https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

Natural Hazards and Earth System Sciences
Discussions

Using machine learning algorithms to identify predictors of social vulnerability in the event of an earthquake: Istanbul case study

Oya Kalaycioglu[1,2], Serhat Emre Akhanli[3], Yahya Emin Mentese[4], Mehmet Kalaycioglu[5], and Sibel Kalaycioglu[6]

[1]Department of Statistical Science, University College London, London, WC1E 6BT, United Kingdom

[2]Depratment of Biostatistics and Medical Informatics, Bolu Abant Izzet Baysal University, Bolu, 14030, Turkey

[3]Department of Statistics, Mugla Sitki Kocman University, Mugla, 48000, Turkey

[4]Kandilli Observatory and Earthquake Research Institute, Bogazici University, Istanbul, 34684, Turkey

[5]Tomorrow's Cities Research Group, City and Regional Planning Div., Middle East Technical University, Ankara, 06800, Turkey

[6]Department of Sociology, Middle East Technical University, Ankara, 06800, Turkey

Correspondence to: Oya Kalaycioglu (oya.kalaycioglu.09@ucl.ac.uk)

Abstract. For an effective disaster risk mitigation plan and for building a society more resilient to natural disasters, it is essential to understand the factors that are related to social vulnerability as an important dimension to social risk. This study aims to identify the associations between socio-economic and socio-demographic household characteristics and earthquake related social vulnerability using survey data collected from 41,093 households in Istanbul. Machine learning models, namely: logistic regression, classification tree, random forest, support vector machine, naive bayes, artificial neural network, and K-nearest neighbours, were employed to classify households according to their social vulnerability status. Due to the disparity of class size for the outcome variable, subsampling strategies were applied for dealing with imbalanced data. Artificial Neural Network (ANN) was found to have the optimal predictive performance when random majority under sampling was applied (AUC: 0.813). The results from the ANN method indicated that not having social security, living in a squatter house and having high risk of job loss after an earthquake were among the most important predictors for increasing social vulnerability risk. Additionally, the level of education, the ratio of elderly persons in the household, owning a property, household size, ratio of income earners, and having savings were associated with vulnerability. An open access R-

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

shiny web application was developed to visually display the performance of ML methods, important variables for the social vulnerability risk classification and the spatial distribution of the variables across Istanbul neighbourhoods. The machine learning methodology and the findings that we present in this paper can serve as a guidance for decision makers in identifying and prioritising action towards target groups to reduce their vulnerability risk prior to earthquakes.

1 Introduction

The United Nations Office for Disaster Risk Reduction describes disasters as events that exceed the capacity of states and/or communities to cope with the consequences of a hazard (UNISDR, 2009). However, not all hazards result in a disaster. The evolution of an earthquake event into a disaster is typically studied through the lenses of geoscientists, civil engineers and earthquake engineers, since the most tangible results or causes of a disaster are physical. However, it is often forgotten or ignored that the human consequences of disasters are in part derived from the composition of the population and society prior to the event. Therefore, we posit that a more comprehensive understanding of disasters is possible by looking at both physical and social aspects.

Istanbul, which is the 13th most populated city in the world with a population of more than 15 million (WUP, 2021), is exposed to earthquake hazards due to the North Anatolian fault which lies across the southern border of the city. Historical records show that in approximately every 100, 250 and 500 year periods, a severe earthquake hits Istanbul and causes significant casualties and damage to infrastructure. A recent study suggests that Istanbul and other nearby areas in the Marmara region are at a significantly high risk of devastating earthquakes, with magnitudes between 7.1 and 7.4 (Lange et al., 2019). It is estimated that an earthquake with a magnitude of 7.5 in the North Anatolian Fault across Istanbul would cause tens of thousands of deaths, and catastrophic damage to buildings and infrastructure (IMM and KOERI, 2019), leading to the destruction of human life and wellbeing, along with economic and social devastation.

There are various studies in the fields of earthquake engineering and geosciences that address the earthquake hazard and the physical vulnerability of infrastructure and buildings in the Istanbul metropolitan area (IMM and KOERI, 2019; Parsons et al., 2000; Parsons, 2004; JICA and IMM, 2002; Erdik et al., 2003; Ersoy and Koçak, 2016). Such studies are important for interpreting the possible consequences of earthquakes and they support decision makers and public authorities in developing strategies and policies for disaster response. Nevertheless, developing robust and concrete disaster risk reduction measures requires consideration of social aspects as well as physical ones. Among the social aspects, that of "social vulnerability" has become an increasingly popular topic in natural disaster research (Shen et al., 2018). Social vulnerability differs from physical vulnerability in that it does not regard the number of possible injuries or fatalities that may occur due to, for example, the collapse of buildings. It is, rather, a measure of the capacity of an individual or household to anticipate, cope with, resist, and recover from the impact of an earthquake (Blaikie et al., 2014; Wisner et al., 2012). Additionally, social vulnerability increases the social risks of different social groups in relation to a set of socioeconomic

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

conditions and is needed to be determined before a particular hazard hits the society (Cannon, 2008). Social risk refers to expected human and economic losses resulting from a particular natural hazard in a given time and place (Cutter, 1996). Hence, assessing the possible social risks of vulnerable groups can perform a predictive role towards improved preparedness and ability to recover from natural disasters (Ogie and Pradhan, 2019). The identification of the factors that contribute to social vulnerability is therefore crucial for effective disaster risk management and for building a more resilient society (Aksha et al., 2019).

In the literature, there are many different aspects to the assessment of social vulnerability, and they can be based on the location of the research, the hazard type, the scale, and the temporal focus. One of the first studies that describes vulnerability from a social perspective is Cutter's research (Cutter, 1996) that focuses on social aspects of the vulnerability concept based on a detailed "vulnerability" literature review. Another study by Cutter et al. (2003) is analytical and includes county scale social vulnerability analysis from census data. Social vulnerability is assessed based on 42 different variables that were reduced to 11 significant indicators to construct a social vulnerability index. Various studies thereafter, assessed the indicators that could be used to measure the social vulnerability for a certain location and time frame (Holand et al., 2011; Bergstrand et al., 2015; Fatemi et al., 2017; Rufat et al., 2019; Spielman et al., 2020; Rahman et al., 2022). It can be suggested that there is almost a consensus between those studies where social vulnerability is defined as a function of gender, health status and access to healthcare, poverty, age, ethnicity, property ownership, and socio-economic indicators (Kalaycioglu et al., 2006). In addition to such individual characteristics, space related metrics such as rural/urban status (Cutter et al., 2000; Cova and Church, 1997; Mitchell, 2000) and population density (Cutter et al., 2000; Morrow, 1999; Puente, 1999) as well as public service-related indicators, such as infrastructure quality and the proximity of healthcare facilities (Cutter et al., 2000; White, 2000; Bolin and Stanford, 1991; Duzgun et al., 2011), are used to form different types of approaches for assessing the social vulnerability on larger scales such as across a district or a city. Measuring social vulnerability on a regional scale can give important insights into the variation within and overall vulnerability of a country, region or population group. Nevertheless, it may still be informative to measure vulnerability at a household level, using all dimensions and indicators, to represent better individual vulnerabilities (Debesai, 2020).

Numerous studies have examined the factors relating to social vulnerability in the event of an earthquake, which have used either descriptive statistics (Yücel and Görün, 2010; Walker et al., 2019), or traditional data analysis tools, such as linear or logistic regression (Noriega and Ludwig, 2012; Syed and Kumar Routray, 2014; Llorente-Marrón et al., 2020). While the former lacks the incorporation of the relationships between the vulnerability indicators, the latter relies heavily on data assumptions. In contrast, machine learning (ML) algorithms allow for a larger number of predictors, can handle complex interactions between predictors, can model nonlinear relationships and they do not make any distributional assumptions regarding the data (Ryo and Rillig, 2017). Due to these advantages, there is an emerging interest in using ML methods for making predictions or classifications for large scale survey data (Buskirk et al., 2018). A relatively small number of researchers have opted to use ML methodology over regression techniques in vulnerability research (Dwyer et al.,

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

2004; Alizadeh et al., 2018; Yoon and Jeong, 2016; Abarca-Alvarez et al., 2019), and indeed a detailed model-based assessment of the predictors of social vulnerability to earthquakes is lacking.

Several applications of machine learning methods that relate to social vulnerability in natural hazards are present in the literature. Dwyer et al. (2004) used decision tree methodology for identifying individuals at social risk to natural hazards in Perth City, Australia, and found 11 decision rules that determine high social vulnerability to natural hazards. By collecting data with questionnaires, they investigated the relative importance of 13 indicators related to demographic and economic household attributes in contributing to the prediction of social vulnerability. Other studies were based on larger sampling units such as districts, neighbourhoods or communities, in contrast to our study which was based on household survey data. Alizadeh et al. (2018) used Artificial Neural Networks (ANN) for deriving a social vulnerability to earthquake map in the Tabriz city, Iran, for detecting the most and the least vulnerable zones from a spatial perspective. Yoon and Jeong (2016) assessed the vulnerability to natural disasters at a community level in South Korea using 12 vulnerability variables including social, economic, natural environment and built environment aspects. They examined the important vulnerability indicators by the use of traditional linear regression as well as two machine learning techniques, Random Forests and Cubist. They showed that machine learning techniques have better model performances compared to traditional regression methodology. In another vulnerability study, Abarca-Alvarez et al. (2019) identified deprived areas that are more prone to social vulnerability in Andalusia using decision trees. They used variables related to socio-demography, socio-economy, community, and public infrastructure. Their dataset was taken from the Andalusian Population Census available at a regional scale. All three aforementioned studies are related to the idea of comparing different and larger settlement units for social vulnerability to natural hazards (Alizadeh et al., 2018; Yoon and Jeong, 2016; Abarca-Alvarez et al., 2019).

In this study, we attempt to give further contribution to social vulnerability research in natural disasters by identifying the most important factors that contribute to the prediction of social vulnerability of households in Istanbul in the event of an earthquake, using machine learning methodology. Accordingly, we address the following research questions: (1) What is the best performing ML method for classifying the risk status of social vulnerability? (2) What are the most influential predictors associated with social vulnerability? We posit that the application of a broad conceptual model, developed based on ML algorithms, leads to a better understanding of households that would be socially vulnerable in the event of an earthquake.

The research presented in this paper acts as an addition or phase two of a previous study. The first phase covers the calculation of the social vulnerability score for more than 40,000 households in Istanbul, as explained in Menteşe et al. (2019). It considers the concept of social vulnerability as a state that arises from intrinsic characteristics of society, such as the perception of risk and the measures taken against risk, as well as cultural values and socio-economic status. Then, in the second phase, presented in this paper, we assessed to what extent household characteristics can predict the severity of social vulnerability risk via ML methods. The predictors we use in this study have been restricted to quantifiable variables as they present tangible information for modelling and measuring social vulnerability. This type of household information is available in metropolitan / district municipalities, neighbourhood mukhtars, city governorship and from Address Based

Population Registration System. At present, such data has not been used for risk reduction policies by public decision makers. Hence, our model can serve as guidance for the decision makers for identifying and prioritising action towards target groups in the interests of risk mitigation.

