Building a predictive model for poverty in Palestine using Machine Learning classification tools

Prepared by:

Ali Algharabeh

1185387

Committee:

Dr. Hassan Abu Hassan	Supervisor
Dr. Tareq Sadeq	Member
Dr. Mustafa Jarrar	Member

Submitted in partial fulfillment of the requirements for the "Master Degree in Applied Statistics and Data Science" from the faculty of Graduate Studies at Birzeit University - Palestine

May 2020

جامعة بيرزيت
BIRZEIT UNIVERSITY

Building a predictive model for poverty in Palestine using Machine Learning classification tools

Prepared by:

Ali Algharabeh

1185387

Committee:

Name	Signature
Dr. Hassan Abu Hassan	.....
Dr. Tareq Sadeq	.....
Dr. Mustafa Jarrar	.....

Acknowledgements

I would like to thank my family and friends who gave me the unconditional support and encouragement to complete this research along with my supervisor. Special thanks to Professor Rajai Khanji (Professor of English Linguistic at University of Jordan) and Dr. Sadi El-Krunz (Professor of Statistics at Al-Quds University) for their help in this thesis.

Table of Contents

List of Tables

Abstract

This study focuses on comparing competive models in predicting poverty status in Palestine using data provided by The Palestine Expenditure and Consumption Survey, PECS 2017, which is carried by the Palestinian Central Bureau of Statistics (PCBS). It expands on demographic factors by utilizing them in regression models and a machine learning classifier (decision tree) to predict the poverty status of a household. The study finds numerous demographic variables for the head of household, housing, and household conditions that can be used in predicting the poverty status of a household. The study showed that among the four models used, the decision tree provided the highest accuracy and validity indicators.

ملخص

تركز هذه الدراسة على مقارنة نماذج للتنبؤ بحالة الفقر في فلسطين باستخدام البيانات المقدمة من مسح الإنفاق والاستهلاك 2017 والذي أجراه الجهاز المركزي للإحصاء الفلسطيني. تستخدم هذه الدراسة العوامل الديموغرافية في بناء نماذج الانحدار ومصنف تعلم الآلة (شجرة القرار) للتنبؤ بحالة الفقر في الأسرة. وجدت الدراسة العديد من المتغيرات الديموغرافية لرب الأسرة ومتغيرات تتعلق بظروف الأسرة والمسكن التي يمكن استخدامها للتنبؤ بحالة الفقر للأسرة. وأظهرت الدراسة أن شجرة القرار قدمت أعلى قيم في مقاييس الدقة والصدق بين جميع النماذج الأربعة المستخدمة.

Chapter 1

Introduction

1.1 Rationale

“There is perhaps no better test of the social progress of a nation than that which shows what proportion are in poverty; and for watching the progress the exact standard selected as critical is not of great importance, if it is kept rigidly unchanged from time to time.” Arthur Bowley (1915, p.213). In order to measure poverty and reduce it, two main approaches are used. Absolute measures are the first approach which use poverty lines with constant real value. Relative measures are the second approach where poverty line varies.

Atikson and his colleagues (2001, p 102) argued that “it is scientifically impossible to determine an accurate and valid poverty line: i.e. a financial threshold below which a person is defined as being poor” since poverty is relative, multidimensional and changes over time. However, this study will build a predictive poverty model using regression models and decision tree.

Poor people often lack adequate food, education, shelter and health. The changing extent of poverty is a subject of importance nowadays due to recession. This research studies the demographics of Palestinian society to predict poverty. Wealth and poverty have different measures and indicators. Historically, income is used as the main measure to predict poverty and wealth. However, income is not sufficient to predict poverty and wealth as more other factors should be taken into account that will be discussed later in this study. Income and wealth are two different

terminologies and the differences between them will be distinguished later in this study. A better indicator than income to measure poverty is expenditure for several reasons. One of the main reasons is that people tend to reflect their true expenditure rather than their income. Another important reason for developing countries is self-production, which is not calculated in income but it can be approximated in consumption. Another reason is that income tends to vary due to seasonality, however, expenditure has less variability.

This thesis adopts a quantitative methodology in the form of secondary data which is taken from The Palestine Expenditure and Consumption Survey; PECS 2017, which is considered a tool to measure poverty in Palestine. It aims to analyze poverty in terms of the number of poor people and identify their main demographic features, such as locality, social and economic statuses. Data collected in the survey provide us with indicators of poverty percentage across Palestine. Studying poverty features is an analysis that allows us to compare poor households/individuals to non-poor households/individuals. Studying the features of poverty is the first step to analyze poverty indicators and causes. This research aims to pinpoint the factors of poverty by analyzing multiple variables. Hence, machine learning algorithms such as regression models and decision tree will be utilized to identify poverty factors in Palestine for the year 2017.

1.2 Research objectives

Given the lack of investigation and prediction of poverty and to study and predict poverty in Palestine and shaping a better understanding of factors related to it, this study aims to meet the following objectives:

Research Objectives:

1. Comparing competive models in predicting poverty status in Palestine using data provided by PECS 2017 survey.
2. Building a predictive model for poverty using data from PECS 2017 survey.
3. Identifying factors and demographics of poverty in the Palestinian society.

In order to meet the above objectives, the following questions will be addressed:

1- What are the accuracy and validity measures for the used models and how they compare.

2- What are the features and factors that can be used to build a predictive model for poverty in Palestine.

3- What are the demographics of Palestinian society, how to understand them, and how they are related to poverty.

Such an examination will allow us to predict poverty using the best possible model, and know the reasons behind it from the available data that will be analyzed. Dealing with poverty is part of economic development. According to Fosu (2015), he found that economic growth played a vital role in reducing poverty in Sub-

Saharan African countries. Hence, understanding the factors of poverty and building a model to predict will help in economic development.

1.3 Structure of the Thesis

The thesis begins by presenting the literature covering poverty definition, concepts, causes, and factors related to it. Moving to the methodology section stating the rationale behind using regression models and decision tree in predicting poverty; the following chapter includes the findings and discussions before wrapping up this study with conclusions along with the limitations and future research.

Chapter 2

Literature review

Poverty can be classified as two types absolute and relative. Absolute poverty is usually based on nutritional requirements and refers to what is socially acceptable for living conditions. Relative poverty, on the other hand, compares the highest segments of a population (rich people) with the lowest segments (poor people). It is not necessary for relative and absolute poverty to move in same direction. For instance, when the gap between rich and poor people decreases and a decline in the number of rich people will cause relative poverty to decrease and absolute poverty to increase. Absolute and relative poverty can move in the same direction. For instance, when prices rise faster than incomes, the status of some households in terms in terms of population classes may decrease, and also the absolute poverty (living standards) have decreased equivalently. Hence, relative and absolute poverty has decreased and moved in the same direction. (Renata Lok-Dessallien, 2000)

Many developmentalists prioritize reducing absolute poverty since it is related to malnutrition and starvation. On the other hand, many advocates of the rights-based place the highest priority on relative poverty since it's related to socioeconomic classes.

There are two approaches to measure poverty: means/ends and quantitative/ qualitative. Means indicate inputs placed to achieve a result; ends are the outcome.

That being said, there are many means to choose from. Ends, on the other hand, tend to correlate more closely to the outcome we are going to study. The second approach is quantitative/qualitative. Quantitative data contains numerical data and can be aggregated as opposed to qualitative information (Renata Lok-Dessallien, 2000). This study will utilize both means and ends indicators and quantitative and qualitative indicators.

Objective and subjective perspectives are another approach for poverty. The objective perspective involves normative judgments or value-based approach which is a popular approach used by Economists. On the other hand, the subjective approach is based on people's preferences. The poverty line, under the subjective method, is defined by society itself which makes it a socially realistic method. Poverty measurement has been dominated by the objective approach. Supporters of this approach claim that people are not always the best judge. (Renata Lok-Dessallien, 2000)

Historically, poverty has been measured using the income approach also known as money-metric. Households are considered poor if their income or consumption falls below a certain threshold. This threshold varies in each country and is defined as the minimum living standard and what is socially accepted in a population. A popular indicator for this approach is the headcount index. (Renata Lok-Dessallien, 2000)

The basic needs approach is another popular approach to study and measure poverty. It defines poverty as lack of material requirements needed to meet human basic needs such as: food, shelter, sanitation and health services.

Income-based and basic needs approaches are predominately use quantitative indicators. (Renata Lok-Dessallien, 2000)

Several researchers, use discrete choice models in poverty analysis, such as: Amuedo_Dorantes (2004) for Chile, Geda et al (2001) for Kenya; Kabubo-Mariara (2002) for Kenya; Charlette-Gueard and Mesple-Somps (2001) for Cote d`voire, Goaed and Ghazouani (2001) for Tunisia; Roubaud and Razafindrakoto (2003). The analysis then proceeds by utilizing binary logit or probit model to estimate poverty. Predicting poverty using machine learning algorithms utilizing decision tree methodology has been conducted on demographic data in Nicaragua (Källestål et al., 2019). Variables used in this study such as: house walls material type, water availability, floor type, electricity in the house, and education level were used in the decision tree. Results from the study showed that having access to water and education level are highly significant. Households that have no access to water have a high chance of being poor while those who have high-level education are less likely to be poor. Moreover, the authors emphasized on the fact that setting a decision tree before having classical regression models is important because it enabled them assessing the importance of variables from a large set of explanatory variables before being utilized in a regression model. It also includes and evaluates interactions between predictor variables automatically. The purpose of having a decision tree is to find important predictors and their interactions from a large set of variables and then use them in regression models.

Another study carried out in Kenya, utilized regression modeling to predict poverty. Mwabu et al. (2001) had a comprehensive study that deals with poverty and

identifying its factors. The authors of the paper identified important determinants of poverty such as: age, place of residence (rural, urban), size of household and level of schooling. The paper used two regression models (discrete and continuous), it used overall and food expenditures as dependent variables. The authors used this approach rather than using the logit/probit model because transforming dependent variable into binary leads to a loss of information. The authors have identified important factors of poverty: region, level of education, age, size of household, and place of residence (rural versus urban). The importance of these variables is that they don't change when the total expenditure is set as the dependent variable.

Another study on poverty, by Oyugi (2000), used probit model regression. Discrete and continuous indicators of poverty are used as dependent variables. The explanatory variables include: the number of household members who are able to read and write holding area, livestock unit, household size, working sector and source of water for household use. Results showed that almost all of these variables are significant determinants of poverty.

We get interesting results when comparing results from both models in the above two studies. After conducting regression models, household size, education level, age, residence type, and ability to read and write are the top five important factors of poverty at the national level. In the probit model the key factors of poverty are: household size, source of water, ability to read and write, employment in off-farm activities, engagement in agriculture, and owning a side-business. Region appears to be a common important factor in determining poverty for both approaches.

Rodriguez and Smith (1994) utilized a logistic regression model to estimate the effect of demographical variables on the predicting poverty of a household in Costa Rica. The main result showed that poverty was higher for head of households for those with lower level of education.

Asif (2007) analysis on poverty shows that household size, age of head of household, educational attainment, the dependency ratio, the male ratio of workers, livestock population, land ownership, tend to be significant factors of poverty. He also concluded, the dependency ratio and household size are found to be positively and significantly associated to the probability of the household being poor. On the other hand, age of the head of household, educational attainment of household members, and land ownership are found to be negatively and significantly associated to the probability of the household being poor.

Mok et al. (2007) analysis on poverty, in Malaysia, shows that an increase of one year of formal education reduces the chance of a household falling into poverty, while households with children under 15 years of age and female adults tend to have a higher chance of being poor. The number of children has generally been found to be positively correlated with poverty in studies across the developing world (Carter and May, 1999; Dreze and Srinivasan, 1997; Ray, 2000).

