Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-021-01026-y OPEN

Identifying the roots of inequality of opportunity in South Korea by application of algorithmic approaches

Seungwoo Han [1]✉

This study identifies the roots of inequality of opportunity in South Korea by applying algorithmic approaches to survey data. In contrast to extant studies, we identify the roots of inequality of opportunity by estimating the importance of variables, interpreting the estimated results, and analyzing the importance of individual variables, instead of measuring inequality of opportunity. We apply a decision tree classification algorithm, light gradient boosting machine, and SHapley Additive exPlanations to estimate the importance of the studied variables and interpret the estimated results. According to the estimated results, the region where the individuals grew up, their gender, and their father's job during their childhood were the main factors contributing to inequality of opportunity. This study proves that the considerable regional disparity and social environment perpetuate gender inequality in South Korean society. It argues that an individual's socio-economic achievements are strongly influenced by their father's background, thus, outweighing other family background-related factors. Individuals receive unequal opportunities owing to a combination of region, father's background, and their own gender, thereby, affecting their socioeconomic achievements. If these factors remain influential from birth to adulthood, removing the conditions that structure them would be one way to achieve equality of opportunity.

[1] Division of Global Affairs, The State University of New Jersey, Rutgers, NJ, USA. ✉email: seungwoo.han@rutgers.edu

Introduction

As a result of an increase in inequality after the Asian financial crisis in 1997, several studies from various perspectives on inequality have been conducted on South Korean society (Birdsall, 2000; Koo, 2007; Kang and Yun, 2008; Lee et al., 2012; An and Bosworth, 2013; Kang and Rudolf, 2016; Koh, 2019). Inequality has recently been the subject of Korean cultural content, for example, the movie *Parasite* and the TV series *Squid Game* (Jang, 2021; Jin, 2021; Shin, 2020). Among all the issues related to inequality, the one that has become a sensitive topic for young Koreans is social awareness regarding the unfair distribution of opportunities due to recent political issues (Ban and Kang, 2021; Lee, 2019; Shim, 2019).

To formulate related policies, it is necessary to analyze the origin of inequality of opportunity. This study aims to identify the roots of inequality of opportunity in South Korea by applying algorithmic approaches to survey data. Specifically, it applies a decision tree classification algorithm, light gradient boosting machine (LightGBM), and SHapley Additive exPlanations (SHAP) to estimate the importance of the studied variables and to interpret and analyze the results.

According to Rawls' (1971) definition of justice, equality of opportunity is an ethical value that enables members of society to pursue their interests through fair and equal opportunities. According to Roemer (1998), equality of opportunity begins with a discussion of factors that individuals cannot be held responsible for, called *circumstances*, and factors that individuals have control over and take responsibility for, called *effort*. Generally, there are two approaches to measure inequality of opportunity. One is the ex-ante utilitarian perspective, in which the value of opportunity sets is indicated by the average outcome within the specific *type*. Individuals sharing the same *circumstances* are regarded as a *type*, and if socio-economic disparities between *types* arise from *circumstances*, then the result is considered a consequence of inequality of opportunity. This *between-type* inequality corresponds to a weak criterion of ex-ante inequality of opportunity (Ferreira and Peragine, 2015; Fleurbaey and Peragine, 2009). The other is the ex-post view that focuses on individual outcomes, conditional on *effort* exertion (Fleurbaey, 1998; Fleurbaey and Peragine, 2009). According to this perspective, equality of opportunity would be satisfied if individual outcomes are equalized within groups exerting the same effort. Individuals with equal levels of effort exertion realize the same outcomes. This study focuses on inequalities between social groups defined by the set of *circumstances* from an ex-ante utilitarian perspective.

Various approaches have been used in empirical studies to estimate the inequality of opportunity and measure its impact. Among these, the parameter-based approaches, which depend on statistical assumptions of variables, have been limited by bias and model selection (Balcazar, 2015; Roemer and Trannoy, 2016; Brunori et al., 2019a). Moreover, nonparametric test approaches, which partition the sample into each *type*, have been criticized for arbitrary segmentation (Brunori et al., 2019a). However, this nonparametric approach of the decision tree classification algorithm is free from bias and model selection because it does not make assumptions about parameters and linearity. Furthermore, it partitions the sample using a machine learning algorithm instead of arbitrary segmentation.

The decision tree approach was mentioned in Brunori et al.'s (2018) study, which is consistent with the theory proposed by Roemer (1998). Several studies used this method to measure inequality of opportunity in sub-Saharan African countries (Brunori et al., 2018, 2019b), European countries (Brunori and Neidhofer, 2020), and India (Lefranc and Kundu, 2020), and compared the results with classical parametric and nonparametric approaches.

In contrast, this study aims to identify the roots of inequality of opportunity by estimating the importance of variables, interpreting the estimated results, and analyzing the importance of individual variables, instead of merely measuring inequality of opportunity. Moreover, unlike existing studies that use a regression tree to measure inequality of opportunity across all ages, this study uses a classification tree to group people based on a specific criterion. To identify the roots of inequality of opportunity, a specific group that has lived in a similar era and social environment must be analyzed, and a criterion necessary to define the socio-economic disparity between *types* must be identified. For instance, if the specific group is *millennials* and the criterion for the disparity is minimum wage, then the importance of circumstance variables can be estimated based on the minimum wage in the binary classification process of that specific group.

This study utilizes the decision tree classification algorithm in analyzing data, which is known to have overfitting and instability problems; it also utilizes the LightGBM, which overcomes the drawbacks of the decision tree classification algorithm, through the boosting learning method.

While the tree-based models internally calculate the importance of the values of variables, these values may vary depending on how the importance of the variable was computed. SHAP, an algorithm based on game theory, allows consistent estimation of the importance of variables (Lundberg and Lee, 2017a, 2017b). Although LightGBM makes it difficult to interpret the results without knowledge of the process between the input and output of data, like a black box, SHAP overcomes this limitation by allowing interpretation of the estimated results and analysis of the importance of individual variables.

This study also utilizes country-specific circumstance variables provided by the Youth Panel Survey, which reflect *circumstances* and economic activities of the youth in South Korea. Python 3.7 and Scikit-learn 0.22.2 are used to analyze the data.

The remainder of this paper is organized as follows: section "Background and empirical approaches" describes the background and empirical approaches; section "Methodology" describes the methodology; section "Results and discussion" reports the results of the analysis; and section "Conclusion" concludes the paper.

Background and empirical approaches

Background. According to Rawls (1958, 1971), in an egalitarian theory, justice can be understood as an endeavor to replace equality of results with equality of opportunity. This political philosophy of favoring equality of opportunity can be explained using metaphors, such as "levelling the playing field" or "equality at the starting gate." The just society envisioned by Rawls is a society in which members are given fair and equal opportunities to pursue their interests. Rawls' theory of justice is not directly related to an individual's welfare level but is about the conditions that structure it.

Other philosophical contributions to this discourse were provided by Sen (1980), Dworkin (1981a, 1981b), Arneson (1989), and Cohen (1989). Rawls' emphasis on primary goods, Sen's capability approach, Dworkin's view of equitable resources, and Arneson and Cohen's individual responsibility and equal opportunity take a slightly different view of equality. Nevertheless, they all value equality of opportunity. In other words, the equality they seek guarantees equal opportunity for each member of society to achieve their desired results.

