Risk Factors of Becoming NEET Youth in Iran: A Machine Learning Approach

Hamed Rahmani, Wim Groot

Abstract**—**The term "youth not in employment, education or training (NEET)" refers to a combination of youth unemployment and school dropout. This study investigates the variables that increase the risk of becoming NEET in Iran. A selection bias-adjusted Probit model was employed using machine learning to identify these risk factors. We used cross-sectional data obtained from the Statistical Center of Iran and the Ministry of Cooperatives Labor and Social Welfare that are taken from the labor force survey conducted in the spring of 2021. We look at years of education, work experience, housework, the number of children under the age of 6 years in the home, family education, birthplace, and the amount of land owned by households. Results show that hours spent performing domestic chores enhance the likelihood of youth becoming NEET, and years of education, years of potential work experience decrease the chance of being NEET. The findings also show that female youth born in cities were less likely than those born in rural regions to become NEET.

Keywords**—**NEET youth, probit, CART, machine learning, unemployment.

I. Introduction

CONCERNS about the increase in the number of unemployed people, especially among youths with an academic degree, is high as its negative impact on youths' lives and their families are clearly seen. What is concerning is that those who have dropped out of school and are not working appear to be more vulnerable to, for example, criminal activity and drug addiction [3].

In the spring of 2021, the rate of joblessness among the population aged above-15 years in Iran was 11.8%, while the rate of joblessness among youths from 15 to 29 years old was 24.4% [1]. Iran ranks 11th (29.2%) in terms of the joblessness rate in the world [2]. Therefore, youths' joblessness is one of Iran's most important social and economic issues.

The word "NEET", was initially introduced at the political level in England, and refers to youths ranging from 16-24 who have dropped out of school early, are not involved in training, and do not work anywhere [4]. This concept was introduced and applied in most developed countries and some developing countries. Fourlong [5] considered that one of the main reasons for using this concept was the complex increase in youths' employment process, the weakening of youths' spirit for education and skills acquisition, and the increase in part-time work patterns.

Given the effects associated with NEETs, it is important to identify the causes and roots of being NEET. This will enable the development of appropriate strategies to ensure the productive involvement of youth in the economy and society at large. The National Survey of Socioeconomic Characteristics (CASEN) is the foundation for most studies on young unemployment and NEET. Several studies have presented prediction models that may evaluate the chance of becoming NEET based on a variety of criteria [6], [7]. In addition, some studies have delved into public intervention policies targeting this specific demographic [8], whereas certain analyses explored these policies and critiqued their limitations, such as a restricted scope, insufficient interdisciplinary considerations, and lack of a gender perspective [9].

Studies show that men and women have different reasons for being outside the school system and the labor market. Young women make decisions based on their domestic responsibilities, which is a pattern that has persisted over time, whereas for their male peers, it is, among others, because they lack interest in motherhood and fatherhood [10].

While other countries have a history of research on NEET [11]-[15], there is little evidence available for Iran. Studies in European countries have shown that not all circumstances affecting young people during the transition from school to work can be included in NEET. For this reason, various studies have emphasized the significance of creating typologies to address the variety of this phenomenon [12]-[14], [16].

Many NEET dimensions have been studied. However, no academic study has been conducted on NEET in Iran. Because of Iran's high youth unemployment rate, analyzing the factors associated with NEET is relevant. We apply a machine-learning approach to explore the multiple variables associated with NEET. A classification tree-based algorithm (CART) is being used in our data-driven approach. The CART identifies significant variables, the ranges of those variables, and their higher-order interactions of them that together are predictive of the outcome (in this case, being NEET). Classification tree-based algorithms are a type of supervised machine learning that apply a logic of recursive partitioning. By predicting the outcome and comparing it to the observed values, a second subset of data are used to validate the patterns derived from algorithms trained with a subset of data [17]. Classification trees are also easy to describe and comprehend since they aesthetically and accurately mirror the human decision-making process [18]. Thus, classification trees offer an informative and

Mr.Hamed Rahmani is with School of Business and Economics of Maastricht University, Maastricht, The Netherlands (corresponding author, phone: 021620691801; e-mail: hamed.rahmani@ maastrichtuniversity.nl).

Prof. Dr. Wim Groot is with School of Business and Economics of Maastricht University, Maastricht, The Netherlands (e-mail: w.groot@maastrichtuniversity.nl).

intuitive understanding of the problem's structure [19].

This paper aims to analyze the profile of Iranian NEETs based on family and living situation factors and their previous work and education experience. Due to the large number of concurrent processes at work, it is difficult to identify the most significant ones analytically. Most quantitative research has relied on a deductive method [11]-[14], in which hypotheses are constructed based on theory and then statistically verified to determine the theory's generalizability. Instead, this study uses a more inductive, supervised machine learning approach. This method can give ideas about the problem being studied or confirm what has already been learned about unemployment as a social problem.

The remaining part of the paper is organized as follows: In the data and methodology section, we discuss the data sources and describe the methodologies were used. In the next section, we present the analyses of findings with the results, and at the end conclusions are presented.