Litchfield and McGregor (2008) analyzed the factors of household poverty in Tanzania. Results showed that individuals living in a female-headed household have higher odds of being poor and lower standards of living compared to their counterpart's male-headed households. The authors showed that educated parents have a lower probability of being poor compared to uneducated headed households.

An increase in household size increases the chance of being poor and causes a reduction in living standards. In Indonesia, as the size of household increases the probability of a household being poor increases (Widyanti et al., 2009). However, in Pakistan, larger households are less likely to be poor (Tareen et al., 2008). This may be because large households utilize economies of scale and may reduce poverty through higher engagement in the workforce and consuming fewer sources.

Age is another key determinant of poverty prediction. Kitov (2006) observed that an individual reaches maximum income at some age between 45 and 55 years, and then drops. (Wagle, 2006), a study in Nepal, found that as age of the household head increases the probability of a household being poor increases. Empirical evidence has showed that young individuals tend to have lower income compared to older individuals (Higgins and Williamson, 2003). A study by Meng and Gregor (2007) indicates that as the age of head of household increases the probability of being poor decreases.

Female-headed households are particularly vulnerable to poverty as shown in several studies. Maitra (2002) showed in 1993 that female-headed households in South Africa were more vulnerable of being poor compared to their male-headed households' counterparts. Same result is found where female-headed households are more vulnerable of being poor in a study conducted in India (Meenakshi and Ray, 2000), South Africa (Aliber, 2001), and Kenya (Muyanga, 2008).

Bertranou and Khamis (2005), studied the relationship between poverty and labor market sectors. Head of households working in manufacturing, hotels, retail trade, construction and restaurants have higher probabilities of being poor. Authors relate

that because even though these industries are dynamic and growing but with low benefits.

A study carried out by World Bank (2007) on poverty in Sri Lanka showed that poverty is strongly associated with household determinants, such as: educational attainment, employment status, and family size. The authors also found out that large households, with children, are more vulnerable of being poor. A study carried by Hippolyte Fofack (2002) on poverty factors in Burkina Faso indicated that age is significant when predicting poverty and is more important in rural areas compared to urban. According to the author, age remains the strongest predictor of rural poverty. Moreover, household asset ownership is a significant poverty determinant where the probability of poverty decreases as household asset ownership increases.

Generally, salaries and wages represent income, but it can also include other variables such as: tax-exempt interest, taxable interest, dividends, and trusts. Wealth is defined as a household's net worth (Keister & Moller, 2000), that is total assets minus total liabilities. In developing countries, there is a negative correlation between a household size and consumption (or income) (Atkinson, Anthony B., 1987). This means that people living in a large household tend to be poorer than those who live in a smaller household. However, Virola and Martinez (2007) argued that a bigger household size might be good for certain countries. The reasoning for that is driven by economies of scale for certain goods which allow possibilities for sharing. This means that larger households could attain the same level of welfare on a lower per capita expenditure compared to smaller households. According to

Perlman (1976), demographics factors such as: level of education, household size and structure, health, race, age, gender, and labor force determine to large extent poverty status. According to Mullahy and Wolfe (2000), inadequate basic needs which leads us to locality type (rural, urban) may lead households to act inefficiently which would lead to poverty eventually. One of the findings by Baiyegunhi and Fraser (2012) indicated that better-educated households have lower probability to be poor compared to uneducated people. In developed countries, the number of enrollments in primary education is very high as compared to underdeveloped countries. UNESCO recommends spending about 4% of GDP on education for underdeveloped countries. Health is another important feature. In general, developing countries suffer more from health conditions compared to developed countries. This is due to a higher level of diseases, limited resources to social and health protection. Gender inequality tends to be higher in developing countries compared to developed countries which lead to lower chances for women to be productive and hence have a higher chance of being poor.

In Indonesia, better-educated households have lower probability of being poor (Widyanti et al., 2009). Huang (1999), in his study of poverty in Taiwan, concluded that education and wages are correlated and is statistically significant. Head of household years of education significantly reduce the probability of being poor (Meng and Gregory, 2007; Mok et al., 2007).

Servaas van der Berg, in his book “Poverty and Education”, stated that better-educated people have a higher chance of being employed, are more productive, and earn higher incomes than lower educated people. He also argues that the labor

market affects the impact of education on earnings and thus on poverty, education can also affect other areas, such as farming (Orazem, Glewwe & Patrinos, 2007: 5). In the labor market, higher wages for higher educated people may be traced to higher productivity or due to the fact that employers employed better-educated people to obtain good paying jobs.

Zoe Oxaal found out in his study about "Education and Poverty" that females in developing countries typically receive less education than males do. In general, developed countries have greater educational equality for males and females. Whereas among poor countries, there is a significant variation, both in overall levels of enrolment and in female/male enrolment ratios.

In the literature review, (Källestål et al., 2019), setting a decision tree before having classical regression models is important because it enabled assessment of the important variable from a large set of explanatory variables. It also includes and evaluates interactions between explanatory variables. The purpose of having a decision tree is to identify important predictors before utilizing random explanatory variables in classical regression models. In Korea, earned wage has the highest effect on elderly poverty using a decision tree (Park, 2018).

The Palestinian Central Bureau of Statistics (PCBS) conducted a survey called Palestinian Expenditure and Consumption Survey (PECS) in 2017 and used it to estimate poverty in West Bank and Gaza Strip. The following table shows the sample distribution by selected variables.

Table 2.1: Sample Distribution by Selected Variables.

Variables	Household % distribution	Average household size	# of households in the sample
Palestine	**100.0**	**5.5**	**3,739**
WB	64.5	5.2	2,411
Gaza	35.5	6.1	1,328
Type of Locality			
Urban	73.1	5.5	2,732
Rural	17.4	5.4	652
Camp	9.5	5.9	355
Sex of Head of Household			
Male	89.9	5.8	3,363
Female	10.1	3.3	376

Source: PCBS, 2020. PECS 2017

The results of this study showed that individuals in the Gaza Strip have a higher chance of being poor than individuals in West Bank as shown in the table below:

Table 2.2 Poverty Distribution by Region:

Region	Poverty Gap		Poverty Severity	
	Consumption	Income	Consumption	Income
West Bank	2.8	7.9	0.9	4
Gaza	15.7	30.8	6.5	18.1
Palestine	**7.9**	**16.9**	**3.1**	**9.5**

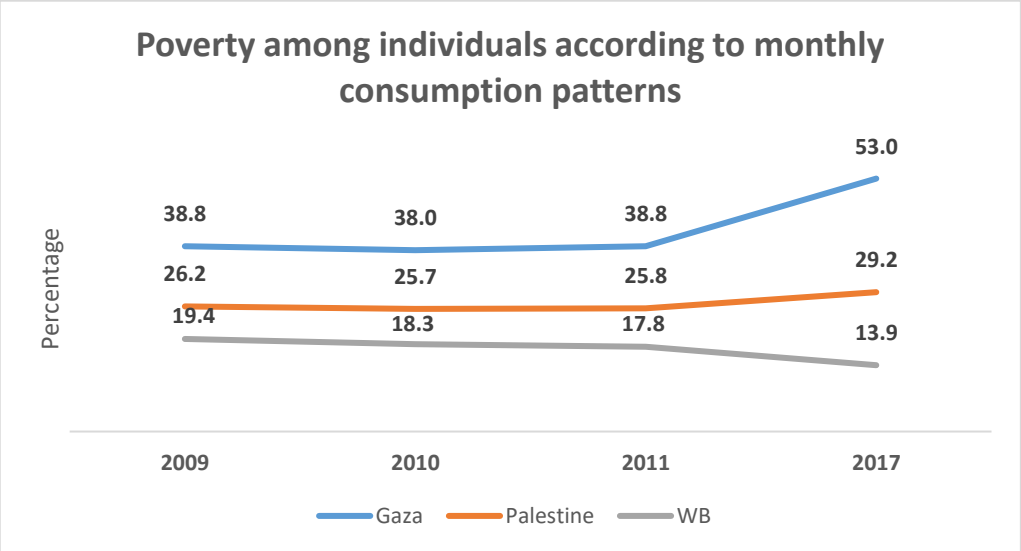
Source: PCBS, 2020. PECS 2017

Poverty gap represents the total amount required to raise consumption levels of the poor to the poverty line. Poverty severity represents mean of total relative squares of poverty gap for all the poor. (PECS, 2017).

On the other hand, the consumption data indicated that poverty spread percentage among Palestinian individuals was 29.2% in 2017: 13.9% in the West Bank and 53.0% in Gaza Strip. While 41.1% of individuals' income was less than the national poverty line: 24.0% in the West Bank and 67.6% in Gaza Strip (PECS,2017).

16.8% of individuals in Palestine were suffering from deep poverty in 2017 (5.8% in the West Bank and 33.8% in the Gaza Strip), while 30.3% of individuals earned an income that was less than the deep poverty line in 2017 (15.1% in the West Bank and 53.9% in Gaza Strip) (PECS,2017)..

According to consumption, data indicated that the overty percentage increased by 13.2% in 2017 compared to 2011.



After conducting the PECS survey, the most vulnerable groups were among the people of the refugee camps. The poverty percentages for refugee camps residents

in 2017 were 45.4%, 29.4% among urban residents, and 18.7% among rural residents. As for deep poverty, 29.3% among refugee camps residents, 16.7% among urban residents, and 9.7 among rural residents.

As anticipated and according to previous literature, household size affects the probability of being poor. Households with more than 10 individuals have the highest poverty percentage of 61.1% vs. households that consist of 2 to 3 individuals with 11.8%. These numbers support the conclusion that poverty rates are higher for bigger households compared to smaller households.

77% of Palestinian households have children. Households that have more children are more vulnerable of being poor. 20.3% of households that consist of 2-3 children were vulnerable to poverty versus 29.6% with a household that consists of 3-4 children and reaches to 60.6% for households with 7-8.

As for the sex of the household head, the study showed that 10.1% of Palestinian households are headed by females. Households headed by females have a 30.6% poverty rate compared to 29.2% for households headed by males. However, in 2011, female headed households have a 25% poverty rate compared to 25.9% for male headed households.

Finally, the study showed that 42.1% of heads of households that are not part of the labor force are vulnerable to being poor versus 25.8% for heads of households enrolled in the labor force. As anticipated and in alignment with the literature, unemployed heads of households had a 59.6% poverty rate while employed head of households had a 24.2% poverty rate.

Chapter 3
Methodology

3.1 Research Approach

This piece of research will predict poverty in Palestine using four different and competitive models. In order to do that, the study is based mainly on The Palestine Expenditure and Consumption Survey; PECS 2017 which is carried by the Palestinian Central Bureau of Statistics (PCBS). The survey data is collected for the period between October 2016 and September 2017. The survey covered urban, rural, and refugee camps populations. Demographic and detailed household information related to asset ownership and housing characteristics. The survey questionnaire covered demographic and social questions about the household, characteristic of the labor force, housing characteristics such as: residence type, number of rooms, social assistance, and income generation.

Initial sample was 5,612, and 3,739 households responded.

A two-stage stratified cluster sample is used as following:

a. A PPS random sample of 391 enumeration areas.
b. A systematic random sample of 12 households were selected from each enumeration area selected in the first stage.

Consumption expenditure or a deprivation index is often used to measure a low standard of living (Pantazis, Gordon, and Levitas 2006). A low standard of living is often associated with a low consumption expenditure. This study will utilize

consumption expenditure to identify the poverty line of a household, poverty status computation is stated in detail in section 3.2.4.

