the labour force survey, the typical practice of using other survey variables to aid imputation is not feasible. However, Statistics Iceland possesses a significant amount of administrative data regarding all residents in Iceland. Some of this data could signal people's status on the labour market, and therefore be used as predictive variables in imputation. This study takes the first step in assessing the possibility of imputation in the Icelandic labour force survey, specifically, by focusing on prediction accuracy of different methods. The prediction accuracy of the methods explored in this study demonstrates that utilizing administrative data as predictor variables allows for effective imputation of labour force status. Furthermore, the study identified specific considerations to keep in mind before imputation, emphasizing the importance of balancing the factor distribution before imputation. Additionally, the results indicate that imputation is not suitable for a particular age group.

Útdráttur

Lykilstærðir varðandi íslenskt hagkerfi og samfélag eru unnar úr niðurstöðum Vinnumarkaðsrannsóknar Hagstofu Íslands (VMR). Svarhlutfall í rannsókninni hefur þó farið minnkandi frá árinu 2011 og ákveðnir hópar líklegri til þess taka ekki þátt í rannsókninni. Því er hætta til staðar á að rannsóknin gefi bjagaðar niðurstöður. Til þess að koma í veg fyrir bjaga í dag er notuð vog með því markmiði að úrtakið lýsi þýðinu betur. Hins vegar er eðlilegt að kanna hvort hægt sé að grípa til annara aðferða eins og tilreiknunar til þess bæta niðurstöður í VMR. Þar sem helsta tegund brottfalls í VMR er fullkomin villa (*e. unit-nonresponse*) er oftast ekki hægt að nota önnur svör úr rannsókninni fyrir tilreiknun. Hins vegar býr Hagstofa Íslands yfir miklu magni af skráargögnum varðandi íbúa Íslands. Aðgengileg skráargögn gætu því mögulega gefið vísbendingu varðandi stöðu fólks á vinnumarkaði. Í þessari rannsókn verður tekið fyrsta skrefið í að kanna hvort skráargögn geti verið notuð til þess að tilreikna svör í VMR. Til þess að meta það var nákvæmni mismunandi aðferða til þess að spá fyrir um atvinnustöðu skoðuð. Niðustöður rannsóknarinnar sýna að vissulega er hægt að nota skráargögn til þess að spá fyrir um atvinnustöðu. Niðurstöður gefa einnig til kynna ákveðna þætti til þess að hafa í huga fyrir tilreiknun. Þ.e.a.s. mikilvægi jöfnunar flokka fyrir tilreiknun kom greinilega í ljós og einnig að tilreiknun er mögulega ekki viðeigandi fyrir ákveðinn aldurshóp.

Table of Contents

List of Figures

List of Tables

Abbreviations

MI	Multiple imputation
MICE	Multiple imputations by chained equations
EU-LFS	European Union labour force survey
IS-LFS	Icelandic labour force survey
ILO	Labour force status variable
MCAR	Missing completely at random
MAR	Missing at random
MNAR	Missing not at random
CART	Classification and regression trees
MSE	Mean squared error
MDA	Mean decrease in accuracy
MDG	Mean decrease in Gini index
ILO_skrá	Derived variable for ILO representation. Calculated from taxable income and unemployment-benefits data. Created prior to the research and was accessible in the Statistics Iceland database

1 Introduction

1.1 Labour force survey

Since the year 1991, Statistics Iceland has carried out a comprehensive labour force survey which serves as the primary source for key statistics regarding Iceland's labour market status (Statistics Iceland: Labour-market, n.d.). These statistics include crucial indicators such as activity rate, employed persons, unemployment and working hours. The survey's findings hold significant importance in Iceland as they are used by many, for example ministries, media, scholars as well the Central Bank of Iceland. Accurate labour market statistics serve as the cornerstone for informed policy decisions, economic planning, and for understanding changes in Icelandic society, making them essential. The survey became a part of the European Union labour force survey (EU-LFS) in the year 1995 following agreements with the European Economic Area (EEA). Since then, Statistics Iceland has followed international definitions and guidelines when conducting the survey. Results can therefore be used on an international scale by foreign institutions like Eurostat, and comparison between countries is possible (Sigurðsson, 2022). Clearly, the Icelandic labour force survey (IS-LFS) holds significant importance. However, surveys in general face an issue.

An international trend analysis covering the years 1980-1997 revealed a decline in response rates over time (De Leeuw & De Heer, 2002). Building upon this research, Luiten, Hox, and De Leeuw (2020) conducted a more recent analysis in the time period 1998 to 2015 and confirmed a consistent decline in survey response rates on an international scale. Except for later onset, IS-LFS is not an exception to this trend. Since 2011, the response rate has been rapidly declining. In 2011, the response rate reached its peak at 86.2%, but by 2020 it had dropped to 66% (Sigurðsson, 2022). To summarize, the sample size in IS-LFS is decreasing. It is often thought that having a large sample size is preferable to a smaller one. This idea is supported by the fact that, when all other factors are held constant, enlarging the sample size in a study results in a decrease of the standard error and narrowing of the confidence interval (Meng, 2018). However, it is essential to recognize that more data is not always better, and it should be approached with caution. Blindly accepting the notion that more data is better without considering other factors could potentially harm the accuracy of results. This has been clearly demonstrated by the big data paradox phenomenon.
The big data paradox is a phenomenon we can observe in the real world. It suggests that having a large biased sample can be particularly problematic as it leads to biased estimates and overly narrow confidence intervals, making us more sure of our biased results (Meng, 2018). Furthermore, research on the big data paradox phenomenon has shown that a small random sample can produce results just as accurate as those from a significantly larger but biased sample (Bradley et al., 2021). Hence, quality over quantity.
With all that said, the decrease in response rates in IS-LFS alone might not necessarily be a cause for concern. However, Sigurðsson found evidence of a systematic difference between responders and non-responders (2022). That is, subjects who decline to participate exhibit

systematic differences compared to participants. Issues arise from this as it can introduce non-response bias and limit the generalizability of the results to the entire population (Groves, 2006). Due to this, certain adjustments must be made to ensure that the sample in the IS-LFS survey accurately mirrors the intended population.

There are several strategies available to minimize the effect of nonresponse in surveys and a common way is to use nonresponse weighting. Weights are assigned after data collection, primarily aiming to modify the data, making it a better reflection of the intended study population. Thus, the main purpose of the weighting method is adjusting the data to make it more representative of the group we want to study. This helps minimizing potential biases. Currently, this method is applied to the sample in the IS-LFS. However, weighting procedures are an imperfect solution to address problems associated with non-response, as they come with some issues themselves. One known issue with weights is increased variance in survey estimates (Kish, 1992; Little et al., 1997). Respondents with higher weights have greater influence on the estimates, making the results more sensitive to the specific responses of these individuals. This leads to an increased variability in the estimates, in simpler terms, the study's findings might not be as accurate or reliable. Another issue arises from having a small number of individuals representing large groups, the lack of diversity in the respondent group persists even after weighting. For instance, if only one respondent represents a particular country, the results cannot be effectively stratified or broken down by other characteristics such as age, sex, or other factors. This limits the ability to provide detailed insights about different subgroups within the larger population. These limitations suggest the consideration of supplementary tools.

An increasingly popular method for enhancing survey data is imputation. Imputation essentially involves using statistical techniques to fill in or replacing missing values for non-respondents based on the responses provided by others. One of the early implementations of imputation was documented in 1957 in a research conducted by the U.S. Census Bureau, where a univac computer was used to impute missing items (Buuren, 2018). This technique ensures that all cases are considered, by estimating missing data using available information. Once all the gaps are filled, the dataset can be analysed using regular methods for complete data.
In surveys, imputation is immensely beneficial as it helps us achieve a more comprehensive and precise collection of responses (Perneger & Burnand, 2005). With a potentially richer dataset, there is a possibility of gaining deeper insights, and being less likely to miss crucial information about the intended study population. Moreover, employing imputation can help by including larger and a more diverse group of participants in the survey allowing for a better stratification of the results. Additionally, imputation can also reduce the effort needed from survey respondents (Rässler, Schnell, 2004). As administrative and unconventional data become more accessible, respondents can be required to provide less information since certain data is either already known or can be imputed using the available information. This can be particularly valuable in situations where response rates are declining like in the IS-LFS.

1.2 Research questions

In summary, imputation can clearly be a valuable tool for addressing non-response in surveys. In the IS-LFS, the most common type of non-response is when individuals do not participate. In these cases, we can not use additional survey data to fill in the missing information.

However, Statistics Iceland possesses a significant amount of administrative data regarding all residents in Iceland. This fact leads to the question whether this administrative data can be utilized to impute values for non-respondents in the IS-LFS. This could result in larger sample size, enable better stratification of estimates, and potentially reduce the variance in survey estimates. Nevertheless, before looking into the above possibilities it must be asked if imputation is appropriate in this setting. During our research, we take the initial step of this assessment through the following research questions:

- In regard to prediction accuracy, can imputation for labour force status be done with available administrative data as predictor variables?
- Additionally, how do different imputations methods compare in terms of prediction performance?

2 Background

2.1 Weights

Non-response in surveys is generally considered as either item-nonresponse or unit-nonresponse. Item non-response occurs in a survey when respondents answer some questions but leave others unanswered or provide incomplete or partial responses to specific items or questions. On the other hand, unit-nonresponse occurs when a sampled individual does not participate in the survey or does not provide answers to any of the questions (Grossmann, 2020). A typical example of this is when subjects are not reachable through the phone. In the IS-LFS, unit-nonresponse is the prevalent type of non-response.