The layout of the paper is as follows. First, the household survey data is introduced. Secondly, ML methods including subsampling strategies for the imbalanced class variables are explained, along with the model assessment criteria. Then we train, validate, and compare predictive performances of ML models for social vulnerability and find the best performing ML model. Finally, the importance of the predictors and their effects on the response for estimating social vulnerability were discussed for the best performing model.

2 Materials and Methods

2.1 Social vulnerability survey

A large-scale household survey was carried out by the Directorate of Earthquake and Ground Research of Istanbul Metropolitan Municipality between the period of 2017 and 2018, after approved by institutional review board. The authors of this study were given the permission to use this survey data after the data were fully anonymized. N=41,093 households in 955 sub-districts/neighbourhoods, with residential occupation covering the whole jurisdiction boundaries of the metropolitan municipality of Istanbul, were included in the study of social vulnerability (IMM, 2018). The households were randomly selected from the Address Based Population Registration System Database of the Turkish Statistical Institute using the proportionate stratified sampling method. All 955 neighbourhoods within 39 districts of Istanbul were taken as strata, then households were randomly selected from each neighbourhood. The number of households in each neighbourhood taken is proportional to the neighbourhood population.

The survey data was obtained via face-to-face interviews with one household member, who is between 18 and 70 years of age and who is able to give relevant and accurate information about the household. The verbal and written informed consents were obtained from the participants during the data collection stage. The survey included questions related to socio-demography, socio-economy, duration lived in an urban environment, access to health services, social solidarity, risk perception, actions taken to reduce risk and cultural beliefs.

2.2 Assessment of social vulnerability

In the first phase of the study, a social vulnerability index score was calculated for each household from the survey data (IMM, 2018; Menteşe et al., 2019). The households were then defined in clusters as households with "severe risk of social vulnerability" and all others as "non-severe risk of social vulnerability", by dichotomising the vulnerability score using the established cut-off points. Thus, a binary variable (with an imbalance ratio of 1/5 in favour of non-severe risk) was generated as an indication of risk of social vulnerability level, which in turn was used as the primary outcome for all the further analyses presented in this paper. The three main reasons for defining the social vulnerability as a binary outcome were: (1) To find out the key vulnerability predictors that discriminates between the households that requires the most urgent action and all others, (2) to increase the accuracy of predictions obtained by machine learning methods, and (3) to ease the interpretation of the results.

2.3 Selection of predictors and data preparation

Using the social vulnerability risk level as the binary outcome, household characteristics were taken as predictors to build a model. The predictors chosen have been selected following extensive literature reviews, discussions with experts and with the aim of exploring quantitative variables in predicting the risk of social vulnerability of a household in the event of an earthquake.

Prior to model development, the predictors were prepared in terms of data representation, standardization and feature selection. As the predictors represent household characteristics, they were sought at household level. As made clear by Akhanli and Hennig (2020), data representation is about enabling better interpretation of the relevant information. Therefore, the predictors which are measured at household level, such as the number of women, men, <5 years olds, >65 years olds and the number of income earners were taken in proportion to the given household's size (HhS). Then, in order to make the variation of continuous variables comparable, these variables were standardized into the same scale with unit variance standardization (Hennig and Liao, 2013). For the final step, we used feature selection prior to process the data and we identified the predictors with near zero variance, as the predictors which take only one value may cause numerical problems during resampling (Kuhn, 2008). The set of 26 variables used for model building are presented in Table 1, along with their relevance in relation to the objectives of our study.

Table 1. Predictors used in model building for the classification of social vulnerability risk.

Themes	**Variable**	**Definition of a variable or survey question**
Socio-Demographic	Household size	Number of people living in the house (HhS) (Range: 1-14)
	Average age	Average age of the household members in years (Range: 8.8-85)
	Number of women/HhS	Ratio of women in the household (Range:0-1)
	Number of men/HhS	Ratio of men in the household (Range:0-1)
	Number of <5 year olds/HhS	Ratio of <5 years old children in the household (Range:0-0.67)
	Number of >65 years of age/HhS	Ratio of over 65 years old individuals in the household (Range:0-0.1)
	Average education	Average years of education of the household members who are over 15 years old (Range:0-17)
	Social security	Are there any household members with social security? (yes/no)
Health	Health insurance	Are there any household members with health security or insurance? (yes/no)
	Disability	Are there any disabled or elderly persons who needs care in the Hh? (yes/no)
	Health access	Do you have any healthcare facilities nearby to your home? (yes/no)

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

Socio-Economic	Number of income earners/HhS	Ratio of the number of income earners in the household (Range:0-2)
	Regular salary income	Are there any household members who have regular salary income? (yes/no)
	Pension income	Are there any household members who earn pension income? (yes/no)
	Rent/interest income	Are there any household members who earn income from rent or from interest? (yes/no)
	Income support from public authorities	Are there any household members who receive income support from public authorities? (yes/no)
	Risk of job loss	Are there any household members with the risk of job loss in an earthquake? (yes/no)
	House ownership	Do any of the household members own the house of your residence? (yes/no)
	Type of the house	What is the type of the home of your residence? (apartment flat, squatter house, detached house, gate keepers lodge)
	Natural gas heating	Do you have natural gas heating at the home of your residence? (yes/no)
	Own house in Istanbul	Are there any household members who own a house in Istanbul, other than the home of residence? (yes/no)
	Own land in Istanbul	Are there any household members who own land in Istanbul? (yes/no)
	Own house out of Istanbul	Are there any household members who own house outside Istanbul? (yes/no)
	Own land out of Istanbul	Are there any household members who own land outside Istanbul? (yes/no)
	Saving	Are there any household members who have savings to use for emergency situations? (yes/no)
	Dept	Are there any household members who have dept to third parties (inc. bank, relatives, friends, etc.)? (yes/no)

2.4 Machine learning methods

We developed models for classification of households in terms of their social vulnerability risk in the event of an earthquake using seven supervised machine learning (ML) algorithms: logistic regression (LR), classification and regression tree (CART), random forest (RF), artificial neural network (ANN), support vector machine (SVM), Naïve Bayes (NB), K-Nearest Neighbours (KNN). Supervised ML adopts an algorithm to learn the mapping function from the input variables to the output variable and it is suited well to classification problems. Models were developed using the variable set in Table 1 as the input variables, while a binary social vulnerability risk status of each household was the output variable. We developed a prediction model using 90% of the social vulnerability dataset to train the underlying algorithm, while 10% of the dataset was held back as independent testing data for evaluating the performance of the models. We note that these algorithms have different tuning parameters. For different tuning parameter alternatives, the choice of the optimal tuning parameter was determined by the largest area under the curve (AUC) value of the receiver operating characteristic (ROC) curve. The workflow for the model building is shown in Fig. 1.

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

Natural Hazards and Earth System Sciences
Discussions

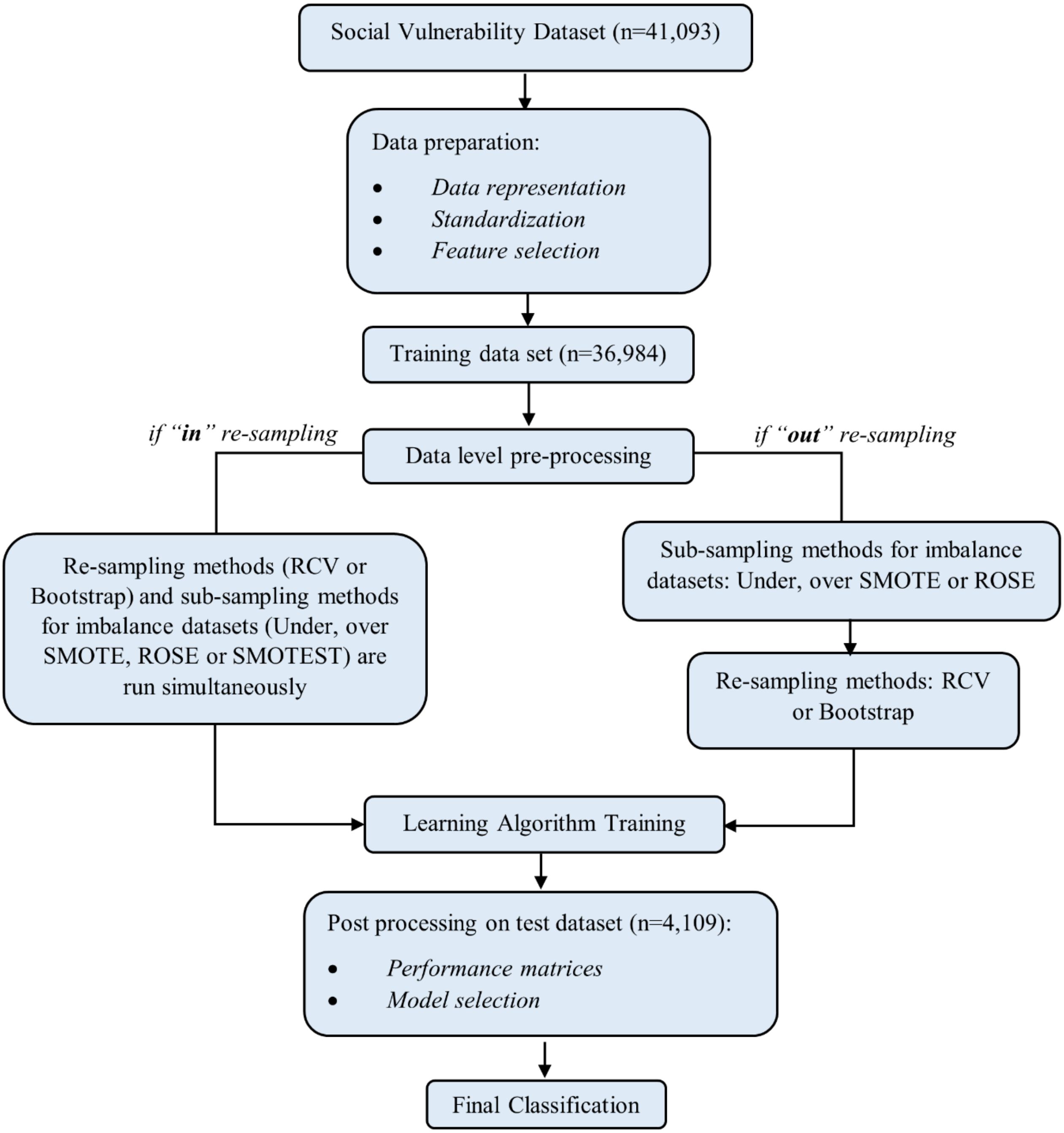

Fig 1. Machine learning flowchart for data processing and model development.

2.5 Data level pre-processing

2.5.1 Resampling techniques

Repeated cross validation (RCV) and bootstrap resampling procedures were used to draw multiple subsamples from the original data to build machine learning models on the training data and to validate the models, in each instance, on the data that were excluded from the subsample. The tuning parameters were selected as 5-fold with 4 repetitions for repeated cross-validation and 20 repetitions for bootstrap, resulting in the same amount of resampling. The number of resampling repetitions was kept low to diminish the computational time burden for large data sets.

