ETÜSBED

Araştırma Makalesi | *Research Article* DOI: 10.29157/etusbed.1292006

Estimation of Tax Loss and Evasion in Turkey With Data Mining Process

Veri Madenciliği Süreci ile Türkiye'de Vergi Kayıp ve Kaçakların Tahmini

Derya ŞENCAN

T.C. Hazine ve Maliye Bakanlığı, Gelir İdaresi Başkanlığı, Isparta-Türkiye
e-mail: sencanderya80@gmail.com

Geliş Tarihi/Received: 4.5.2023
Kabul Tarihi/Accepted: 24.8.2023

Sorumlu Yazar/Corresponding Author:
Derya ŞENCAN
sencanderya80@gmail.com

Atıf /Cite this article:
Şencan, D. (2023). Estimation of Tax Loss and Evasion in Turkey With Data Mining Process. *Erzurum Teknik Üniversitesi Sosyal Bilimler Enstitüsü Dergisi,* 17, 11-24.

Öz

Vergi kayıp ve kaçakları tüm dünyada olduğu gibi ülkemizde de en büyük sorunlardan biridir. Denetimlerin yanı sıra, istatistiksel teknikler ve makine öğrenimi algoritmaları da vergi kayıp ve kaçaklarının tespitinde büyük önem taşımaktadır. Bu çalışmada vergi kayıp ve kaçak oranı; enflasyon oranı, işsizlik, vergi yükü, cari açık, ekonomik büyüme (GSYİH), devletin büyüklüğü gibi faktörlere bağlı olarak veri madenciliği süreci ile tahmin edilmiştir. Veri madenciliği sürecinde on iki modelleme tekniği kullanılmıştır. Her modelden elde edilen sonuçlar karşılaştırılmış ve bazı istatistiksel göstergeler kullanılarak en iyi model belirlenmiştir. Buna göre, vergi kayıp ve kaçak oranı tahmininde en başarılı sonucu R2, MAE ve RMSE değerleri sırasıyla 0,931, 0,2356 ve 0,2473 olan Gaussian processes modeli vermiştir. Vergi kayıp ve kaçak oranını etkileyen değişkenlerin ağırlık değerleri duyarlılık analizi ile belirlenmiştir. Vergi kayıp ve kaçaklarında pozitif etkisi en yüksek olan faktörlerin işsizlik ve enflasyon oranları olduğu görülmüştür. Bu faktörleri vergi yükü ve GSYİH değerleri izlemektedir. Devletin büyüklüğü ve cari açık faktörlerinin ise vergi kayıp ve kaçak oranı üzerinde negatif etkiye sahip olduğu görülmüştür. Çalışmadan elde edilen sonuçların ülkemizdeki vergi kayıp ve kaçak oranının tahmin edilmesine katkı sağlayacağı düşünülmektedir.

Anahtar Kelimeler: Vergi kaybı, Vergi kaçağı, Veri madenciliği.

Abstract

Tax losses and evasion are one of the biggest problems in our country as well as all over the world. In addition to audits, statistical techniques and machine learning algorithms are also of great importance in detecting tax losses and evasion. In this study, the tax loss and evasion rate has been estimated by the data mining process depending on factors such as inflation rate, unemployment, tax burden, trade openness, economic growth (GDP), and the size of government. Twelve modeling techniques were used in the data mining process. The results obtained from each model were compared and the best model was determined using some statistical indicators. Accordingly, the Gaussian processes model gave the most successful result in estimating tax loss and evasion rate, with R2, MAE and RMSE values of 0.931, 0.2356 and 0.2473, respectively. The weight values of the variables affecting the tax loss and evasion rate were determined by sensitivity analysis. It has been observed that the factors with the highest positive effect on tax losses and evasion are unemployment and inflation rates. These factors are followed by tax burden and GDP values. It was seen that the size of government and the trade openness factors had a negative effect on the tax loss and evasion rate. It is thought that the results obtained from the study will contribute to the estimation of the tax loss and evasion rate in our country.

Keywords: Tax loss, Tax evasion, Data mining.