The discussion that emerged after Rawls presented an ethical justification for equality of opportunity changed the perception of equality and contributed to the philosophical debate surrounding egalitarianism and development. Roemer (1993, 1998) and

Fleurbaey (1994, 2008) systematized the measurement of inequality of opportunity through a more precise definition of equality of opportunity. According to their theory, the measurement of inequality of opportunity begins by defining the factors that fall under individual responsibility, and those that do not (Roemer, 1998).

According to Roemer's (1998) theory, if opportunities are evenly distributed, the consequences of an individual's choice may be outside the influence of social justice. In other words, the level of equality must be measured not only by the degree of inequality that is currently observed but also by information on the source of its results. Given the same level, this inequality may be the result of personal responsibility or the result of factors beyond personal responsibility, also known as *circumstances*, such as gender or parental background, which are not considered personal responsibility. Thus, the socio-economic gap resulting from differences in *circumstances* can be interpreted as resulting from inequality of opportunity.

The Korean context. South Korean society can be divided into several generations: people who were in extreme poverty in the 1950s and the 1960s immediately after the Korean War; people who experienced rapid economic growth in the 1970s and the 1980s; people who were affected directly by economic shocks during the Asian financial crisis in the 1990s; people who entered the labor market after the crisis; and finally, the *millennials*. Based on both external factors and their respective social environments, each generation is bound to be significantly different from the others (Scitovsky, 1985; Cho and Kim, 1991; Fields, 1994; Lim and Jang, 2006; Kim, 2007; Koo, 2013; Lee, 2017). *Millennials* share a similar social environment, and the influence of other external factors is relatively small, considering their short period of experience in the labor market.

According to Deloitte (2021), in a survey of 45 countries, 73% of the South Korean *millennials* who were surveyed answered that wealth was distributed "not fairly equally" or "not at all equally," which was higher than the global level (69%). These South Korean *millennials* are witnessing the emergence of a new status order in their society. As a result, the terms "the dirt spoon" and "the gold spoon" are now widely used (Kim, 2017). The English idiom "born with a silver spoon in one's mouth" has been adopted by South Korean society, which led to the spoon class theory discourse. This theory refers to the idea that an individual's socio-economic achievement is determined by one's parents' income and family background, regardless of one's efforts.

Since 2015, the spoon class theory has been primarily used by *millennials* in online communities in South Korea (Kim, 2017). Despite its ambiguous origin and implications, the theory clearly indicates a widespread social perception, particularly among young people, that opportunities are not equally available for everyone. Such self-ridiculing discourse reveals the depth of the younger generation's animosity toward society. In South Korean society, aside from inadequate compensation for effort exertion, the younger generation is highly dissatisfied with extreme disparities in *circumstances* between social groups.

In this study, the wage is set as a socio-economic achievement, and any level below the minimum wage is considered the most adverse socio-economic condition. Minimum wage refers to the minimum remuneration paid to wage earners to sustain themselves in society (Starr, 1981; Neumark and Wascher, 2008). It is directly linked to constitutional values, such as human freedom and quality of life. Therefore, living below the minimum wage implies living under the most unfavorable socio-economic conditions. By analyzing the *circumstances* of this adverse condition, the roots of inequality of opportunity can be identified, thereby enabling an understanding of the *circumstances* that produce the highest inequalities.

Empirical approaches. This study considers that the wage level of an individual is determined by *circumstances*, *effort*, *social policy*, and *luck* (Arneson, 1989; Cohen, 1989; Fleurbaey, 1994; Roemer, 1998, 2008; Lefranc et al., 2009; Ferreira and Peragine, 2015; Roemer and Trannoy, 2016). According to Roemer (1998), an individual's achievement is determined by *effort*, depending on their responsibilities and *circumstances*, which are beyond their control. This study analyzes *circumstances* that create such socio-economic disparities and assumes the following theoretical conditions.

The ultimate wage level of individual y_i is created as a function of *circumstances* C_i, *effort* e_i, social policy φ_i, and *luck* l_i, in Eq. (1). This considers wages and socio-economic achievement as a vector of *circumstances*, $y_i \in Y$, including a finite number of elements, $c_i \in C_i, C_i \in C$, responsibility characteristics, denoted as *effort*, $e_i \in e$, social policy related to individual wage level, $\varphi_i \in \phi$, and *luck*, l.

$$y_i = f(C_i, e_i, \varphi_i, l_i). \tag{1}$$

If equality of opportunity is achieved, *circumstances* do not affect the wage level, according to the condition presented in Eq. (2):

$$\frac{\partial f(C_i, e_i, \varphi_i, l_i)}{\partial C_i} = 0, \forall C_i. \tag{2}$$

According to the condition presented in Eq. (3), the *effort* is distributed independently from the *circumstances*:

$$G(e_i|C_i) = G(e_i), \forall e_i, \forall C_i. \tag{3}$$

In the case of *luck*, suppose that $y_i = f(C_i, e_i, l_i)$ is the function of an individual's wage-generating process. When *circumstances* and *effort* are given, the distribution of wages can be identified, where $H(y|C_i, e_i)$ and Fc_i, e_i would be the distribution of *luck*, as shown in Eq. (4). If the condition presented in Eq. (2) holds, individuals face similar prospects according to their *efforts*, regardless of *circumstances* (Lefranc et al., 2009; Roemer and Trannoy, 2016). In other words, the distribution of *luck* is even-handed, irrespective of *circumstances*. This allows the distribution of *luck* to be dependent on *effort* and independent of *circumstances*, as shown in Eq. (5).

$$H(y|C_i, e_i) = F_{C_i, e_i}\left(f^{-1}(y, C_i, e_i)\right), \forall C_i, \forall e_i. \tag{4}$$

$$H(.|C_i, e_i) = H\left(.\middle|C_j, e_i\right) = K(.|e_i), \forall e_i, \forall\left(C_i, C_j\right). \tag{5}$$

Under the condition presented in Eq. (6), *policy* φ_i is independent of C_i. Assuming that C_{j-1} is the most disadvantaged *circumstance* in society and C_j is one level higher, they do not have the same distribution of wages under *policy* φ_i as presented in Eq. (7).

$$P(\varphi_i|C_i) = P(\varphi_i). \tag{6}$$

$$J\left(\cdot\middle|C_j, \varphi_i\right) \neq J\left(\cdot\middle|C_{j-1}, \varphi_i\right). \tag{7}$$

If the above conditions are established and the outcome generating function of *circumstances* and *effort* is f: $C_i \times e_j \rightarrow \mathbb{R}_+$, the ultimate wage level of an individual can be rewritten as presented in Eq. (8):

$$y_i = f(C_i, e_i). \tag{8}$$

This results in a sample of individuals, each of whom is characterized by *effort* e_i and the vector of *circumstances* C_i.

Table 1 *n* circumstance-by-*m* efforts matrix.

	e_1	e_2	e_3	e_4	…	e_m
C_1	y_{11}	y_{12}	y_{13}	y_{14}	…	y_{1m}
C_1	y_{21}	y_{22}	y_{23}	y_{24}	…	y_{2m}
C_3	y_{31}	y_{32}	y_{33}	y_{34}	…	y_{3m}
C_4	y_{41}	y_{42}	y_{44}	y_{44}	…	y_{4m}
…	…	…	…	…	…	…
C_n	y_{n1}	y_{n2}	y_{n3}	y_{n4}	…	y_{nm}

Assuming *n circumstances* and *m efforts*, the *n-by-m* matrix can be expressed as shown in Table 1.