2.5.2 Subsampling for the imbalanced class variables

A dataset is said to be imbalanced when the classification categories are not represented equally (Lin and Nguyen, 2020). In our study, social vulnerability data set consists of imbalanced class variables, in which the "severe risk of social vulnerability" class has a lower frequency compared to the "non-severe" class. The imbalance ratio of these two classes was approximately 1/5. The main challenge of the imbalance problem in standard machine learning algorithms is that the minority classes can be overlooked and weighed down by the majority one (Ramyachitra and Manikandan, 2014). In order to address this issue, we used various subsampling approaches during the data pre-processing steps as explained below:

(i) *Random majority under-sampling* (*Under*): Under-sampling randomly samples from the majority class and returns a sub sample which has the same size as the minority class, thus ensuring the majority class prevalence is equal to that of minority one for subsequent modelling (Batista et al., 2004). For instance, assume a binary class variable in which 90% of training set samples belongs to the majority class, while the remaining 10% are in the minority class. Under-sampling will randomly sub sample from the majority class such that its prevalence is 10%. As a result, only 20% of the total training set will be used for the classification model. While balancing the class variable, however in some cases this approach may remove many important or otherwise influential data points prior to modelling.

(ii) *Over-sampling:* Three different over-sampling strategies were applied:
Random minority over-sampling (Over): It aims to balance the distribution of the class variable by taking random replicates of the minority class (Batista et al., 2004). Although it helps to improve the accuracy of classification in imbalanced datasets, it is prone to overfitting and computational problems when the data set is large (Maheshwari et al., 2017).
Synthetic Minority Over-sampling Technique (SMOTE): It creates artificial minority examples by interpolating between randomly selected examples of the minority class and their nearest neighbours

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

(Chawla et al., 2002). It attempts to avoid overfitting problem by using new synthetic minority class examples instead of replicating minority samples.

Random Over-Sampling Examples (ROSE): It generates artificial balanced samples according to a smoothed bootstrap approach and aids in the phases of estimation and accuracy evaluation of a classification algorithm in the presence of an imbalanced class variable (Menardi and Torelli, 2014).

The above procedures are independent of resampling methods such as the repeated cross-validation and the bootstrap. On the other hand, these subsampling procedures can also be performed for the resampling techniques, so that subsampling is conducted inside of resampling. In this paper, when subsampling procedures performed outside of resampling techniques it is referred as “out sampling”, otherwise it is expressed as “in sampling”.

One could also consider creating a custom-made subsampling procedure. In this respect, we also apply the transformed version of SMOTE that use 10 nearest neighbours instead of the default of 5 by adopting a simple wrapper function, which we call as the “SMOTEST”. Note that the SMOTEST function is only performed inside of the resampling (Kuhn and Johnson, 2013).

2.6 Statistical analysis and model performance assessment

The characteristics of the study population was summarised using descriptive statistics. Pearson’s chi-square tests were used to compare categorical variables, and independent samples t-tests or non-parametric Mann Whitney U tests were used to compare continuous variables between the non-severe and severe risk groups depending on the data distribution. In studies with large sample sizes, in addition to p-values, it is also relevant to provide effect sizes as it can help deciding whether the difference found is meaningful or not (Bakker et al., 2019). Thus, we have reported effect sizes in the univariate comparisons that measures the strength of the relationship between two variables along with the p-values to assess whether the effect of a variable is real and large enough to be useful or not. Cohen’s d statistic with sample size adjustment was used for normally distributed continuous variables, Cohen’s r value which is calculated by dividing the z value obtained from the Mann Whitney test to the square root of the sample size was used for non-normally distributed variables, and Cramer's V is used for categorical variables (Fritz at el., 2011).

For various machine learning applications confusion matrices were generated. Sensitivity, specificity and accuracy with 95% Confidence Intervals (CIs) were calculated for each ML algorithm using different resampling and subsampling techniques. The models were fitted with two different resampling strategies and eight subsampling techniques. In addition, we fitted the models to the raw data without any subsampling, and thus we obtained results for 18 combinations of various sampling strategies for each ML algorithm.

In line with the objective of the study, we compared the methods in terms of their success in identifying the households with severe risk of social vulnerability, which is the minority class with smaller prevalence in our study. Therefore, we used sensitivity (true positives / (true positives + false negatives)) as the primary measure for assessing the

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

model performance and balanced accuracy ((sensitivity + specificity) / 2). We identified the best performing method as the one with the highest sensitivity and balanced accuracy, provided that the AUC of the ROC curve is greater than 0.7 and model could be considered as acceptable to discriminate households with severe risk from those with non-severe risk (Hosmer et al., 2013).

The sensitivity and specificity of the best performing method with those of other methods were compared with pairwise comparisons using McNemar's chi-square test (Kim and Lee, 2017). In addition, AUC comparisons were performed using DeLong chi-square statistics (DeLong et al., 1988). Bonferroni adjustment was applied in these pairwise comparisons of ML methods and $\alpha<0.05/7=0.007$ was considered as an indication of statistically significant difference in terms of performance metrics between two methods.

For the final step of the analysis, the important variables of each model were assessed. The identification of the important predictors is either based on the contribution of each variable to the model, or by an ROC curve analysis conducted on each predictor (Kuhn, 2008).

2.7 Open-access R-shiny web application

An open-access R-shiny web application was created for visualising summary statistics and predictive performances of the ML methods for the classification of households in terms of their social vulnerability risk. Users are able to examine the distribution of the characteristics of the households with severe and non-severe risk of social vulnerability, compare the performances ML and subsampling methods based on a user defined evaluation criteria, assess variable importance rankings for each ML method and obtain the area-based calculations of the variables in the Istanbul map. The R-shiny web application is freely available online and can be accessed at https://oyakalaycioglu.shinyapps.io/Social_Vulnerability/. The components of this R-Shiny application are presented in detailed in Fig. 2. All analyses were performed in the statistical programming environment R version 4.0.3 (R Core Team, 2013) and the machine learning model development was carried out using the R caret package (Kuhn, 2008). The spatial distribution of the important predictors within the city scale were expressed via the 3.10 version of QGIS software (QGIS, 2021).

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

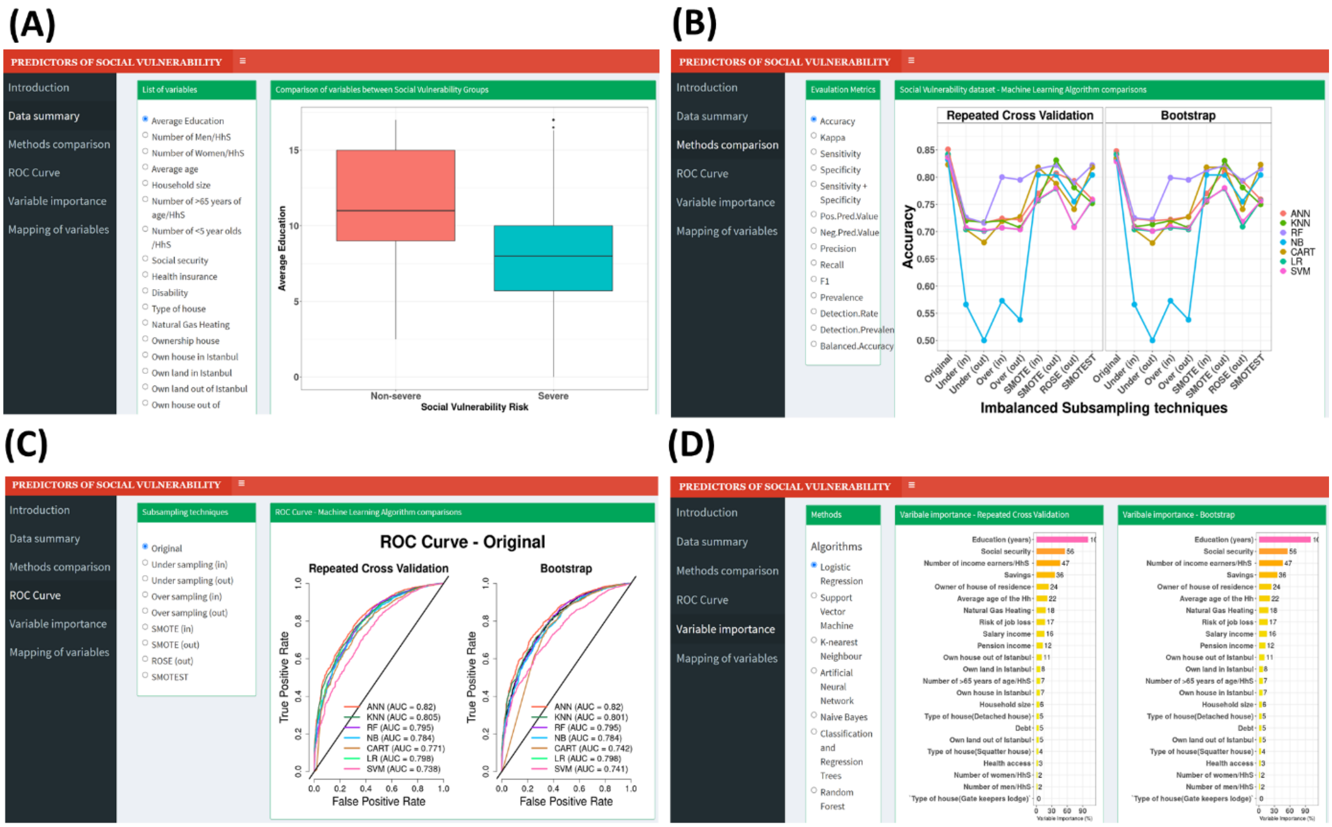

Fig. 2. The components of open-access web application created in R-shiny interface. (can be accessed from https://oyakalaycioglu.shinyapps.io/Social_Vulnerability/). The left side commands allow the user to choose which analysis to activate. (A) Summary statistics of the variables are visually compared across social vulnerability risk groups. Box plots and bar plots were used for continuous and categorical variables, respectively. (B) The performance metric chosen by the user (y-axis) in comparison to subsampling method (x-axis). The ML methods are displayed in different colours. Two separate plots are generated for RCV and bootstrap resampling techniques. (C) For the chosen subsampling method, ML methods are compared in terms of the AUC of the ROC curve. Different coloured lines represent different ML methods. (D) For the chosen ML method and subsampling techniques, variable importance plots are displayed.

3 Results

3.1 Descriptive statistics

The prevalence of households that were identified as having a severe risk of social vulnerability to a possible earthquake in Istanbul was 7,052 (%17.2) among 41,093 households. The median household size was 3, with values ranging

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

from 1 to 14 residents, and the median of the average age of the households varied between 8.8 to 85 years with the median being 35.5. The median of the average education was 8 years (Range: 0-17 years) in the entire survey sample, while it was 8.8 years (Range: 0-17 years) in those households with non-severe risk and 6 (Range: 0-16.3 years) in those households with severe risk. Additional comparisons between social vulnerability risk groups in terms of socio-demographic, health and socio-economic information are demonstrated in Table 2. In particular, households with severe risk were often overcrowded, less educated, older, had a low number of income earners, had low levels of savings and had less access to social security and health insurance compared to the non-severe risk group. The statistically significant variable with the largest effect on social vulnerability was the average education of the household (Cohen's d = 0.947), followed by the ratio of income earners (Cohen's d = 0.366) and the ratio of over 65 years olds in the household (Cohen's r = 0.120), having social security (Cramer's V = 0.211), having health security or insurance (Cramer's V = 0.226), having natural gas heating at home (Cramer's V = 0.152), the presence of anyone with a disability or who is elderly and needs care at home (Cramer's V = 0.142) and having savings for emergency situations (Cramer's V = 0.135).