Introduction

In order for a society to develop and increase its welfare level, essential needs such as health, education, security, transportation and communication must be met by the state. In order for the state to provide these services, it needs tax, which is the main source of financing. Tax is the money that the government collects from taxpayers to meet its financing needs. However, taxes are collected incompletely or cannot be collected at all due to reasons such as high tax rates or lack of tax awareness, as it causes a decrease in taxpayers' income. This situation has caused tax loss and evasion to become one of the biggest problems in our country as well as all over the world.

Tax loss is the tax effect of both the revenues that the state gave up on its own will and the revenues that could not be reached to the treasury without its own will. The state, at its own will, may forgo the income it will generate through exceptions, discounts and other incentives to companies through laws. In some cases, businesses can take their income out of taxes by taking advantage of legal deficits, which is called tax avoidance, but corporations may not declare their income to the state through illegal means. In this case, the concepts of tax evasion and informal economy appear. Regardless of the reason, the state incurs tax losses due to these situations.

There are many reasons for tax losses and evasion in our country, including economic, financial, administrative, political, legal, social-psychological. In the 1950s, when the Turkish Tax System was established, taxpayers, who were in a closed mixed economic structure, sought ways to avoid taxes in this structural transformation in the tax system, since they were not yet prepared for the tax system based on declaration (Armağan, 2016). Economic problems (inflation, budget deficits, and balance of payments deficits), political problems (frequent political changes, implementation of populist policies) and social problems (low level of welfare and living standards of the society) in Turkey in the post-1980 period have reduced the effectiveness of taxation (Demircan, 2004). Among the main causes of tax losses and evasion, there is primarily the "shadow economy" (Çomaklı, 2008). Tax burden is seen among the most important causes of informality and tax losses and evasion in our country. In addition, studies in the literature show that inflation is the most important determinant of the tax burden (Ay et al., 2014). Other causes of tax loss and evasion include the excess of exceptions and exemptions, lack of effectiveness in auditing, lack of deterrence in penalties, not clear enough tax laws, frequent tax amnesties, lack of tax awareness, and inability to provide voluntary tax compliance (Gerçek and Uygun, 2022). Although there are many studies on the estimation of tax evasion with different methods for some countries, there is no study in the literature on the estimation of tax evasion with data mining method in Turkey, which is one of the developing countries. Some studies on tax evasion estimation are given in Table 1. As can be seen in Table 1, different methods are used in tax evasion estimation studies. Madžarević-Šujster (2002), Petanlar et al. (2011), Amoh and Adafula (2019), Dunem and Arndt (2009), Angour and Nmili (2019), Dell'Anno and Davidescu (2019), Athanasios et al. (2020), Albarea et al. (2020), Levin and Widell (2014), Uyar et al. (2021) successfully used econometric methods to investigate the relationship between macroeconomic variables and tax evasion. Tabandeh et al. (2012), Xiangyu et al. (2018), Tabandeh and Tamadonnejad (2015), Raikov (2021), Rahimikia (2017), Faúndez-Ugalde et al. (2020), Hemberg et al. (2016), Warner et al. (2015), Zumaya et al. (2021), Shakil and Tasnia (2022) used artificial intelligence and machine learning methods in estimating tax loss and evasion. The obtained results from these studies indicate that artificial intelligence and machine learning models can achieve great accuracy. So, these methods can be used successfully in the estimation of tax loss and evasion.

Table 1: Summary of previous studies on tax evasion estimation

Authors	Year	Country	Variables	Methodology
Madžarević-Šujster	2002	Croatia	GDP, adjusted employment, income tax	Causality method