Poverty statistics used in this study is based on the official definition developed by PCBS in 1998. It combines absolute and relative features and is based on the basic needs of a household. Two poverty lines which will be disclosed afterwards are based according to the actual spending patterns of Palestinian households. Deep poverty line which reflects the budget for food, housing and clothing. Poverty line adds other necessities including education, health care, personal care, transportation, and housekeeping supplies. Poverty line and deep poverty have been adjusted based on household size and the number of children in order to reflect the different consumption needs of households. Both terms are based on a budget of basic needs for a family of five persons (two adults and three children). In 2017, the poverty line and deep poverty line for the reference household (two adults and three children) was 2,470 NIS (671 USD) and 1,974 NIS (536 USD) respectively. (The dollar exchange rate during 2017 was 3.68 NIS) (PECS,2017). Another two terms related to poverty are poverty severity and poverty gap. The poverty gap is the gap between the income of the poor and poverty line; poverty severity represents the variation and differentials among the poor (which equals to the mean of the total relative squares of poverty gaps for all the poor). (PECS, 2017).

When identifying households' poverty status based on their consumption expenditure, the poverty status of 429 households (11% of the sample) was adjusted by the researcher from originally being identified as not poor to poor since their

income is found to be lower than the poverty line and their consumption expenditure.

Poverty was applied at the household level and could be presented at either an individual or household level.

3.2 The Models

3.2.1 Data Examination

The research study identifies what variables affect and correlate to poverty and builds prediction models utilizing R software. The study aims to examine what demographical predictors are best to predict poverty. Variance Inflation Factor (VIF) will be used to test for the existence of multicollinearity among the predictor variables. The data gathered from PCBS is applied at the head of the household level, housing, and household conditions. A machine learning classifier, decision tree, will be used by assigning the binary variable *poor* (which takes the value of one if the household is classified as poor and the value zero if the household is classified as non-poor) as the dependent variable. In addition to the decision tree, two logistic regression models and a linear regression model will be applied to the data to identify which model provides the highest accuracy in predicting poverty. In the logistic regressions, the binary variable is used as the dependent variable (outcome) and a set of predictors as independent variables, these predictors vary from model to model and are shown under each model. For a detailed description of variables used, refer to table 6.7 in the appendix (Chapter 6).

Before building the models and during data examination, the researcher used the interquartile method to identify outliers for continuous predictors. Table 3.1 below shows predictors that have outliers and the percentage of outliers from the whole data. Due to their insignificant number, the researcher ignored the outliers after examining their effect on data except for the total expenditures variable where 81 records (2%) were removed from the data set.

Table 3.1: Outlier Statistics for Selected Variables

	Age	**Number of Children**	**Number of rooms**	**Total Expenditures**
Q1	37	2	3	2,456
Q3	56	5	5	6,987
IQR	19	3	2	4,531
Q1 - 1.5 * IQR	8.5	-2.5	0	-8,024
Q3 + 1.5 * IQR	84.5	9.5	8	15,011
Lower than lower limit	0	0	0	0
Greater than the upper limit	24	10	7	81
% of data	1%	0%	0%	2%

3.2.2 Decision Tree

The decision tree is one of the machine learning classifiers used in this study to predict poverty status and to use its output as inputs for model IV. According to Plapinger (2017), decision tree is a non-parametric method meaning that there are no pre-assumptions that needs to be set for the distribution of the errors or the data. Unlike regression models, such as linear regression, which assumes that errors are normally distributed, have a mean of zero, a constant variance and that the observations are independent of each other. Classification Trees uses discrete set of values as target variables. In a decision tree, each node/leaf, represent a class label, non-leaf nodes are features. Regression Trees decision trees use a continuous value for target value. Both, classification and regression trees are commonly known as CART (Classification and Regression Tree).

The goal of a decision tree is to have the optimal choice at the end of each node. An algorithm known as Hunt's algorithm which is both greedy and recursive is used in a decision tree to get optimal choice. It is greedy since it makes the most optimal decision at each node and it is recursive since it splits the larger chunks into smaller chunks. Purity is the decision to split at each node. 100% impure is said when a node is split evenly 50/50 and is said to be 100% pure when all of its data belongs to a single class. A decision tree makes decisions by splitting nodes into sub-nodes and this process is performed multiple times during training process until only homogeneous nodes are left. The best cut-off point makes the two resulting subsets as different as possible with respect to the target outcome. The algorithm continues the search and split recursively until the criteria is met.

Our aim is to optimize our model by reaching maximum purity. In order to measure purity, we use entropy and information gain. Entropy tells us how much a set of data is pure/impure; information gain looks at all nodes together and the expected drop in entropy after the split and is calculated as in the below equation:

Information Gain (IG) = Entropy(parent) - Weighted Sum of Entropy (Children).

The goal is to have it reach zero in order to reach maximum purity.

Over fitting could happen when continuous search by decision tree which increases the chances of finding specific apparent patterns within data which will reduce error in training data at the cost of an increased error in testing data (Quinlan & Cameron-Jones, 1995). To solve the overfitting issue, controls will be applied to decision tree such as pruning and early stopping. Pruning avoids the problem by terminating the splitting process that is likely to overfit the data (Frank, 2000). Early stopping control will be used to have a simplified and easy reading decision tree. To get a good accuracy ten decision trees are constructed. In our model, ten iterations of the decision tree are constructed.

3.2.3 Logistic Regression

A Probit or Logit model is usually used for studying poverty in regression models. A dichotomous dependent variable is used which represents whether a household is poor or not.

A Logit model is a binary model in which the dependent variable Yi (poor) has two values either one or zero and a collection of continuous or categorical explanatory variables Xi, that is:

$$\log\left(\frac{Pi}{1-Pi}\right) = \alpha + \beta Xi$$

Where:

α: Intercepts

β: Parameters

Xi: Vector of predictors

Pi: The probability that the dependent variable equals one.

The dependent variable is known as log odds or Logit and has this form $\log\left(\frac{Pi}{1-Pi}\right)$.

A positive sign of estimated coefficients from the above ratio means that the chance of being poor is higher than that for the reference category, for categorical variables, and also means that this probability increases with the independent variable for continuous predictors keeping other explanatory variables constant. Putting it in other words, a positive coefficient signifies a positive correlation between independent and dependent variables; a negative coefficient signifies a negative correlation.

3.2.4 Building the Models

In order to identify key factors of poverty, the researcher first set a dichotomous variable indicating whether the head of household is poor or not using the poverty line threshold (based on expenditure), identified by PCBS, and is computed relative to the household size. That is:

$$\textbf{Poor} = \begin{cases} 1\ if\ household\ is\ poor \\ 0, otherwise \end{cases}$$

The researcher then conducted a logistic regression model, a linear regression model, a machine learning classifier (decision tree) and another logistic regression model which is based on the results of the decision tree results to identify whether predictor factors were associated with poverty. The study aims to compare the two logistic regression models (model I and model IV), the linear regression model (model II), and the decision tree (model III), and then decide which one has the highest accuracy measures in predicting poverty.

The first model conducted is a logistic regression which used predictors based on previous literature reviews, mainly, the Multidimensional Poverty Index (MPI). MPI, developed by Oxford University and United Nations Development Program, covers people living under minimum internationally agreed living standards, such as food, education, having access to clean water and a sanitation system. Predictors used in the logistic regression are: gender of head of household, age, refugee status, marital status, the highest level of school education, type of school/university, employment status, labor status, whether the household received remittances from abroad, whether the household took a car loan, occupancy type of the dwelling, type of dwelling walls, type of dwelling ceiling, whether the household has a sanitation system, source of energy used for heating and internet availability, the dependent variable is the logit value, $logit\ (p) = \ln\left(\frac{p}{1-p}\right)$, p is the probability that the head of household is poor. The following formula summarizes the first model.

$$logit\ (p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + B_{16} X_{16}$$

The second model is a linear regression model which used the predictors from first model: gender of head of household, age, refugee status, marital status, the highest

level of school education, type of school/university, employment status, labor status, whether the household received remittances from abroad, whether household took a car loan, occupancy dwelling, type of dwelling walls, type of dwelling ceiling, whether the household has a sanitation system, source of energy used for heating and internet availability. Since linear regression is used here, the dependent variable in this regression model is the household monthly expenditure (continuous) rather than the poverty status.

$$Yi = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + B_{16} X_{16} + \epsilon i$$

Where ϵi is the error term with the assumption that it has to be normally distributed with mean zero and constant variance and the explanatory variables are assumed to be independent of each other.

The fitted values from the above linear regression model will then be classified as poor and as non-poor according to the same criteria applied for logistic regression models, the aim of this post classification is to be able to produce the classification table and the accuracy measures.

The third model is the decision tree where it utilized most of the relevant variables found in PECS 2017 Survey. The variables found important predictors of poverty in the decision tree are used by the fourth model.

The forth model is a logistic regression model which used the predictors of the region, whether any family member is unable to visit other family members due to Israeli restrictions, households that work at Israeli working sector, the subjective perspective of household on optimal income required to meet their basic needs, how far is actual income from optimal income, households that receive aid from

international organizations, households that receive social aid, subjective perspective of household on their dwelling, number of rooms excluding bathroom and kitchen, households that took a bank loan to pay their debts; the dependent variable is the logit value, $logit\ (p) = \ln\left(\frac{p}{1-p}\right)$, where p is the probability that the household is poor.

The independent variables in this logistic regression (model IV) will be built upon the results of the decision tree. This logistic regression model uses the enter method for independent variables where all predictors studied are entered into the regression model.

$$logit\ (p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + B_9 X_9$$

All of the models will be applied to data related to the head of the household and housing conditions since we are interested mainly in poverty at the household level. Regression models, unlike a decision tree, indicate the significant relationships between dependent and independent variables and allow us to assess the importance of these variables by interpreting odds ratio, for logistic regression, and coefficient estimates, for linear regression.

3.3 Model Validity

To test the validity of the regression models and the decision tree, data will be divided into a training (80% of data) and a testing (20% of data) sets. Training data will be used to train the model and to estimate its parameters; testing data will be used for validating the model and computing its accuracy. Four indicators: recall

(sensitivity), specificity, precision, and accuracy will be produced to test the validity of each model. Variance inflation factor (VIF) will be used to test multicollinearity between independent variables in the regression models. A rule of thumb for interpreting VIF is variables are said to have no correlation when the value is one, variables are said to be moderately correlated for values between 1 and 5 and are highly correlated for values greater than 5.

Chapter 4
Research Findings

4.1.1 Logistic Regression Model I

The following table shows the variables found significant in this model. Also, it shows the model coefficients estimates, their standard errors, the value of the test statistic (z value), the p-value, and the exponent of the estimated coefficient (odds ratio). The regression model formula is found in table 6.1.