Table 2. Univariate analysis of the study population characteristics.

Variables	Risk of Social Vulnerability		Effect size (Cohen's d[a] or Cohen's r[b] or Cramer's V[b)])	P
	Non-severe (n=34,041)	**Severe (n=7,052)**		
Socio-Demographics				
Household Size (HhS)			d = 0.178	<0.001
mean±sd	3.28±1.40	3.54±1.72		
median(min-max)	3 (1-13)	3 (1-14)		
Average education (years)			d = 0.947	<0.001
mean±sd	9.11±3.22	6.11±2.9		
median(min-max)	8.8 (0-17)	6 (0-16.3)		
Average age of the HH			d = 0.107	<0.001
mean±sd	38.28±14.49	39.87±16.65		
median(min-max)	35.5 (10.3-85.0)	36.4 (8.8-84.0)		
No. of women / HhS			d = 0.130	<0.001
mean±sd	0.48±0.23	0.51±0.23		
median(min-max)	0.5 (0-1)	0.5 (0-1)		
No. of men / HhS			d = 0.130	<0.001
mean±sd	0.52±0.23	0.49±0.23		
median(min-max)	0.5 (0-1)	0.5 (0-1)		

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

No. of <5 years old children / HhS			r = 0.010	<0.001
mean±sd	0.037±0.099	0.039±0.088		
median(min-max)	0 (0-0.7)	0 (0-0.7)		
No. of >65 years old individuals/HhS			r = 0.120	<0.001
mean±sd	0.09±0.24	0.15±0.30		
median(min-max)	0 (0-1)	0 (0-1)		
Number of income earners / HhS			d = 0.366	<0.001
mean±sd	0.53±0.28	0.43±0.24		
median(min-max)	0.5 (0-2)	0.3 (0-2)		
Social security	30956 (90.9)	5118 (72.6)	V = 0.211	<0.001
Membership to a non-governmental organisation	872 (2.6)	70 (1.0)	V = 0.040	<0.001
Health				
Health insurance	33563 (99.9)	6206 (88.0)	V = 0.226	<0.001
Any disabled or elderly who needs care in the Hh	1112 (3.3)	789 (11.2)	V = 0.142	<0.001
Health access	28309 (83.2)	5682 (80.6)	V = 0.026	<0.001
Socio-Economic				
Regular salary income	27342 (80.3)	4899 (69.5)	V = 0.100	<0.001
Pension income	11283 (33.1)	2320 (32.9)	V = 0.002	0.688
Rent/interest income	1794 (5.3)	180 (2.6)	V = 0.048	<0.001
Income support from public authorities	646 (1.9)	470 (6.7)	V = 0.111	<0.001
Risk of any job loss in Hh in an earthquake	11808 (34.7)	2790 (39.6)	V = 0.038	<0.001
Ownership of the house of residence	22105 (64.9)	4057 (57.5)	V = 0.058	<0.001
Status of the house of residence			V = 0.087	<0.001
Apartment flat	30453 (89.5)	5797 (82.2)		
Squatter house	912 (2.7)	379 (5.4)		
Detached/semi-detached house	2578 (7.6)	851 (12.1)		
Gate keepers lodge	98 (0.3)	25 (0.4)		
Natural gas heating at home	31164 (91.5)	5580 (79.1)	V = 0.152	<0.001
Ownership of any other house in Istanbul	5667 (16.6)	585 (8.3)	0.088	<0.001

Land ownership in Istanbul	2669 (7.8)	282 (4.0)	V = 0.056	<0.001
House ownership outside Istanbul	4210 (12.4)	491 (7.0)	V = 0.078	<0.001
Land ownership outside Istanbul	7092 (20.8)	889 (12.6)	V = 0.064	<0.001
Savings for emergency situation	5499 (16.2)	260 (3.7)	V = 0.135	<0.001
Any dept of Hh members	11009 (32.3)	2728 (38.7)	V = 0.051	<0.001

[a]0.2 = a small effect, 0.5 = a medium effect, 0.8 = a large effect. [b]0.1 = a small effect, 0.3 = a medium effect, 0.5 = a large effect. HhS: Household size. No: Number

3.2 Comparison of machine learning methods

The comparison of the machine learning models in terms of their sensitivity, specificity, balanced accuracy, and AUC under different subsampling methods are presented in Fig. 3. The additional comparisons of models using other evaluation metrics (e.g. positive prediction value, negative prediction value, accuracy, F1 score, Recall, Precision, etc.) can be found in the R-shiny application Within these comparisons, no substantial differences were observed in the model performance indicators of different ML strategies between RCV and bootstrap resampling methods. Therefore, we present the results that were obtained with 5-fold cross-validation.

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

Natural Hazards and Earth System Sciences
Open Access
Discussions

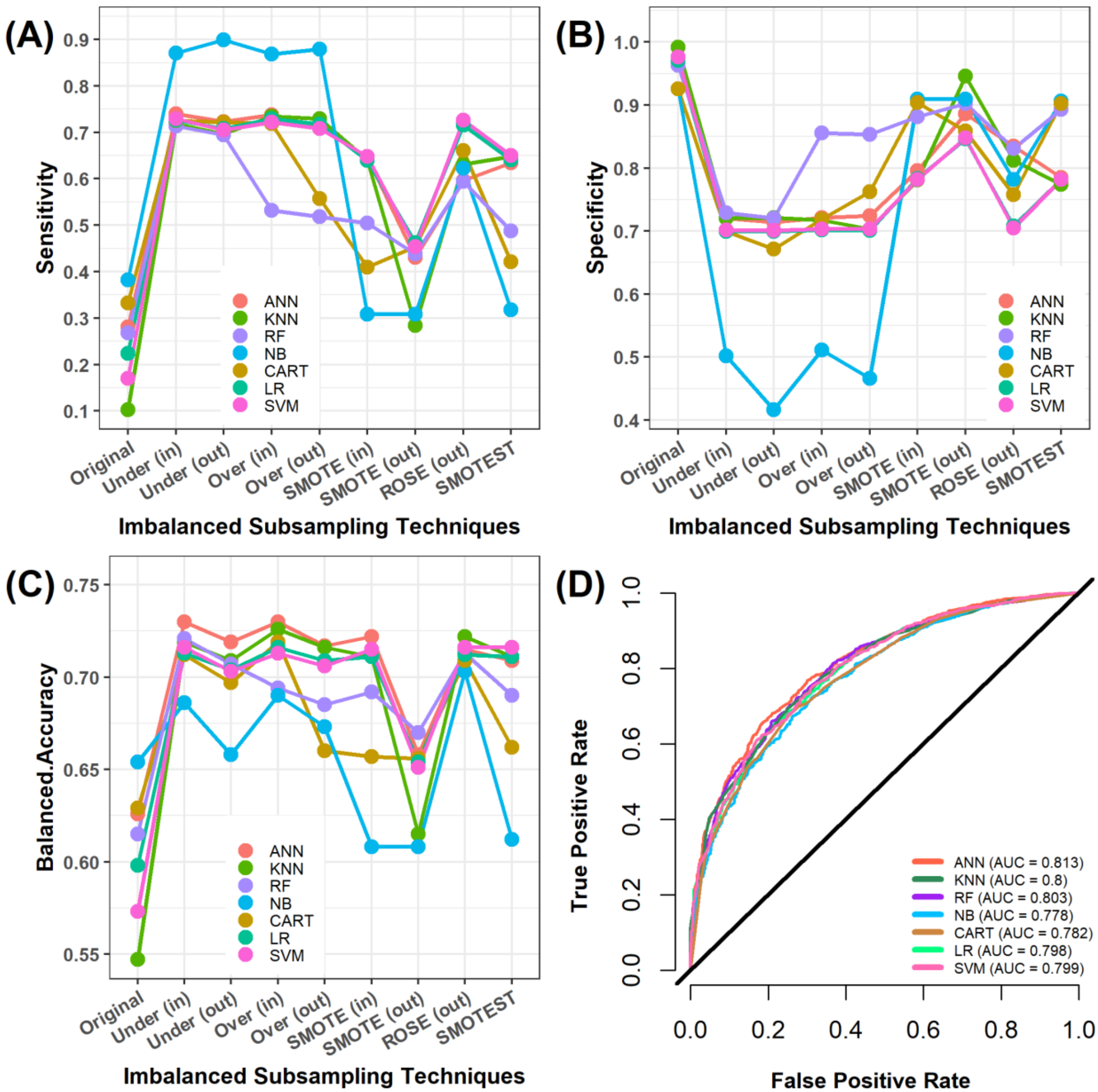

Fig 3. Model performance comparisons. ML methods are visualized in different colours in all figures. (A) Sensitivity (y-axis) in comparison to subsampling technique (x-axis). (B) Specificity (x-axis) in comparison to subsampling technique (y-axis). (C) Balanced accuracy ((sensitivity + specificity) / 2) (x-axis) in comparison to subsampling technique (y-axis). (D) Using the under(in) imbalanced subsampling technique, ML methods are compared in terms of the AUC of the ROC curve.

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

Natural Hazards and Earth System Sciences
Discussions

As mentioned earlier, the data set suffered from imbalanced class variables, particularly the outcome variable, and as such significant differences were observed when subsampling strategies were applied. Using the standard algorithm without subsampling (referred as “Original”) resulted in poor sensitivity (Fig. 3A), but inflated specificity (Fig. 3B) rates. Based on the criteria that AUC>0.7, overall, the methods fitted with under subsampling inside the resampling procedure (referred as under(in)) performed better in terms of model performance metrics when compared to other subsampling methods. The highest balanced accuracy for each method was also obtained with under(in) subsampling (Fig 3C).

In Table 3, all ML methods using under(in) subsampling were compared to their counterpart using the original data without imbalanced subsampling. When the subsampling strategy was not employed, all ML methods had inflated specificity due to the imbalance present in the studied study sample, where the negative class is dominant. Here we remind the reader that the priority in this study was to assess the performance of the models in terms of their success in identifying the households with severe risk of social vulnerability, which is the minority class, but therefore also the positive class. Using under(in) subsampling strategy demonstrated superior sensitivity and balanced accuracy rates compared to using original data and other subsampling strategies. Therefore, the results obtained with under(in) subsampling are considered for further comparisons between ML methods. Classification results for the ML models using under(in) subsampling are presented with ROC curves in Fig. 3D. The ROC curves for all other subsampling strategies with all other methods can be found in the R-shiny web application.

Table 3. Comparison of the model performances of ML methods using raw data and under(in) subsampling.