A common way to deal with unit-nonresponse in surveys is to use nonresponse weighting. Currently, member states of the European Union (EU), three EFTA countries (Iceland, Norway, and Switzerland) and four candidate countries (Montenegro, North Macedonia, Serbia and Turkey all employ this approach for their EU-LFS data (Eurostat, 2022).
In nonresponse weighting, weights are applied posterior to the data collection and groups less likely to participate in the survey get higher weights than they would otherwise. For example, if a response from an older individual is considered as a baseline of 1, we could assign a weight of 1.5 to address the lower response rate within this demographic. This allows for adjustments that make the sample more representative of the target population, thus, reducing potential bias (Holt & Elliot, 1991). As mentioned earlier, Statistics Iceland possesses a significant amount of administrative data regarding all residents in Iceland. This data ranges from year 2003 to the present and can be linked to the subjects in the dataset at each time they are selected to participate in the survey. This available data provides information about the characteristics of both responders and non-responders in the survey, forming the foundation for calculating the weights. Commonly used characteristics in weights are demographic variables (e.g. age, sex, race, education), geographic variables and other relevant information. The characteristics currently used in weights for the Icelandic labour survey are: age, sex, education, a three level origin variable classifying individuals as Icelandic, immigrants for less than 10 years or immigrants for more than 10 years. Finally, ILO_skrá, a derived variable designed to estimate an individual's labour force status, is used. A more detailed description of the creation of this variable can be found in Section 3.1.3. These variables are then used as independent variables in logistic regression model, with the dependent variable being the response (yes or no). The regression model can be expressed through the following formula:

$$\log\left(\frac{P(answer=1)}{1-P(answer=1)}\right) = \beta_0 + \beta_1 \times origin + \beta_2 \times sex + \beta_3 \times education + \beta_4 \times age + \beta_5 \times ILO\_skrá \quad (1)$$

In the formula above, the left-hand side represents the log-odds ratio, which is the natural logarithm of the odds of the event answer=1 happening compared to it not happening. On the right-hand side β_0 is the intercept term, and the remaining terms are the predictor variables (origin, sex, education, age, and ILO_skrá), each multiplied by their respective coefficients

(β_1,β_2,β_3,β_4,β_5). This logistic regression model is then employed to estimate the probability of responding (response propensity) and thereby determining the likelihood of an individual participating in the survey based on their characteristics mentioned above. The mean response propensity of the survey is then divided by the response propensity for each individual to calculate their non-response weight. Through this method, groups less likely to participate in the survey get higher weights than they would otherwise, allowing for adjustments that make the sample more representative of the target population. Considering the variables used to construct the Icelandic weights, Ólafsson's research illustrated that individuals within the age range of 21-35 exhibit lower participation rates in the IS-LFS. Moreover, sex, education, ILO_skrá and origin were highlighted as influential factors affecting participation (2022). Thus, if imputation is to be employed, it is crucial for the imputation process to perform effectively for these specific demographic groups.

2.2 Imputation

While the majority of research has focused on imputation for item non-response, some have proposed its applicability in unit non-response situations as well (Rässler & Schnell, 2004). Multiple imputation (MI) methods such as Multiple imputation by chained equations (MICE) have been shown to decrease variance in survey estimates when compared to weighting adjustments (Peytchev, 2012). MI involves generating multiple sets of plausible values for missing data and combining them to create more accurate and robust estimates. Specifically, MICE is a sophisticated MI method employed to address missing data. A more detailed description of MI will be provided in Section 2.2.2, and for a comprehensive understanding of MICE, please refer to the detailed explanation in Section 3.2.1. Alanya et al. (2015) also employed MI to handle unit non-response and contrasted its efficiency with weighting. Their investigation showed that MI performed superiorly in certain scenarios but not universally. Today, data imputation plays a crucial role in handling missing data, and various imputation techniques are used spanning from simple methods such as mean or median imputation to more complex methods, such as regression-based MI. Furthermore, the selection of imputation method is context-dependent, and it can vary not just between surveys, but also within a single survey based on specific conditions. This means that distinct imputation methods may be employed for different variables within the same dataset.

2.2.1 Applications

Multiple organizations responsible for national statistics employ imputation methods to enhance data quality and improve survey estimates. In their quality report for the EU-LFS some organizations mention using data imputation for non-response. For instance, in the United Kingdom an imputation method known as “roll-forward” has been applied to the labour force survey data. When a previous response is followed by a non-response, the data from the previous response is carried forward and used in the current period. However, this roll-forward imputation is limited to a single period (The Office for National Statistics, 2021). In Romania, the INCDECIL variable, which represents monthly take-home pay from the main job, is imputed using hot-deck imputation. In this method missing values are filled using responses from similar subjects (donors) within the same dataset (Eurostat, 2018).

When looking outside of Europe, the United States Census Bureau (USCB), which is responsible for producing statistics about the American population and economy, also uses

imputation to fill in missing values in some cases. In the National Survey of Children's Health (NSCH), imputation is used to fill in missing data for several demographic variables. For variables such as sex and race, the hot-dec method is used. The USCB also employs a more complex imputation approach for the family poverty ratio. Specifically, MI with a regression-based imputation is used to generate six plausible values for the family poverty ratio (U.S Census Bureau, 2021).

Statistics Canada also uses imputation in their survey statistics. Statistics Canada has designed an imputation system referred to as CANCEIS, which stands for the Canadian Census Edit and Imputation System. This system is used to fill in missing data within census datasets and employs a nearest-neighbour imputation approach. This method is similar to hot-deck imputation, but with a distinction: imputations are derived from a set of potential donors (the "nearest neighbours") closely matching the characteristics of the subject requiring imputation, rather than a single donor. Statistics Canada has also used imputation to reduce response burden by utilizing financial information about respondents from the Canada Revenue Agency. However, these administrative data frequently contain missing or inconsistent values. To effectively utilize the data, some imputation systems have been employed to enhance data quality prior to preceding to the next stage (Statistics Canada, 2022).

2.2.2 Multiple imputation (MI)

Generally, imputation methods can be divided to single imputation and MI, and with time, MI has gained popularity. Single imputation methods estimate the potential value for each missing entry and replace it with a single value in the dataset. Methods that impute a single value include mean imputation, last observation carried forward and hot deck imputation. However, these methods are considered suboptimal as these approaches overlook the uncertainty associated with imputing values. Imputed values are treated as true values, and therefore, standard techniques for complete data are applied to generate results. This leads to a variance estimate that notably underestimates the actual variances and less accurate confidence intervals (Shao & Sitter, 1996).

The creation of MI was prompted by the criticism of single imputation methods not addressing this uncertainty adequately. MI, which estimates and replaces missing values multiple times, is considered a more effective approach in handling missing data, because employing multiple plausible values accounts for the uncertainty associated with estimating the possible missing values (Li et al, 2015). The fundamental concept behind MI is creating multiple plausible estimates for missing values based on the available information. Instead of relying on a single imputed value, the method generates multiple imputations and creates equivalent number of complete datasets. The imputed datasets are then analysed separately and the results from all datasets combined using specific rules or formulas to appropriately account for the imputation uncertainty into the final analyses (Buuren, 2018).
To visually and clearly illustrate the fundamental steps of MI, we present Figure 1. As illustraded in Figure 1, we begin with a dataset containing missing data. Imputation involves generating three distinct values for each data point, resulting in three complete datasets. Subsequently, analyses are conducted independently for each dataset, and finally, the outcomes from these analyses are combined to form pooled results.

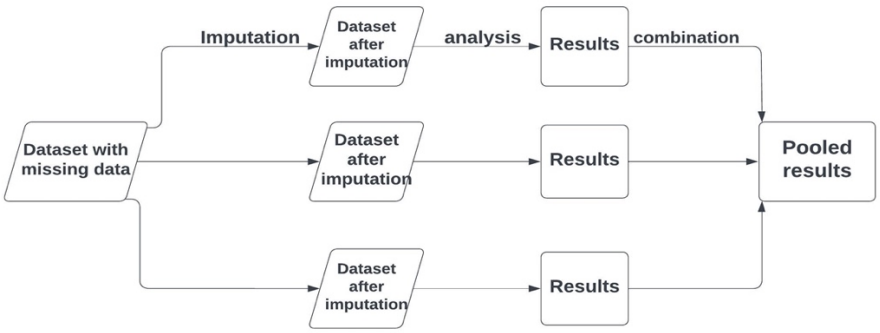


Figure 1. Flowchart showing MI process.

A crucial assumption to consider before using MI regards the missing data mechanism. Missingness is to be either missing completely at random (MCAR) or missing at random (MAR). MCAR means that missingness is completely at random, that is missingness has no systematic pattern and is unrelated to other information. MAR means that missingness is related to only observed values, not unobserved values. When the two assumptions above are not met, missingness is said to be missing not at random (MNAR). The idea behind the assumptions is that missingness is not due to some hidden pattern that could bias the results (Buuren, 2018). As Sigurðsson showed (2022) the missingness in the IS-LFS is indeed not MNAR but systematic related to variables that are observed. For instance, we are aware that specific age groups are less likely to participate in the study and we have information about the age of each individual selected for the study. Thus, the missingness is indeed MAR and an important assumption regarding MI is therefore fulfilled in our situation.

MI has demonstrated its effectiveness as a superior imputation technique. Additionally, considering the collaboration with Statistics Iceland for this project, where their objective was to utilize MI techniques for imputing values in the IS-LFS, we opted to use MI methods into the present study. In Section 3.2 we will go into more technical detail regarding the imputation methods used in the study.

3 Methodology and data

3.1 Participants and data gathering

The present study involved respondents from the IS-LFS. The survey is a telephone survey, where interviewers contact respondents via telephone to conduct interviews. During these interviews, responders are presented with a series of questions from a questionnaire. The sample is chosen with a simple random sample from individuals aged 16 to 74 who are registered residents of Iceland. Data about participants spanning from the years 2003 to 2022 was utilized in the current analysis. Total number of subjects in the research was n = 59.914.

Essentially, the data utilized in the research was a constructed dataset, meaning it was created from various sources for the current research. This dataset included only one variable from the IS-LFS, which was the variable we indented to predict (target variable). The remaining variables in the dataset, referred to as auxiliary variables, consist of administrative data obtained from various other sources available to Statistics Iceland. These variables were used to predict our target variable.

3.1.1 Target variable: IL0

As previously noted, the objective was to estimate individuals labour force status. Thus, the labour force status variable is our target variable. For the rest of the paper, we will refer to this variable as ILO.

ILO is a derived variable calculated using responses from three questions from the IS-LFS. The labour force consists of individuals who are employed or unemployed. The rest, those who do not have a job and are not looking for one are categorized as: ‘‘not part of the labour force“. Those who are not part of the labour force are often students, retired or unable to work (International Labour Organization, 2020). Hence, ILO is a three-level categorical variable with the levels: “employed”, “unemployed” and “not part of the labour force”. The current format of the IS-LFS has been conducted since 2003, thus, we have ILO values from respondents spanning back to that year.

3.1.2 Auxiliary variables (independent variables)

As mentioned earlier, Statistics Iceland holds a significant amount of administrative data regarding all residents in Iceland. When referring to administrative data, we mean data that does not originate from the IS-LFS itself, but from external sources accessible to Statistics Iceland. Importantly, this administrative data is not exclusively tied to individuals selected to participate in the research, but available for all Icelandic residents. The administrative data we used comes from many different sources accessible to Statistics Iceland concerning a variety of factors spanning from age to taxable income.
Importantly, the administrative data does not consist of a single measurement from a single time point; rather, it is accessible over a timespan. In other words, it is for example possible to track an individual’s monthly salary, extending back to the year 2003. This enables the option to link the administrative data to subjects in a dataset at certain times using individual

identifier and date. Hence, like ILO, we have administrative data for respondents back to the year 2003. From this administrative data, auxiliary variables were generated.

3.1.3 Building the dataset

The goal of the data gathering process was to create our dataset containing ILO (target variable) and the corresponding auxiliary variables (*independent variables*). Here we explain the process of creating the dataset in two steps:

1. We gathered data from respondents in the IS-LFS from the year 2003. The survey is a longitudinal study and therefore we had multiple measurements for some respondent in ILO. However, given the magnitude of the data and lack of empirical support for imputation methods to deal with repeated measure design, it was decided to delete multiple rows for subjects and have only one row for each subject in our dataset. For subjects that had more than one record, a single record was selected at random. At this point the dataset contained the following four variables:
   1: "ILO":
   2: "Date-variable": indicating the time of participation
   3: "Haxid ": Individual identifier.
   4: "Haxid2": Functioning as an individual identifier based on the time of participation. Meaning, subjects that participated in the survey more than once were assigned distinct "haxid2"values for each instance of participation.