Table 4.1: Logistic Regression Model I Output

Variable	Odds Ratio	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept	1.2830	0.2492	0.2338	1.0660	0.2865
D4Female	1.5828	0.4592	0.2858	1.6070	0.1081
D6	0.9730	-0.0274	0.0048	-5.7390	9.50e-09 ***
D7.1	1.5173	0.4169	0.1071	3.8940	9.86e-05 ***
D7.2	1.1617	0.1499	0.4154	0.3610	0.7183
D8.2	0.3767	-0.9762	0.3819	-2.5560	0.010576 *
D8.3	0.1550	-1.8645	1.2875	-1.4480	0.1476
D8.4	0.8469	-0.1662	0.2999	-0.5540	0.5795
D8.5	1.2170	0.1963	0.4918	0.3990	0.6897
D18.0	2.0723	0.7286	0.2197	3.3160	0.000913 ***
D18.12	0.9933	-0.0067	0.1431	-0.0470	0.9627
D18.13	0.6794	-0.3865	0.2286	-1.6910	0.090858 .
D18.14	0.5609	-0.5782	0.1848	-3.1280	0.001758 **
D18.16	0.3145	-1.1569	0.5277	-2.1920	0.028363 *
D22.2	1.0827	0.0795	0.1578	0.5040	0.6144
D22.3	1.1750	0.1613	0.2367	0.6810	0.4957
D22.4	1.1599	0.1483	0.2287	0.6490	0.5165
L5.1	0.8282	-0.1885	0.2351	-0.8010	0.4229
L5.2	1.2262	0.2039	0.1453	1.4040	0.1603

Variable	Odds Ratio	Estimate	Std. Error	z value	Pr(>|z|)
L5.5	1.6504	0.5010	0.2294	2.1840	0.028980 *
LabourStatus2	4.9701	1.6034	0.2569	6.2420	4.32e-10 ***
LabourStatus3	2.7183	1.0000	0.1586	6.3040	2.90e-10 ***
C03.1	1.0564	0.0549	0.1892	0.2900	0.7717
C07_3.1	0.4189	-0.8702	0.5220	-1.6670	0.095536 .
H2.2	0.6012	-0.5088	0.3637	-1.3990	0.1619
H3.2	0.8092	-0.2118	0.2058	-1.0290	0.3034
H3.3	0.0000	-12.5915	409.8675	-0.0310	0.9755
H3.4	1.5382	0.4306	0.1559	2.7610	0.005758 **
H3.5	0.0000	-13.7530	341.4285	-0.0400	0.9679
H3.6	467,240	13.0546	484.3923	0.0270	0.9785
H12.1	0.5375	-0.6208	0.1377	-4.5100	6.50e-06 ***
H12.2	0.5268	-0.6410	0.2394	-2.6770	0.007423 **
H12.3	0.9392	-0.0627	0.2424	-0.2590	0.7958
H12.5	0.7007	-0.3556	0.2485	-1.4310	0.1523
H13.2	2.9399	1.0784	0.3800	2.8380	0.004543 **
H13.3	0.2889	-1.2415	1.1475	-1.0820	0.2793
H13.4	2.8083	1.0326	0.2730	3.7830	0.000155 ***
H16_7.2	2.8621	1.0515	2.0239	0.5200	0.6034
H16_8.2	0.7065	-0.3474	1.1318	-0.3070	0.7589
H22_2.0	3.9727	1.3795	0.1409	9.7920	< 2e-16 ***
H22_2.1	0.8558	-0.1557	0.1375	-1.1330	0.2573
H22_2.2	0.8057	-0.2161	0.5036	-0.4290	0.6679
H22_2.4	1.6554	0.5041	0.1261	3.9970	6.42e-05 ***
H22_2.6	1.5445	0.4347	0.2843	1.5290	0.1263
H31.1	0.38859	-0.9452	0.0979	-9.6560	< 2e-16 ***
Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**					

To test for multicollinearity, VIF values were produced for the predictors, and shown in table 6.2 in the appendix, no values are higher than 5 therefore there is no multicollinearity between the independent variables.

To test the validity of the model, the coefficients of the prediction model produced using the training data set were applied to the testing data set and several measures of accuracy are produced and put in the following table.

Table 4.2: Logistic Regression Model I Validity Indicators

Recall (Sensitivity)	50%
Specificity	92%
Precision	72%
Accuracy	80%

The above model, as mentioned before, was based on factors and variables stated in the Multidimensional Poverty Index (MPI).

The results of the above logistic regression model showed that MPI variables are important in predicting poverty reflected by a good accuracy rate of 80%, 50% for recall, 92% for specificity, and 72% for precision.

Recall percentage (sensitivity) refers to the ratio of the correctly positive labeled (poor) by prediction model to all who are poor in actual data, therefore our model was able to correctly classify 50% of the poor households as poor. Precision refers to the ratio of actual poor households to households classified as poor by the prediction model. In other words, 72% of households classified by the model as poor are really poor and 28% of them are not. Specificity refers to the ratio of households correctly classified as non-poor by the model to all households who are non-poor in the testing dataset. Our model had 92% for specificity meaning that our model miss-classified only 8% of non-poor households as poor. The model has an

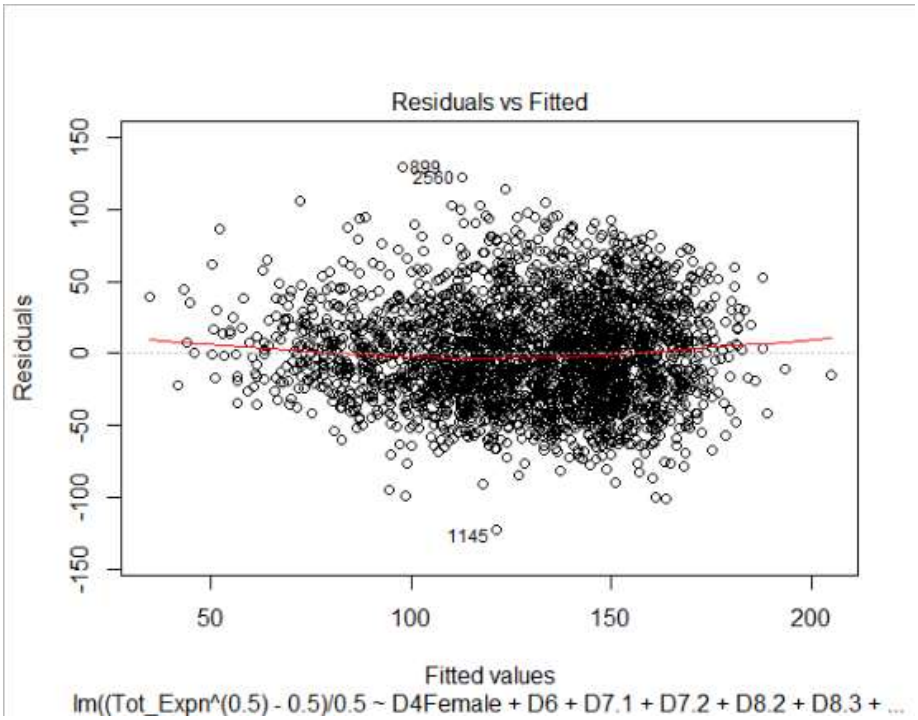
80% accuracy, meaning that it miss-classified 20% of households either as poor or non-poor.

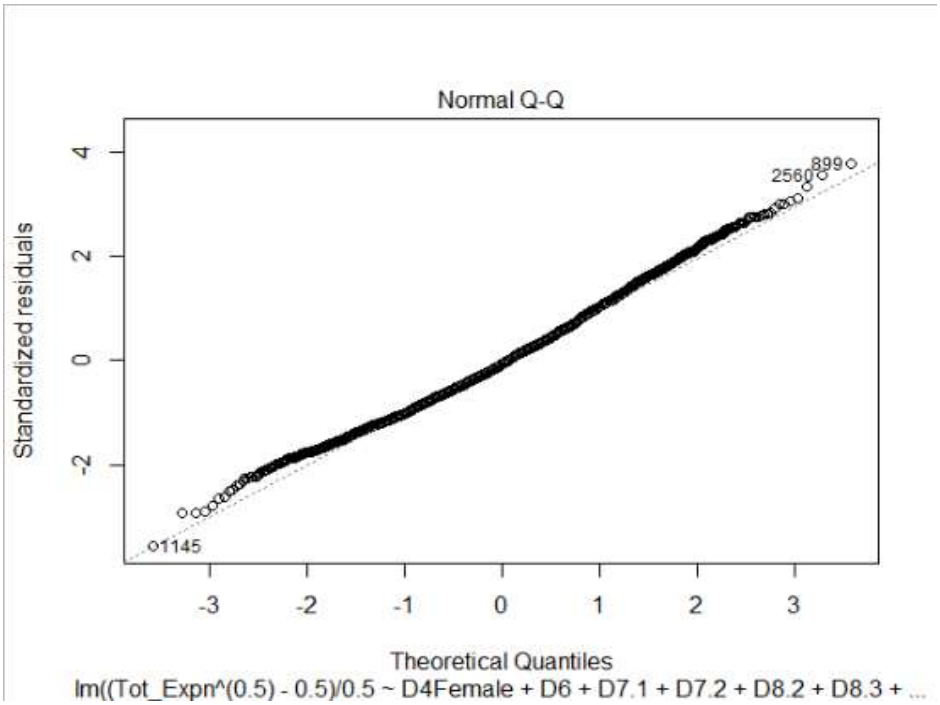
4.1.2 Linear regression Model II:

The following table shows the variables found significant in this model. Also, it shows the model coefficients estimates, their standard errors, the value of the test statistic (t value), and the p-value. The regression model formula is found in table 6.1 in the appendix. The results below are based on a transformed independent variable using power transformation of power 0.5 since the error term in the original regression model was not normally distributed.

To test for multicollinearity, VIF values were produced for the predictors, and shown in table 6.2 in the appendix, no values are higher than 5 therefore there is no multicollinearity between the independent variables.

The chart below shows the residual plot. Residual values and observed values are randomly distributed. The red line indicates that the plot has no unique trend between residual and observed values indicating that the assumptions of linearity and constant variance are met. The Q-Q Plot and the histogram charts below shows that residuals are normally distributed.


Residuals vs Fitted
Residuals
Fitted values
lm((Tot_Expn^(0.5) - 0.5)/0.5 ~ D4Female + D6 + D7.1 + D7.2 + D8.2 + D8.3 + ...
899
2560
1145
150
100
50
0
-50
-100
-150
50
100
150
200


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm((Tot_Expn^(0.5) - 0.5)/0.5 ~ D4Female + D6 + D7.1 + D7.2 + D8.2 + D8.3 + ...
899
2560
1145
4
2
0
-2
-3
-2
-1
0
1
2
3

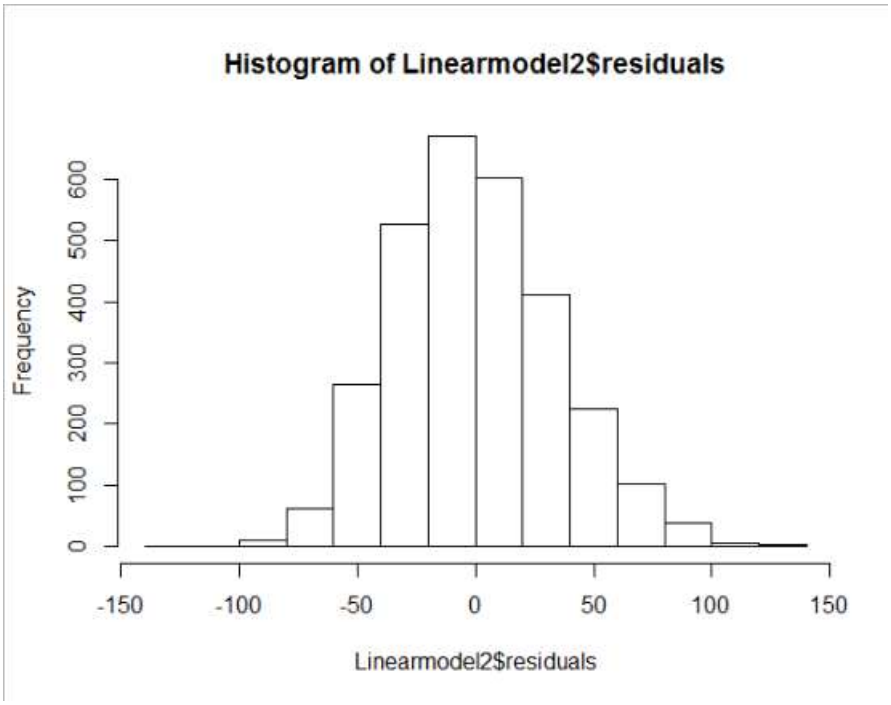


Table 4.3: Linear Regression Model II – Coefficients Estimates

Variable	Estimated Coefficients	Std. Error	t value	p-value
(Intercept	103.78	3.27519	31.687	< 2e-16 ***
D4Female	(2.29)	4.17042	-0.55	0.58218
D6	0.44	0.06463	6.825	1.07e-11 ***
D7.1	(7.02)	1.49186	-4.704	2.67e-06 ***
D7.2	3.01	5.88725	0.51	0.60974
D8.2	(21.30)	5.13861	-4.145	3.49e-05 ***
D8.3	(24.22)	14.24294	-1.7	0.08915 .
D8.4	(12.58)	4.4117	-2.851	0.00438 **
D8.5	(9.51)	7.28226	-1.305	0.19186
D18.0	(24.35)	3.32217	-7.33	2.98e-13 ***
D18.12	(0.13)	1.94928	-0.068	0.94545
D18.13	(1.46)	2.85986	-0.512	0.60876
D18.14	1.77	2.29204	0.772	0.44018
D18.16	15.12	5.04513	2.996	0.00276 **
D22.2	(2.36)	2.28812	-1.03	0.30298
D22.3	1.45	2.94494	0.493	0.62209
D22.4	2.48	2.82263	0.879	0.37923