ML Models	AUC (95%CI)	Accuracy (95% CI)	Balanced Accuracy (95% CI)	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	Diff sens* (%) (95% CI)
Original data (no subsampling)						
LR	0.798 (0.776-0.820)	0.842 (0.830-0.853)	0.598 (0.573-0.623)	0.224 (0.194-0.257)	0.971 (0.965-0.976)	NA
CART	0.771 (0.752-0.790)	0.823 (0.811-0.835)	0.629 (0.610-0.649)	0.332 (0.297-0.368)	0.926 (0.916-0.934)	NA
RF	0.795 (0.775-0.815)	0.842 (0.830-0.853)	0.615 (0.598-0.632)	0.268 (0.236-0.303)	0.963 (0.955-0.969)	NA
SVM	0.738 (0.709-0.767)	0.836 (0.825-0.848)	0.573 (0.560-0.586)	0.170 (0.144-0.200)	0.976 (0.970-0.981)	NA
NB	0.784 (0.767-0.801)	0.832 (0.820-0.843)	0.654 (0.635-0.673)	0.382 (0.346-0.419)	0.926 (0.917-0.935)	NA

K-NN	0.805 (0.772-0.838)	0.838 (0.826-0.849)	0.547 (0.535-0.559)	0.102 (0.081-0.127)	0.992 (0.989-0.995)	NA
ANN	0.820 (0.801-0.839)	0.851 (0.840-0.862)	0.626 (0.609-0.643)	0.281 (0.248-0.316)	0.971 (0.964-0.976)	NA
Using Under (in) subsampling						
LR	0.798 (0.785-0.811)	0.704 (0.690-0.718)	0.713 (0.689-0.737)	0.726 (0.691-0.759)	0.699 (0.683-0.715)	0.502 (0.483-0.520)
CART	0.782[a] (0.768-0.796)	0.704 (0.690-718)	0.712 (0.690-0.734)	0.725 (0.690-0.757)	0.699 (0.684-0.715)	0.393 (0.373-0.413)
RF	0.803 (0.790-0.816)	0.722 (0.708-736)	0.713 (0.692-0.734)	0.711 (0.676-0.744)	0.724 (0.709-0.738)	0.443 (0.421-0.465)
SVM	0.799 (0.786-0.812)	0.707 (0.693-721)	0.715 (0.693-0.737)	0.729 (0.694-0.761)	0.702 (0.687-0.718)	0.559 (0.541-0.576)
NB	0.778[b] (0.763-0.793)	0.566[a] (0.550-0.581)	0.690 (0.671-0.710)	0.871[a] (0.843-0.894)	0.502[a] (0.485-0.519)	0.489 (0.471-0.507)
K-NN	0.800 (0.786-0.814)	0.720 (0.705-0.733)	0.719 (0.697-0.742)	0.719 (0.684-0.752)	0.720 (0.704-0.735)	0.617 (0.600-0.633)
ANN	0.813[a,b] (0.800-0.826)	0.724[a] (0.710-0.737)	0.730 (0.709-0.752)	0.740[a] (0.706-0.772)	0.720[a] (0.705-0.735)	0.459 (0.440-0.478)

*Diff sens: The difference in sensitivity between the same ML method with and without subsampling strategy for imbalanced problem. Same superscript letters indicate statistically significant difference in a performance measure between two methods, at α<0.05/7=0.007 significance level. CI: Confidence Interval. NA: Not Applicable.

The best performing method in terms of AUC, accuracy, balanced accuracy and sensitivity was artificial neural network (ANN) using under(in) subsampling strategy (AUC: 0.813 (0.800-0.826), Accuracy: 0.724 (0.710-0.737), Balanced accuracy : 0.730 (0.790-0.752), Sensitivity: 0.740 (0.706-0.772), Specificity: 0.720 (0.705-0.735)). The sensitivity + specificity was equal to 1.46 for ANN using under (in) and is considered to be in an acceptable range as the value is halfway between 1, which is useless, and 2, which is perfect (Power et al., 2013). Naïve Bayes (NB) also produced a high sensitivity rate of 0.871 (0.843-0.894), however it resulted in significantly lower specificity (0.502 (0.485-0.519)) and overall accuracy 0.566 (0.550-0.581) compared to ANN (p=0.003 and p<0.001, respectively). While ANN balances sensitivity (0.740) and specificity (0.720), NB emphasizes sensitivity (0.871) over specificity (0.502). All other methods using under(in) sampling provided similar sensitivity rates between the range of 71.9% and 72.9%, and specificity rates between 69.9% and 72.4%. When AUC was considered, CART was also significantly worse than ANN (0.782 (0.768-0.796) vs. 0.813 (0.800-0.826), p

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

= 0.005). Logistic regression, random forest, support vector machine and k-nearest neighbours did not show significant differences from ANN in terms of performance metrics.

3.3 Important predictors for the machine learning methods

In Fig. 4, a visual summary of the average relative importance of the predictors as indicated by the ML methods using under(in) sampling is presented. The most important variable for every model is given a score of 100%, followed by the next important variable which takes a relative value between 0 and 100. The variables which appeared in top ten most influential variables in all seven models were education, having social security, the ratio of income earners in the household and having savings for emergency situations (Fig. 4A). Of these variables, the variable with the highest average importance was education.

.

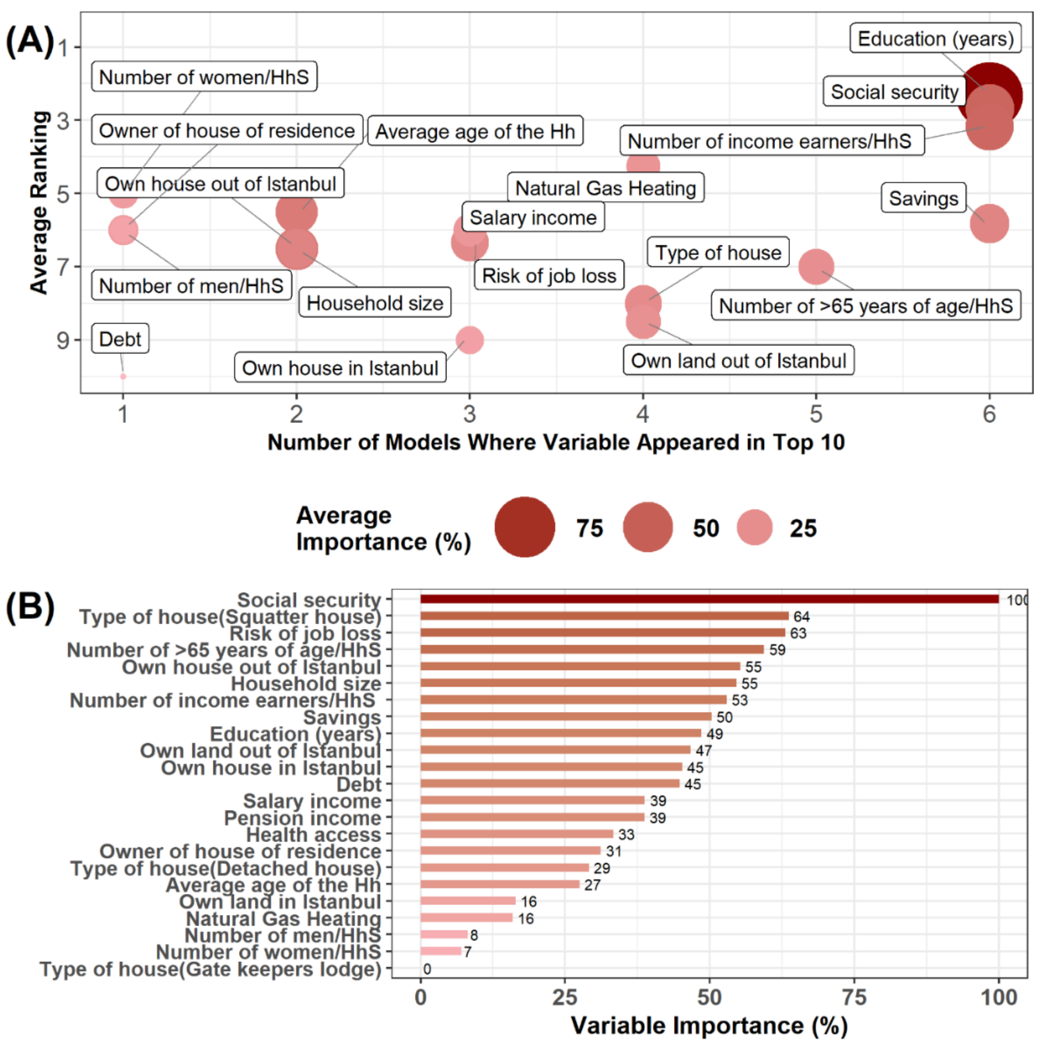

Fig. 4. Important predictors for the assessment of social vulnerability. (A) the average relative importance of the predictors obtained with ML methods using under(in) sampling. Average ranking of the predictor across all models (y-axis) in comparison to number of models that the predictor appeared in top ten most important variables (x-axis). (B) Variable importance for the ANN-under(in) model

In Fig. 4B we investigated the relative importance of the independent variables within the top performing model, ANN-under(in), using the approach suggested by (Garson, 1991). Based on this model, the most important variable for the classification of households' social vulnerability appeared to be having social security. The other predictors with over 50% of relative importance were a mixture of demographic and economic variables including living in a squatter house, risk of job loss in a possible earthquake, ratio of the over 65-year-olds in the household, owning a house outside of Istanbul, household size, ratio of income earners in the household and having savings for emergency situations

3.4 Spatial distribution of the important predictors of the ANN model

Based on the variable importance analysis with the top performing model, ANN-under(in), we performed area-based calculations to compare the neighbourhood characteristics in Istanbul. For categorical variables, the prevalence in the neighbourhood was calculated, while neighbourhood averages were used for the continuous variables. The three most important predictors of social vulnerability were subsequently displayed as a five-category map in Fig. 5.

For Fig. 5A, the areas represented with dark red colours, below 70%, indicates those neighbourhoods with the lowest social security and these areas are prevalent in the outer regions of the metropolitan area. On the other hand, those neighbourhoods close to the central region mostly cover households with higher prevalence of social security benefits. The number of neighbourhoods with high-density of squatter housing (>20%) was 27 (Fig. 5B). These neighbourhoods are scattered throughout the city and are not concentrated in any specific region. The households in which persons at risk of losing their jobs in a possible earthquake live, are mainly located in the central region of the city (Fig. 5C). The distribution of all other variables across neighbourhoods of Istanbul can be found the R-shinny web application.