2. We added the auxiliary variables into the dataset. Auxiliary variables within tables possessing "Haxid2" were connected to our dataset using that particular variable. In instances where tables lacked "Haxid2", the date-variable and "Haxid" served as a foreign key, that is, they were used as a link between other tables containing auxiliary data.

3. At this point we had a complete dataset with ILO-values and values for auxiliary variables at the time subjects responded in the survey, ranging from the year 2003 to 2022.

Table 1 lists all the variables in our complete datasets. The table provides sufficient description for all variables, except for ILO_skrá. ILO_skrá is a variable computed using available auxiliary data. This variable was created prior to the research and was accessible in the Statistics Iceland database. The intention behind creating the variable was to make a single variable that would closely represent The International labour Organization's definition of ILO. The variable is calculated using taxable income and unemployment benefits data. Originally the variable had 4 levels: "employed", "unemployed", "both employed and unemployed" and "Out of the labour force". Due to the limited data in the "unemployed" level, a decision was made to combine the "both employed and unemployed" level into the "unemployed" level. After making this adjustment, ILO_skrá also featured the same levels as ILO. An important point regarding the ILO_skrá variable is that there has not yet been an evaluation of how well ILO_skrá represents ILO.

Table 1. Variables in our Dataset.

	Variable name	Description	Type of variable
Target variable	ILO	Labour force status: 1: *Employed* 2: *Unemployed* 3: *Not part of the labour force*	Nominal
Auxiliary variables	ILO_Skrá	Labour force status: 1: *Employed* 2: *Unemployed* 3: *Not part of the labour force*	Nominal
	Sex	1: *Male* 2: *Female*	Binary
	Age	Age at participation in study	Numeric
	Origin2	Origin status: 1: *Immigrant for less than 10 years* 2: *Immigrant* 3*: Icelandic background*	Nominal
	Education	Current education level: 0: *Non* 1: *Primary* 2: *Vocational education(L-level)* 3: *Vocational education(H-level)* 4: *Certified trades* 5: *Upper/post-secondary* 6: *Short cycle tertiary education* 7: *Bachelor's or equivalent* 8: *MA/MS/PHD*	Nominal
	Working	Current working status according to taxable income: 0: *no* 1: *yes*	Binary
	Unemployed	Indicating if got paid from Unemployment insurance fund or not the same month as selected to participate: 0: *no* 1: *yes*	Binary
	Compensation	Indicating if got paid from social insurance administration the same month as selected to participate: 0: *no* 1: *yes*	Binary

	Pension	Indicating if got paid pension the same month as selected to participate: 0: *no* 1: *yes*	Binary
	Pension 2	Pension in ISK for the month of participation	Numeric
	Disabled	Indicating if disabled or not: 0: *no* 1: *yes*	Binary
	Real estate	Total real estate value in ISK	Numeric
	Salary	Total salary in ISK for the month of participation	Numeric
	Salary +1	Total salary in ISK for 1 month after participation	Numeric
	Salary -1	Total salary in ISK for 1 month previous participation	Numeric
	Independent contractor	independent contractor or not: 0: *no* 1: *yes*	Binary
	Unemployment benefits	Unemployment benefits in ISK for the month of participation	Numeric
	Unemployment benefits +1	Unemployment benefits in ISK for 1 month after participation	Numeric
	Unemployment benefits -1	Unemployment benefits in ISK for 1 month previous participation	Numeric
	Total income	Total income in ISK for the month of participation	Numeric

3.2 Statistical methods

In this section we list and describe all the statistical methods and techniques used in the current research.

3.2.1 Multiple imputation by chained equation (MICE)

MICE was the MI technique chosen to implement in the current research. MICE provides a variety of method settings suitable for different variable types. The ”MICE” package in ”R was used for the imputation (Buuren, 2011).

The process of MICE can be divided into the following four steps:
For simplicity we will give an example with a data set with three variables: age, income, and gender.

- Step 1: Missing values in the three variables are replaced with a “placeholder” which is derived from the available non-missing values in the variable intended to impute. An example would be replacing the missing value with the mean.
- Step 2: The “placeholders” for one of the three variables, age for example, is set back to missing value.
- Step 3: a prediction model is trained, where age is the dependent variable and income and gender the independent variables. Only data where age is not missing is used to train this model.
- Step 4: Next, we replace the missing age values with predictions from the prediction model trained in step 3. Later, when age serves as an independent variable in the prediction for income and gender, we will use both predicted and observed values for age. Thus, age will not have any missing values when employed as an independent variable.

These 4 steps are then repeated for the other two variables income and gender. Completing step 1 to 4 for each of the variables completes one cycle. A specific number of cycles are performed, and the predictions are updated after each cycle. After finishing the specific number of cycles, we have one data set with no missing values. The entire imputation process is then repeated for as many datasets we want (Azur et al., 2011). In the current research, m=5 datasets were generated, as it is recommended for the initial imputation phase (Buuren, 2018).

3.2.2 CART (classification and regression trees):

Classification and regression trees or CART, are a class of machine learning algorithms. CART is recognized for its simplicity, interpretability, and effectiveness. In essence, CART is an algorithm that creates decision trees, that divide the data using a selected criterion. The algorithm starts by selecting the variable that can best split the data in two subsets. Commonly used criterion for classification is Gini impurity. The equation for calculating Gini impurity is given by (2).

$$\text{Gini}\,(p) = 1 - \sum_{i=1}^{J} p_i^2 \qquad 2$$

Here, Gini (p) represents the Gini impurity for a node, J is the number of classes, and p_i is the probability of randomly selecting an instance of class i from the node.

Essentially, Gini impurity measures the impurity of a dataset. A high Gini impurity suggests a diversity or impurity among the datapoints, indicating a mixture of different classes within the dataset. Within a CART algorithm, the goal is to split the data in a way that reduces the Gini impurity, leading to more accurate classifications.
In the case of regression, mean squared error (MSE) is often used, its calculation is defined by (3).

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \textbf{3}$$

Here, n represents the number of observations, y_i is the actual value for observation i, and $\hat{y}_i$ is the predicted value for observation i.
After the initial split, the algorithm proceeds to split the resulting subsets into additional subsets, repeating this process until a predefined stopping criterion is met (Breiman et al. 1984). Finally, CART is tuned to find the optimal size. This is done using cross-validation, dividing the data for both training and testing. The goal is to find the right tree size, avoiding overfitting to training data and ensuring it works well with new, unseen data.
Figure 2 illustrates how CART operates with a decision tree diagram. Decision trees have two types of notes: The decision node and the leaf node. Decision nodes serve as points where decisions are made and branch into multiple paths, while leaf nodes represent the outcomes resulting from those decisions and do not branch out any further. Thus, the leaf nodes represent the prediction made.

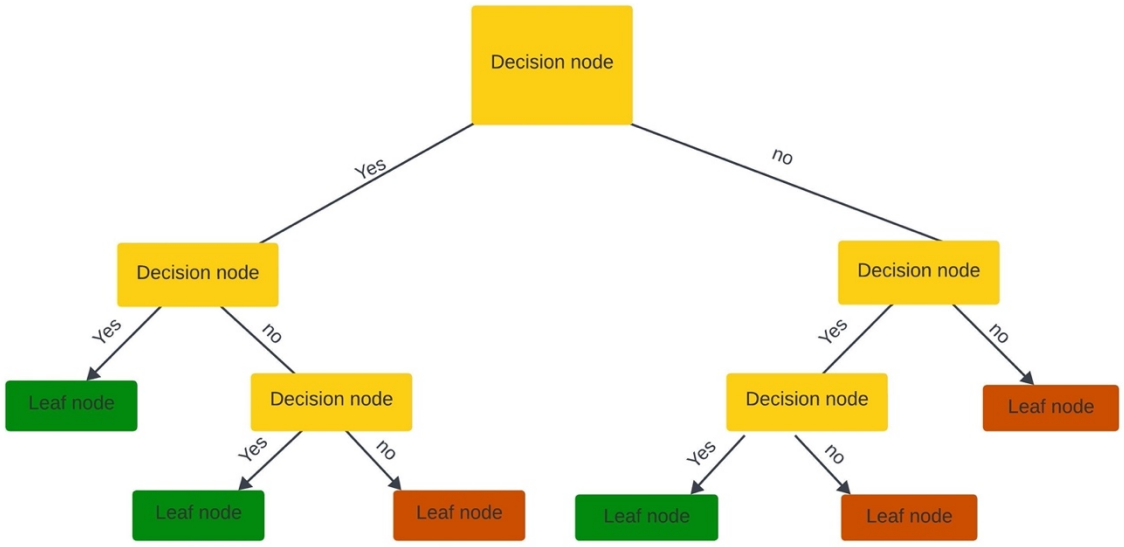


Figure 2. A Decision tree diagram.

3.2.3 Random forest

Random forest is a popular machine learning algorithm introduced by Leo Breiman in 2001. Random forest can be used both for classification and regression tasks. Thus, applicable in many situations. The key idea behind Random forest is to combine predictions from multiple decision trees to create a more accurate prediction.

1) The first step is to create a bootstrapped dataset from the training set. A dataset created through bootstrapping involves resampling data by randomly choosing samples with replacement from the original dataset.
2) The second step is to create a decision tree with the bootstrapped dataset. At each note of the decision tree, a random subset of variables is considered for splitting. This ensures diversity among the trees and prevents the model from depending heavily on single variables.
3) Next, we repeat step one and two depending on how many trees you want in your forest.
4) Finally, each decision tree in the forest has its own prediction. In classification the final prediction is determined by a majority vote, whereas in regression, the mean of prediction from all the trees is used.

Furthermore, the model is tuned to find how many variables to use at each split. This is determined using the out of bag error. This metric estimates how well the model is likely to perform on unseen data, by evaluating the model on data points that were not used during its training. In essence, multiple random forests are constructed, each using varying size of subsets of variables for splitting at each node. The out of bag errors are calculated for these forests. By comparing the out of bag errors across the forests, we select the subset size of the forest with the lowest out-of-bag error (Breiman, 2001).

3.2.4 Polytomous regression

Polytomous regression, also known as multinomial regression, is a traditional technique within categorical analysis. It enables modelling association between predictors and non-ordered categorical outcomes with more than two levels (Agresti, 2012). Polytomous regression extends the concept of logistic regression. In logistic regression, the dependent variable is represented by a logarithmic transformation of the odds, which is known as the logit. In Polytomous regression the analysis treats the outcome variable as a pairwise comparison between two categories. For instance, in the context of three level variable (A, B and C), the analysis would compere A vs B and A vs C. , alternatively we could also compare B vs A and B vs C, or C vs A and C vs B (Hua et al, 2021). The equation for multinomial regression is given by (4)

$$Pr(Y = j \mid X) = \frac{e^{\boldsymbol{X}\beta_j}}{\sum_{k=1}^{K} e^{\boldsymbol{X}\beta_k}} \quad \textbf{4}$$

Here, the left-hand side of the equation represents the probability of the response variable Y belonging to category j, given the predictor variables X. K is the total number of categories, βj is the vector of coefficients for category j, and e is the base of the natural logarithm.