In the matrix, the sample can be divided into *types* T_i, which share the same level of *circumstances* and *trenches* T^i, which share the same degree of *effort*. *Types* T_i and *trenches* T^i are at the core of the two approaches for measuring the inequality of opportunity: The first is the ex-ante approach that focuses on inequality between *types* categorized by the conditions of *circumstances*, and the second is the ex-post approach that focuses on inequality in the outcomes of individuals based on *effort* exertion (Fleurbaey and Peragine, 2009). In other words, the ex ante approach focuses more on inequalities between social groups defined by the set of *circumstances* shared by members of each group. Conversely, the ex-post approach focuses more on outcome inequalities among individuals who exert the same effort. This study aims to examine the differences between a specific group in the most adverse social condition and others, and to analyze the *circumstances* that structurally produce inequality in society. Accordingly, in the context of current study, this study restricts itself only to an *ex-ante* utilitarian perspective, where an individual $x_i \in \{1,...,N\}$ constitutes a *type* that shares the same *circumstances*, and its society is classified into finite *types*, $T_i = \{t_1, t_2, ., t_n\}$. These classified *types* are mutually exclusive, and each t has a distribution of p^t in the sample.

The problem with an empirical approach that only considers observable circumstances is that the residual domain is not included in the model (Lefranc et al., 2009). If the condition presented in Eq. (3), $G(e_j|C_i) = G(e_j), \forall e_j, \forall C_i$ is valid, the condition presented in Eq. (9) is also established. Therefore, a consistent approach is only possible under observable *circumstances*.

$$F\left(.\middle|\widehat{C}_i\right) = F\left(.\middle|C_i\right), \widehat{C}_i \subseteq C_i. \tag{9}$$

Observable variables of *circumstances* are generally a subset of the actual variables ($\widehat{C} \subseteq C$), thereby affecting an individual's socio-economic achievement. Hence, an estimation can be made only with the observed variables, even if not all the variables in the *circumstances* are considered. This serves as an advantage when using the nonparametric decision tree classification algorithm, which will be explained in more detail in the section "Methodology".

Classification of types. This study uses the *between-type* inequality approach (Bourguignon et al., 2007; Ferreira and Gignoux, 2011; Lefranc et al., 2009; Kanbur and Snell, 2017). Assume that there are *n circumstances*, $i \in \{I,...,n\}$, and *m* $effort_j \in \{1,...,m\}$, the *between-type* inequality approach calculates first the mean $Y_\mu = \{\mu_1, \mu_2,...,\mu_n\}$ of the values of each *typet*, as shown in Eq. (10). This eliminates inequality within each *type* and maintains the inequality between *types*.

$$\mu_i = \frac{1}{N_m} \sum\nolimits_{x_i \in t_i} y_{ij}, \forall x_i, \forall t_i. \tag{10}$$

If there is a gap between *types*, there is an inequality of opportunity because of the difference in *circumstances* ($c_i \in C_i$, $C_i \in C$). Applying this to the approach of this study, every μ is classified as a binary based on minimum wage.

$$k_i = \mu_i < \text{min\_}wage \text{ or } \mu_i \geq \text{min\_}wage, \forall k_i \in \{0, 1\} \tag{11}$$

Every μ has a value of 0 or 1, depending on the condition presented above in Eq. (11). In other words, every *type* is classified as 0 or 1, which can be considered the result of the difference in *circumstances*.

Methodology

Methodological background. The empirical literature on measuring the inequality of opportunity chooses either parametric or nonparametric tests. Equation (12) shows the reduced form of the regression model proposed by Bourguignon et al. (2007), Trannoy et al. (2010), Ferreira and Gignoux (2011), and Singh (2011), which measures the impact of observable *circumstances* through parametric tests. Let y_i be the socio-economic achievement of an individual i and C_i be the vector of *circumstances*.

$$y_i = \beta C_i + \varepsilon. \tag{12}$$

In this model, both the direct and indirect effects of *circumstances* on y_i are captured by the regression coefficients through their effects on *effort* (Ferreira and Peragine, 2015). The problem is that such a model cannot include every circumstance variable. Thus, the model has a downward bias, which can undermine the inequality of opportunity (Balcazar, 2015; Brunori et al., 2018). Therefore, the coefficient cannot be considered causal.

To solve this problem, researchers generally attempt to reduce downward bias by including more variables, such as interaction variables and higher-order polynomials, into the equation. However, this increases variance and causes an upward bias (Ferreira and Peragine, 2015; Brunori et al., 2019a). Moreover, researchers have to choose a model to deal with the above issues, which can be an important factor in determining the outcome when testing and measuring inequality of opportunity.

In contrast, nonparametric models are free from these issues. For nonparametric tests, inequality of opportunity can be estimated by considering only the observed circumstance variables (Checchi and Peragine, 2010). Researchers can divide the sample into mutually exclusive *types* based on all the variables being considered. Therefore, the advantage here is that a study does not have to make assumptions about the interaction of variables when analyzing the results.

However, the limitations of nonparametric tests arise when a small sample size is divided into mutually exclusive *types*, thereby causing an overestimation of results. Therefore, it is necessary to split enough observations into each *type* (Brunori et al., 2018, 2019a) to ensure the reliability of the estimates. However, in reality, individuals are not evenly distributed among the *types*. Therefore, during the process of dividing the entire *type* arbitrarily, the number of circumstance variables should be extremely limited while considering the balance between variables.

The decision tree analysis has been proposed as a data-driven method to overcome the limitations described above (Brunori et al., 2018; Brunori and Neidhofer, 2020). It is classified as a nonparametric machine learning method because it is not based on a probability density function (Murthy, 1998). Moreover, it does not make statistical assumptions about parameters and does not imply that the underlying relationship between variables is linear (Hastie et al., 2009; Murphy, 2012; Hegelich, 2016). These advantages enable the use of only observed circumstance variables in the analysis (Brunori et al., 2018; Brunori and Neidhofer, 2020). Therefore, it is free from potential endogeneity in the model. In addition, it is not necessary to limit the number of variables in the partitioning sample because partitioning is possible through an algorithm, rather than arbitrary judgment.

The decision tree algorithm is divided into classification and regression trees, also known as CART (Hastie et al., 2009). Identifying which method is more appropriate depends on the purpose and empirical approach of the study. The decision tree classification algorithm predicts or explains the response of a categorical dependent variable (Hastie et al., 2009). It divides the sample into mutually exclusive regions based on a specific classification condition. This study aims to estimate the importance of variables in a binary classification according to the ex-ante approach, based on the most unfavorable condition in society. To analyze the inequality of opportunity based on a particular group, a specific criterion that defines the socio-economic disparity with other groups is necessary. If the disparity is based on minimum wage, then estimating the importance of variables in the binary classification based on it is logically consistent with the purpose of this study.

Tree-based classification approaches. In this section, the process by which the decision tree classification algorithm learns the classification rules is examined. The importance of variables in partitioning the sample into mutually exclusive *types* can be computed. Further, LightGBM can overcome the drawbacks of the decision tree classification algorithm.