Petanlar et al.	2011	Iran	Tax burden, GDP, bugdet deficit	Currency Demand
Amoh and Adafula	2019	Ghana	Tax burden, unemployment, mobile money activities	Autoregressive distributed lag model
Tabandeh et al.	2012	Malaysia	Tax burden, the size of governments, inflation rate income, trade openness	Artificial Neural Network
Dunem and Arndt	2009	Mozambique	Tax rates	Baseline and Augmented Model
Angour and Nmili	2019	Morocco	Unemployment, inflation, urbanization, tax burden, openness rate, public spending index	MIMIC model
Dell'Anno and Davidescu	2019	Romania	Direct and indirect taxes, unemployment, GDP	MIMIC model
Athanasios et al.	2020	Greece	Total tax revenue, GDP	Currency Demand
Albarea et al.	2020	Italy	Personal income tax	Microsimulation
Xiangyu et al.	2018	China	VAT Tax Burden, maintenance margin, agent insurance	Neural Network
Levin and Widell	2014	Kenya and Tanzania	The tax data (import duty rates, VAT rates and excise duty rates)	Baseline and Augmented Model
Tabandeh and Tamadonnejad	2015	Malaysia	Unemployment	Artificial Neural Network
Uyar et al.	2021	Different countries	Innovation capacity	Granger causality
Raikov	2021	Russia	Tax evasion	Neural networks
Rahimikia	2017	Iran	Financial variables	Hybrid intelligent

Faúndez-Ugalde et al.	2020	Latin American countries	Characterization of taxpayers	Artificial intelligence
Hemberg et al.	2016	USA	Tax evasion	Genetic algorithms
Warner et al.	2015	USA	Tax evasion	Genetic algorithms
Zumaya et al.	2021	Mexico	Tax evasion	Machine learning
Shakil and Tasnia	2022	Asia and Pacific	Tax evasion	Artificial intelligence

Causing tax evasion in order to take advantage of tax avoidance and tax evasion causes the state to experience tax loss. There are various factors in the literature that can have a direct impact or cause tax evasion. In this study, unlike the literature, tax loss and evasion rates were estimated using data mining method, depending on the basic factors such as inflation rate, unemployment, tax burden, economic growth (GDP), size of government and trade openness, which lead to tax evasion.

Data mining

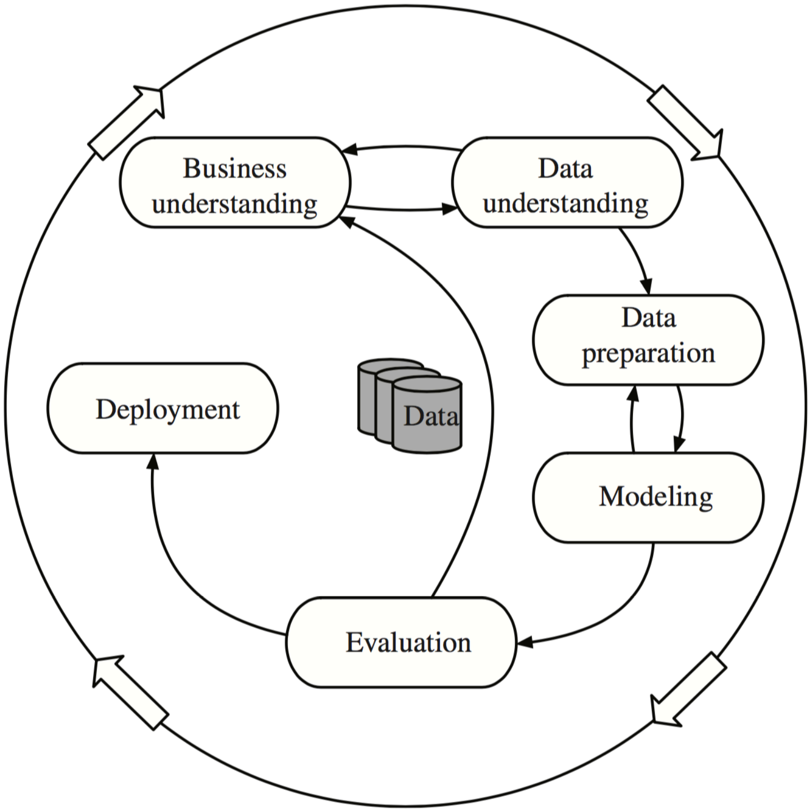
Data mining is a multidisciplinary area that combines computer science and statistics to extract information from a dataset and convert it into a comprehensible form for future use. This technique is widely used in various fields, including banking, finance, healthcare, communications, medicine, and engineering. Data mining involves various operations such as association, clustering, prediction, and classification. The data mining process consists of six main stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, as shown in Fig. 1.