II. Data and Methodology

The analysis is based on cross-sectional data from the labor force survey conducted in the spring of 2021 in Iran. Data were obtained from the Statistical Center of Iran and the Ministry of Cooperatives Labor and Social Welfare. The sampling for the Labor Force Survey of Iran was based on 1173 primary sampling units (PSUs) or enumeration areas (EAs). Each PSU or EA was defined as a geographical region of non-overlapping contiguous territory with discernible borders. The country has five geographical regions: the north, south, east, west, and center. Each division was further subdivided into city corporations, urban localities, and rural. As a result, the country was split into 15 strata. Housing prices categorized households as 'expensive', 'semi-expensive,' or 'not expensive'.

Each PSU or EA contained around 213 households, and the total sampling frame included 117,000 households. PSUs or EAs were established nationwide and included people from all socioeconomic levels to generate a representative sample of Iran's entire population. Multistage cluster random sampling was utilized as the sampling method. The procedure was as follows:

- Stage one: 1173 PSUs or EAs are randomly chosen from all districts and 15 regional strata, ensuring that all three kinds of households are represented. At this phase, a total of 117,000 households are chosen.
- Stage two: Clusters of 24 houses were randomly selected from each 1173 PSUs or EAs, with no replacement of non-responding families allowed. This stage of the process involves selecting a total of 28,112 households.

The formula used to determine the sample size for subpopulations was:

$$n=\left[\frac{(1-p)}{p}\times\left(\frac{z(\alpha/2)}{r}\right)^2\right]\times deff \tag{1}$$

where p = the proportion of a priori required characteristics of the population; z(α/2) = value of the standard normal variate allowing 100(1–α)%p confidence; r = margin of error allowed, N = size of population; and assuming α = 0.005, deff = 2, p = 0.042 [2].

From the Quarterly Labor Force Survey 2021 dataset [2], NEET refers to people who meet three requirements: (i) youth aged 16 to 24 years old, (ii) they must be without a job; and (iii), they must have missed education or training during the past four weeks preceding the survey. Here, education refers to formal education only; informal education and short-term educational activities are not included. All instances of at least one hour of compensated work per week prior to the survey are included in the definition of employment. Therefore, the term "unemployment" refers to a situation where there has been less than one hour of paid work in the previous week. Based on our three criteria for selecting NEET people from data, NEET status was the dependent variable, which was equal to one if the individual was NEET, and zero otherwise. We use three separate NEET variables: one for all youth, one for males, and one for females. This makes it easier to perform the study independently for all male and female youth and draw on the results afterwards. Gender, education, work experience, and time spent doing housework are the explanatory variables used. We selected these variables because they were common among NEET youth and many previous studies [6]-[8], [10], and we could select only NEET youth from the population by them.

The number of years in school was used to determine education level. Three assumptions were made regarding individuals and their educational qualifications: (i) it was assumed that no individual had completed more than one diploma, bachelor's or master's degree; (ii) it was assumed that individuals who completed a bachelor's or master's degree did not complete a diploma before obtaining their higher qualification.; and (iii) diploma is equivalent to 1 year of schooling, bachelor's to 4 years, and master's to 1 year. PhD graduates are removed from the study since there is no agreement on the number of years required to get this degree. This dataset included 0.08% PhD graduates; eliminating it did not significantly lower the sample size. Our dataset did not directly measure work experience. Thus, prospective experience replaced labor market experience. According to Mincer [20], potential experience is the gap between age and the number of years of education. Time spent on household work was the sum of all household activities, production of commodities and services for own use, and total hours worked each week.

In addition to these explanatory variables, several instrumental variables are also used in the model. These factors reflect the collection of underlying variables that impact a person's education, work, or training choice. We use these factors to account for unexpected behavior between variables. These factors include birthplace, the quantity of land held by households, the number of children under the age of 6 years, and the total number of years of education for all family members. The type of birthplace is used to verify if an early childhood environment or education influences the NEET status. Youth born in cities are less likely to become NEETs because of the advantages in their early childhood. All household members' cumulative years of education were considered to verify whether family education affects NEET youth status. In this study, household wealth was measured by the amount of land owned by households. It is expected that NEET youth status is positively correlated with land ownership. This is because a wealthier household can afford to indulge in more 'conspicuous leisure' than a poor household [10]. Female youth may have to stay at home to care for young children, and male youth may be forced to get a job to help support their families if there are young children in the home. The variables and their definitions used in this analysis are described in Appendix Table VII.

The analysis aims to classify the factors affecting NEET youth status in Iran. Our dependent variable is binary, so we use a probit model in our estimations. Two desirable characteristics of Probit led to its selection: (i) the estimated conditional probability is always between zero and one, and (ii) the relationship between the independent and dependent variables is non-linear. Consequently, as the estimate of the conditional probability approaches zero or one, the rate of change diminishes as the estimated conditional probability approaches either one or zero [21].

The Probit models for this analysis are given as follows:

$$\Pr(\text{NEET youth} = 1|X1,X2,X3) = \Phi(\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \varepsilon i) \quad (2)$$

$$\Pr(\text{Male NEET youth} = 1|X1,X2,X3)= \Phi(\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \varepsilon i) \quad (3)$$

$$\Pr(\text{Female NEET youth} = 1|X1,X2,X3) = \Phi(\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \varepsilon i) \quad (4)$$

where, X1 = education, X2 = experience, X3 = housework and Φ = cumulative standard normal distribution function.