3.2.5 Performance assessment and comparison of methods.

To assess performance and compare different methods for the imputation of ILO, we examined the performance metrics accuracy and sensitivity. In the current research specificity was not reported duo to its constantly high values. The primary focus was directed towards enhancing sensitivity.
Accuracy in machine learning is a metric used to measure the performance of a predictive model in classification. Accuracy can be derived from (5). It reports the proportion of correct predictions out of the total predictions made. In sum, accuracy tells us how well the model can make correct predictions.
Sensitivity is a more detailed metric for evaluating the performance of a classification prediction model. Equation (6) shows the basic formula for sensitivity in a situation where we have a binary dependant variable. When we have a binary dependant variable, where values are classified as either positive or negative, sensitivity measures the ability of the model to correctly identify positive instances from the actual positive ones. The idea behind sensitivity can be extended to scenarios with a dependent variable with more than two levels. In this context it focuses on the model's ability to correctly predict instances belonging to a specific class. If we were for example to find the sensitivity in class A out of classes A, B, C the sensitivity formula would extend to (7). Same approach can be applied to compute specificity.

$$accuracy \quad = \frac{\text{Number of Correct Predictions}}{\text{Total Number of predictions}} \tag{5}$$

$$sensitivity = \frac{\text{True Positives}}{\text{True Positives } + \text{False Negatives}} \tag{6}$$

$$sensitivity(Class\ A) = \frac{\text{True Positives for class A}}{\text{True Positives for class A } + \text{False Negatives for Class A}} \tag{7}$$

3.2.6 Variable importance

When predictions were made without utilizing MICE, interpretability in understanding how predictions were achieved became feasible. To enhance interpretability of how predictions were made, we looked at variable importance for the Random forest model and created a decision tree diagram for the decision tree. Variable importance in Random forest models can be assessed using both mean decrease in Gini index metric (MDG) and mean decrease in accuracy (MDA). MDG index indicates how each variable influences the homogeneity of the nodes and leaves in the forest. In other words, if the variable is important, it helps the model distinguish between classes. MDG measures the decrease in the Gini index when a particular variable is used in the model, therefore, the higher the MDG, the more important the variable

is. MDA measures the average decrease in accuracy when a particular dependant variable is removed from the model during the out-of-bag process in Random forest (Han, 2016). For the decision tree we showed a decision tree diagram. A decision tree diagram provides a visual representation of how the decision-making process is structured. The picture is structured like a flowchart that guides the reader to the final decision.

3.2.7 Random under-sampling

Random under-sampling is technique used in the context of imbalanced dataset. Having an imbalanced dataset means that the distribution of levels in variables are not equal. This situation arises, for instance, when we have a factor variable where one level has a significantly higher count of cases compared to the other two levels. In random under-sampling the size of the majority level is reduced intentionally by randomly removing instances from that level. This balances the level distribution, making the count in the majority level closer to the number in the minority level (Ganganwar, 2012).

3.3 Protocol

In this section we will provide clear step-by-step overview of the research process. We break the process down into four distinct parts, which correspond to the sub-sections. The first part describes the process when testing and compering different methods within MICE to predict for ILO. Next, we describe the process where we tested and compared different prediction model's ability to predict values for ILO. Thirdly we analysed the distribution of correct predictions. Lastly, we will explain our simple attempt to create a simulation mimicking the process of predicting ILO as if it were done today. To achieve this, we generated a test dataset that mirrors the group for which predictions would be conducted today.

3.3.1 Prediction with MICE

In this section, we present a straightforward breakdown of the process of using different MICE methods to predict for ILO. Figure 3 gives an additional visual description of the process.

- First, we created our dataset, containing values for both ILO and auxiliary variables, from respondents in the IS-LFS. The creation of the dataset is described in more detail in the Section 3.1.3.
- Our dataset had *n=59914* subjects, with no missing values in ILO or the auxiliary variables.
- The next step was to generate missing values in ILO. This was done using the "missMethods" package in R (Rockel, 2020). The missing completely at random (MCAR) mechanism was selected, and both 20% and 40% of the data was made missing in ILO.
- At this step we have two datasets with missing values in ILO, one with 20% and the other with 40% missing.
- Next we used the "MICE" package in R (Buuren, 2011) to impute the missing values in ILO in our dataset. The chosen methods within MICE where cart, Polytomous regression and Random forest. M was set to 5, meaning that five complete data sets were created after imputation.

- Finally, performance assessment of prediction accuracy was done. Knowing both the predicted and the actual value of ILO for each subject made the calculation of our performance metrics possible.
- To obtain the performance metrics (i.e. accuracy and sensitivity) for the this approach, a confusion matrix was generated from each imputed dataset. With a total of *m=5* datasets, we created five individual confusion matrices. These five matrices where then used to calculate a single mean matrix. Using this mean matrix, we calculated both accuracy and sensitivity for each MICE method by the proportion of missing data generated.

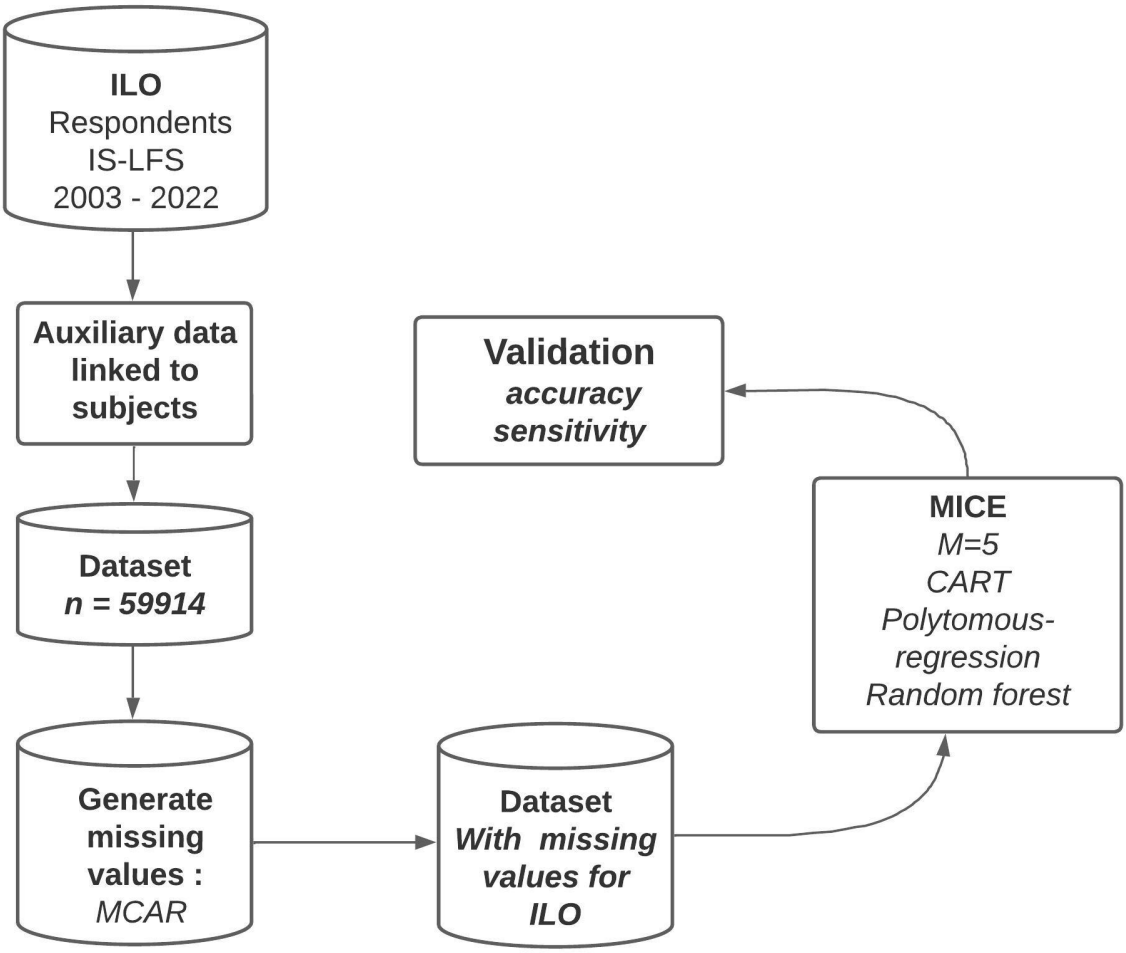


Figure 3. Flowchart showing process of predicting ILO with MICE.

3.3.2 Prediction with standard prediction models.

When assessing the ability to predict ILO with standard prediction models (without using MICE), we adopted a different strategy. We now could approach the problem from the standpoint of a standard prediction task. Thus, the dataset was divided into a test and training set and how well we could predict ILO was evaluated on the test-set. This enabled us to correct for the class imbalance that we discovered in the ILO variable. That is, our data had way more individuals classified as "working" compared to "not working" or ''not part of the labour force ". This class imbalance was adjusted using random under-sampling. In addition, using this approach allowed tuning of the prediction models to improve performance. For the Random forest model, tuning was used to determine the optimal number of variables to use at each split. Similarly, for the CART model, tuning was performed to identify the optimal size of the CART.
Here, we present a straightforward breakdown of the process of using standard prediction models to predict ILO. Figure 4 gives an additional visual description of the process:

- We created our dataset, containing data from both the ILO and auxiliary variables from respondents in the IS-LFS since 2003 to 2022. The creation of the dataset is described in more detail in the Section 3.1.3.
- Our dataset had *n=59914* subjects, with no missing values in ILO or the auxiliary variables.
- The Dataset was split into test and training data, where participants were randomly assigned to either the test or training data set.
- Class imbalance in the ILO variable within the training-set was adjusted using random under-sampling.
- Next, we built a prediction model using the train-data.
- Prediction model is tuned to find the optimal settings for best performance.
- The model's ability to predict for ILO on unseen data (the test-data) was then assessed by analysing the evaluation metrics accuracy and sensitivity.
- Lastly, to enhance the interpretability of how the predictions were made, we looked at variable importance for the Random forest model and created a decision tree diagram for the decision tree.