First, the process by which the decision tree classification learns classification rules is examined. It finds if/else statements in each variable area and splits it repeatedly to create rules in the entire area. The process is explained as follows (Hastie et al., 2009):

$$\widehat{p}_{mk} = \frac{1}{N_m} \sum\nolimits_{x_i \in R_m} 1_{y_i \in R_m}. \tag{13}$$

The data consist of $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and each observation is $(x_i, y_i) \in \mathbb{R}^{p+1}$, $(i = 1,...,n)$. When the target classification result has two classification values, the decision tree classification method as the predicted value of the dependent variable y as a function of the explanatory variables, $I = \{I^1,...,I^p\}$. The method partitions the sample into mutually exclusive regions R_1, R_2, R_m using explanatory variables I. In node m, representing a region R_m with N_m observations, Eq. (13) above indicates the proportion of class k. Depending on the proportion of k, node m has a classification value of 0 or 1 as shown in Eq. (14).

$$k(m) = arg\ max_k \widehat{p}_{mk} \tag{14}$$

Considering the partitioning process, based on the growing-splitting rules, the decision tree algorithm repeats the growing and splitting of categories into two regions. The rules were created based on the splitting criteria of the data. Depending on how well the distribution of the target variable (dependent variable) is distinguished, purity or impurity of data is used. In the two-child tree nodes, the variable that maximizes the sum of purity or minimizes the sum of impurity is selected as the splitting criterion (Hastie et al., 2009).

Specifically, the model finds the best condition for partitioning a dataset that results in either the largest sum of purity or the smallest sum of impurity; where a low degree of *cross-entropy*, as shown in Eq. (15), represents the degree of congestion of the data, or where a high degree of *Gini index*, as shown in Eq. (16), indicates the uniformity of the data. After splitting iteratively across child nodes, if all data belong to a specific classification, the partitioning stops, and the classification is determined (Hastie et al., 2009).

$$\text{Cross} - \text{entropy}: -\sum\nolimits_{k=1}^{K} \widehat{p}_{mk} \log \widehat{p}_{mk}. \tag{15}$$

$$\text{Gini index}: \sum\nolimits_{k \neq k'} \widehat{p}_{mk} \widehat{p}_{mk'} = \sum\nolimits_{k=1}^{K} \widehat{p}_{mk} \left(1 - \widehat{p}_{mk}\right). \tag{16}$$

According to the classification principle, this algorithm performs internal variable selection, which is an integral part of the procedure. Further, this algorithm splits the entire region into mutually exclusive regions using the explanatory variables of I, which are the splitting criteria. In other words, the decision tree classification algorithm can compute the importance values of variables in the process of dividing the entire region into mutually exclusive regions.

When the condition presented in Eq. (3), $G(ej|C_i) = G(e_j)$, $\forall e_j$, $\forall C_i$ is valid, $F\left(.\middle|\widehat{C}_i\right) = F\left(.|C_i\right)$, $\widehat{C}_i \subseteq C_i$ is established, $F(.|I) = F\left(.\middle|\widehat{C}\right) = F(.|C)$, $I \subseteq \widehat{C} \subseteq C$ is established, and the classification value k in Eq. (11) is linked to the value k classified in Eq. (14). When these conditions are met, finite $T_i = \{t_1, t_2, .., t_n\}$ is partitioned into mutually exclusive *types* by *circumstances* and has an analogous meaning to the finite $R = \{R_1, R_2, ..., R_m\}$, which is split into mutually exclusive regions by I using the decision tree classification algorithm. In the analysis, the explanatory variables represent *circumstances*, and the region represents the *type*. *Types* that are divided into finite small spaces (regions) have a classification value of 1 or 0, and this algorithm calculates the importance values of circumstance variables I, which is "a subset of C," in the process of classification.

There are several advantages of the decision tree classification algorithm. In the analysis, it was not necessary to convert categorical variables into dummy variables. Thus, the decision tree classification algorithm can handle continuous and categorical explanatory variables simultaneously without this conversion (Hastie et al., 2009; Murphy, 2012). Besides, this algorithm has the advantage of being insensitive to the monotone transformation of variables (Timofeev, 2004; Murphy, 2012).

Meanwhile, the biggest drawback of the decision tree algorithm is overfitting, which makes it unstable (Li and Belford, 2002; Murphy, 2012; James et al., 2013). This is partly due to the greedy nature of tree splitting. In this regard, small changes to the input data can greatly affect the structure of the tree because of the hierarchical nature of the tree growth process. Therefore, an error at the top can affect the rest of the tree (Li and Belford, 2002; Murphy, 2012).

Ensemble learning is a process designed to overcome shortcomings such as overfitting (Hastie et al., 2009). It refers to the process of generating several decision trees, which are multiple weak learners, and combining them to derive a more accurate and stable final prediction (Kuncheva and Whitaker, 2003).

This study proposes LightGBM as a model to overcome the shortcomings of the decision tree classification algorithm. LightGBM learns using a boosting method, a method of learning that reduces errors by assigning weights to incorrectly predicted observations so that multiple weak learners can predict more accurately while learning sequentially (Hastie et al., 2009; Chen and Guestrin, 2016). Equation (17) explains the key point of forwarding stage-wise additive modeling, which is the fundamental approach of the boosting algorithm.

$$G(x) = \sum\nolimits_{m=1}^{M} \alpha_m G_m(x). \tag{17}$$

The boosting algorithm produces a sequential weak classifier, $G_m(x)$, $m = (1,2,...,M)$, where $G(x)$ has the final classification value. $\alpha m, m = (1,2,...,M)$, the weight for each weak classifier, is constantly updated to allow better classification in the next step. In other words, successive classifiers are sequentially created and updated from $G_1(x)$ to $G_M(x)$ by minimizing the loss function to obtain the final classification. Among all boosting models, the models that minimize the loss function of the entire system through the gradient descent method are called gradient boosting decision trees (GBDT). LightGBM belongs to this family of boosting models (Chen and Guestrin, 2016; Ke et al., 2017).

Describing the method of LightGBM in detail, LightGBM's splitting method is called Gradient-based One-Side Sampling (GOSS). The basic approach of GOSS is analogous to Eq. (17) (Ke et al., 2017). LightGBM is designed in such a way that it inherits the advantages of existing boosting models and compensates for their shortcomings (Ke et al., 2017). Its biggest advantage is that, unlike other GBDT models, it uses the leaf-wise tree growth method. While other GBDT models use the level-wise tree growth method to reduce the depth of the tree, the LightGBM method does not balance the tree but deeply splits the leaf nodes with the maximum delta loss, thereby resulting in an asymmetric tree. As it repeats learning, the tree, generated by continuously dividing the leaf node with the maximum delta loss, reduces greater loss than the level-wise algorithm.

SHapley additive exPlanations. When tree-based models internally compute the importance values of the variables, the values may vary depending on the method of calculation. The importance values of variables can be calculated for a single prediction (individualized) or for an entire dataset to explain a model's overall behavior (global). Global importance values can be calculated for an entire dataset in different ways, thereby resulting in inconsistent results (Lundberg and Lee, 2017a, 2017b). If the importance values of the variables differ depending on the method of calculation, the reliability of the results inevitably decreases. Furthermore, it is difficult to make meaningful comparisons between the variables. In addition, the machine learning algorithm is limited, thereby making it difficult to interpret the estimated results without knowledge of the process between the input and output of data, like a black box (Burrell, 2016; Ribeiro et al., 2016). With only results and no interpretation, it is difficult to explain and understand the social phenomenon convincingly.