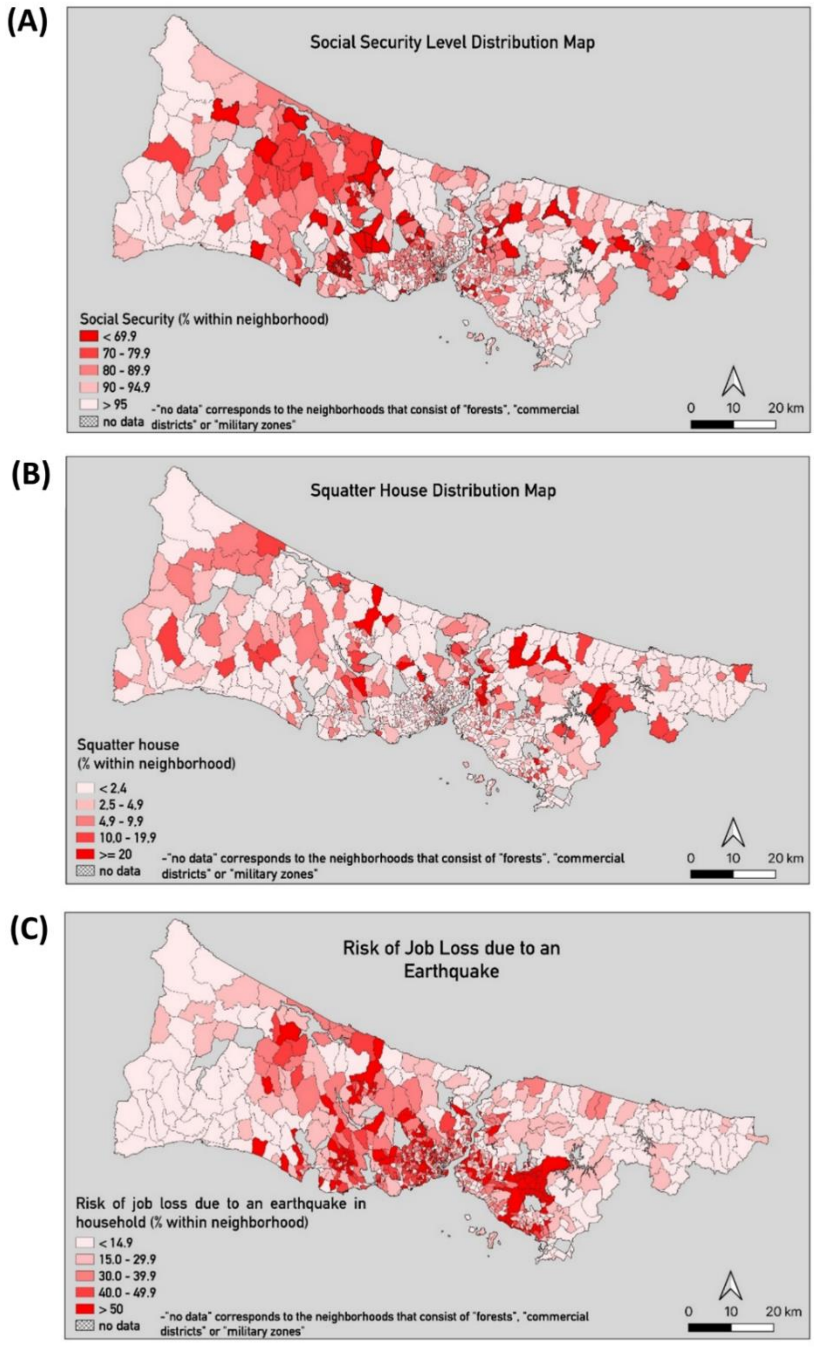

Fig. 5. The five-category neighbourhood map of the three most important predictors of social vulnerability. (A) Neighbourhood prevalence of having social security (B) Neighbourhood prevalence of living in squatter houses (C) Neighbourhood prevalence of risk of job loss of any household member in a possible earthquake

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

4 Discussion

4.1 The selection of the optimal ML method

Based on our classification results, the best performing ML method for identifying households with severe risk was ANN using under subsampling within the resampling procedure to address the problem of class imbalance (sensitivity: 0.740, balanced accuracy: 0.740, AUC:0.813). An AUC of 0.813 for ANN model indicated a good ability to discriminate households with severe risk of social vulnerability in the event of an earthquake in Istanbul from those with non-severe risk. For many decades, data analysis in the social sciences has focused on identifying causal links between a set of empirically derived variables (Di Franco and Santurro, 2021). Interrelated social relations between the variables in our data set may be best handled by ANN.

The methodology of the Artificial Neural Network drew inspiration from networks of biological neurons found in the nervous system (McCulloch and Pitts, 1943). For modelling purposes, it is often represented as interconnected groups of nodes (i.e. the predictors), in which subsequent processing between the nodes occurs according to their interconnections. This structure enhances the capacity for handling complex nonlinear relationships between dependent and independent variables in large data sets (Hornik et al., 1989). In quantitative social research, relationships between socio-demographic and socio-economic variables cannot be ignored (Meade et al., 1970). The use of ANN is therefore an effective tool for identifying hidden nonlinear relationships that arise in social research (Di Franco and Santurro, 2021). We note that apart from CART and NB, all methods provided similar AUC results with no significant differences. There was no significant difference between ML methods except with NB in terms of the performance of identifying households at risk of severe social vulnerability (i.e. sensitivity).

4.2 The importance of subsampling for imbalanced class variable

An important aspect of our study was to find the most viable solution for the imbalance problem in our dataset, as the imbalance ratio between the two groups was around 1/5. When no subsampling strategy was applied for the imbalance problem, we obtained poor sensitivity rates. A 39.3% to 61.7% gain in sensitivity was achieved when under(in) subsampling was applied, and therefore the imbalance was being addressed, compared to using the original raw data without subsampling. In our study, when ML models without subsampling strategies were used, the overall accuracy were higher due to the inflated specificity compared to the models using subsampling strategies. ML models are trained to maximize the overall accuracy and therefore, if trained on imbalanced data they are prone to over predict the class with higher frequency, which is non-severe vulnerability risk group in our data set (Esposito et al., 2021). Therefore, the models based on the original imbalanced data failed to identify households with severe risk of social vulnerability, and they failed to meet our aims in this study.

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

Among subsampling methods, random majority under-sampling approach resulted in the best performance for all ML methods. This method discards data points from the majority class at random until a more balanced distribution is reached. Our data set was sufficiently large not to be negatively affected by the discarding of data. Our results obtained with random under sampling are consistent with the ML literature, in the sense that if the size of the dataset is large, then it is better to employ an under-sampling method (Durahim, 2016).

4.3 Important variables and their theoretical implications

A favourable property of ML methods is that the importance of the independent variables in the models can be obtained. This can be done by two means: by computing the contribution of the variable to the model, as per the standard regression model, or by computing its contribution to the AUC of ROC curve. The variable importance rankings tend to differ between different ML models, as they use a different algorithm or weighting scheme. Saarela and Jauhiainen (2021) compared the variable importance measures obtained by different ML methods used for classification and showed that the most important features differ depending on the technique. In the present study, performance measures were not very different between ML methods, so first we averaged the rankings of the variables across seven ML models using under(in) subsampling. On average, education was found to be the most important variable in all ML methods, followed by having social security, the ratio of the income earners in the household and having savings to be used in emergency situations.

When we assessed the top performing model, which was ANN, the most important variable was found to be social security, followed by living in a squatter house and risk of job loss in a possible earthquake. Social security, meaning the right to have the guarantee of unemployment benefits, retirement pensions and public protection from job injuries, is gained through regular work and employment. In Turkey, the rate of unregistered labourers who are not affiliated with social security institution in total employment was recorded as 27.4% (Turkish Statistics Institute, 2021), while the most unregistered sectors are agriculture and service (Ocal and Senel, 2021). Unregistered employment means that no social insurance premiums will be paid by the employer, thus the employees cannot have the benefits of social security (Turkoglu, 2013). On the other hand, people in agriculture are mostly self-employed and do not have social security since they cannot regularly afford to pay the social security premiums. Hence the map we have presented on social security indicates the significance of having social security in the case of Istanbul households, also representing the general situation in Turkey (Turkoglu, 2013). The neighbourhoods towards the North-West of Istanbul represent mostly agricultural areas as well as households with lower social security. On the other hand, those neighbourhoods close to the centre of the metropolitan area cover mostly people employed in services and industrial sectors having a higher prevalence of social security benefits. Having social security is a significant indicator of social welfare and for the guarantee of general well-being of citizens, as explained above. The lack of social security and of insurance, particularly in a demonstrably unstable economy such as Turkey, increases vulnerability to many kinds of crises, including natural disasters and health emergencies such as pandemics. In the data presented, the prevalence of social security in the severe risk group is around 72% whereas in the non-severe risk group it is as high as 91%.

Natural Hazards and Earth System Sciences
Open Access
EGU
Discussions

Based on our findings, living in a squatter house was found to be the second most important variable in increasing the risk of social vulnerability using the ANN method. Squatter housing comprises houses that are assembled quickly and that do not conform to technical and legal standards, and as such represent grave vulnerability in the event of an earthquake and are more likely to result in building collapse. However, previous academic studies of earthquake engineers for Istanbul inform that a large proportion of buildings in Istanbul are not earthquake-resistant (IMM and KOERI, 2019; Parsons, 2004; JICA and IMM, 2002; Erdik et al., 2003; Ersoy and Koçak, 2016). In particular, squatter houses are very low-quality buildings, when taken together with the poor socio-economic characteristics of their residents, represent high social vulnerability for these households. Hence the building type indicator representing the illegal construction (with a distinct value called "*gecekondu*" as the Turkish name for poor squatter settlements) can be used for representing at-high-risk buildings, in the event of an earthquake, as they have not been built to withstand such an event. A study by Abarca-Alvarez et al. (2019) in Andalusia, which used a decision tree analysis, showed the importance of dwelling variables on social vulnerability, such as average age of constructions and the density of housing buildings in a census section urban area. In our study, age of the buildings was not available, however, the type of the house was found to be an important predictor of social vulnerability.

With the ANN method, the third highest ranked variable was the risk of job loss in a possible earthquake. Here, as mentioned above in social security indicator, the labour market opportunities in Turkey are highly dominated by the informal sector (Ocal and Senel, 2021). A recent study showed that informal employment increases social vulnerability to natural hazards (Mavhura and Manyangadze, 2021). These may be either in the form of casual, seasonal employment or self-employment, where social security and social insurance registrations are not provided by the employers. Most of those working in the informal sector are unregistered within the social security scheme. Most small and self-employed businesses are without security since they could not afford to pay their premiums regularly. These types of employees and small businesses mostly fall below the poverty line even if they may be observed as working (Adaman et al., 2015). These households depending on unregistered labour and small businesses in the informal sector have a high probability of experiencing vulnerability when a disaster strikes. In an earthquake, their workplaces may be damaged or closed which means a vulnerability risk for them due to loss of jobs or income. In the COVID-19 Pandemic, when small workplaces have been required to close or to restrict their services for a long period of time, most of these working people suffered severe job and income losses, hence severe vulnerability emerged (Bartik et al., 2020; Gray et al., 2022). One related factor for job loss is the duration of unemployment. Thus, the persons who live through long period of unemployment are also prone to severe vulnerability. As the third map indicates, the neighbourhoods in the centre of Istanbul are populated with small and informal workplaces, mostly with unregistered employees. When a disaster occurs, these groups living in neighbourhoods close to the city centre are at high risk of severe social vulnerability.