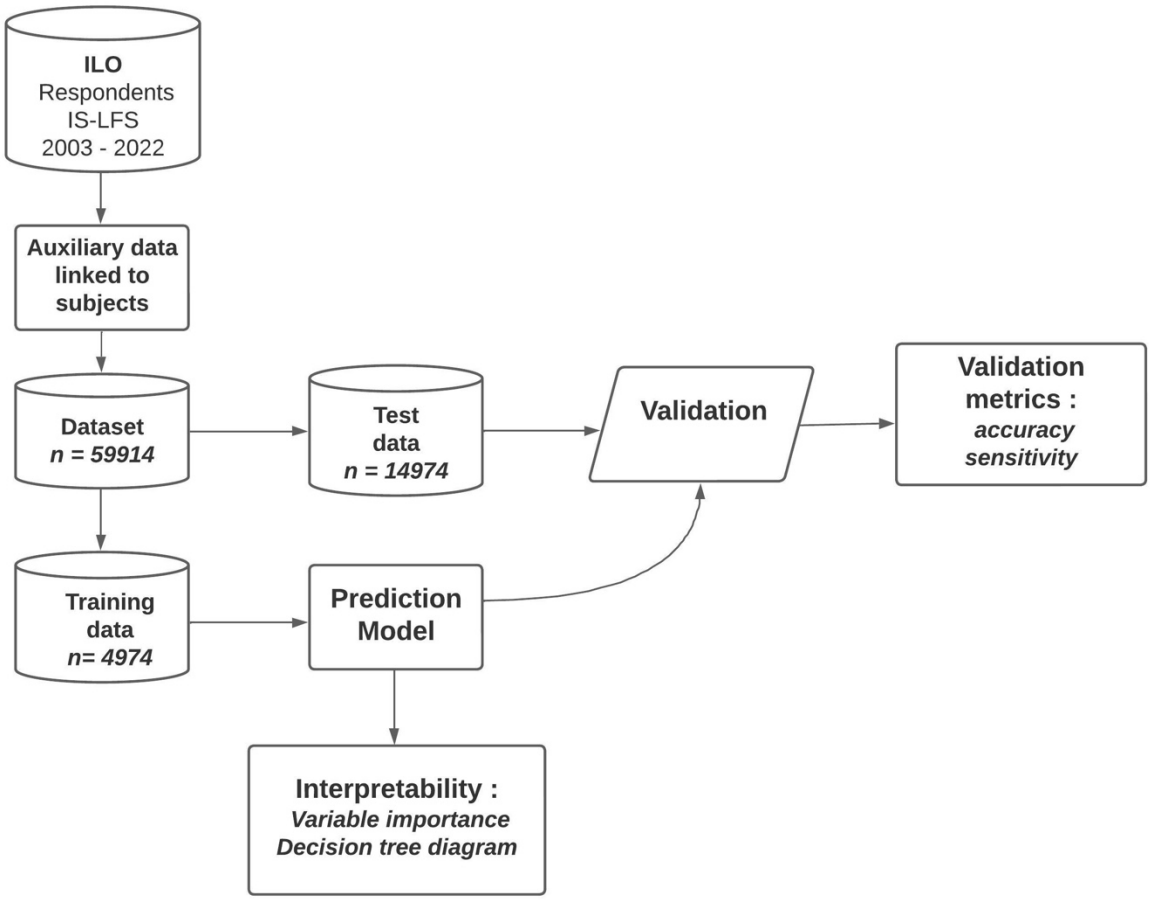


Figure 4. Flowchart showing process of predicting ILO with standard prediction models.

3.3.3 Distribution of correct predictions

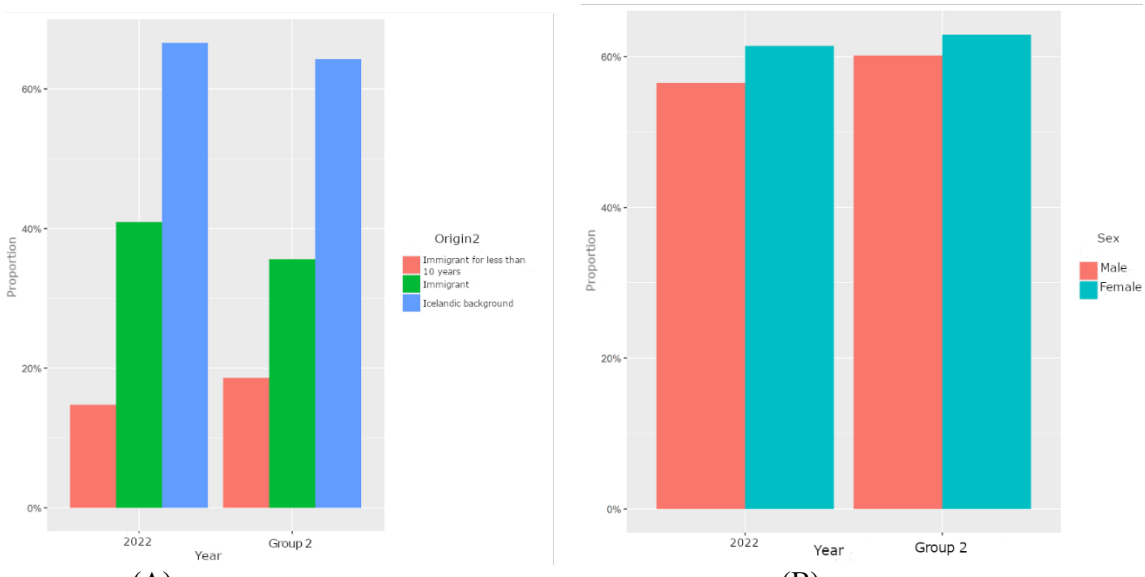
If we intend to employ a method to predict for ILO it is crucial that the model's prediction performance remains consistent across various groups within the sample. Further, it is especially important that the model can predict well for the groups that have higher likelihood of not taking part in the survey. For this reason, we analysed the distribution of correct predictions by selected variables. The variables we chose were education, origin and age, as these variables have been shown to be related to non-response in the IS-LFS (Sigurðsson, 2022). To analyse the distribution of corrected predictions we used predictions from the most effective method. Simple bar plots were made for each variable, and the proportion of correct predictions by levels in each variable were analysed.

3.3.4 Approach to usage today, creation of Group 2.

Finally, to further evaluate our most effective method, we introduced some adjustments considering both the results in Section 4.3 and if the method would be employed today. For this we adopted a simple approach in an attempt to simulate the scenario if the most effective prediction method would be employed in present time. This essentially involved the creation of a group (test-set) designed to mimic the today's non-responder group. This group will be called Group 2. Finally, predictions for Group 2 were made using the most effective method.

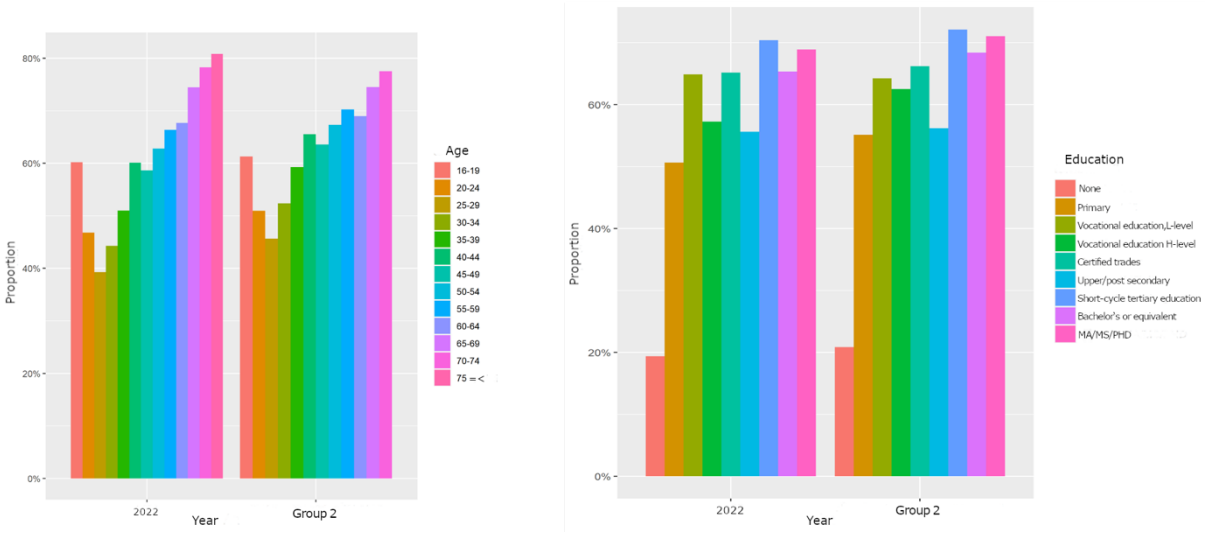
For this task, we used data from the year 2022 to represent present time and used the years 2011-2012 to create Group 2. We chose these years because of a high response rate at that time. Based on results from the distribution of correct predictions in Section 4.3 Group 2 only consisted of individuals 18 years and older. Further to approach today's scenario, we attempted to simulate the systematic difference between responders and non-responders as seen today. Here, we provide a clear breakdown of the process:

- First, we trained today's response model using data from 2022. The model is shown equation (1).
- Next, we predicted response propensity for subjects in the responder's group from the years 2011 to 2012 (Group 2). According to their response propensity subjects were labelled responder or non-responder. That is, if a subject had the response propensity of 0.7, it had a 70% probability of being labelled responder and 30% probability being labelled non-responder.
- At this step Group 2 should have similar systematic difference between responders and non-responders as seen in the year 2022.
- To evaluate the above assumption, we looked at Figure 5 and 6, were we compared the systematic difference between responders and non-responders in the year 2022, with our data in Group 2 after the simulation.
- Figures 5 and 6, give evidence that ours attempt to mimic the systematic difference between responders and non-responders in the year 2022, on Group 2 was successful, that is we see similar patterns in the response rate by our chosen variables, for the year 2022 and Group 2.
- Next, we removed the subject labelled as responders in Group 2 leaving us with a group that should resemble today's non-response group.
- Finally, we predicted ILO on Group 2 using the currently most effective method.



(A) (B)

Figure 5. Proportion of people that respond. (A) by origin. (B) by sex. In both graphs, left columns show values for the year 2022, and the right Group 2.



(A) (B)

Figure 6. Proportion of people that respond. (A) by age. (B) by education. In both graphs left columns show numbers for the year 2022, and the right Group 2.

4 Results

The goal of the research was to assess weather prediction is possible for the ILO variable originating from the IS-LFS, additionally we aimed to compere different methods in terms of performance of prediction. Sections here will follow the same order as the protocol Section 3.3 to ensure simplicity.

4.1 Comparison of MICE methods

In this section, an evaluation and exploration are carried out to identify the best imputation method within the MICE package. This is done by assessing prediction accuracy in predicting ILO.