SHAP borrows the concept devised by Shapley (1953) and is based on game theory. As proposed by Lundberg and Lee (2017a), the SHAP value is a measure of the contribution of each variable to the output that interprets the estimated results. This approach estimates the importance of variables based on a solid theoretical foundation consistently and analyzes how each variable affects the output.

SHAP is based on an additive feature attribution method, which explains a model's output as the sum of real values attributed to each explanatory variable. The goal of SHAP is to estimate the attribution of each variable and explain the results. This approach can be explained by the explanation model, which is a linear function of the binary variables as shown in Eq. (18).

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{18}$$

Here, $g(z')$ is the local surrogate model of the original model, which helps interpret the original model, where $z' = \{0,1\}^M$. M is the number of explanatory variables, and $\phi \in \mathbb{R}$ (Lundberg and Lee, 2017b). z'_i equals 1 when a variable is observed; otherwise, it equals 0, and ϕ_is are the variable attribution values. Focusing on ϕ_i, the equation to estimate it is presented as Eq. (19) (Shapley, 1953; Lundberg and Lee, 2017a, 2017b).

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} \left(f_x(S \cup \{i\}) - f_x(S)\right). \tag{19}$$

In this equation, N is the set of all explanatory variables and S is defined as the subset of variables from N, $S \subset N$, not including i. $\frac{|S|!(M-|S|-1)!}{M!}$ is the weighting factor that counts the number of permutations of the subset S, and $f_x(S)$ is the expected output given the variable subset S, which is like the marginal average of all variables other than the subset S. Since it is necessary to know global importance, the absolute SHAP values per variable across the data are averaged as shown in Eq. (20).

$$I_i = \frac{1}{N} \sum_{j=1}^{N} \left| \phi_i^{(j)} \right|. \tag{20}$$

The importance values of variables can be determined through Eq. (20), but the results cannot be interpreted and their importance cannot be analyzed. They represent neither the range and distribution of impacts that the variable has on the output nor the relation of the variable's value to output. However, the SHAP summary plot can be utilized, which uses $\phi_i^{(j)}$ to convey all aspects of the importance of variables while remaining visually concise (Lundberg and Lee, 2017b).

Evaluation. In the literature on machine learning algorithms, the importance of model performance varies depending on the objective. Research focuses either on accurate prediction or understanding the relationships between variables (Celiku and Kraay, 2017; Hegelich, 2016). This study estimates the importance of variables in the process of dividing the sample into mutually exclusive *types* and interpreting the results, instead of making accurate predictions. However, the study reports the evaluation results in the form of a comparison between the performance of the decision tree classification and LightGBM. Data are divided into training and test data (8:2), and the performance of the model is evaluated using test data (Bonaccorso, 2018). In addition, accuracy, precision, recall, F1, and ROC-AUC are utilized as evaluation metrics for predictive performance evaluation (Powers, 2011) (see Appendix A).

Data. Since variables related to *circumstances* may differ depending on the socio-cultural characteristics of the society to which the individual belongs, they should be collected based on a sufficient understanding of each society (Roemer, 1993). The *Youth Panel Survey* conducted by the Korea Employment Information Service, affiliated with the Ministry of Employment and Labor, is analyzed. The Youth Panel Survey provides country-specific *circumstances* of the *millennials* of South Korea and their current wages. The population of the Youth Panel Survey included males and females from all over the nation, between the ages of 15 and 29, in 2007. The sample was extracted using the multi-stage area probability sampling method, and the survey was conducted in 2017 using a person-to-person interview method. In 2007, respondents were asked questions in a multiple-choice questionnaire about their *circumstances* around the age of 14, and in 2017, the same respondents were surveyed about their current wages. These data make it possible to analyze the socio-economic status of the survey respondents after 10 years. This study does not apply the approach of converting choices into dummy variables, considering the characteristics of the decision tree algorithm discussed in the previous sections (see B2 in Appendix B).

Parameters such as parental education, jobs, and other family background are widely used in the empirical literature on the inequality of opportunity (Brunori et al., 2018; Brunori and Neidhofer, 2020; Ferreira and Gignoux, 2011; Palomino et al., 2019; Roemer and Trannoy, 2016). Family structures, such as single-parent families, or living without parents, can affect a broad set of outcomes at a particular point in a child's life (Conway, 2012; Martin, 2006; Smock and Manning, 1997). In South Korea, Choi and Min (2015) proved that parents' education and income levels affect their offspring's educational achievement and performance in the labor market after graduation, through linear estimations. Oh and Ju (2017) revealed that there is significant inequality of opportunity in income acquisition between advantaged and disadvantaged *circumstances*, such as a father's education and occupation, through nonparametric tests.

In South Korea, there are significant gaps among regions in terms of the level of economic development and public and educational services (Kim and Jeong, 2003; Byun and Kim, 2010; Jeon, 2012). While gender inequality in the labor market of South Korea has continuously improved in terms of labor force participation rates, gender wage gap, and the proportion of regular workers, inequality of opportunity remains an existing social phenomenon (Park, 2007; Kim et al., 2016).

Regarding the tenancy status of the house and housing types in South Korea, the proportion of people living in condominiums increases as economic status increases. In contrast, the proportion of people living in multiplex housing units tends to increase with a lower economic status (Ha, 2002, 2004, 2007; Kim, 1997). In addition, the tenancy status of the house is divided into "owned," "lease on a deposit basis," and "monthly rent," depending on the economic status (Ha, 2002, 2004, 2007; La Grange and Jung, 2004).

Hence, the following parameters can be treated as *circumstances* for this analysis: the region where the respondent lived, the respondent's gender, whether the respondent lived with their parents, the job and occupational position, level of education, and physical presence of the respondent's parents, the number of parents working for economic activities, the number of siblings the respondent has, and the tenancy status and housing type of the respondent, all around the age of 14. The dependent variable has a binary classification value based on the minimum wage level of 6470 Korean won (KRW) per hour in 2017, which is equivalent to a monthly salary of 1,352,230 KRW and 209 h (see Appendix B).

Results and discussion

First, the estimation results of the decision tree classification and LightGBM are interpreted, and then the importance of individual variables is analyzed. The SHAP summary plot in Fig. 1 shows the estimated results of decision tree classification. Based on this, the region where the respondents lived is estimated to be the most important variable in the classification. This is followed by the respondents' gender, father's job, mother's job, father's position, housing type, number of siblings, tenancy status, father's education, mother's education, mother's occupational position, the physical presence of their parents, number of working parents, and living with their parents. The summary plot shows that the influence and intensity of these parameters gradually decrease from the top to the bottom. However, it is difficult to clearly distinguish between the variables in the plot, making it difficult to interpret and analyze the results. This is due to the unstable nature of the decision tree algorithm, which limits interpretation.