The determinants of NEET status for all youth, as well as the determinants of NEET status for each gender, will be investigated using three independent models. Youth do not make decisions about employment, education, or training on the spur of the moment. In the real world, youth make deliberate and logical decisions about whether or not to work, study, or pursue the training. As a result, they self-select into the NEET population by making these decisions. When the errors are connected to the independent variables, the Probit model becomes endogenous, and unbiased estimates are no longer possible [22]. However, these models may still be estimated using a two-step estimating approach devised by Heckman [23]-[25]. Heckman argued that if unobserved heterogeneity could be handled independently and the resultant information included in the main model, the issue of sample selection bias might be eliminated. As a result, Heckman argued that the selection bias caused by non-random sample selection might be seen as a simple specification mistake. He proposed a two-step estimating approach to address the issue [25]. Heckman's sample selection bias correction approach is comprised of the following steps:

Step1. Using a Probit model, estimate the underlying determinants that impact an individual's choice to self-select into the sample and get the inverse Mills ratio.

Step2. Re-estimate the previous model using the inverse Mills ratio as an extra regressor. The selection equation is calculated in the first stage, while the main equation is obtained in the second phase. This approach was dubbed 'Probit model with sample selection'. The general form of the Probit model with sample selection is explained in Appendix 2.

The sample selection bias correction model incorporates the underlying elements that determine a person's decision to self-select for NEET status and the factors that immediately affect young people's NEET status.

Our study also employs classification and regression tree (CART) analysis. Classification trees visually reflect and forecast effects of a qualitative answer model for a binary dichotomous dependent variable. Classification trees are decision trees in which an underlying construct model determines the options. The roots of classification trees are shown at the top, with the branches downward. Classification trees rely on three essential building blocks: A set of binary questions must be formulated that can partition the data with each response; second, principles must guide where to place splits between branches to organize the observations; finally, a criterion determines the optimal tree structure based on question order and split points [25]. There is a general definition of a model, and the model’s independent variables provide the binary questions that divide the tree into its divisions. Branches of the tree are arranged so that the descending branches are 'purer' than the parent branches, where purity refers to the degree to which a node consists of only one homogeneous class. The classification tree method examines all variables at each node to determine the optimal split for each variable. In order for a tree to be considered right-sized, it must not significantly reduce impurities with additional branch separation [26]. At the start of the CART analysis, the dataset is divided into two parts using a ratio of 60:40, with 60% of the data are labeled as training or learning data, while 40% is marked as test data. Observations are automatically allocated to one of two data groups: training or testing. The classification tree uses training data to train the algorithm and test data to validate the model's predictions. Classification necessitates a systematic procedure to anticipate which class a categorical variable will fall into, depending on the model's familiarity with training results [27]. A recursive partitioning algorithm is utilized [28], [29]. The method involves segmenting the prediction space into specific unchanging areas and predicting a given observation based on the training observation mode in the area to which that observation belongs [30]. If the relationship between a model's dependent and independent variables is non-linear and dynamic, decision trees based on recursive partitioning algorithms will outperform conventional approaches such as linear regression [31]. This approach is well adapted to grow a simple tree of classification based on the Probit model. More complex CART analysis methods, such as bagging or a random tree, cannot be used for ease of understanding. The primary reason for using the CART study is to explore the paths that could be pursued to lead youth out of the NEET status.

III. Result

The Quarterly Labor Force Survey spring 2021 dataset included 573,475 observations from youth aged 16 to 24 years. A subset of 192,182 observations containing data on NEET status was used for this study. From this sample, 44% (84560) were male and 56% (107622) were female.

The Probit model results in Table I indicate that for the model of all youth, male, and female, the connection between education and the probability of obtaining NEET status is positive and statistically significant. However, education has a greater influence on the NEET status of males than on females. Hours spent performing home chores are also linked to the likelihood of having NEET status in all youth, all male youth, and female youth models.

However, when all other variables are kept constant, females are more likely than males to have a NEET status owing to housekeeping. For all of the models, the Wald chi-square test value is significant. As a result, the null hypothesis that all regression coefficients are zero can be rejected. Consequently, all three models are statistically significant when aggregated. The coefficients of the Probit model cannot be regarded as marginal effects due to its non-linearity [32], [33]. Woolridge [34] recommends calculating the average marginal effects of these models to determine the magnitude of the impact of each independent variable on the dependent variable [35].

TABLE I
Results of Probit model estimation

Model	Probit model		
Variables	NEET Youth	Male NEET youth	Female NEET youth
Education	0.1292*** (0.0031)	0.1435*** (0.0048)	0.1312*** (0.0032)
Experience	0.2072*** (0.0016)	0.1899*** (0.0035)	0.2108*** (0.0183)
Housework	0.0236*** (0.0003)	0.0091*** (0.0004)	0.0178*** (0.0004)
Constant	-4.8986*** (0.0391)	-5.0632*** (0.0819)	-4.2934*** (0.0451)
Wald chi-squared	31760.42***	4461.53***	18448.52***

Table II indicates that, when all other variables are kept constant, each extra year of education decreases being NEET by 2.4% among male youth, while for female youth there is a 2.1% decrease. For males, years of potential work experience reduce the likelihood of becoming a NEET by 3.5%, while for females, it decreases by 3.6%. Domestic duties are almost twice as important in predicting NEET status for female adolescents than for male adolescents. For every additional hour spent on household tasks, the likelihood of becoming a NEET increases by 0.31% for females, but only by 0.18% for males.