Table 2 displays the evaluation metrics for all the MICE methods by the proportion of missingness. When assessing metrics, it was observed that the accuracy for predicting ILO remained consistently high across all methods within MICE, regardless of whether the missingness was set at 20% or 40%.
The imputation method within MICE with the highest accuracy showed to be Random forest with an accuracy of 0.87. Following closely was the CART method with an accuracy of 0.86, while Polytomous regression exhibited the lowest accuracy at 0.84. Notably, the accuracy for all methods remained relatively consistent, regardless of the percentage of missingness. However, when examining the sensitivity values for each class in ILO, notably high values were observed in one level while being lower in the others. Specifically, the employed level exhibited a sensitivity around 0.9. Conversely, sensitivity was quite low in the other two levels. In the "Not part of the labour force "level, the sensitivity ranged from 0.67 to 0.71, and was even lower in the "Unemployed "level or between 0.07 and 0.15.

Table 2. Accuracy and sensitivity by MICE-method and proportion of missingness

			ILO levels		
			Not part of the labour force	Unemployed	Employed
MICE method	Missingness	Accuracy	Sensitivity	Sensitivity	Sensitivity
CART	20%	0.86	0.67	0.15	0.93
	40%	0.86	0.70	0.14	0.92
Polytomous-regression	20%	0.83	0.63	0.11	0.91
	40%	0.84	0.65	0.11	0.92
Random forest	20%	0.87	0.71	0.07	0.92
	40%	0.86	0.69	0.11	0.91

4.2 Comparison of standard prediction model methods.

In this section, an analysis and search for the most accurate prediction model for ILO forecasting is conducted. The distinction here from Section 4.1 is that prediction is not carried out using MICE but instead uses the standard approach of constructing prediction models, enabling modifications as outlined in Section 3.3.2.

When metrics for the prediction model approach were examined, similar accuracy values were observed as those obtained using MICE. That is, accuracy was quite high for our decision tree and our Random forest model, or between 0.85 to 0.87. Sensitivity in the employed level was also in line with MICE, around 0.9. However, in contrast to the results from using MICE, our sensitivity values in the "Not part of the labour force" and "Unemployed" levels were not as low. In the "Not part of the labour force" level sensitivity ranged from 0.68 to 0.72. In the "Unemployed" level sensitivity ranged from 0.73 to 0.78, which is significantly higher when compared to the MICE methods. In the "Not part of the labour force" level sensitivity was similar to results from using MICE, or 0.68 and 0.72. From these metrics it was observed that the Random forest 1 model had the best overall performance when predicting values for ILO.

Table 3. Accuracy and sensitivity by prediction model.

		ILO levels		
		Not part of the labour force	Unemployed	Employed
Prediction model	Accuracy	Sensitivity	Sensitivity	Sensitivity
CART	0.87	0.68	0.73	0.93
Random forest 1	0.85	0.72	0.78	0.89

4.2.1 Variable Importance and decision tree diagram

In this section, we will present the variable importance plots for Random forest 1 and the decision tree for the CART model discussed in Section 4.2.
The variable importance plot by MDG is shown in Figure 7. The plot shows the 8 most important variables according to MDG, in decreasing order. The ILO_skrá variable was the most important variable according to MDG. As previously mentioned, ILO_skrá is a derived variable for ILO representation, calculated from taxable income and unemployment-benefits data. The second most important variable according to MDG was the salary for the month of participation in the survey. The third most important variable was the salary for the next month.

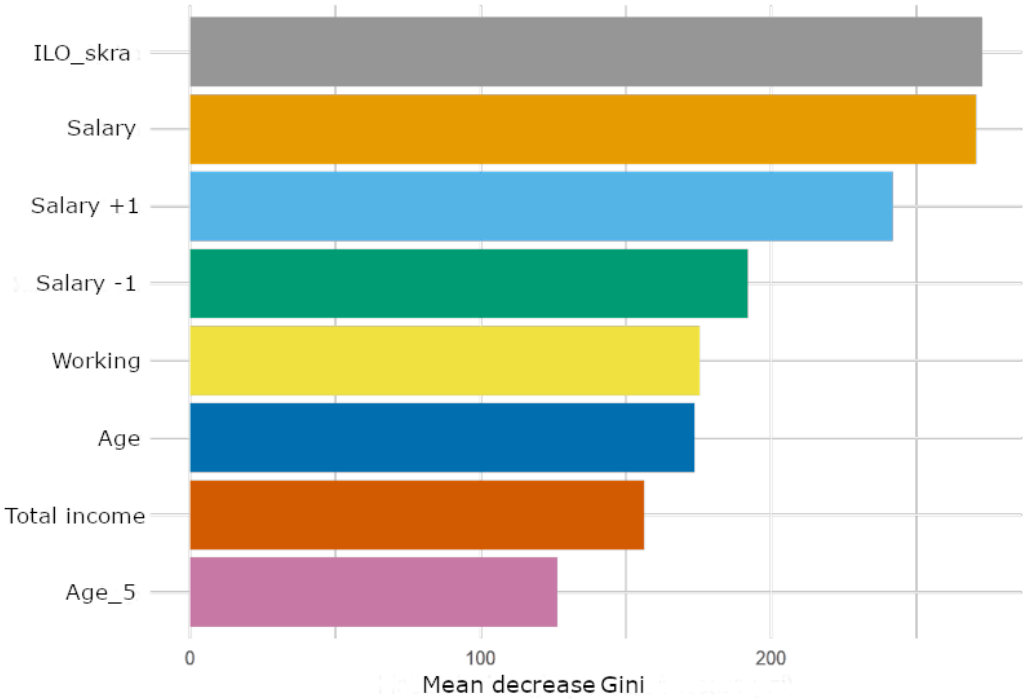


Figure 7. Variable importance plot by MDG.

Figure 8 shows the variable importance in the Random forest 1 model according to MDA. Salary for the next month was the most important variable, followed by age and then salary for the previous month. Interestingly, ILO_skrá did not even rank as the eight most important variable based on mean decrease in accuracy.

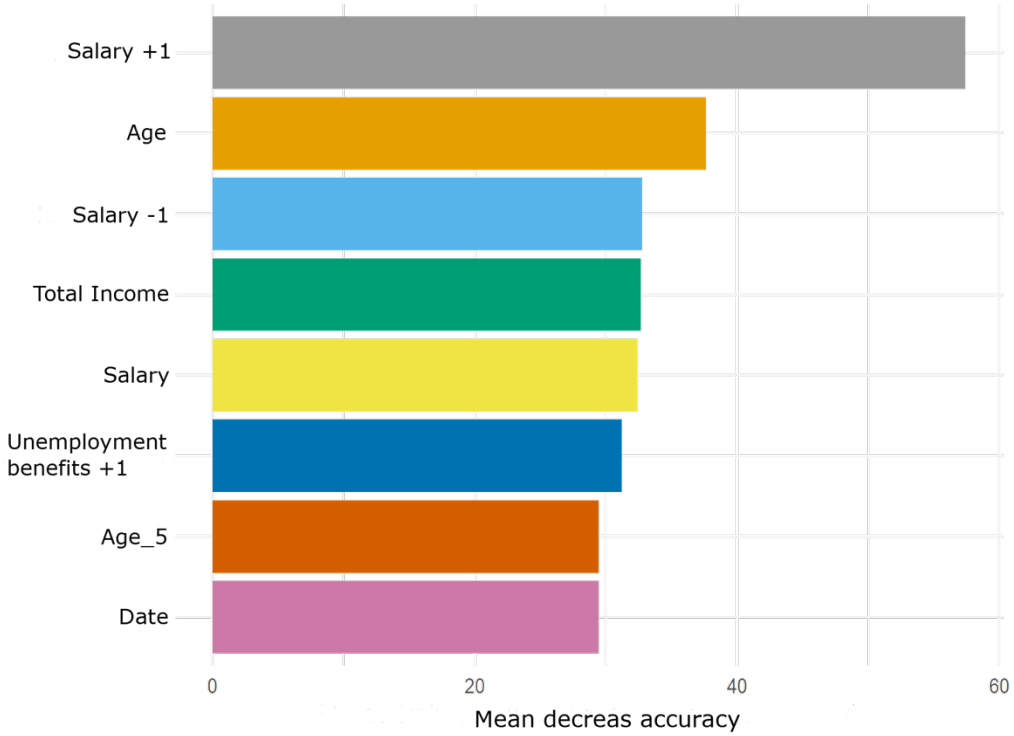


Figure 8. Variable importance plot by MDA.

Figure 9 precisely illustrates the process how the CART model made its classifications with a decision tree diagram. As we can see ILO_skrá is an important variable in the model and is heavily relied upon for making predictions. The CART model can be broking down in the following steps:

1. Individuals who take the value employed in ILO_skrá are categorized as such.
2. Those who take the value unemployed in ILO_skrá are categorized "unemployed".
3. Now the remaining group consists of individuals labelled "not part of the labour force" in ILO_skrá. This subgroup undergoes the further classification:
    a. Individuals who received pension the same month they were selected to participate in the study are categorized as "not part of the labour force".
    b. Individuals who did not receive pension the month they were selected to participate but receive pension for more than or equal to 255.000 ISK the month of participation were also classified "not part of the labour force".
    c. Lastly, individuals who did not receive pension for more than or equal to 255.000 ISK the month of participation and fall outside the age range of 40 to 74 years old were classified unemployed.

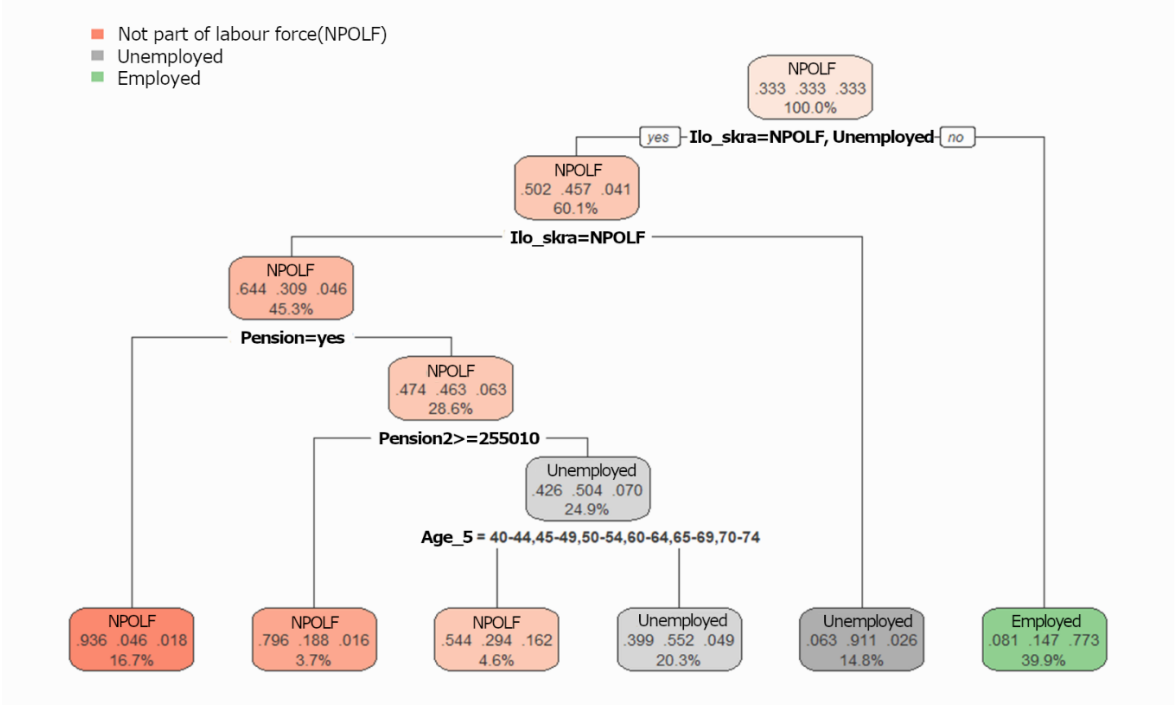


Figure 9. CART model decision tree. If condition = no than proceed to right. If condition = yes than proceed to left.