The other variables among the top ten most important predictors that contribute to the model performance of the ANN model were a mixture of demographic and economic variables. These included the ratio of over 65-year-olds in the household, owning a house outside of Istanbul, household size, the ratio of income earners in the household, having savings for emergency situations, owning land outside of Istanbul, and the level of education of the occupants. It is known that poorer people are more vulnerable to natural hazards as they settle in buildings at higher risk as they are more affordable to them. Furthermore, the associations between income and level of education are strong and consistent; that is children from poorer family backgrounds have a tendency of achieving a lower level of education (West, 2007). Also, the poor have less access to resources that reduce risks and therefore cannot take as many precautions to cope with a disaster when it occurs (Hallegatte et al., 2020).

5 Limitation and recommendations

We have found that socially, economically, and environmentally vulnerable communities are more likely to suffer disproportionately from disasters. However, our analysis was based solely on quantifiable household data, and variables related to environmental factors and building infrastructures were not available in our survey-based data set. Furthermore, the predictors in the survey data are specific to earthquake risk but not necessarily relevant to multiple disaster risks. Another important limitation regards the fact that we are using social vulnerability scores that are predefined in the previous survey study (phase one). As the urbanization process is always live in a vibrant city like Istanbul, the regeneration and renewal processes in Istanbul may cause possible changes in the location of residents and their socio-economic positions both upward and downward, which differs according to each urban project scheme. This may result in a continuous change and dynamic social vulnerability of households and neighbourhoods which needs to be studied in further research.

For future studies, we recommend using household data along with community level spatial predictors to enhance the predictive ability of the models. In addition, the spatial distribution of social vulnerability risk can further be detected along with the fault lines. We note that we could not perform a validation of the ML models using a separate and independent dataset due to the unavailability of such survey data derived from another source. Although the models were tested using an independent testing data from our survey data, the model predictions may benefit from validation studies which could be conducted using an additional dataset.

6 Conclusion

This research presents a new and alternative decision-making support tool for public authorities to develop ideas for future governance mechanisms based on interdisciplinarity. To address the first research question on determining the best performing ML method, we compared seven different supervised ML techniques which can be employed for binary classification with imbalanced class variables. We demonstrated that an ANN using majority under sampling was the optimum method in terms of sensitivity, AUC, and other relevant performance metrics. Results from the variable importance analysis fulfil the second research question regarding the most influential predictors of social vulnerability risk of households. The variable importance results showed that economically deprived households which do not have social security, those that have a high risk of job loss in the event of an earthquake, live in squatter houses and are less educated are at the highest risk of social vulnerability to earthquakes. We stress strongly and have demonstrated that our research outcomes have potential to support decision makers to develop more effective policies through prioritising the vulnerable target groups, understanding the perspective and preference of communities, considering urgency and high risk in exceptional locations and developing more sensitive and effective projects for the needs of the people expected to be affected. Furthermore, the local authorities, mainly Municipalities, can benefit from the results of this study in accordance with their disaster risk reduction activities in urban transformation processes, training, and awareness raising events. This study made use of machine learning methodology and assessed their performances on social data based on an interdisciplinary collaboration where the statistics, urban planning and sociology disciplines intersect, to understand disaster risk mitigation and how to build a society more resilient to natural disasters.

Code Availability: R codes can be obtained by contacting Oya Kalaycioglu at her e-mail address: oya.kalaycioglu.09@ucl.ac.uk

Data Availability: Data are available from the authors with the permission of Istanbul Metropolitan Municipality, Directorate of Earthquake and Ground Research.

Author Contribution: OK and YEM planned the initial concept of the study. OK led the writing of the manuscript with contributions from all the co-authors. OK and SEA implemented the data analysis, trained ML models, and designed R-shiny web application. YEM obtained the data and designed Fig. 5. MK and SK contributed to literature review on social vulnerability and interpret the findings. All authors critically reviewed the manuscript

Competing interests: The authors declare that they have no conflict of interest.

References

Abarca-Alvarez, F. J., Reinoso-Bellido, R., and Campos-Sánchez, F. S.: Decision Model for Predicting Social Vulnerability Using Artificial Intelligence, ISPRS Int. J. Geo-Inf., 8, 575, https://doi.org/10.3390/ijgi8120575, 2019.

Adaman, F., Aslan, D., Erus, B., and Sayan, S.: ESPN Thematic Report on in-work poverty in Turkey, European Comission, Brussels, 2015.

Akhanli, S. E. and Hennig, C.: Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes, Stat. Comput., 30, 1523–1544, https://doi.org/10.1007/s11222-020-09958-2, 2020.

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

Aksha, S. K., Juran, L., Resler, L. M., and Zhang, Y.: An Analysis of Social Vulnerability to Natural Hazards in Nepal Using a Modified Social Vulnerability Index, Int. J. Disaster Risk Reduct., 10, 103–116, https://doi.org/10.1007/s13753-018-0192-7, 2019.

Alizadeh, M., Alizadeh, E., Asadollahpour Kotenaee, S., Shahabi, H., Beiranvand Pour, A., Panahi, M., Bin Ahmad, B., and Saro, L.: Social Vulnerability Assessment Using Artificial Neural Network (ANN) Model for Earthquake Hazard in Tabriz City, Iran, Sustainability, 10, 3376, https://doi.org/10.3390/su10103376, 2018.

Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., and Van Dooren, W.: Beyond small, medium, or large: points of consideration when interpreting effect sizes, Educational Studies in Mathematics, 102, 1–8, https://doi.org/10.1007/s10649-019-09908-4, 2019.

Bartik, A. W., Bertrand, M., Cullen, Z., Glaeser, E. L., Luca, M., and Stanton, C.: The impact of COVID-19 on small business outcomes and expectations, Proceedings of the national academy of sciences, Proceedings of the National Academy of Sciences, 117, 17656–17666, https://doi.org/10.1073/pnas.2006991117, 2020.

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C.: A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data, ACM SIGKDD Explorations Newsletter, 6, 20–29, https://doi.org/10.1145/1007730.1007735, 2004.

Bergstrand, K., Mayer, B., Brumback, B., and Zhang, Y.: Assessing the Relationship Between Social Vulnerability and Community Resilience to Hazards, Soc. Indic. Res., 122, 391–409, https://doi.org/10.1007/s11205-014-0698-3, 2015.

Blaikie, P., Cannon, T., Davis, I., and Wisner, B. (Eds.): At risk: Natural hazards, people's vulnerability and disasters, 2nd ed., Routledge, London, UK, 2014.

Bolin, R. and Stanford, L.: Shelter, Housing and Recovery: A Comparison of U.S. Disasters, Disasters, 15, 24–34, https://doi.org/10.1111/j.1467-7717.1991.tb00424.x, 1991.

Buskirk, T. D., Kirchner, A., Eck, A., and Signorino, C. S.: An Introduction to Machine Learning Methods for Survey Researchers, Surv. Pract, 11, https://doi.org/10.29115/SP-2018-0004, 2018.

Cannon, T.: Reducing People's Vulnerability to Natural Hazards: Communities and Resilience, The United Nations University World Institute for Development Economics Research (UNUWIDER), http://hdl.handle.net/10419/45089, 2008.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: Synthetic Minority Over-sampling Technique, JAIR, 16, 321–357, https://doi.org/10.1613/jair.953, 2002.

Cova, T. J. and Church, R. L.: Modelling community evacuation vulnerability using GIS, Int. J. Geogr. Inf. Syst., 11, 763–784, https://doi.org/10.1080/136588197242077, 1997.

Cutter, S. L.: Societal responses to environmental hazards, Int. Soc. Sci. J., 48, 525–536, https://doi.org/10.1111/1468-2451.00053, 1996.

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

Cutter, S. L., Mitchell, J. T., and Scott, M. S.: Revealing the Vulnerability of People and Places: A Case Study of Georgetown County, South Carolina, Ann. Assoc. Am. Geogr., 90, 713–737, https://doi.org/10.1111/0004-5608.00219, 2000.

Cutter, S. L., Boruff, B. J., and Shirley, W. L.: Social Vulnerability to Environmental Hazards, Soc. Sci. Q., 84, 242–261, https://doi.org/10.1111/1540-6237.8402002, 2003.

Debesai, M. G.: Factors affecting vulnerability level of farming households to climate change in developing countries: evidence from Eritrea, IOP Conf. Ser.: Mater. Sci. Eng., 1001, 012093, https://doi.org/10.1088/1757-899x/1001/1/012093, 2020.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L.: Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, Biometrics, 44, 837–845, https://doi.org/10.2307/2531595, 1988.

Di Franco, G. and Santurro, M.: Machine learning, artificial neural networks and social research, Qual. Quant., 55, 1007–1025, https://doi.org/10.1007/s11135-020-01037-y, 2021.

Durahim, A. O.: Comparison of sampling techniques for imbalanced learning, YBS Dergi, 2, 181–191, 2016.

Duzgun, H. S. B., Yucemen, M. S., Kalaycioglu, H. S., Celik, K., Kemec, S., Ertugay, K., and Deniz, A.: An integrated earthquake vulnerability assessment framework for urban areas, Nat. Hazards, 59, 917, https://doi.org/10.1007/s11069-011-9808-6, 2011.

Dwyer, A., Zoppou, C., Nielsen, O., Day, S., and Roberts, S.: Quantifying social vulnerability: a methodology for identifying those at risk to natural hazards, Geoscience Australia Record, 2004.

Erdik, M., Aydinoglu, N., Fahjan, Y., Sesetyan, K., Demircioglu, M., Siyahi, B., Durukal, E., Ozbey, C., Biro, Y., Akman, H., and Yuzugullu, O.: Earthquake risk assessment for Istanbul metropolitan area, Earthq. Eng. Eng. Vib., 2, 1–23, https://doi.org/10.1007/BF02857534, 2003.

Ersoy, Ş. and Koçak, A.: Disasters and earthquake preparedness of children and schools in Istanbul, Turkey, Geomat. Nat. Haz. Risk, 7, 1307–1336, https://doi.org/10.1080/19475705.2015.1060637, 2016.

Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., and Riniker, S.: GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning, J. Chem. Inf. Model., 61, 2623–2640, https://doi.org/10.1021/acs.jcim.1c00160, 2021.

Fatemi, F., Ardalan, A., Aguirre, B., Mansouri, N., and Mohammadfam, I.: Social vulnerability indicators in disasters: Findings from a systematic review, Int. J. Disaster Risk Reduct., 22, 219–227, https://doi.org/10.1016/j.ijdrr.2016.09.006, 2017.

Fritz, C., Morris, P., and Richler, J.: Effect Size Estimates: Current Use, Calculations, and Interpretation, J. Exp. Psychol. Gen., 141, 2-18. doi:10.1037/a0024338, 2011.

Garson, G. D.: Interpreting Neural-Network Connection Weights, AI Weights, 6, 46–51, 1991.

Gray, B. J., Kyle, R. G., Song, J., and Davies, A. R.: Characteristics of those most vulnerable to employment changes during the COVID-19 pandemic: a nationally representative cross-sectional study in Wales, J. Epidemiol. Community Health, 76, 8-15, doi:10.1136/jech-2020-216030, 2022

Hallegatte, S., Vogt-Schilb, A., Rozenberg, J., Bangalore, M., and Beaudet, C.: From Poverty to Disaster and Back: a Review of the Literature, EconDisCliCha, 4, 223–247, https://doi.org/10.1007/s41885-020-00060-5, 2020.