TABLE II
Average Marginal Effects of Probit Model Estimation

Model	Probit model		
	NEET youth	Male NEET youth	Female NEET youth
Education	0.0263*** (0.0003)	0.0252*** (0.0006)	0.0211*** (0.0004)
Experience	0.0387*** (0.0002)	0.0361*** (0.0004)	0.0364*** (0.0003)
Housework	0.0044*** (0.0000)	0.0016*** (0.0001)	0.0028*** (0.0000)

Appendix Table VIII provides additional information on the selection bias adjusted Probit model. The selection bias-adjusted Probit model produced similar results compared to the previous Probit model. In models of all NEET youth, male NEET youth, and female NEET youth, the coefficients for education, experience, and housekeeping variables are positive and statistically significant. Nevertheless, each case's coefficients are smaller than the Probit model predicts. Probit models did not take into account underlying factors that might affect whether or not a young person is NEET. Consequently, education, experience, and housekeeping variables were overestimated. For example, owning land has a negative and statistically significant link to the number of female teens who do not work or go to school. Also, the strength of this connection grew along with the amount of land the family owned. As a result, female youth from affluent households are less likely to become NEETs than those from low-income families. On the other hand, land ownership has no statistically significant impact on the NEET status of male adolescents in the majority of instances.

The number of children in the house makes it less likely for male youth to be NEET, while it makes it more likely for female youth to be NEET. This finding fits with the traditional roles of men and women in Iranian culture, in which men are the breadwinners and women are the caretakers.

Youth from well-educated families, regardless of gender, were less likely to become NEET. Female youth born in cities had a lower risk of becoming NEET than female youth born in rural areas, while male youth born in cities and rural areas had the same chance of becoming NEET.

The Wald test of independent equations is based on the null hypothesis that the selection and main equations are independent. The Wald test of independent equations' chi-squared test statistic is statistically significant for all three models. The average marginal effects show that the selection bias-adjusted Probit model underestimated the influence of education, experience, and housekeeping factors (Table III).

TABLE III
Average Marginal Effects of Selection Bias Adjusted Probit Model

Variables	NEET Youth	Male NEET youth	Female NEET youth
Education	0.0093*** (0.0001)	0.0114*** (0.0003)	0.0105*** (0.0003)
Experience	0.0142*** (0.0001)	0.0159*** (0.0001)	0.0179*** (0.0005)
Housework	0.0017*** (0.0000)	0.0006*** (0.0000)	0.0013*** (0.0000)

In machine learning, classification issues are often graphically summarized using a confusion or error matrix. The confusion matrix shows the actual variable classes, and the other dimension displays the expected variable classes [36].

A confusion matrix may evaluate the raw number or rate of successful and failed predictions [37]. In the confusion matrix, TP stands for true positives, TN stands for true negatives, FP stands for false positives, and FN stands for false negatives. The following quality evaluation criteria can be derived from the evaluation criteria:

(i) TPR: True Positive Rate or Sensitivity or Recall: {TPR = (TP)/(TP + FN)},

(ii) FPR: False Positive Rate: {FPR = (FP)/(FP + TN)}, TNR: True Negative Rate or Specificity: {TNR = (TN)/(TN + FP)}, FNR: False Negative Rate: {FNR = (FN)/(FN + TP)};

(iii) Prevalence = Σ(Condition positive)/Σ(Total population), Accuracy = [Σ(True positive) + Σ(True negative)]/Σ(Total population);

(iv) PPV: Positive Predictive Value or Precision: {PPV= (TP)/(TP + FP)}, NPV: Negative Predictive Value: {NPV = (TN)/(TN + FN)}, FDR: False Discovery Rate: {FDR = (FP)/(FP + TP)}, FOR: False Omission Rate: {FOR = (FN)/(FN) + (TN)};

(v) LR+: Positive Likelihood Ratio: {LR + = TPR/FPR}, LR−:Negative Likelihood Ratio: {LR− = FNR/TNR};

(vi) DOR: Diagnostic Odds Ratio {DOR = (LR +)/(LR−)};

(vii)F_1 score = 2/[(1/Recall) + (1/Precision)];

A basic confusion matrix with several frequently related performance metrics is presented in Table IV. Confusion matrices aid in analyzing the effectiveness of machine learning techniques and assessing the degree of fit of the Probit model, which includes a binary dichotomous dependent variable [38]. Appendix Table IX-XI show the confusion matrices for all NEET youth, male NEET youth, and female NEET youth from the Probit model.

TABLE IV
A GENERIC CONFUSION MATRIX AND THE PERFORMANCE INDICATORS ASSOCIATED

		True condition			
		Condition Positive	Condition Negative	Prevalence	Accuracy
Predicted condition	Positive	TP	FP	PPV	FDR
	Negative	FN	TN	FOR	NPV
		TPR	FPR	LR+	DOR
		FNR	TNR	LR-	

Three indications of the fit quality may be generated using the data from a standard confusion matrix. These are the percentage of properly anticipated events or accuracy, the DOR, and the F_1 score. The percent properly predicted, often known as accuracy, refers to how well the predicted values of a model match the actual values [34]. Accuracy is expressed as a percentage, with 100% being the model's most possible match [35]. The DOR is the second indication of the quality of fit, and DOR is unaffected by the number of classes in a binary dichotomous variable, it is preferred above accuracy when one class has a disproportionately high frequency.