4.3 Distribution of correct predictions

In the following section, the distribution of correct predictions by the Random forest 1 model will be examined.
First, the proportion of correct predictions by level of education was analyzed. As Figure 10 shows, the proportion of correct predictions was quite similar for the different levels of

education. However, we see that the model performs worst when predicting ILO for individuals with primary school level of education.

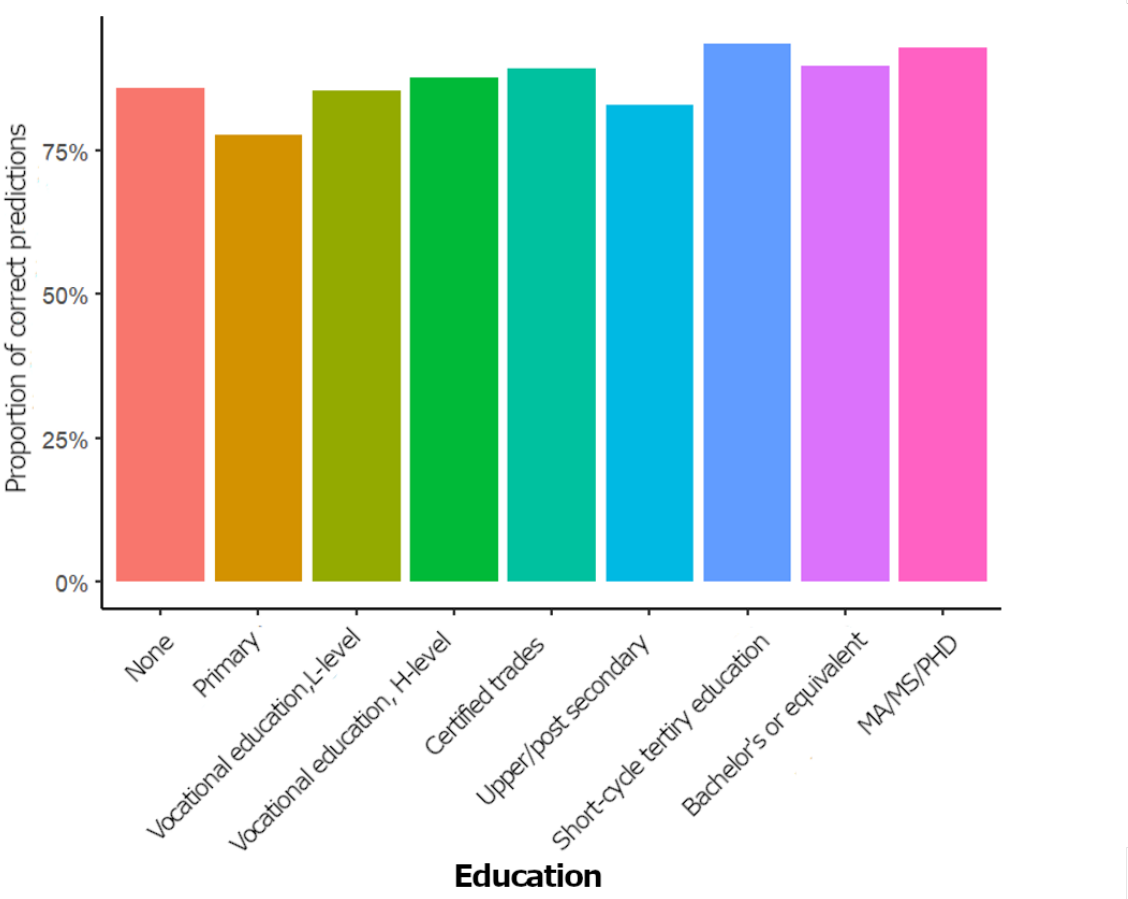


Figure 10. The proportion of correct predictions by education.

Next, the proportion of correct predictions by origin2 was analyzed. No major differences in the proportion of correct prediction by origin was observed, indicating that the model worked similarly well for people with Icelandic background and immigrants.

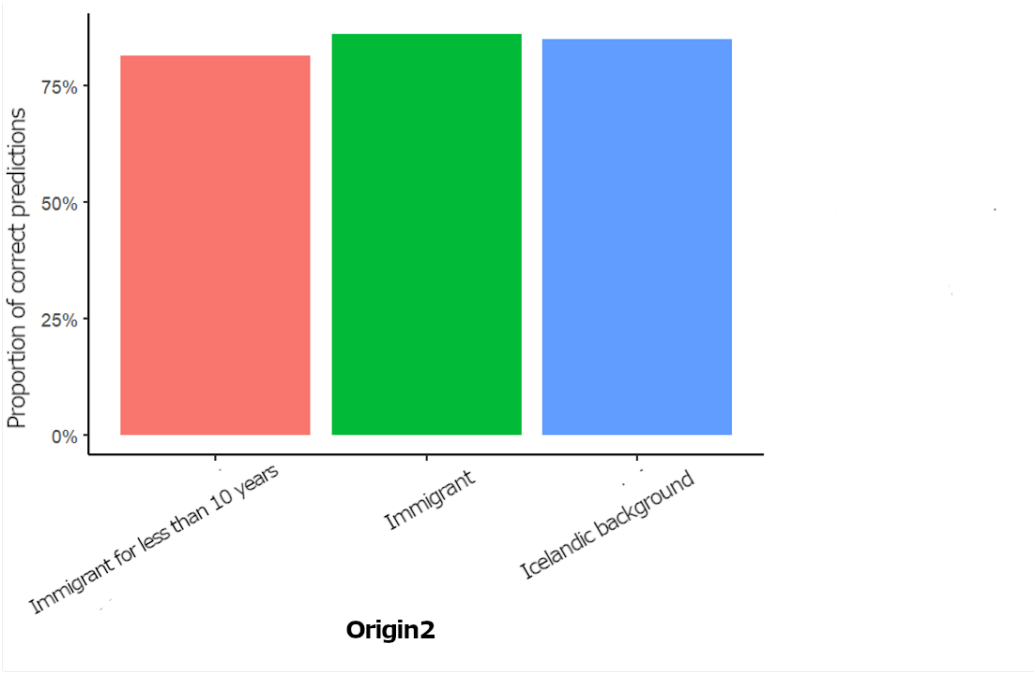


Figure 11. The proportion of correct predictions by origin.

Next, proportion of correct predictions by age was examined. The model's performance varied among different age groups. The model's accuracy remained more consistent for the older age groups from 35-75 but started to decline for individuals younger than 35. Individuals between the ages of 20 and 75 were not really a concern, as the proportion of correct responses within that age group was considered sufficiently high. However, the age group 16 to 19 raised concerns, having only around 50% of correct prediction rate. Hence, we analyzed the proportion of correct prediction in the age group further.

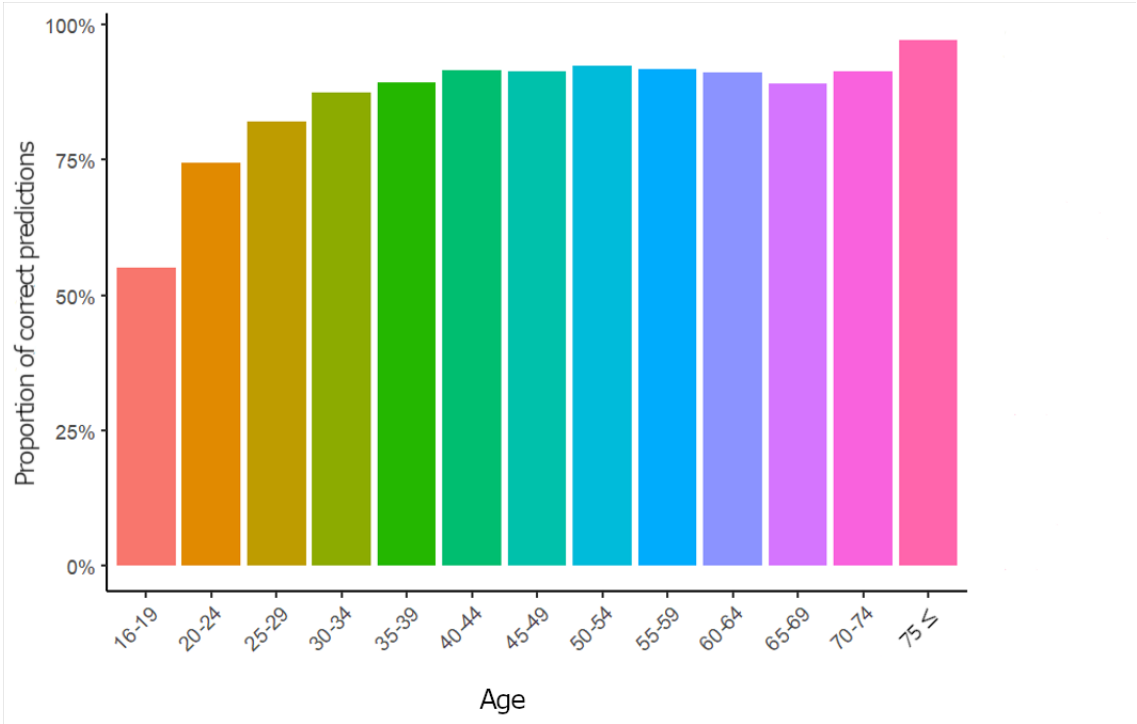


Figure 12. The proportion of correct prediction by age.

When examining the distribution of correct predictions for the ages 16 to 19, as shown in Table 4, we observed that the proportion of correct predictions was notably low for both 16- and 17-year-olds. Only 41.3% of 16 years old were given right prediction, and 58.5% of the 17 years old.

Table 4. The proportion of correct prediction for 16 to 19 years old.

Age	Proportion of correct predictions
16	0.413
17	0.585
18	0.647
19	0.657

4.4 Comparison of test-groups

Here, a comparison will be made between the current best results achieved using the Random forest 1 model in Section 4.2 (Original Group), and the results obtained using the same model for Group 2 described in Section 3.3.4. This group consists of individuals aged 18 years and older, and it is designed to resemble the non-response group observed today, that is certain groups of people were more likely to belong the this test-set. Essentially, we are comparing predictions made by the Random forest 1 model on two different test groups. When comparing results obtained when predicting ILO on the new group the accuracy increased from .85 to .89. The sensitivity also improved for all the levels in ILO. Thus, we have an overall better performance when making predictions for Group 2.