Hennig, C. and Liao, T. F.: How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification, R. Stat. Soc. C: Appl. Stat., 62, 309–369, https://doi.org/10.1111/j.1467-9876.2012.01066.x, 2013.

Holand, I. S., Lujala, P., and Rød, J. K.: Social vulnerability assessment for Norway: A quantitative approach, Nor Geogr Tidsskr., 65, 1–17, https://doi.org/10.1080/00291951.2010.550167, 2011.

Hornik, K., Stinchcombe, M., and White, H.: Multilayer feedforward networks are universal approximators, Neural Netw., 2, 359–366, https://doi.org/10.1016/0893-6080(89)90020-8, 1989.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (Eds): Applied Logistic Regression, 1st ed., Wiley & Sons, New Jersey, US, 2013.

IMM (Istanbul Metropolitan Municipality), and KOERI (Kandilli Observatory and Earthquake Research Institute): İstanbul İli Olası Deprem Kayıp Tahminlerinin Güncellenmesi Projesi (Updating The Earthquake Loss Estimation for Istanbul), Istanbul, 2019.

IMM (Istanbul Metropolitan Municipality), Directorate of Earthquake and Ground Research: Afetler Karsisinda Sosyal Hasargörebilirlik Sonuç Raporu (Final Report of Survey Study for Social Vulnerability to Natural Disasters), Istanbul, 2018.

JICA (Japan International Cooperation Agency) and IMM (Istanbul Metropolitan Municipality): The study on a disaster prevention/mitigation basic plan in Istanbul including seismic microzonation in the Republic of Turkey, Istanbul, 2002.

Kalaycioglu, S., Rittersberger, H., Çelik, K., and Gunes, F.: Integrated natural disaster risk assessment: The socio-economic dimension of earthquake risk in the urban area, Proceedings Geohazards Engineering Conferences International, Lillehammer, Norway, June 18-21, 2006.

Kim, S. and Lee, W.: Does McNemar's test compare the sensitivities and specificities of two diagnostic tests?, Stat. Methods Med. Res., 26, 142–154, https://doi.org/10.1177/0962280214541852, 2017.

Kuhn, M.: Building Predictive Models in R Using the caret Package, J. Stat. Sotfw., 28, https://doi.org/10.18637/jss.v028.i05, 2008.

Kuhn, M. and Johnson, K.: Applied predictive modeling, Springer, New York, US, 2013.

Lange, D., Kopp, H., Royer, J.-Y., Henry, P., Çakir, Z., Petersen, F., Sakic, P., Ballu, V., Bialas, J., Özeren, M. S., Ergintav, S., and Géli, L.: Interseismic strain build-up on the submarine North Anatolian Fault offshore Istanbul, Nat. Commun., 10, 3006, https://doi.org/10.1038/s41467-019-11016-z, 2019.

Lin, H. I. and Nguyen, M. C.: Boosting Minority Class Prediction on Imbalanced Point Cloud Data, Appl. Sci., 10, 973, https://doi.org/10.3390/app10030973, 2020.

Llorente-Marrón, M., Díaz-Fernández, M., Méndez-Rodríguez, P., and González Arias, R.: Social Vulnerability, Gender and Disasters. The Case of Haiti in 2010, Sustainability, 12, 3574, https://doi.org/10.3390/su12093574, 2020.

Maheshwari, S., Jain, D. R., and Jadon, D. S.: A Review on Class Imbalance Problem: Analysis and Potential Solutions, IJCSI, 14, 43-51, https://doi.org/10.20943/01201706.4351, 2017.

Mavhura, E. and Manyangadze, T.: A comprehensive spatial analysis of social vulnerability to natural hazards in Zimbabwe: Driving factors and policy implications, Int. J. Disaster Risk Reduct., 56, 102139, https://doi.org/10.1016/j.ijdrr.2021.102139, 2021.

McCulloch, W. S. and Pitts, W.: A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biol., 5, 115–133, https://doi.org/10.1007/BF02478259, 1943.

Meade, J. E., Wrigley, E. A., Brass, W., Boreham, A. J., Glass, D. V., and Grebenik, E.: Demography and Economics, Popul. Stud., 24, 25, https://doi.org/10.2307/2172399, 1970.

Menardi, G. and Torelli, N.: Training and assessing classification rules with imbalanced data, Data Min. Knowl. Discov., 28, 92–122, https://doi.org/10.1007/s10618-012-0295-5, 2014.

Menteşe, E. Y., Kalaycıoğlu, S., Çelik, K., Türkyılmaz, A. S., Çelen, Ü., Kara, S., Kılıç, O., Baş, M., and Uğur, C.: Understanding Social Vulnerability Against Disasters in Istanbul, Geophysical Research Abstracts, 21st EGU General Assembly 2019, Vienna, Austria, 7-12 April 2019, 2019.

Mitchell, J.: Crucibles of Hazard: Mega-Cities and Disasters in Transition, Journal of Government Information, 27, 916-918, https://doi.org/10.1016/S1352-0237(00)00249-5, 2000.

Morrow, B. H.: Identifying and Mapping Community Vulnerability, Disasters, 23, 1–18, https://doi.org/10.1111/1467-7717.00102, 1999.

Noriega, G. R. and Ludwig, L. G.: Social vulnerability assessment for mitigation of local earthquake risk in Los Angeles County, Nat. Hazards, 64, 1341–1355, https://doi.org/10.1007/s11069-012-0301-7, 2012.

Ocal, M. and Senel, D.: Türkiye'de Kayıt Dışı İstihdamın Bölgesel Analizi (Regional Analysis of Informal Employment in Turkey), Çalışma ve Toplum Dergisi, 2, 1201–1232, 2021.

Ogie, R. I. and Pradhan, B.: Natural Hazards and Social Vulnerability of Place: The Strength-Based Approach Applied to Wollongong, Australia, Int. J. Disaster Risk Reduct., 10, 404–420, https://doi.org/10.1007/s13753-019-0224-y, 2019.

Parsons, T.: Recalculated probability of M ≥ 7 earthquakes beneath the Sea of Marmara, Turkey, J. Geophys. Res., 109, B05304, https://doi.org/10.1029/2003JB002667, 2004.

Parsons, T., Toda, S., Stein, R. S., Barka, A., and Dieterich, J. H.: Heightened Odds of Large Earthquakes near Istanbul: An Interaction-Based Probability Calculation, Science, 288, 661–665, http://doi: 10.1126/science.288.5466.661, 2000.

Power, M., Fell, G., and Wright, M.: Principles for high-quality, high-value testing, Evid. Based Med., 18, 5–10, https://doi.org/10.1136/eb-2012-100645, 2013.

Puente, S.: Social vulnerability to disasters in Mexico City: An assessment method, in: Crucibles of Hazards. Mega-cities and Disasters in Transition, edited by: United Nations University, Tokyo, Japan, 295–334, 1999.

QGIS Development Team.: QGIS Geographic Information System, Open Source Geospatial Foundation Project, http://qgis.osgeo.org, 2021.

R Core Team.: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www. R-project. org/, 2013.

Rahman, Md. M., Rahman, Md. M., and Jerin, T.: Social vulnerability to earthquake disaster: insights from the people of 48th ward of Dhaka South City Bangladesh, Environ. Hazards, http://doi: 10.1080/17477891.2022.2085075, 2022.

Ramyachitra, R. and Manikandan, P.: Imbalanced dataset classification and solutions: a review, IJCBR, 5, 1–29, 2014.

Rufat, S., Tate, E., Emrich, C. T., and Antolini, F.: How Valid Are Social Vulnerability Models?, Ann. Am. Assoc. Geogr., 109, 1131–1153, https://doi.org/10.1080/24694452.2018.1535887, 2019.

Ryo, M. and Rillig, M. C.: Statistically reinforced machine learning for nonlinear patterns and variable interactions, Ecosphere, 8, e01976, https://doi.org/10.1002/ecs2.1976, 2017.

Saarela, M. and Jauhiainen, S.: Comparison of feature importance measures as explanations for classification models, SN Appl. Sci., 3, 272, https://doi.org/10.1007/s42452-021-04148-9, 2021.

Shen, S., Cheng, C., Yang, J., and Yang, S.: Visualized analysis of developing trends and hot topics in natural disaster research, PLoS One, 13, e0191250, https://doi.org/10.1371/journal.pone.0191250, 2018.

Spielman, S. E., Tuccillo, J., Folch, D. C., Schweikert, A., Davies, R., Wood, N., and Tate, E.: Evaluating social vulnerability indicators: criteria and their application to the Social Vulnerability Index, Nat. Hazards, 100, 417–436, https://doi.org/10.1007/s11069-019-03820-z, 2020.

Syed, A. and Kumar Routray, J.: Vulnerability assessment of earthquake prone communities in Baluchistan, IJDRBE, 5, 144–162, https://doi.org/10.1108/IJDRBE-12-2010-0053, 2014.

Turkish Statistics Institute, Labour Force Statistics 2021 : https://data.tuik.gov.tr/Bulten/Index?p=Labour-Force-Statistics-February-2021-37487&dil=2, last access: 21 March 2022, 2021.

Turkoglu, İ.: Sosyal devlet bağlamında Türkiye'de sosyal yardım ve sosyal güvenlik (Social Aid And Social Security in Turkey Within Social State Context), Akademik İncelemeler Dergisi, 8, 275–305, 2013.

United Nations International Strategy for Disaster Reduction (UNISDR), UNISDR Terminology on Disaster Risk Reduction: https://unisdr.org/files/7817_UNISDRTerminologyEnglish.pdf, last access: 21 April 2021, 2009.

Walker, T., Kawasoe, Y., and Shrestha, J.: Risk and Vulnerability in Nepal, Findings from the Household Risk and Vulnerability Survey, World Bank, Washington DC, US, 2019.

https://doi.org/10.5194/nhess-2022-198
Preprint. Discussion started: 20 July 2022

West, A.: Poverty and educational achievement: why do children from low-income families tend to do less well at school?, J. Poverty Soc. Justice, 15, 283–297, 2007.

White, G. F.: The hidden costs of coastal hazards: Implications for risk assessment and mitigation, The H. John Heinz III Center for Science, Economics, and the Environment, Island Press, 2000.

715 Wisner, B., Gaillard, J. C., and Kelman, I.: Framing disaster: theories and stories seeking to understand Hazards, vulnerability and risk, in: Handbook of Hazards and Disaster Risk Reduction, edited by: Wisner, B., Gaillard, J., and Kelman , I., Routledge, London, UK, 18–34, 2012.

World Urbanization Prospects (WUP): United Nations population estimates and projections of major Urban Agglomerations, https://worldpopulationreview.com/world-cities , last access: 16 April 2022, 2021.

720 Yoon, D. K. and Jeong, S.: Assessment of Community Vulnerability to Natural Disasters in Korea by Using GIS and Machine Learning Techniques, in: Quantitative Regional Economic and Environmental Analysis for Sustainability in Korea, edited by: Kim, B. and Kim, B. H. S., Springer, Singapore, 123–140, 2016.

Yücel, G. and Görün, A.: Earthquake and Physical and Social Vulnerability Assessment for Settlements: Case Study Avcılar District, Megaron, 5, 23-32, 2010.