The DOR is defined as the ratio of the chances that a predicted value will be positive if the real value is positive to the odds that a predicted value will be negative if the real value is negative. An increased DOR indicates a better fit, with values ranging from 0 to infinity. The F_1 score, a third indicator of fit quality, is also used. The harmonic average of accuracy, sensitivity, or recall is used to generate the F_1 score [39].

Precision is defined as the percentage of true positives (TP) predicted by a model to total positives [36]. The proportion of correctly predicted positive occurrences by a model, or the actual positive rate, is referred to as sensitivity or recall. F_1 score values range from 0 to 1, with higher values indicating better model fit. According to Table V, male NEET youth models exhibit higher accuracy and DOR than female NEET youth models. Male NEET youth models, on the other hand, have lower F_1 scores than female NEET youth models, it is difficult to tell whether the male or female young NEET models have a better goodness of fit.

TABLE V
THE PERFORMANCE OF THE SELECTION BIAS ADJUSTED PROBIT MODEL

Dependent variables	NEET youth	Male NEET youth	Female NEET youth
Accuracy	84.91%	89.69%	84.69%
DOR	31.8186	197.3206	18.1436
F_1 score	0.8649	0.4821	0.9013

Due to the insufficient explanation of goodness of fit provided by the confusion matrix indicators, receiver operating characteristics (ROC) graphs for the Probit model were created and evaluated. ROC graphs may be used to visualize, organize, and choose classifiers depending on their performance [40]. ROC graphs are two dimensional plots in which the true positive (sensitivity) rate is plotted on Y-axis against the false positive rate (specificity) on X-axis. A good prediction model is one achieving the right balance between sensitivity and specificity. Statistically, this corresponds to ROC $\geq$ 0.7 [41].

ROC curves are commonly found above the diagonal line in ROC space. A ROC curve closer to the top-left corner of the ROC space suggests better performance than a curve closer to the center diagonal line. A model's area under the ROC curve, which can be calculated using integral calculus, is a better way to judge its performance than just looking at it. The area under a ROC curve in the ROC space above the center diagonal line ranges from 0.5 to 1.0, with 1.0 being the best model [40]. As shown in Table VI, the area under the ROC graphs for each model is calculated. Models representing young female NEETs performed better than models representing young male NEETs. But test results cannot be used to figure out if the real specification of a model is good enough. Calculation of performance of the selection bias adjusted Probit model in Table V and ROC $\geq$ 0.7 for all models as Table VI again confirms that of the selection bias adjusted Probit model estimated the effect of the education, experience, and housework variables. So, the selection bias adjusted Probit model for all NEET youth used education, experience, and housework as independent variables for training and testing the classification tree of NEET youth.

TABLE VI
AREA UNDER ROC CURVE

Variables	Youth NEET	Male youth NEET	Female youth NEET
Area under ROC Curve	0.8963	0.7218	0.8721

Using a classification tree to analyze the causes of NEET youth status offers a new perspective and shows possible ways out. NEETs who spent less than 40 hours per week on housework had 78% chance of leaving the NEET status. They had 82% chance of leaving the NEET status if they had less than 20 years of potential labor-force experience. Those who worked more than 40 hours per week on household duties, on the other hand, had an 85% probability of becoming NEETs. There was a 90% chance that these youth would remain unemployed if they also had 12 years or more of potential labor market experience. A young NEET classification tree suggests that housekeeping is the major barrier preventing youth from leaving their NEET status. This is primarily because housework affects young women more than young men.

IV. Conclusion

The purpose of this study is to analyze NEET youth in Iran and to examine the factors that contribute to NEET youth. In order to explain NEET youth in Iran, Probit and selection bias-adjusted Probit models were used. In this study, we additionally employed an implementation of the CART analysis to complement the econometric analysis. This study used data from the Quarterly Labor Force Survey 2021. According to the findings of this study, years of schooling, years of potential labor market experience, decrease the probability of a young person becoming NEET. The results also indicate that having small children in the house and hours spent doing household duties all increases the probability of female youth being NEET while reducing the likelihood of male youth becoming NEET. The geographical location of female youth influences their NEET status. Female youth born in cities were less likely to enter NEET status than female born in rural regions [42]. According to the classification tree based on the selection bias adjusted Probit model, home activities have a significant impact on employment. In order for adolescents to transition out of NEET status, reducing the burden of housekeeping was the most important factor [43]. The results of this study shed light on the fundamental causes of NEET in Iran.

One drawback of this study is that the results only apply to Iran and cannot be generalized across nations or time. Nonetheless, the general approach might be reproduced in other countries to investigate the issue of young unemployment. Furthermore, in this research, a basic application of CART analysis was employed, which may be regarded as a preliminary exploratory exercise. The results of the study could be made more reliable by making more improvements to this method. Furthermore, the small sample size used by the participants is not typical of all Iranian NEETs. As a result, the results of the participants should be taken with caution. Much bigger samples, subject to time and finances, may be undertaken to incorporate the views of jobless youngsters, academics, employers, and policymakers. An analysis like this could shed light on the skills gap among Iran's youth as well as the real needs of businesses.