Variable	Estimated Coefficients	Std. Error	t value	p-value
L5.1	8.81	2.7501	3.204	0.00137 **
L5.2	0.23	1.99596	0.113	0.91017
L5.3	(70.62)	35.41793	-1.994	0.04625 *
L5.5	(7.16)	3.32174	-2.155	0.03125 *
LabourStat	(19.92)	3.71605	-5.361	8.93e-08 ***
LabourStat	(14.05)	2.2018	-6.382	2.03e-10 ***
C03.1	(5.01)	2.60016	-1.925	0.05433 .
C07_3.1	22.28	5.03501	4.425	9.99e-06 ***
H2.2	(1.87)	4.63905	-0.403	0.68688
H3.2	(0.61)	2.63653	-0.231	0.81704
H3.3	0.65	17.36264	0.037	0.97011
H3.4	(10.79)	2.27048	-4.751	2.12e-06 ***
H3.5	4.99	15.83872	0.315	0.75268
H3.6	2.93	20.22637	0.145	0.88493
H12.1	14.25	1.69693	8.397	< 2e-16 ***
H12.2	7.77	2.93427	2.648	0.00815 **
H12.3	(2.38)	3.43101	-0.694	0.48800
H12.5	5.08	3.53636	1.437	0.15096
H13.2	(11.24)	5.18214	-2.169	0.03016 *
H13.3	27.57	14.4458	1.909	0.05641 .
H13.4	(10.02)	3.83953	-2.61	0.00910 **
H16_7.2	22.83	26.7864	0.852	0.39405
H16_8.2	(17.10)	17.52097	-0.976	0.32929
H22_2.0	(23.46)	2.09393	-11.204	< 2e-16 ***
H22_2.1	1.17	1.72633	0.676	0.49930
H22_2.2	(2.02)	5.89323	-0.343	0.73169
H22_2.4	(3.70)	1.79935	-2.056	0.03990 *
H22_2.6	(0.99)	4.43208	-0.224	0.82253
H31.1	26.26	1.43112	18.348	< 2e-16 ***
Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**				
Residual standard error: 34.56 on 2880 degrees of freedom Multiple R-squared: 0.3764, Adjusted R-squared: 0.3667 F-statistic: 38.63 on 45 and 2880 DF, p-value: < 2.2e-16				

Findings:

All of the findings are applied to the head of the household.

- Age (D6): as the age of head of household increases by one year, household monthly expenditure increases by 1 NIS keeping other factors constant.
- Refugee status (D7): being a refugee compared to being not a refugee household head would have a lower expenditure by 9 NIS keeping other factors constant.
- Marital status (D8): a single head of household (never married) or widowed have lower expenditures compared to a married head of household by 103 and 34 NIS respectively keeping other factors constant.
- Education level (D18): head of households that have never been educated have lower expenditures than those that went to school up until eleventh grade by 136 NIS keeping other factors constant. However, head of households with a master's degree have higher expenditures than those that went to school up until eleventh grade by 65 NIS keeping other factors constant.
- Employment status (L5): head of households that are employers tend to have higher expenditures than those with regular wages by 24 NIS, while unpaid and irregular waged head of households have lower expenditures than those with regular wages by 1,212 and 10 NIS respectively keeping other factors constant.
- Labor Force Status: unemployed and out of labor force head of households tend to have lower expenditures than employed head of households by 90 NIS and 43 NIS respectively keeping other factors constant.

- Households that received remittances (C03): Households who received remittances during the time of survey tend to have lower expenditures than those who did not receive remittances by 4 NIS keeping other factors constant.
- Type of loan-Car loan (C07): head of households who took a car loan tend to have higher expenditures than those who did not have a car loan by 135 NIS keeping other factors constant.
- Dwelling Type (H3): households that have a rented dwelling without paying have lower expenditures than those that have an owned dwelling by 24 NIS keeping other factors constant.
- The material of Dwelling Walls (H12): households that have their dwelling's walls made of stone or stone and cement have higher expenditures than those that have their dwelling's walls made of concrete and bricks by 58 and 19 NIS respectively keeping other factors constant.
- The material of Dwelling Celling (H13): households that have their dwelling's ceiling made of fiber cement or metallic have lower expenditures than those that have their dwelling's celling made of concrete by 26 and 20 NIS respectively keeping other factors constant.
- The main source of energy used for heating (H22_2): households that don't have any type of heating energy or use wood compared to those households that use electricity as a main source of energy have lower expenditures by 126 and 2 NIS respectively keeping other factors constant.

- Access to the internet (H31): households that have access to the internet have higher expenditures than households that don't have access to the internet by 186 NIS keeping other factors constant.

As shown above, R-square has a low value of 37% indicating that the linear regression model is not a very good model to apply for predicting poverty. As shown in table 4.4 below, root mean square error (RMSE) has a value of 35 NIS, which measures the difference between values predicted by the model and values observed has a value that is lower than standard deviation of dependent variable. Similar to RMSE, root mean squared prediction error (RMSPE) summarizes the predictive ability of a model and measures the mean of the predicted values and observed value squared has a value of 34.7. The ideal value for the difference between predicted value and true value is zero meaning that the model predictions reflect the true values. The values of RMSE and RMSPE are lower than standard deviation of transformed dependent variable. The RMSPE is very close to the RMSE indicating a good predictive ability and validity of the model.

Table 4.4: Error Measures

RMSE	**RMSPE**	**Standard Deviation of transformed dependent variable**
35	34.7	41

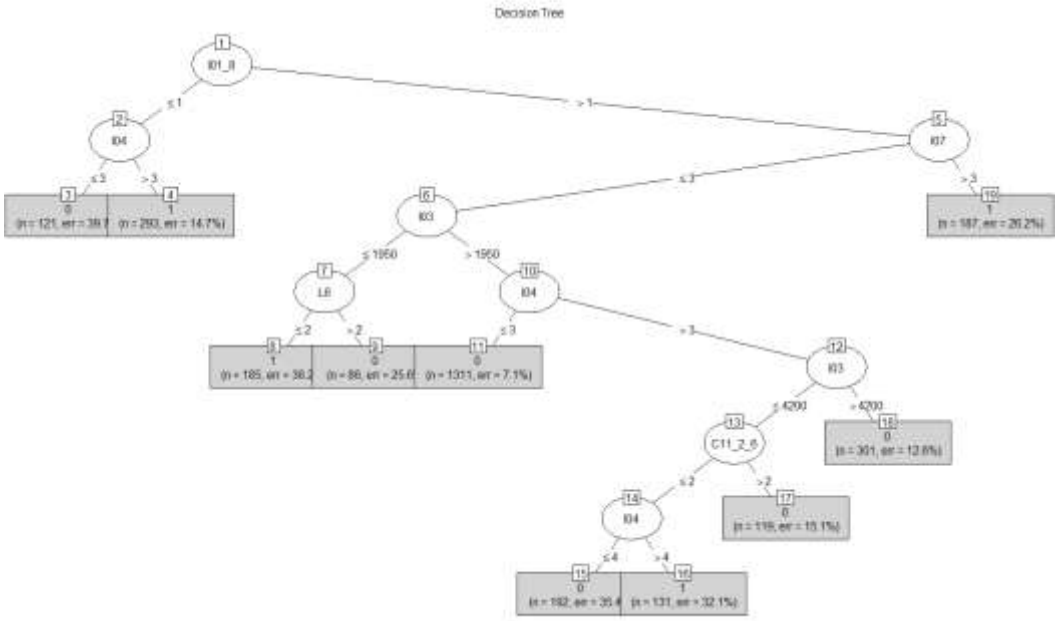
To test the validity of the model, the coefficients of the prediction model produced using the training data set were applied to the testing data set and several measures of accuracy are produced and put in the following table.

Table 4.5: Linear Regression Model II Validity Indicators:

Recall (Sensitivity)	57%
Specificity	96%
Precision	90%
Accuracy	80%

Recall percentage (sensitivity) refers to the ratio of the correctly positive labeled (poor) by prediction model to all who are poor in actual data, therefore our model was able to correctly classify 57% of the poor households as poor. Precision refers to the ratio of actual poor households to households classified as poor by the prediction model. In other words, 90% of households classified by the model as poor are poor and 10% of them are not. Specificity refers to the ratio of households correctly classified as non-poor by the model to all households who are non-poor in the testing dataset. Our model had 96% for specificity meaning that our model miss-classified 4% of heads of non-poor households as poor. The model has an 80% accuracy, meaning that it miss-classified 20% of households either as poor or non-poor.

4.1.3 Decision Tree Model III:



Findings:

As shown in Table 4.7 below, the above tree shows that we can correctly classify 89% of the sample and 71% of the poor and 96% of the non-poor using just the following variables, namely, households that receive aid from international organizations, household's opinion on how much money is needed to meet their basic needs and how far is the current actual income to what they expect it to be, household's opinion on their dwelling, if any female or child in the family exposed to verbal violence; place of the head of household work.

The decision tree diagram does not show all variables showed in the attribute usage table due to the controls applied: pruning and early stopping.

Attribute percentages used in Table 4.6 below represent weight used from each variable to predict the poverty status (binary variable) of a household. Variables used by the decision tree to predict poverty status are found in Table 4.11.

Table 4.6: Decision Tree Attribute Usage

(a) (b)	<-classified as---- ----
1942 102	(a): class 0
230 652	(b): class 1
Attribute usage	**Predictor**
100.00%	Region
100.00%	D10_4
100.00%	I01_5
100.00%	I01_8
100.00%	I01_11
100.00%	I03
100.00%	I04
100.00%	I09
99.42%	I07
90.29%	Children
76.56%	H32_1
74.20%	I01_14
61.21%	H14
57.55%	H10_2
54.78%	H31
51.47%	ID00
43.03%	I05
40.57%	H22_2
38.48%	H6
36.71%	C13_6
35.99%	I01_12
32.64%	C10_9
31.65%	C11_3
29.36%	H3
28.23%	Occupation
27.89%	ID08
24.64%	H23_9

24.20%	Industry
16.13%	H32_2
16.03%	H19_2
15.38%	C11_2_2
15.11%	C11_2_6
9.26%	L6

To test the validity of the model, the coefficients of the prediction model produced using the training data set was applied to the testing data set and several measures of accuracy are produced and put in the following table.