Table 5. Prediction made by Random forest 1 on different groups.

		ILO levels		
		Not part of the labour force	Unemployed	Employed
	Accuracy	Sensitivity	Sensitivity	Sensitivity
Original Group	0.85	0.72	0.78	0.89
Group 2	0.89	0.75	0.79	0.92

5 Discussion

The research aim was to assess the possibility of using administrative data as predictor variables for imputation in the ILO variable from the IS-LFS, specifically focusing on prediction accuracy. Additionally, we aimed to compere different imputation methods in terms of prediction performance.
Given the high prediction accuracy, we can conclude that it is possible to impute peoples working force status (ILO) using administrative data accessible to Statistics Iceland. This is the key result from the study. In more detail Random forest model (Random forest 1) proved to be the most successful among all the methods. The model's performance did not seem to vary significantly between groups. We can therefore have confidence that the model can make accurate predictions for ILO, for most individuals selected to participate in the IS-LFS. Importantly, the model seemed to predict well for groups less likely to participate in the survey, such as individuals within the age range of 21-35. Also, performance did not vary within the variable's origin and sex, which have also showed to be factors affecting participation. However, we did observe lower accuracy for individuals under the age of 18. This is likely to the fact that today individuals younger than 18-year-old cannot receive unemployment benefits. Therefore, we conclude that predicting ILO for individuals younger than 18 is not suitable.
Another critical discovery from the study is that it is not effective to simply apply MI with MICE by directly inputting the data, as is commonly done, without prior data processing.
In the MICE methods we saw an accuracy paradox. That is, when looking only at the accuracy, all the methods within MICE seemed to perform well. However, inspecting the sensitivity gave another picture. Among all the MICE methods, predictions were not accurate when attempting to predict ILO for individuals who were not part of the labour force or unemployed individuals. When further inspected this issue seemed to be most likely due to the imbalanced distribution of ILO levels in the data. In simple terms, the dataset had a much higher number of employed individuals making the MICE methods better at predicting for employed individuals and worse in the other two levels. Due to no previous examples of balancing data before using MICE that possibility was ignored. However, we did random under sampling before training both the CART and the Random forest model. Furthermore, as the auxiliary variables consistently have complete data and ILO is the only variable with missing values, the chained equations component within the MICE becomes useless. Utilizing MICE under these circumstances essentially boils down to employing the prediction method iteratively. That is, when we use Random forest within MICE it become the same as running a standard Random forest multiple times. Therefore, the primary distinction between our Random forest with MICE and the Random forest from the standard prediction model method lies in the pre-processing steps of, level balancing before training the standard Random forest model and tuning the model. Thus, the observed differences in sensitivity results are likely attributable to these pre-processing steps. In addition to this drawback, the MICE method lacked interpretability, essentially functioning as a black-box. If data manipulation like imputation were to be used on official statistics data, having a clear description of the process can be important. In comparison to the MICE methods the standard prediction models outperformed in this aspect by providing a decision tree diagram and variable importance plots.

5.1 Further research

As stated in the introduction, this study is an initial step in exploring the possibility of imputing ILO. Further research is necessary to gain a more thorough understanding and make informed decisions regarding the matter. For further research on imputation of ILO the following aspects should therefore be considered:

- Consider the impact of an imbalanced distribution of levels within variables intended for imputation. As demonstrated in the case of ILO, just as standard prediction methods, MI methods like MICE are influenced by class imbalance. Future research should proceed with caution in this regard. Although it is not a common practice, one potential approach could be to balance levels before utilizing imputation methods like MICE.
- As individuals younger than 18-year-old cannot receive unemployment benefits, it would likely be more suitable for future research to identify a distinct imputation method for this age group. In this method, the ILO variable would be constrained to the categories "employed" or "not part of the labour force.".
- An important role of imputation is to preserve the relationships between variables in a dataset. One recommended approach for evaluating MI methods involves calculating estimate measures like β. Subsequently, simulate missing data, perform imputation, and recompute β again to assess the proximity of the estimate calculated with imputed values to the original (Buuren, 2018). However, because our study focused on prediction accuracy, this aspect was not investigated. Therefore, for future research, it would be crucial to explore various imputation methods in this regard to identify the most suitable one for the IS-LFS.
- Like stated before, Statistics Iceland holds a significant amount of information regarding residents in Iceland. The process of navigating through the available data and finding potential useful variables for predicting ILO was a time-consuming task itself. Moreover, the data was only accessible within a specific timeframe, making it impossible to ensure the inclusion of all available and useful variables. For example, information about whether subjects are presently enrolled in educational program was not included. In future research, including this information could notably improve the accuracy of predictions. This crucial data is accessible to Statistics Iceland and could act as a strong determinant to differentiate those who are unemployed or not part of the labour force. As mentioned earlier, a significant portion of individuals outside the labour force are often students. Thus, incorporating a binary variable with levels such as "currently a student" or "not" is likely to enhance predictions for the ILO variable. Thus, including more predictor variables is possible for further research.

Although the specific implementation of how imputation would be used is not yet clear and requires further decision making, we have demonstrated its possibility for imputation of ILO. Given that we demonstrated the suitability of the imputation of ILO, there are potential benefits. For example, this opens the possibility to impute ILO values for individuals who were not reachable to take part in the survey. This could result in larger sample size and a more diverse group with values in ILO, thus enabling better stratification of estimates. For example, results for a particular group, can be more effectively stratified or broken down by characteristics such as age, sex, or other factors after imputation. In addition, this could help

counteract the systematic difference between the responders and non-responders by imputing values for groups with particularly low response rates, potentially reducing the variance in survey estimates. The possibility of imputation can also be thought of in the context of response-burden. We observe a declining response rate over time and considering today's fast-paced life and smartphones ability to quickly identify calls, people may choose not to answer. Perhaps there's an opportunity to avoid requiring individuals to respond to questions for ILO classification. Only Administrative data might be used to classify individuals based on the ILO variable. This approach could make the questionnaire shorter and ease the response burden. However, it is important to carefully consider the EU-LFS guidelines in this regard. In summary, knowing the imputation of ILO with administrative variables is possible, gives Statistics Iceland additional tools in its arsenal.

References

Alanya, A., Wolf, C., & Sotto, C. (2015). Comparing multiple imputation and propensity-score weighting in unit-nonresponse adjustments: a simulation study. *Public Opinion Quarterly*, *79*(3), 635-661.

Agresti, A. (2012). *Categorical Data Analysis* (Vol. 792). John Wiley & Sons.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, *600*(7890), 695–700. https://doi.org/10.1038/s41586-021-04198-4

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth Publishing.

Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). CRC Press.

Cheng Hua, Dr. Y.-J. C. (2021). Companion to BER 642: Advanced regression methods. Retrieved September 19, 2023 from https://bookdown.org/chua/ber642_advanced_regression/

De Heer, W., & De Leeuw, E. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. *Survey nonresponse*, *41*, 41-54

Eurostat. (2018). Reference Metadata in ESS Standard for Quality Reports Structure (ESQRS). European Commission. Retrieved September 19, 2023, from https://ec.europa.eu/eurostat/cache/metadata/EN/employ_esqrs.htm

Eurostat. (2022). Quality report of the European Union Labour Force Survey 2020. European Commission. Retrieved March 30, 2022, from https://ec.europa.eu/eurostat/en/web/products-statistical-reports/-/ks-ft-22-003

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, *2*(4), 42-47.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646-675.

Grossmann, W. (2020, March 25). Non-response. CROS - European Commission. https://cros-legacy.ec.europa.eu/content/non-response_en

Han, H., Guo, X., & Yu, H. (2016, August). Variable selection using mean decrease accuracy and mean decrease Gini based on random forest. In *2016 7th IEEE International Conference on Software Engineering and Service Science* (ICSESS) (pp. 219-224). IEEE.

Holt, D., & Elliott, D. (1991). Methods of weighting for unit non-response. *The Statistician*, *40*(3), 333-342. https://doi.org/10.2307/2348286

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*, *8*(2), 183.

Luiten, A., Hox, J., & de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. *Journal of Official Statistics*, *36*(3), 469-487.

Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple imputation: A flexible tool for handling missing data. *JAMA*, *314*(18), 1966-1967.

Little, R. J. A., Lewitzky, S., Heeringa, S., Lepkowski, J., & Kessler, R. C. (1997). Assessment of weighting methodology for the National Comorbidity Survey. *American journal of epidemiology*, *146*(5), 439-449.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, Big Data Paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, *12*(2). https://doi.org/10.1214/18-aoas1161sf

Msaouel, P. (2022). The big data paradox in clinical practice. *Cancer Investigation*, *40*(7), 567–576. https://doi.org/10.1080/07357907.2022.2084621

Rässler, S., & Schnell, R. (2004). Multiple imputation for unit-nonresponse versus weighting including a comparison with a nonresponse follow-up study (No. 65/2004). Diskussionspapier.

Shao, J., & Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, *91*(435), 1278-1288.

Statistics Canada. (2022, September 25). Imputation. Statistics Canada. Retrieved September 20, 2023, from https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch3/imputation/5214784-eng.htm

Statistics Iceland. (n.d.). Labour-market. Statistics Iceland. Retrieved from https://www.statice.is/statistics/society/labour-market/

Sigurðsson, Ó. M. (2022, February). Saga af brottfalli Vinnumarkaðsrannsókn Hagstofu Íslands (2003 til 2020).

The Office for National Statistics. (2021, November 2). Labour Force Survey: alternative imputation during the coronavirus (COVID-19) pandemic. Office for National Statistics. Retrieved November 2, 2021, from https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemploye

etypes/methodologies/labourforcesurveyalternativeimputationduringthecoronaviruscovid19pandemic

Tobias Rockel. (2020). missMethods'(0.4.0). Retrieved from https://cran.r-project.org/web/packages/missMethods/missMethods.pdf

U.S Census Bureau. (2021). National Survey of Children's Health Guide to Multiply Imputed Data Analysis. www2.census.gov/programs-surveys/nsch/technical-documentation/methodology/NSCH-Analysis-with-Imputed-Data-Guide.pdf

van Buuren S, Groothuis-Oudshoorn K. (2011). mice: Multivariate Imputation by Chained Equations in R(3.16.0). Retrieved from https://cran.rproject.org/web/packages/mice/index.html

Perneger, T. V., & Burnand, B. (2005). A simple imputation algorithm reduced missing data in SF-12 Health Surveys. *Journal of Clinical Epidemiology*, *58*(2), 142–149. https://doi.org/10.1016/j.jclinepi.2004.06.005

Peytchev, A. (2012). Multiple imputation for unit nonresponse and measurement error. *Public Opinion Quarterly*, *76*(2), 214–237. https://doi.org/10.1093/poq/nfr065