Appendix 1

TABLE VII
Definitions of Variables

Variable	Definition
NEET youth	Variables are binary dichotomous for all individuals suggesting NEET status Between the ages of 16 to 29 0 = Not NEET 1 = NEET youth
Male NEET youth	A binary dichotomous variable which indicates NEET status only for male people between the ages of 18 and 35 0 = Male (Not NEET) 1 = Male (NEET)
Female youth NEET	A binary dichotomous variable that indicates NEET status only for female people between the ages of 18 and 35 0 = female (Not NEET) 1 = female (NEET)
Education	Amount of years of study, measured in terms of the top class passed 0 = No Class Passed 1 = Class 1 2 = Class 2 3 = Class 3 … 13 = Diploma 16 = Bachelor's degree 17 = Master's degree
Experience	As a labor market experience variable, potential experience is utilized.
Housework	The number of hours spent each week on domestic chores = (total number of hours spent performing all household chores such as cooking, washing clothing, cleaning, shopping, caring for children and the elderly, and any other miscellaneous job) + (total number of hours spent per week producing products for personal use) + (number of hours spent per week in the production of services for own consumption)
Family education	The cumulative number of years of education of all members of the household
Birthplace	A binary dichotomous variable indicating the kind of birthplace 0 = Born in a rural place 1 = Born in an urban place
Children	Total number of children, aged less than 6 years old, per household
Land	A set of dummy variables indicating the total amount of house owned by the household, measured in m2 Land 1 = 1 if household does not own any house and equal to zero otherwise Land 2 = 1 if household owns 20-50 m2 of the house and equal to zero otherwise Land 3 = 1 if household owns 50–100 m2 of the house and equal to zero otherwise Land 4 = 1 if household owns 100–199 m2 of the house and equal to zero otherwise Land 5 = 1 if household owns 200 or more m2 of the house and equal to zero otherwise. Land 1 and 2 = 'not expensive', Land 3 = 'semi-expensive', Land 4 and Land 5 = 'expensive'

Appendix 2. The General Form of the Probit Model with Sample Selection

$$YiProbit = \alpha'Xi+\varepsilon 1i \text{ [Main equation]}$$

$$YiSelect = \beta'Zi+\varepsilon 2i \text{ [Selection equation]}$$

where,

$$Z = (\gamma + \beta 1W1 + \beta 2W2 + \cdots ,+\beta nWn)$$

The regression function for the subsample of complete observations (YiSelect ≥0) is given by

$$E(YiProbit|Xi, YiSelect \geq 0) = \alpha' Xi + E(\varepsilon 1i|Xi, YiSelect \geq 0)$$

It is assumed that ε1 and ε2 are bivariate standards normally distributed with correlation coefficient ρ, so that

$$E(\varepsilon 1i|Xi, YiSelect \geq 0) = \rho\lambda i$$

with

$$\lambda i = \frac{h(Ai)}{H(-Ai)} = \mathrm{h(Ai)}$$

and

$$\mathrm{Ai} = -[\gamma + \beta 1W1 + \beta 2W2 + \cdots ,+\beta nWn]$$

where h = standard normal probability distribution function; H = standard normal cumulative distribution function and λi = inverse Mills ratio. Hence, the selection bias-corrected Probit model becomes:

$$YiProbit = \alpha' Xi + \rho\lambda i + \varepsilon 1i$$

where,

$$E(\varepsilon 1i|Xi, YiSelect \geq 0) = 0$$

and

$$E(\varepsilon^2 1i|YiSelect \geq 0) = \tau_i^2$$

with

$$\tau_i^2 = 1 + \rho^2\lambda i(Ai - \lambda i)$$

Therefore, for this specific analysis, the selection bias-corrected Probit model is given as:

$$Y_iProbit = \alpha' Xi + \rho\lambda i + \varepsilon 1i \text{ [Main equation]}$$

$$Y_iSelect = \beta' Z_i + \varepsilon_{2i} \text{ [Selection equation]}$$

where, $X_i = \alpha_0 + \alpha_1$ education + α_2 experience + α_3 housework, $Z_i = \beta_0 + \beta_1$ land 2 + β_2 children + β_3 family education + β_4 Birthplace.