The following SHAP summary plot in Fig. 2 shows the estimated results of LightGBM. Based on this, the region where the respondents lived, their gender, their father's job, their mother's job, and the tenancy status of their houses are the five most important variables. In the case of a region, like the decision tree classification algorithm, the region's impact on the output and degree of intensity is much greater than that of the other variables. In other words, the region where the individual lived contributes the most in making a difference in their levels of achievement.

Regarding gender, the division of color in LightGBM is more uniform in both directions as compared to the decision tree classification. Generally, males work in the positive direction, whereas females work in the negative direction. The father's job has more impact than the mother's job, except when a high-value housewife or retired mother works in a positive direction. When considering the parents' education as a determining factor, the father's educational background is relatively higher than the mother's. Considering the impact of the father's job and educational background, and the respondent's gender on the output, the overall impact of gender is socially significant.

Regarding the tenancy status of the house, the category with high value acts strongly in the positive direction, and the category with a low value affects the output negatively. When the status of the house is "owned," which indirectly reflects the level of wealth, a positive output is indicated. In the case of the number of parents working, there is a mixture of high and low values, but the engagement of both parents has a slightly greater positive impact.

Considering the father's educational background, a value slightly greater than the average (dark pink) works in the positive direction, and a very high value (red) works in the negative direction. Thus, it can be assumed that the father's educational background works in a positive direction at the college level. In addition, it can be observed that the number of siblings, when few, acts in the positive direction. Considering that the average number of siblings is 2.3, it can be concluded that the output is positive when the number is below 2.

Table 2 lists the evaluation results according to the evaluation metrics for each model. When comparing the two models, the evaluation values significantly increase in all evaluation metrics of

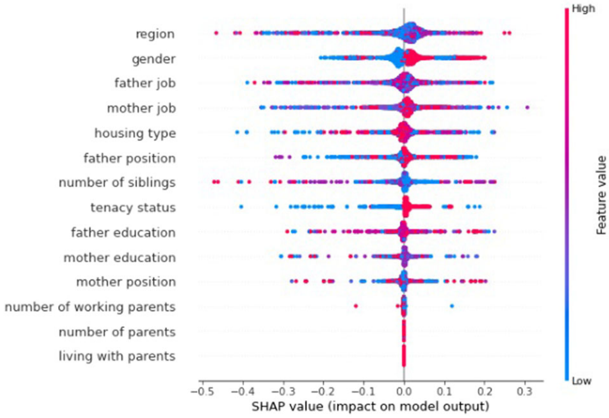

Fig. 1 Decision tree classification SHAP summary plot. Note: (1) The order from the top vertically indicates the importance of the variable. (2) Red color indicates a high value and blue color indicates a low value of the variable. (3) The horizontal axis denotes the impact of the value of the variable on the output. (4) The density achieved by the dots indicates their intensity.

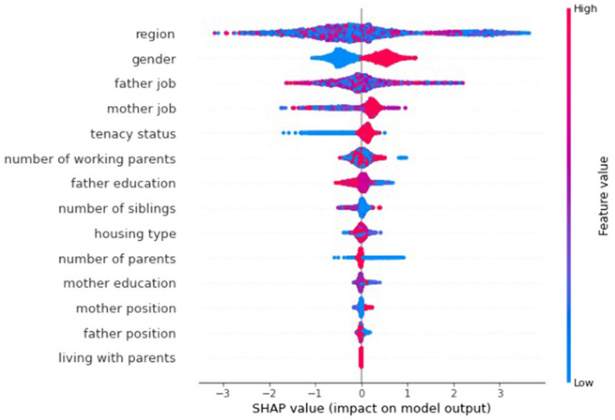

Fig. 2 LightGBM SHAP summary plot. Note: (1) The order from the top vertically indicates the importance of the variable. (2) Red color indicates a high value and blue color indicates a low value of the variable. (3) The horizontal axis denotes the impact of the value of the variable on the output. (4) The density achieved by the dots indicates their intensity.

Table 2 Evaluation results.

	Accurate	Precision	Recall	*F*1	ROC-AUC
Decision tree	0.8636	0.9386	0.9144	0.9263	0.5049
LightGBM	0.9360	0.9406	0.9947	0.9669	0.6654

the LightGBM compared to the decision tree classification (see Fig. A1 in Appendix A for ROC-AUC plot).

The interpretation and analysis of the SHAP are summarized as follows: For the decision tree classification, although the importance of variables can be estimated through SHAP, it is difficult to interpret the results accurately because of the unstable characteristics of the algorithm; in contrast, LightGBM is more stable as it sequentially updates multiple classification learners, which is apparent in the SHAP summary plot and evaluation results. Thus, for the interpretation and analysis of the importance of individual circumstance variables, LightGBM seems to be more reliable and appropriate.

Using the application of tree-based models, SHAP, and surveys, the roots of inequality of opportunity in South Korea can be identified. Region, gender, and father's job are the most important circumstance variables. Among them, in terms of impact and intensity of output, the region ranks higher. Evidently, the socio-economic achievement of an individual is greatly influenced by the region where the respondents lived during childhood. On another note, among the parents, both the impact and intensity of the output of the father's job are stronger than those of the mother's job. Furthermore, considering the educational background of both parents, the influence of the father on the younger generation has a greater impact. Interpreting this together with the impact of the respondents' gender, it becomes evident that in South Korean society, there is a social environment in which gender inequality exists.

Conclusion

This study identifies the roots of inequality of opportunity by applying algorithmic approaches and using survey data. The combination of tree-based classification models and SHAP estimates the importance of circumstance variables consistently and analyze which variables strongly influence output and how the values of variables affect it.

The main factors of inequality of opportunity commonly estimated in this study, through the application of SHAP, decision tree classification, and LightGBM are as follows: (1) region, (2) gender, and (3) father's job. Considering the SHAP summary plot and evaluation results, between the two models, LightGBM provides more stable and reliable results for interpretation and analysis.

Region, gender, and father's job are the main factors that form the most unfavorable socio-economic conditions for *millennials*. Region has an enormous impact on an individual's socio-economic achievement, and gender plays a significant role in contributing to the inequality of opportunity. The results of this study suggest that females may have fewer equal opportunities. Based on the factors related to parents' background, the father's job and educational background are considered more important variables than the mother's: the father's background strongly influences an individual's socio-economic achievements. Considering both the effects of the father's background and respondents' gender, the overall effects of males are socially significant.

It is worth noting that this study proves that a huge regional disparity exists in South Korean society. Phrases that represent specific spaces, such as the capital metropolitan area versus rural provinces, in-Seoul versus out-of-Seoul, and Gangnam (rich, south of the Han River) versus Gangbuk (poor, north of the Han River), reflect an individual's identity, social status, and class (Bae and Joo, 2020; Park and Jang, 2020; Yang, 2018). Phrases that define regions in specific ways mean that the opportunities available to individuals vary depending on where they grew up. Whether an individual grew up in the Seoul metropolitan area, in a rural area, or in Gangnam within Seoul, affects one's achievements in South Korean society in many ways. Inequality can be structurally reproduced if a certain group of people living in a certain area monopolizes opportunities, or if some people are spatially excluded from opportunities provided by society (Soja, 2010). The results of this study provide evidence to partially prove this.