Table 4.7: Decision Tree Validity Measures:

Recall (Sensitivity)	71%
Specificity	96%
Precision	90%
Accuracy	89%

Recall percentage (sensitivity) refers to the ratio of the correctly positive labeled (poor) by prediction model to all who are poor in actual data, therefore our model was able to correctly classify 71% of the poor households as poor. Precision refers to the ratio of actual poor households to households classified as poor by the prediction model. In other words, 90% of households classified by the model as poor are poor and 10% of them are not. Specificity refers to the ratio of households correctly classified as non-poor by the model to all households who are non-poor in the testing dataset. Our model had 96% for specificity meaning that our model miss-classified 4% of non-poor households as poor. The model has an 89%

accuracy, meaning that it miss-classified 11% of households either as poor or non-poor.

As seen in table 6.4 in the appendix, the ten decision trees trials conducted have boosted our tree by having a low error rate of 11.3%. Information gain, table 6.5, shows a sample of variables since it is a long list. Variables with high values are used first for the split in the decision tree.

4.1.4 Logistic regression Model IV:

The following table shows the variables found significant in this model. Also, it shows the model coefficients estimates, their standard errors, the value of the test statistic (z value), the p-value, and the exponent of the estimated coefficient (odds ratio). The regression model formula is found in table 6.1 in the appendix.

Table 4.8: Logistic Regression Model IV Output

Variable	Odds Ratio	Estimate	Std. Error	z value	P-Value
(Intercept	1.2642	2.34E-01	2.42E-01	0.967	0.333456
RegionGaza	2.6545	9.76E-01	1.60E-01	6.107	1.01e-09 ***
C10_9.1	0.5330	-6.29E-01	2.10E-01	-3.003	0.002671 **
C10_9.88	1.1156	1.09E-01	1.93E-01	0.566	0.571368
I01_5.1	0.3209	-1.14E+00	2.29E-01	-4.963	6.95e-07 ***
I03	0.9994	-6.17E-04	4.64E-05	-13.302	< 2e-16 ***
I04.1	0.1997	-1.61E+00	3.01E-01	-5.359	8.37e-08 ***
I04.2	0.5371	-6.22E-01	1.82E-01	-3.413	0.000642 ***
I04.4	1.8887	6.36E-01	1.46E-01	4.358	1.31e-05 ***
I04.5	6.7966	1.92E+00	1.67E-01	11.488	< 2e-16 ***
I01_8.1	1.7464	5.58E-01	1.86E-01	2.996	0.002738 **
I01_9.1	1.3400	2.93E-01	1.45E-01	2.019	0.043537 *
I07.1	0.3504	-1.05E+00	2.77E-01	-3.79	0.000150 ***

Variable	Odds Ratio	Estimate	Std. Error	z value	P-Value
I07.2	0.8609	-1.50E-01	1.27E-01	-1.18	0.237916
I07.4	3.5075	1.26E+00	2.12E-01	5.911	3.40e-09 ***
I07.5	18.8182	2.94E+00	1.06E+00	2.761	0.005758 **
H15_10	1.0563	5.48E-02	4.60E-02	1.191	0.233728
C08_8.1	1.3133	2.73E-01	3.02E-01	0.901	0.367362
C08_8.2	0.6113	-4.92E-01	1.79E-01	-2.755	0.005864 **
Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**					

Findings:

- Region: households that live in Gaza have higher odds to be poor by 165% compared to those living in West Bank keeping other factors constant.
- Inability to visit family/relatives due to Israeli occupation (C10_9): households that face difficulty visiting family/relatives due to Israeli restrictions (checkpoints) have a lower odd to be poor by 47% compared to those that don't face restrictions keeping other factors constant.
- Households that work at the Israeli working sector (I01_5): households that receive their salary from the Israeli work sector have a lower odd to be poor by 68% compared to those that don't keep other factors constant.
- How far is household actual income versus optimal income (I04): households with an actual income much higher than optimal income and slightly higher have lower odds to be poor by 80% and 46% respectively whereas households with an actual income that is slightly less or significantly less than optimal income have higher odds to be poor by 89% and 580% respectively keeping other factors constant.

- Households that receive aid from international organizations (I01_8): households that receive aid from international organizations have higher odds to be poor by 75% compared to those that don't keep other factors constant.
- Households that receive social aid (I01_9): households that receive social aid have higher odds to be poor compared to those that don't by 34% keeping other factors constant.
- Household's opinion about their dwelling (I07): households that answered very good about their dwelling's situation have lower odds to be poor by 65% whereas households that answered poor and very poor about their dwelling's situation have higher odds to be poor by 251% and 1782% respectively keeping other factors constant.
- Loans used to pay debts (C08_8): households that took a loan and did not use it to pay debts have a lower odd to be poor by 39% compared to those that did not take a loan keeping other factors constant.

To test for multicollinearity, VIF values were produced for the predictors, and shown in table 6.2 in the appendix, no values are higher than 5 therefore there is no multicollinearity between the independent variables.

To test the validity of the model, the coefficients of the prediction model produced using the training data set was applied to the testing data set and several measures of accuracy are produced and put in the following table.

Table 4.9: Logistic Regression Model IV Validity Indicators:

Recall (Sensitivity)	66%
Specificity	93%
Precision	78%
Accuracy	85%

Recall percentage (sensitivity) refers to the ratio of the correctly positive labeled (poor) by prediction model to all who are poor in actual data, therefore our model was able to correctly classify 66% of the poor households as poor. Precision refers to the ratio of actual poor households to households classified as poor by the prediction model. In other words, 78% of households classified by the model as poor are poor and 22% of them are not. Specificity refers to the ratio of households correctly classified as non-poor by the model to all households who are non-poor in the testing dataset. Our model had 93% for specificity meaning that our model miss-classified 7% of heads of non-poor households as poor. The model has an 85% accuracy, meaning that it miss-classified 15% of households either as poor or non-poor.

Predictors according to their relative importance are: household opinion about optimal income to meet their basic needs, how far is the actual income to the optimal income, region, household's opinion about their dwelling, households that receive their wages from the Israeli work sector, households that are unable to visit family members due to Israeli restrictions, households that receive aid from international

organizations, households that took a bank loan, households that receive social aid, and number of rooms (excluding bathroom and kitchen)

Table 4.10 below shows a comparison between the regression models and decision tree. The decision tree has the highest percentages for all four indicators.

Table 4.10: Comparison of Validity and Accuracy indicators of the Models

Indicators	Logistic Regression (Model I)	Linear Regression (Model II)	Decision Tree (Model III)	Logistic Regression (Model IV)
Recall (Sensitivity)	50%	57%	71%	66%
Specificity	92%	96%	96%	93%
Precision	72%	90%	90%	78%
Accuracy	80%	80%	89%	85%

Table 4.11: Significant variables for regression models and decision tree

Independent Variables	Logistic Regression (Model I)	Linear Regression (Model II)	Decision Tree (Model III)	Logistic Regression (Model IV)
Gender of the head of household	-	-	N/A	N/A
Age	***	***	X	N/A
Refugee Status	***	***	N/A	N/A
Marital Status	*	***	N/A	N/A
Highest level of school completed	***	***	N/A	N/A
Type of school/university	-	-	N/A	N/A
Employment Status (current\ previous job)	*	*	N/A	N/A
Labor force Status	***	***	N/A	N/A

Independent Variables	**Logistic Regression (Model I)**	**Linear Regression (Model II)**	**Decision Tree (Model III)**	**Logistic Regression (Model IV)**
Households that receive remittances from abroad	-	.	N/A	N/A
What was the type of the loan - Car loan	.	***	N/A	N/A
What is the type of occupancy of this dwelling	**	***	X	N/A
What is the principal material of the walls of this dwelling	***	***	N/A	N/A
What is the principal material of the ceiling of this dwelling	***	**	N/A	N/A
Whether the household has a sanitation system	-	-	N/A	N/A
The main source of energy used for heating	***	***	X	N/A
Internet connection at house	***	***	X	N/A
Region	N/A	N/A	X	***
The inability of any family member to visit parents or relatives or friends because of Israeli procedures	N/A	N/A	X	**
Household opinion on the amount needed to meet their basic needs	N/A	N/A	X*	***
How far is actual income from optimal income	N/A	N/A	X*	***
Households that took a loan to pay debts	N/A	N/A	X	**
Households that receive aid from international organizations	N/A	N/A	X*	**
Households that receive social aid	N/A	N/A	X	*
Number of rooms (excluding bathroom and kitchen)	N/A	N/A	X	-
Households that receive their income from the Israeli working sector	N/A	N/A	X	***
Household opinion about their dwelling	N/A	N/A	X*	***

Independent Variables	**Logistic Regression (Model I)**	**Linear Regression (Model II)**	**Decision Tree (Model III)**	**Logistic Regression (Model IV)**
Head of household holds Israeli health insurance	N/A	N/A	X	N/A
Households that receive national insurance (Jerusalem)	N/A	N/A	X	N/A
Households that receive retirement	N/A	N/A	X	N/A
Monthly rental currency	N/A	N/A	X	N/A
If current income is affected compared to the same period last year	N/A	N/A	X	N/A
Type of bank loan taken from	N/A	N/A	X	N/A
Households that receive aid from family or friends	N/A	N/A	X	N/A
Households that receive property income	N/A	N/A	X	N/A
If any female household member was denied from having their bank account	N/A	N/A	X*	N/A
The main occupation of the head of household	N/A	N/A	X	N/A
How far is the dwelling from a youth club	N/A	N/A	X	N/A
The main industry of head of household work	N/A	N/A	X	N/A
Number of Israeli sim cards owned	N/A	N/A	X	N/A
Number of hours electricity is available	N/A	N/A	X	N/A
If any female or child member exposed verbal violence from family	N/A	N/A	X	N/A
Place of work of the head of household	N/A	N/A	X*	N/A

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ; Not significant '-'**

N/A: variable not used in model

X: variable used in Decision Tree

X*: variables identified important by Decision Tree

Most of the variables found in table 4.11 are in line with previous literature. Variables related to living standards, such as: dwelling type, principal material of dwelling's ceiling and walls, education attainment, and asset ownership support the Multidimensional Poverty Index (MPI) on how these variables can be used to predict poverty. The decision tree output predictors used as inputs in the second logistic regression (model IV) are all significant indicating the importance of setting a decision tree before running a regression model which is in line with the study conducted in Nicaragua (Källestål et al., 2019).

Chapter 5
Conclusion

5.1 Summary

In conclusion, this study focuses on predicting poverty status in Palestine using data provided from The Palestine Expenditure and Consumption Survey; PECS 2017 which is carried by the Palestinian Central Bureau of Statistics (PCBS). It expands on demographic factors by utilizing them in regression models and a machine learning classifier (decision tree) to predict the poverty status of a household. The study finds numerous demographic variables related to the head of household and housing conditions that are good at predicting the poverty status of a household, such as: gender, age, number of children, marital status, education level attainment, principal material of dwelling's walls, principal material of dwelling's ceiling, principal material of dwelling's floor, dwelling sanitation system, labor status, the main source of energy used for heating, and refugee status. The regression model based on decision tree output showed a higher accuracy level than the regression model one based on MPI and came very close to the accuracy of the decision tree. Most of the findings came out in line with previous studies. Regression models and decision tree conducted in this study utilized demographic variables to predict poverty and generated good accuracy levels in predicting poverty. Regardless of not having a very high goodness of fit value, the linear regression conducted in this research showed satisfactory results as shown in table 4.10.

Previous studies conducted by (Achia, Wangombe, & Khadioli, 2010; Fofack, 1990; Geda, Jon, Mwabu, & Kimenyi, 2001; Hashmi, Sial, and Hashmi 2019)

mainly used logistic regression models to predict poverty since the outcome is binary. This research used regression models along with a machine learning classifier (decision tree) which showed high accuracy.