TABLE VIII
RESULTS OF SELECTION BIAS CORRECTED PROBIT MODEL ESTIMATION

Model	Selection equation	Main equation	Selection equation	Main equation	Selection equation	Main equation
Variables	Youth NEET	Youth NEET	Male youth NEET	Male youth NEET	Female youth NEET	Female youth NEET
Education		0.1308***		0.1313***		0.1081***
		(0.0029)		(0.0049)		(0.0034)
Experience		0.1990***		0.1930***		0.1882***
		(0.0026)		(0.0038)		(0.0041)
Housework		0.0243***		0.0095***		0.0165***
		(0.0003)		(0.0005)		(0.004)
Land 2	-0.0922***		0.0106		-0.1141***	
	(0.0062)		(0.0092)		(0.0069)	
Land 3	-0.1592***		0.0316***		-0.2101***	
	(0.0064)		(0.0093)		(0.0072)	
Land 4	-0.1966***		-0.0141		-0.2351***	
	(0.0114)		(0.0159)		(0.0127)	
Land 5	-0.1739***		0.0370		-0.2335***	
	(0.0265)		(0.0348)		(0.0299)	
Children	0.0269***		-0.3311***		0.1707***	
	(0.0032)		(0.0056)		(0.0030)	
Family education	0.0072***		0.0104***		0.0034***	
	(0.0001)		(0.0002)		(0.0001)	
Birthplace	-0.0622***		0.0000		-0.0792***	
	(0.0052)		(0.0072)		(0.0058)	
Constant	-0.9624***	-4.4054***	-1.7050***	-4.3763***	-1.1710***	-3.1504***
	(0.0060)	(0.1258)	(0.0088)	(0.1660)	(0.0065)	(0.1584)
Inverse Mills		-0.2868***		-0.2765***		-0.4931***
ratio		(0.0489)		(0.0497)		(0.0554)
Rho		-0.2792		-0.2696		-0.4566608
		(0.0451)		(0.0461)		(0.0438794)
LR chi-squared	3369.93***		8332.86***		4310.89***	
Wald chi-squared		8312.83***		4223.20***		2270.29***
Wald test of independent equation (chi-squared)						

TABLE IX
CONFUSION MATRIX OF PROBIT MODELS

		True condition				
		Condition positive (NEET youth = 1)	Condition positive (NEET youth = 0)	Prevalence (57.04%)	Accuracy (85.21%)	
Predicted condition	Predicted condition positive	TP (43.230)	FP (5697)	PPV (88.36%)	FDR (11.64%)	
	Predicted condition negative	FN (7807)	TN (32.735)	FOR (19.26%)	NPV (80.74%)	
		TPR (84.70%)	FPR (14.82%)	LR+ (5.7152)	DOR (31.8186)	F1 score (0.8649)
		FNR (15.30%)	TNR (85.15%)	LR-(0.1796)		

TABLE X
CONFUSION MATRIX OF PROBIT MODEL FOR MALE NEET YOUTH

		True condition				
		Condition positive (NEET youth = 1)	Condition positive (NEET youth = 0)	Prevalence (14.90%)	Accuracy (87.27%)	
Predicted condition	Predicted condition positive	TP (1345)	FP (57)	PPV (95.93%)	FDR (4.07%)	
	Predicted condition negative	FN (2833)	TN (23.801)	FOR (10.64%)	NPV (89.36%)	
		TPR (32.19%)	FPR (0.24%)	LR+ (134.125)	DOR (197.3206)	F_1 score (0.4821)
		FNR (67.81%)	TNR (99.76%)	LR-(0.6797)		

TABLE XI
CONFUSION MATRIX OF PROBIT MODEL FOR FEMALE NEET YOUTH

		True condition				
		Condition positive (NEET youth = 1)	Condition positive (NEET youth = 0)	Prevalence (63.28%)	Accuracy (81.78%)	
Predicted condition	Predicted condition positive	TP (42.925)	FP (5471)	PPV (88.70%)	FDR (11.30%)	
	Predicted condition negative	FN (3934)	TN (9103)	FOR (30.18%)	NPV (69.82%)	
		TPR (91.60%)	FPR (37.54%)	LR+ (2.4401)	DOR (18.1436)	F_1 score (0.9013)
		FNR (8.40%)	TNR (62.46%)	LR-(0.1345)		

REFERENCES

[1] World Bank (2020). Global unemployment rate from 2002 to 2020. World Bank Blogs.

[2] Statistical center of Iran (SCI). Estimating the population of the youth in Iran, 2021. Iran, Tehran. Available from: https://www.amar.org.ir/.

[3] Ralston, K., Everington, D., Feng, Z., & Dibben, C. (2022). Economic inactivity, not in employment, education or training (NEET) and scarring: the importance of NEET as a marker of long-term disadvantage. Work, Employment and Society, 36(1), 59-79.

[4] Lőrinc, M., Ryan, L., D’Angelo, A., & Kaye, N. (2020). De-individualising the ‘NEET problem’: An ecological systems analysis. European Educational Research Journal, 19(5), 412.

[5] Furlong, A. (2006). Not a very NEET solution: representing problematic labor market transitions among early school-leavers. Work, employment and society, 20(3), 553-569.

[6] Zhang, G., Zhou, S., Xia, X., Yüksel, S., Baş, H., & Dincer, H. (2020). Strategic mapping of youth unemployment with interval-valued intuitionistic hesitant fuzzy DEMATEL based on 2-tuple linguistic values. IEEE Access, 8, 25706-25721.

[7] Su, X., & Wong, V. (2022). Enhancing the career capabilities of NEET youth in Hong Kong: an experience-driven framework. International Journal for Educational and Vocational Guidance, 22(3), 713-738.

[8] Valdemoros-San-Emeterio, M. Á., Sanz-Arazuri, E., & Ponce-de-León-Elizondo, A. (2017). Digital leisure and perceived family functioning in youth of upper secondary education. Comunicar. Media Education Research Journal, 25(1).

[9] Karyda, M., & Jenkins, A. (2018). Disadvantaged neighbourhoods and young people not in education, employment or training at the ages of 18 to 19 in England. Journal of Education and Work, 31(3), 307-319.