The society reflected in the analysis results of this study is different from that of Rawls (1971). This raises the question of how a society with equally distributed opportunities can be created. The answer lies in considering how an individual's socio-economic achievement becomes the outcome of *circumstances*, *effort*, *social policy*, and *luck*. Out of these factors, *circumstances* are not dependent on an individual's choice and cannot be easily changed.

According to the analysis of this study, individuals receive unequal opportunities owing to a combination of region, father's background, and their own gender, thereby affecting their socio-economic achievements. If these factors remain influential from birth to adulthood, removing the conditions that structure them would be one way to achieve equality of opportunity. The ultimate goal of our society is to find *policies* that minimize the impact of *circumstances* and make the results more sensitive to *effort*.

The limitations of this study, along with suggestions for follow-up studies, are as follows: (1) while the current study applies algorithmic methods to the empirical approach of inequality of opportunity, the result is tentative and requires further discussion, particularly on the connection between theory, the empirical approach, and algorithms; (2) although this study selected wages as the criteria of the socio-economic gap, there may be other various criteria, not covered in this study, that may be considered in a follow-up study; (3) while the results suggest that region has the greatest influence on the inequality of opportunity in South Korean society, it is not possible to determine how the regions are stratified and the disparities among them. The analysis of regional disparity is beyond the scope of this study and should be considered by future research.

While this study has a few shortcomings, it still contributes to the development of the analysis of inequality of opportunity based on machine learning algorithms, analyzes the roots of inequality in South Korea, and complements previous studies with the help of a novel approach. Above all, this study contributes to the literature not only by describing social phenomena with data-driven methods, but also by trying to connect classic work, empirical approaches, and machine learning algorithms.

Data availability

The data is publicly available and can be found at: https://survey.keis.or.kr/yp/yp01/yp0101.jsp

Received: 14 June 2021; Accepted: 21 December 2021;
Published online: 12 January 2022

References

An CB, Bosworth B (2013) Income inequality in Korea: An Analysis of Trends, Cases, and Answers. Harvard University Press, Cambridge. https://www.hup.harvard.edu/catalog.php?isbn=9780674073197

Arneson R (1989) Equality of opportunity for welfare. Philosophical Studies 56:77–93

Bae Y, Joo YM (2020) The making of Gangnam: social construction and identity of urban place in South Korea. Urban Aff Rev 56(3):726–757

Balcazar CF (2015) Lower bounds on inequality of opportunity and measurement error. Econ Lett 137:102–105

Ban K, Kang E (2021) Kwak Sang-do's son received 5 billion won for solving a problem of cultural assets? He never did, and he never could. The Kyunghyang Shinmun 30 September 2021. http://english.khan.co.kr/khan_art_view.html?artid=202109301758177&code=710100. Accessed 10 Oct 2021

Birdsall N (2000) The social fallout: safety nets and recrafting the social contract. In: Stephan H (ed.) The political economy of the Asian financial crisis. Institute for International Economics, Washington, pp. 183–216

Bonaccorso G (2018) Mastering machine learning algorithms: expert techniques to implement popular machine learning algorithms and fine-tune your models. Packt Publishing Ltd, Birmingham

Bourguignon F, Ferreira FHB, Menendez M (2007) Inequality of opportunity in Brazil. Rev Income Wealth 53(4):585–618

Brunori P, Hufe P, Mahler DG (2018) The roots of inequality: estimating inequality of opportunity from regression trees. Policy Research Working Paper 8349

Brunori P, Peragine V, Serlenga L (2019a) Upward and downward bias when measuring inequality of opportunity. Soci Choice Welf 52:635–661

Brunori P, Palmisano F, Peragine V (2019b) Inequality of opportunity in sub-Saharan Africa. Appl Econ 51(60):6428–6458

Brunori P, Neidhofer G (2020) The evolution of inequality of opportunity in germany: a machine learning approach. ECINEQ Working Paper Series 2020-514

Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc 3(1):1–12

Byun SY, Kim KK (2010) Educational inequality in South Korea: the widening socioeconomic gap in student achievement. Glob Chang Demogr Educ Chall East Asia 17:155–182

Celiku B, Kraay A (2017) Predicting conflict. Policy research working paper 8075

Checchi D, Peragine V (2010) Inequality of opportunity in Italy. J Econ Inequality 8(4):429–450

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–795. https://doi.org/10.1145/2939672.2939785

Cho LJ, Kim YH (1991) Economic development in the Republic of Korea: a Policy perspective. East-West Center, Honolulu

Choi P, Min IS (2015) A study on social mobility across generations and inequality of opportunity. Soc Sci Stud 22(3):31–56

Cohen GA (1989) On the currency of Egalitarian Justice. Ethics 99:906–944

Conway KS (2012) Family structure and child outcomes: a high definition, wide angle "snapshot". Rev Econ Househ 10:345–374

Deloitte (2021) The Deloitte Global 2021 Millennial and Gen Z Survey. https://www2.deloitte.com/content/dam/Deloitte/kr/Documents/consumer-business/2021/kr_consumer_article_20210706.pdf

Dworkin R (1981a) What is equality? Part 1: equality of welfare. Philos Public Affairs 10:185–246

Dworkin R (1981b) What is equality? Part 2: equality of resources. Philos Public Affairs 10:283–345

Ferreira FGH, Gignoux J (2011) The measurement of Inequality of Opportunity: Theory and an Application to Latin America. Rev Income Wealth 57(4):622–657

Ferreira FHG, Peragine V (2015) Equality of opportunity: theory and evidence. IZA Discussion Paper No. 8994

Fields GS (1994) Changing labor market conditions and economic development in Hong Kong, the Republic of Korea, Singapore, and Taiwan, China. World Bank Econ Rev 8(3):395–414

Fleurbaey M (1994) On fair compensation. Theory Decis 36:277–307

Fleurbaey M (2008) Fairness, responsibility and welfare, 1st edn. Oxford University Press, Oxford

Fleurbaey M (1998) Equality among Responsible Individuals. In: Fleurbaey M, Gravel N, Laslier JF, Trannoy A (eds) *Freedom in Economics: New Perspectives in Normative Economics*, Routledge, London, pp. 206–234

Fleurbaey M, Peragine V (2009) Ex ante versus ex post equality of opportunity. Society for the Study of Economic Inequality Working Paper Series 2009–141

Ha SK (2002) The urban poor, rental accommodations, and housing policy in Korea. Cities 19(3):195–203

Ha SK (2004) New shantytowns and the urban marginalized in Seoul Metropolitan Region. Habitat Int 28:123–141

Ha SK (2007) Housing regeneration and building sustainable low-income communities in Korea. Habitat Int 31(1):116–129

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference and prediction. Springer, New York

Hegelich S (2016) Decision trees and random forests: machine learning techniques to classify rare events. Eur Policy Anal 2:98–120

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York

Jang S (2021) Squide Game shines a light on inequality in South Korea. East Asia Forum, 22 October 2021. https://www.eastasiaforum.org/2021/10/22/squid-game-shines-a-light-on-inequality-in-south-korea/. Accessed 25 Oct 2021

Jeon M (2012) English immersion and educational inequality in South Korea. J Multiling Multicult Dev 33:395–408

Jin YY (2021) Behind the global appeal of 'squid game', a country's economic unease. N Y Times, 18 October 2021. https://www.nytimes.com/2021/10/06/business/economy/squid-game-netflix-inequality.html. Accessed 25 Oct 2021