Generally, this research offers predictive models for poverty using regression models and decision tree. Moreover, it examines and utilizes Palestinian demographics to predict the poverty of a household. This research supports and is in line with poverty studies carried by Ministry of Social Development.

5.2 Limitations

Concerning this piece of research, demographic variables used in the models are the factors found in PECS 2017 survey; other demographic variables not listed in the survey, such as: physiological factors such as war or discrimination that might have faced household, could have been used in predicting poverty and may have generated better results. Future studies are recommended to take this limitation into account when predicting poverty.

References

1- Achia, T. N.O., Wangombe, A., & Khadioli, N. (2010). A Logistic Regression Model to Identify Key Determinants of Poverty Using Demographic and Health Survey Data. European Journal of Social Sciences, 13 (1).

2- Aliber, M. (2001). Study of the Incidence and Nature of Chronic Poverty and Development Policy in South Africa: An Overview. SSRN Electronic Journal. DOI: 10.2139/ssrn.1754544.

3- Asif, M., Muneer, T. and Kelley, R. (2007) Life Cycle Assessment: A Case Study of a Dwelling Home in Scotland. Building and Environment, 42, 1391-1394.

4- Atkinson, A. B., (1987). On the Measurement of Poverty. Econometrica, Vol. 55, No. 4, 749-764.

5- Baiyegunhi, L., & Fraser, G. (2012). Vulnerability and Poverty Dynamics in Rural Areas of Eastern Cape Province, South Africa. Ghana Journal of Development Studies, 8(2). DOI: 10.4314/gjds.v8i2.6.

6- Bertranou, E., and Khamis, M. (2005), The Labour Market and Pro-poor Growth in Argentina, 2001-2004, mimeo, Ibero-America Institute, University of Gôttingen, Gôttingen.

7- Burkhauser, R. V. and G. J. Duncan. 1988. Life Events, Public Policy and the Economic Vulnerability of Children and the Elderly. Pp. 29-54 in The Vulnerable, edited by John L. Palmer, Timothy Smeeding, and Barbara Boyle Torrey. Washington, DC: Urban Institute.

8- Carter, M. and May J. (1999). One kind of freedom: poverty dynamics in post-apartheid South Africa. Unpublished paper, University of Wisconsin.

9- CHARLETTE-GUEARD and MESPLE-SOMPS(2001) Comprehensive System of Social Security for South Africa, Viewpoint, South Africa Foundation, July. Johannesburg.

10- Dorantes, C. A. (2004), Determinants and Poverty Implications of Informal Sector Work in Chile, Economic Development and Cultural Change, 52, (2), 347-68.

11- Dreze J. and Srinivasan P. V. (1997). Widowhood and poverty in rural India: Some inferences from household survey data, Journal of Development Economics, 54, (2), 217-234.

12- Federal Reserve Board, Divisions of Research & Statistics and Monetary Affairs. (2009). Ponds and Streams: Wealth and Income in the U.S., 1989 to 2007.

13- Fofack, H. (2002). The Nature and Dynamics of Poverty Determinants in Burkina Faso in the 1990s. Policy Research Working Paper; No.2847. World Bank, Washington, D.C. World Bank. https://openknowledge.worldbank.org/handle/10986/14795 License: CC BY 3.0 IGO.

14- Fosu. A.K. (2015). Growth, inequality, and poverty in sub-Saharan Africa: recent progress in a global context. Oxford Dev. Stud., 43 (1), pp. 44-59

15- Frank, E. (2000). Pruning Decision Trees and Lists. Hamilton: The University of Waikato.

16- Geda, A., Jon, N.d., Mwabu, G., & Kimenyi, M.S. (2001). Determinants of Poverty in Kenya: A Household level Analysis. Working papers 2005-44.

17- GOAED,S. and M.GHAZOUANI (2001) "The determinants of urban and rural poverty in Tunisia," discussion paper, Laboratoire d'Econométrie Appliquée (LEA), Faculté des Sciences Economiques et de Gestion de Tunis, Tunisia.

18- Hashmi, A. A., Sial, M. H., & Hashmi, M. H. (2019). Trends and Determinants of Rural Poverty: A Logistic Regression Analysis of Selected Districts of Punjab. The Pakistan Development Review, 47(4II), 909–923. DOI: 10.30541/v47i4iipp.909-923.

19- Higgins, M., and Williamson, J.G. (2003). Explaining Inequality the World Round: Cohort Size, Kuznets Curves, and Openness, Journal of Southeast Asian Studies, 2003, Vol.40, No.3, pp. 286-302.

20- Huang, T. (1999), The impact of education and seniority on the male-female wage gap: is more education the answer?, International Journal of Manpower, Vol. 20 No. 6, pp. 361-374. https://doi.org/10.1108/01437729910289710.

21- Kabubo-Mariara J. (2002). Labour Force Participation in Kenya. African Journal of Economic Policy. 2002; Vol 9, No. 2.

22- Källestål, C., Zelaya, E. B., Peña, R., Peréz, W., Contreras, M., Persson, L. Å., Selling, K. E. (2019). Predicting poverty. Data mining approaches to the health and demographic surveillance system in Cuatro Santos,

Nicaragua. International Journal for Equity in Health, 18(1). DOI: 10.1186/s12939-019-1054-7.

23- Keister, L. and Moller, S. (2000). Wealth Inequality in the United States. Annu. Rev. Sociol. DOI: 10.1146/annurev.soc.26.1.63.

24- Kitov, I. O. (2006). Modeling the Age-dependent Personal Income Distribution in the USA, ECINEQ Society for the Study of Economic Inequality, Working Paper No. 17.

25- Grootaert, C. (1997), The Determinants of Poverty in Cote d'Ivoire in the 1980s, Journal of African Economies, 6, (2), 169-96.

26- Litchfield, J. and McGregor, T. "Poverty in Kagera, Tanzania: Characteristics, Causes and Constraints," PRUS Working Papers 42, Poverty Research Unit at Sussex, University of Sussex.

27- Maitra, P. (2002). The Effect of Household Characteristics on Poverty and Living Standards in South Africa, Journal of Development Economics, Vol. 27, No. 1, pp. 75-83.

28- Meenakshi, J. V., and Ray, R. (2000). Impact of Household Size and Family Composition in Rural India, ASARC Working Papers 2000-02, Australian National University, Australia South Asia Research Centre, Canberra.

29- Meng, X., and Gregory, R. (2007). Urban Poverty in China and its Contributing Factors, 1986-2000, Review of Income and Wealth, Vol. 53, No. 1, pp. 167-189.

30- Mok, T. Y., Gan C. and Sanyal A. (2007). The Determinants of Urban Household Poverty in Malaysia. Journal of Social Sciences 3 (4): 190-196.

31- Mullahy J., Wolfe, B. L. (2000). Health Policies for the Nonelderly Poor, Focus 21, 2: 32-37.

32- Muyanga, M., Household Vulnerability to Transient and Chronic Poverty: Evidence from Rural Kenya, paper presented at the ninth annual conference of the Global Development Network, 2008, 29-31 January, Brisbane.

33- Orazem, P., Glewwe, P., Patrinos, H. (2007). The Benefits and Costs of Alternative Strategies to Improve Educational Outcomes. Working Papers 7352, Iowa State University, Department of Economics.

34- Oxaal, Z. (1997). Education and Poverty: A Gender Analysis. Brighton: University of Sussex.

35- Oyugi, Lineth Nyaboke, 2000, 'The Determinants of Poverty in Kenya' (Unpublished MA Thesis, Department of Economics, University of Nairobi).

36- Palestinian Central Bureau of Statistics. (2018). Palestine-Expenditure and Consumption Survey of 2017. Retrieved from http://www.pcbs.gov.ps/Downloads/book2368.pdf?date=7_5_2018

37- Pantazis, C., Gordon, D. and Levitas, R. (Eds.) (2006). Poverty and Social Exclusion in Britain: The Millennium Survey. Bristol: The Policy Press. Pp. 488, pbk. DOI: 10.1017/S0047279407301025

38- Park, M.-Y. (2018). Analysis of determinants of poverty in the elderly using decision tree analysis. Digital Convergence Research, 16 (7), 63–69. https://doi.org/10.14400/JDC.2018.16.7.06

39- Perlman, J. E. (1976). The Myth of Marginality: Urban Poverty and Politics in Rio de Janeiro. Retrieved from https://books.google.com/

40- Plapinger, Thomas. "What Is a Decision Tree?" *Medium*, Towards Data Science, 26 Sept. 2017, towardsdatascience.com/what-is-a-decision-tree-22975f00f3e1.

41- Lok-Dessallien, R. Review of Poverty Concepts and Indicators. Doi: http://kambing.ui.ac.id/onnopurbo/library/library-ref-ind/ref-ind-1/application/poverty-reduction/Poverty/Review_of_Poverty_Concepts.pdf

42- Quinlan, J. R. & Cameron-Jones, R. M. (1995). Oversearching and layered search in empirical learning. In Proc of the 14th Int Joint Conf on Artificial Intelligence (pp. 1019–1024). Montreal, Canada: Morgan Kaufmann, San Francisco, CA.

43- Regression Validation (2019, August 30). In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Regression_validation

44- Rodriguez, A. and Smith S. M. (1994). A comparison of determinants of urban, rural and farm poverty in Costa Rica, World Development, 22, (3), 381-397

45- ROUBAUD, F. and RAZAFINDRAKOTO, M.(2003) "The multiple facet of poverty: the case of Urban Africa, Provisional version.

46- Tareen A, Ahmed M, Sikander S, Tahir K, Mirza I, Rahman A. (2008) Feasibility study of a community-based intervention for mental retardation in rural Pakistan. Pakistan Paediatrics Journal, 32, 200-207.

47- Van der Berg, S. (2008). Poverty and Education. The International Academy of Education and the International Institute for Educational Planning (UNESCO).

48- Virola, R. A. and Martinez A. M. (2007). Population and Poverty Nexus: Does Family Size Matter? Paper presented at 10th National Convention on Statistics (NCS), Manila, Philippines.

49- Wagle, U. (2006). The Estimates and Characteristics of Poverty in Kathmandu: What do Three Measurement Standards Suggest?, The Social Science Journal, Vol. 43, No. 3, pp. 405-42.

50- Widyanti, W., Suyahadi, A., Sumarto, S. & Yumna, A. (2009). The Relation Between Chronic Poverty and Household Dynamics: Evidence from Indonesia. SMERU Research Institute Working Paper, Jakarta. URL http://saber.eastasiaforum.org/testing/eaber/sites/default/files/documents/SMERU_Widyanti_2009.pdf.