[10] Abayasekara, A., & Gunasekara, N. (2019). Determinants of youth not in education, employment or training: Evidence from Sri Lanka. Review of Development Economics, 23(4), 1840-1862.

[11] Bynner, J., & Parsons, S. (2002). Social exclusion and the transition from school to work: The case of young people not in education, employment, or training (NEET). Journal of vocational behavior, 60(2), 289-309

[12] Giret, JF, Guégnard, C., & Joseph, O. (2020). School-to-work transition in France: the role of education in escaping long-term NEET trajectories. International Journal of Lifelong Education, 39 (5-6), 428-444.

[13] European Foundation for the Improvement of Living and Working Conditions. (2012). Young people and NEETs in Europe: First findings. Dublin. Available at: http://www.eurofound.europa.eu/pubdocs/2011/72/en/2/EF1172EN. pdf.

[14] Salvà-Mut, F., Tugores-Ques, M., & Quintana-Murci, E. (2018). NEETs in Spain: an analysis in a context of economic crisis. International Journal of Lifelong Education, 37(2), 168-183.

[15] Tamesberger, D., & Bacher, J. (2014). NEET youth in Austria: A typology including socio-demography, labor market behaviour and permanence. Journal of youth studies, 17(9), 1239-1259.

[16] Tomczyk, Ł., & Selmanagic-Lizde, E. (2018). Fear of Missing Out (FOMO) among youth in Bosnia and Herzegovina—Scale and selected mechanisms. Children and Youth Services Review, 88, 541-549.

[17] Buntine, W. (2020). Learning classification trees. In Artificial Intelligence frontiers in statistics (pp. 182-201). Chapman and Hall/CRC.

[18] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20-28.

[19] Du, X., Xu, H., & Zhu, F. (2021). A data mining method for structure design with uncertainty in design variables. Computers & Structures, 244, 106457.

[20] Mincer, J. (1974). Schooling, experience, and earnings. New York, NY: National Bureau of Economic Research (NBER).

[21] Williams, R., & Jorgensen, A. (2023). Comparing logit & probit coefficients between nested models. Social Science Research, 109, 102802.

[22] Filippini, M., Greene, W. H., Kumar, N., & Martinez-Cruz, A. L. (2018). A note on the different interpretation of the correlation parameters in the Bivariate Probit and the Recursive Bivariate Probit. Economics Letters, 167, 104-107.

[23] Heckman, J. (1974). Shadow prices, market wages, and labor supply. Econometrica, 42(2), 679–694.

[24] Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement, 5(4), 475–492.

[25] Heckman, J. (1979). Sample selection as a specification error. Econometrica, 47(1), 153–161.

[26] Maceika, A., Bugajev, A., Šostak, O. R., & Vilutienė, T. (2021). Decision tree and AHP methods application for projects assessment: a case study. Sustainability, 13(10), 5502.

[27] Questier, F., Put, R., Coomans, D., Walczak, B., & Vander Heyden, Y. (2005). The use of CART and multivariate regression trees for supervised and unsupervised feature selection. Chemometrics and Intelligent Laboratory Systems, 76(1), 45-54.

[28] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456.

[29] Li, T., Lei, L., Bhattacharyya, S., Van den Berge, K., Sarkar, P., Bickel, P. J., & Levina, E. (2022). Hierarchical community detection by recursive partitioning. Journal of the American Statistical Association, 117(538), 951-968.

[30] Therneau, T., Atkinson, B., & Ripley, B. (2018). rPart—Recursive Partitioning and Regression Trees, Version 4.1-13.

[31] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning with applications in R. New York: Springer.

[32] Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4), 140-147.

[33] Gujarati, D. N. (2013). Basic econometrics. New York: McGraw-Hill and Irwin.

[34] Stock, J. H., & Watson, M. W. (2011). Introduction to econometrics. Boston: Pearson Education.

[35] Woolridge, J. M. (2011, March). Thoughts on heterogeneity in econometric models. In Presidential Address at the annual meeting for the Midwest Economics Association (MEA), St. Louis, Missouri, March (pp. 19-20).

[36] Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. Psychological Methods, 23(3), 389.

[37] Ting, K. M. (2017a). Confusion matrix. In C. Sammut & G. I. Webb (Eds.), Encyclopedia of machine learning and data mining. New York: Springer.

[38] Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37–63. Retrieved August 26, 2018

[39] Hu, B., Dixon, P. C., Jacobs, J. V., Dennerlein, J. T., & Schiffman, J. M. (2018). Machine learning algorithms based on signals from a single wearable inertial sensor can detect surface-and age-related differences in walking. Journal of biomechanics, 71, 37-42.

[40] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1), 1-13.

[41] Jia, Z., Lin, S., Gao, M., Zaharia, M., & Aiken, A. (2020). Improving the accuracy, scalability, and performance of graph neural networks with roc. Proceedings of Machine Learning and Systems, 2, 187-198.

[42] Daoud, M., & Mayo, M. (2019). A survey of neural network-based cancer prediction models from microarray data. Artificial intelligence in medicine, 97, 204-214.

[43] Khatun, F., & Saadat, S. Y. (2020). Youth Employment in Bangladesh. Springer Singapore.