Kanbur R, Snell A (2017) Inequality indices as tests of fairness. IZA Discussion Paper No. 10721

Kang BG, Yun MS (2008) Changing in Korean wage inequality, 1980–2005. IZA Discussion Paper No. 3780

Kang SJ, Rudolf R (2016) Rising or falling inequality in Korea? Population aging and generational trends. Singap Econ Rev 61(5):1550089

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) LightGBM: a highly efficient gradient boosting decision tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds.) Proceedings of the conference on Neural Information Processing Systems 30 (NIPS 2017), 4–9 December 2017, Long Beach, CA, USA

Kim AE (2007) The social perils of the Korean financial crisis. J Contemp Asia 34(2):221–237

Kim E, Jeong YH (2003) Decomposition of regional income inequality in Korea. Rev Reg Stud 33:313–325

Kim H (2017) Spoon theory" and the fall of a populist Princess in Seoul. J Asian Stud 76(4):839–849

Kim J, Lee JW, Shin K (2016) Gender inequality and economic growth in Korea. Pac Econ Rev 23(4):658–682

Kim WJ (1997) Economic growth, low income and housing in South Korea. St. Martin's Press, Inc., New York

Koh Y (2019) The evolution of wage inequality in Korea. KDI Policy Study 2018-01

Koo H (2007) The changing faces of inequality in South Korea in the age of globalization. Korean Stud 31:1–18

Koo Y (2013) Evolution of industrial policies and economic growth in Korea: challenges, crisis, and responses. Eur Rev Ind Econ Policy 7. http://revel.unice.fr/eriep/index.html?id=3598

Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 51:181–207

Grange AL, Jung HN (2004) The commodification of land and housing: the case of South Korea. Hous Stud 19(4):557–580

Lefranc A, Pistolesi N, Trannoy A (2009) Equality of opportunity and luck: definitions and testable conditions with an application to income in France. J Public Econ93:1189–1207

Lefranc A, Kundu T (2020) Inequality of opportunity in India society. Hal-02539364

Lee HH, Lee M, Park D (2012) Growth policy and inequality in developing Asia: lesson from Korea. ERIA Discussion Paper Series 2012-12

Lee J (2017) The labor market in South Korea, 2000–2016. IZA World of Labor 405

Lee S (2019) Understanding young Koreans' rage against Cho Kuk. The Korea Times, 28 August 2019. https://www.koreatimes.co.kr/www/nation/2019/08/356_274736.html. Accessed 21 Aug 2021

Li RH, Belford GG (2002) Instability of decision tree classification algorithms. KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 570–575. https://doi.org/10.1145/775047.775131

Lim HC, Jang JH (2006) Neo-liberalism in post-crisis south Korea: social conditions and outcomes. J Contemp Asia 36(4):442–263

Lundberg SM, Lee SI (2017a) A united approach to interpreting model prediction. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R (eds.). Proceedings of the 31st conference on neural information processing systems, 4–9 December 2017, Long Beach, CA, USA. https://dl.acm.org/doi/pdf/10.5555/3295222.3295230

Lundberg SM, Lee SI (2017b) Consistent feature attribution for tree ensembles. In: Kim B, Malioutov DM, Kush R, Weller A (eds.). Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI2017), 10 August 2017, Sydney, NSW, Australia. https://openreview.net/pdf?id=ByTKSo-m-

Martin MA (2006) Family structure and income inequality in families with children, 1976 to 2000. Demography 43(3):421–445

Murphy KP (2012) Machine learning: a probabilistic perspective. The MIT Press, Cambridge

Murthy SK (1998) Automatic construction of decision trees from data: a multi-disciplinary survey. Data Min Knowl Discov 2(4):345–389

Neumark D, Wascher WL (2008) Minimum wages. The MIT Press, Cambridge

Oh SJ, Ju BG (2017) Inequality of opportunity for income acquisition in Korea. Korean J Public Financ 10(3):1–30

Palomino JC, Marrero GA, Rodrguez JG (2019) Channels of inequality of opportunity: the role of education and occupation in Europe. Soc Indic Res 143(3):1045–1074

Park BG, Jang J (2020) The Gangnam-ization of Korean urban ideology. In: Doucette J, Park BG (eds.) Developmentalist cities? Interrogating urban developmentalism in East Asia. Haymarket Books, Chicago

Park H (2007) Inequality of educational opportunity in Korea by gender, socio-economic background, and family structure. Int J Hum Rights 11(1-2):178–197

Powers DMW (2011) Evaluation: from prevision, recall and *F*-measure to ROC informedness, markedness and correlation. J Mach Learn Technol 2(1):37–63

Rawls J (1958) Justice as fairness. Philos Rev 67(2):164–194

Rawls J (1971) A theory of justice. Harvard University Press, Cambridge

Ribeiro MT, Singh S, Guestrin C (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135–1144. https://doi.org/10.1145/2939672.2939778

Roemer J (1993) A pragmatic theory of responsibility for the egalitarian planner. Philos Public Aff 10:146–166

Roemer J (1998) Theories of distributive justice. Harvard University Press, Cambridge

Roemer JE, Trannoy A (2016) Equality of opportunity: theory and measurement. J Econ Lit 54(4):1288–1332

Sen A (1980) Equality of what?. In: The Tanner lecture on human values, I. Cambridge University Press, Cambridge, pp. 197–220.

Scitovsky T (1985) Economic development in Taiwan and South Korea: 1965–81. Food Res Inst Stud XIX(3):215–264

Shapley LS (1953) A value for n-person games. In: Tucker AW, Kuhn HW (eds.) Contributions to the theory of games, vol I. Princeton University Press, Princeton

Shim E (2019) South Korea's 'privileged' politicians scrutinized after Moon aide appointment.UPI, 11 September 2019. https://www.upi.com/Top_News/World-News/2019/09/11/South-Koreas-privileged-politicians-scrutinized-after-Moon-aide-appointment/7381568213270/. Accessed 20 Oct 2021

Shin H (2020) 'Parasite' reflects deepening social divide in South Korea. Reuters, 10 February 2020. https://www.reuters.com/article/us-awards-oscars-southkorea-inequality-idUSKBN20414L. Accessed 20 Oct 2021

Smock PJ, Manning WD (1997) Nonresident parents' characteristics and child support. J Marriage Fam 59(4):798–808

Singh A (2011) Inequality of opportunity in earnings and consumption expenditure: the case of Indian men. Rev Income Wealth 58(1):79–106

Soja EW (2010) Seeking spatial justice. University of Minnesota Press

Starr GF (1981) Minimum wage fixing: an international review of practices and problems. International Labour Organization, Geneva

Timofeev R (2004) Classification and regression trees (CART) theory and applications. Ph.D. Dissertation. Humboldt University, Berlin, Germany

Trannoy A, Tubeuf S, Jusot F, Devaux M (2010) Inequality of opportunities in health in France: a first pass. Health Econ 19(8):921–938

Yang M (2018) The rise of 'Gangnam style': manufacturing the urban middle classes in Seoul, 1976–1996. Urban Stud 55(15):3404–3420

Competing interests

The author declares no competing interests.

Ethical approval

Not applicable.

Informed consent

Not applicable.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-021-01026-y.

Correspondence and requests for materials should be addressed to Seungwoo Han.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.