Appendix

Table 6.1: Regression Models Formulas

Model	Formula
Logistic Regression Model I	$glm(formula = NN{\sim}D4Female + D6 + D7.1 + D7.2 + D8.2 + D8.3 + D8.4 + D8.5 + D18.0 + D18.12 + D18.3 + D18.4 + D18.16 + D22.2 + D22.3 + D22.4 + L5.1 + L5.2 + L5.3 + L5.5 + LabourStatus2 + LabourStatus3 + C03.1 + C07\ 3.1 + H2.2 + H3.2 + H3.3 + H3.4 + H3.5 + H3.6 + H12.1 + H12.2 + H12.3 + H12.5 + H13.2 + H13.3 + H13.4 + H16\ 7.2 + H16\ 8.2 + H22\ 2.0 + H22\ 2.1 + H22\ 2.2 + H22\ 2.4 + H22\ 2.6 + H31.1, family = binomial(\text{logit}), data = trainingdata11., maxit = 100)$
Linear Regression Model II	$lm(formula = (Tot\_Expn\hat{}(0.5) - 0.5)/0.5{\sim}D4Female + D6 + D7.1 + D7.2 + D8.2 + D8.3 + D8.4 + D8.5 + D18.0 + D18.12 + D18.3 + D18.4 + D18.16 + D22.2 + D22.3 + D22.4 + L5.1 + L5.2 + L5.3 + L5.5 + LabourStatus2 + LabourStatus3 + C03.1 + C07\ 3.1 + H2.2 + H3.2 + H3.3 + H3.4 + H3.5 + H3.6 + H12.1 + H12.2 + H12.3 + H12.5 + H13.2 + H13.3 + H13.4 + H16\ 7.2 + H16\ 8.2 + H22\ 2.0 + H22\ 2.1 + H22\ 2.2 + H22\ 2.4 + H22\ 2.6 + H31.1, data = trainingdata11$
Logistic Regression Model IV	$glm(formula = NN{\sim}\ C10\_9.1 + C10\_9.88 + I01\_5.1 + I03 + I04.1 + I04.2 + I04.4 + I04.5 + I01\_8.1 + I01\_9.1 + I07.1 + I07.2 + I07.4 + I07.5 + H15\_10 + D22.4 + L5.1 + L5.2 + L5.3 + L5.5 + LabourStatus2 + LabourStatus3 + C03.1 + C08\_8.1 + C08\_8.2\ family = binomial(\text{logit}), data = trainingdata11., maxit = 100)$

Table 6.2: Model I-VIF

D4Female	**D6**	**D7.1**	**D7.2**	**D8.2**	**D8.3**	**D8.4**	**D8.5**	**D18.0**
4.2785	2.1603	1.3062	1.0394	1.3425	1.0161	3.8615	1.3259	1.4516
D18.12	**D18.13**	**D18.14**	**D18.16**	**D22.2**	**D22.3**	**D22.4**	**L5.1**	**L5.2**
1.1552	1.1486	1.4147	1.0560	1.3206	1.2320	1.2308	1.0798	1.1873
L5.3	**L5.5**	**LabourStatus2**	**LabourStatus3**	**C03.1**	**C07_3.1**	**H2.2**	**H3.2**	**H3.3**
1.0000	1.0815	1.0499	2.4051	1.0681	1.0135	1.0240	1.0730	1.0000
H3.4	**H3.5**	**H3.6**	**H12.1**	**H12.2**	**H12.3**	**H12.5**	**H13.2**	**H13.3**
1.0742	1.0000	1.0000	1.1351	1.0329	1.0923	1.0634	1.0540	1.0142
H13.4	**H16_7.2**	**H16_8.2**	**H22_2.0**	**H22_2.1**	**H22_2.2**	**H22_2.4**	**H22_2.6**	**H31.1**
1.0488	1.4805	1.4910	1.2736	1.2299	1.0275	1.3279	1.0765	1.0944

Table 6.3: Model II-VIF

D4Female	**D6**	**D7.1**	**D7.2**	**D8.2**	**D8.3**	**D8.4**	**D8.5**	**D18.0**
4.149	1.930	1.321	1.032	1.320	1.017	3.684	1.318	1.397
D18.12	**D18.13**	**D18.14**	**D18.16**	**D22.2**	**D22.3**	**D22.4**	**L5.1**	**L5.2**
1.155	1.157	1.467	1.088	1.320	1.252	1.243	1.103	1.164
L5.3	**L5.5**	**LabourStatus2**	**LabourStatus3**	**C03.1**	**C07_3.1**	**H2.2**	**H3.2**	**H3.3**
1.050	1.080	1.052	2.227	1.064	1.022	1.024	1.074	1.008
H3.4	**H3.5**	**H3.6**	**H12.1**	**H12.2**	**H12.3**	**H12.5**	**H13.2**	**H13.3**
1.082	1.048	1.026	1.197	1.051	1.098	1.060	1.083	1.046
H13.4	**H16_7.2**	**H16_8.2**	**H22_2.0**	**H22_2.1**	**H22_2.2**	**H22_2.4**	**H22_2.6**	**H31.1**
1.076	1.800	1.794	1.299	1.254	1.062	1.296	1.061	1.164

Table 6.4: Decision Tree Trials

Trial	Decision tree		
	Size	Errors	
0	10	488	(16.7%)
1	14	538	(18.4%)
2	8	575	(19.7%)
3	14	566	(19.3%)
4	14	580	(19.8%)
5	12	586	(20.0%)
6	13	545	(18.6%)
7	6	562	(19.2%)
8	10	509	(17.4%)
9	9	545	(18.6%)
boost		332	(11.3%) <<

Table 6.5: Decision Tree Information Gain

Attributes	Importance
ID00	0.1100
ID07	0.1087
ID08	0.1007
Locality	0.0995
Region	0.0991
D1	0.0955
D3	0.0941
Children	0.0895
D4	0.0692
D6	0.0665
D7	0.0659
D8	0.0607
D9	0.0606
D10_1	0.0559
D10_2	0.0554
D10_3	0.0516
D10_4	0.0509
D10_5	0.0501
D11	0.0405
D12_1	0.0382
D12_2	0.0356
D12_3	0.0338
D12_4	0.0330
D12_5	0.0328
D13	0.0300
D14	0.0300
D15	0.0289
D16	0.0274
D17	0.0270
D18	0.0268
D19	0.0266
D20	0.0266
D21	0.0251
D22	0.0248
L2	0.0248
L3_A	0.0245
L3_B	0.0239
L3_C	0.0238
L5	0.0238
L6	0.0235

Table 6.6: Model IV- VIF

Region	**C10_9.1**	**C10_9.88**	**I01_5.1**	**I03**	**I04.1**	**I04.2**	**I04.4**	**I04.5**
1.61174	1.07327	1.16911	1.04349	1.26166	1.13516	1.25199	1.40091	1.56761
I01_8.1	**I01_9.1**	**I07.1**	**I07.2**	**I07.4**	**I07.5**	**H15_10**	**C08_8.1**	**C08_8.2**
1.3655	1.0982	1.0867	1.1660	1.1002	1.0090	1.0222	1.0276	1.0476

Table 6.7: All Variables Used

Code	Variable Name	Type	Values
D6	Age	Continuous	20-98
H15_10	Number of Rooms (Excluding kitchen and bathroom)	Continuous	0-12
Region	Region	Categorical	West Bank =1 Gaza =2
D4	Gender	Categorical	Male = 1 Female = 2
D7	Refugee Status	Categorical	Registered=1 Unregistered=2 Not a refugee=3
D8	Marital Status	Categorical	Married= 1 Never Married = 2 Legally Married = 3 Widowed = 4 Divorced = 5
D18	Education Level	Categorical	Never attended school = 0 Grade 1 through 11 = LS 12th grade = 12 Associated diploma = 13 Bachelor's degree = 14 Master's Degree = 16
C10_9	The inability of any family member to visit parents or relatives or friends because of Israeli procedures	Categorical	Yes = 1 No = 2 N/A = 88
C11_2_6	If any child of a family member exposed verbal violence by a family member	Categorical	Yes = 1 No = 2 N/A = 88
C11_3	If a female family member is disallowed to work	Categorical	Yes = 1 No = 2 N/A =88
L5	Employment Status	Categorical	Employer = 1 Self-Employed = 2 Unpaid worker = 3 Regular waged =4 Irregular waged = 5
LaborStatus	Labor Force Status	Categorical	Employed = 1 Unemployed = 2 Out of labor = 3
C07_1	Car Loan	Categorical	Car Loan = 1
C03	Did household receive any remittances during period of survey	Categorical	Yes = 1 No = 2
H23_9	How far is dwelling from the following services usually used by household - Youth club	Categorical	Less than 1 = 1 , 1-5 = 2 , >5 = 3, N/A = 4
H2	Dwelling Occupancy	Categorical	Residence only = 1 Residence & work = 2
H3	Dwelling Type	Categorical	Owned = 1 Rented without furniture = 2 Rented with furniture = 3 Rented without payment = 4 Provided from work = 5

Code	Variable Name	Type	Values
H12	The material of Dwelling Walls	Categorical	Stone = 1 Stone & cement = 2 Old Stone = 3 Concrete blocks = 4 Concrete bench = 5 Mud = 6 Other = 7
H13	The material of Dwelling Ceiling	Categorical	Concrete = 1 Metallic = 2 Wood = 3 Fiber cement = 4
H14	The material of Dwelling Floor	Categorical	Soil = 1 Wood = 2 Tiles = 3 Marble = 4 Cement = 4 Brick/Stone =6 Other =7
H22_2	The main source of energy used for heating	Categorical	None = 0 Gas =1 Kerosene = 2 Electricity = 3 Wood = 4
H7_2	The currency of monthly rental payment	Categorical	N/A = 0 NIS = 1 JOD = 2 USD =3
H31	Internet Connection	Categorical	Yes = 1 No = 2
H6	What is the source of the loan	Categorical	Commercial bank = 1, Islamic Bank = 2, Relative & Friends = 3, Lending institution = 4, Other = 5 N/A = 0
I03	The total amount of money that a household needs to meet its basic needs -household's opinion	Categorical	Yes = 1 No = 2
I04	How far is actual income to optimal income required to meet basic needs	Categorical	1: Much higher than this #, 2: Slightly higher, 3: About same, 4: Slightly less, 5: Much less
I05	If current income is affected compared to the same period last year	Categorical	Much better =1, Somewhat better = 2,: About same = 3, Somewhat worse = 4, Much worse = 5, Don't know = 99
I07	Household's opinion on their dwelling	Categorical	Yes = 1 No = 2
I09	Area of owned land	Continuous	0 - 20 k
C13_6	Households that receive aid from family or friends	Categorical	Yes = 1 No = 2
Occupation	Head of household's work occupation	Categorical	Legislators = 1 Professionals = 2 Service, Shops = 3 Skilled Agriculture = 4 Craft & related work = 5 Plant & Machine operators = 6 Elementary occupations = 7
Industry	The main industry of head of household work	Categorical	Agriculture = 1, Mining =2, Construction = 3, Commerce = 4, Transportation = 5, Services = 6

Code	Variable Name	Type	Values
H32_1	Number of local sim cards owned	Continuous	0 - 15
H32_2	Number of Israeli sim cards owned	Continuous	0 - 15
C11_2_2	If any female household member was denied from having their bank account	Categorical	Yes = 1 No = 2
C08_8	Bank loan used to pay debts	Categorical	Yes = 1 No = 2 0= N/A
L6	Place of work of the head of household	Categorical	Home=1, Same locality = 2, Same governorate = 3, Different governorate = 4, Israel = 5, Settlement = 6, Abroad = 7
H19_2	Number of hours electricity is available	Categorical	Less than an hour = 1, One hour = 2, 2-4 =3, 5-9 = 4, 10-23 = 5, Whole day = 6,
C11_2_4	If any child in a household exposed violence	Categorical	Yes = 1 No = 2
I01_5	Households that receive their income from the Israeli working sector	Categorical	Yes = 1 No = 2
I01_7	Households that receive wages and salaries from Remittances - abroad	Categorical	Yes = 1 No = 2
I01_8	Households that receive aid from international organizations	Categorical	Yes = 1 No = 2
I01_9	Households that receive wages and salaries from Remittances - social aid	Categorical	Yes = 1 No = 2
I01_11	Households that receive wages from Jerusalem (National insurance)	Categorical	Yes = 1 No = 2
I01_12	Households that receive wages from property income	Categorical	Yes = 1 No = 2
I01_14	Households that receive wages from retirement	Categorical	Yes = 1 No = 2
D10_4	Households that have health insurance provided by an Israeli insurance company	Categorical	Yes = 1 No = 2