The Business Understanding stage is essential to define the project objectives and create a preliminary plan. The Data Understanding stage is focused on collecting and understanding the available data. The Data Preparation phase involves selecting, cleaning, creating, integrating, and formatting the data to produce a final dataset. The Modeling phase includes selecting suitable modeling techniques, creating the test design, building and evaluating the models.

Data mining methods are divided into two main categories as supervised and unsupervised. The supervised expression is used when there is a well-defined or precise goal in data mining. If a special definition is not made for the desired result or there is uncertainty, the unsupervised expression is used. Unsupervised methods are mostly used for understanding, recognizing and discovering data. Supervised methods are used to extract information and conclusions from the data. Therefore, supervised data mining techniques, which are detailed below, were used in this study.

In this study, we apply twelve different modeling techniques to estimate the tax loss and evasion rate based on the inflation rate, unemployment, tax burden, economic growth (GDP), size of the government, and trade openness. All data mining analyses were performed using the WEKA 3.9 software (Waikato Environment for Knowledge Analysis).

Fig.1: The data mining process (Witten et al., 2005)



Gaussian processes

The Gaussian process model is a powerful tool used to tackle difficult machine learning problems. It is popular due to its computational simplicity and non-parametric, flexible nature. Essentially, a Gaussian process is a probabilistic model that allows for nonlinear regression by limiting the prior distribution to fit the available training data. Using a Gaussian process provides easy coding, hyperparameter control, and accurate predictions. In summary, the Gaussian process model offers various advantages and is widely used in various applications (Seeger, 2004).

Multi-Layer Perception (MLP)

The Multi-layer perceptron (MLP) is a frequently used artificial neural network architecture. MLPs are feed-forward networks that utilize the back-propagation algorithm to learn. This algorithm involves updating the weights of all layers based on the error level at the neural network's origin. MLP is considered a supervised network, as it learns through examples during training. The training algorithm repeatedly adjusts the synapse weights using the input/output data until convergence is achieved. This process is known as supervised learning, and it enables the network to generalize to unseen data (Ture et al., 2005).

Simple Linear Regression (SLR)

Simple Linear Regression is a mathematical modeling technique that helps to establish a relationship between a dependent variable and one or more independent variables. This method utilizes known or existing values to estimate the necessary parameters. By analyzing this relationship, we can make predictions about the dependent variable's values based on the independent variable's values. Therefore, Simple Linear Regression is a useful tool in many fields, such as economics, social sciences, and engineering (Zou et al., 2003).

Sequential Minimal Optimization Regression (SMOreg)

SMOreg is a software tool that applies the support vector machine (SVM) to regression problems. It uses various algorithms to learn the required parameters. SVM is a powerful approach for addressing both pattern recognition and regression problems, and has recently attracted attention from the neural network and mathematical programming communities. One of the significant benefits of the SMO algorithm is its speed and ease of implementation (Shevade et al., 2000).

KStar

The KStar algorithm is considered a "lazy" algorithm that leverages an entropy-based distance function to transform probabilities from one sample to another by applying all relevant transformations. To perform classification using KStar, new sample probabilities are added to all group members, and this process is repeated for each group to determine the one with the highest probability. This approach is effective in solving classification problems, and it has been studied extensively, as evidenced by recent research by Khosravi et al. (2021).

Additive Regression (AR)

Additive Regression is flexible and non-parametric regression algorithm. The estimation is performed by adding the estimations of each classifier. Additive Regression is advantageous in finding interactions between variables and complex data structure stored in high-dimensional data (Ture et al., 2005).

Random Committee

The random committee algorithm works by creating a group of base classifiers that use distinct random number seed values. During classification, the algorithm generates predictions based on the average probability estimates. To determine the final classification result, the algorithm computes the average probability estimates generated by each base classifier. This approach has been extensively studied and has demonstrated its effectiveness in solving classification problems, as shown by recent research conducted by Niranjan et al. (2018).

Decision Table (DT)

The Decision Table algorithm is an inductive algorithm that can create a classifier from a training set of labeled examples. To create a more accurate dataset model, the algorithm eliminates attributes that have little or no contribution to the model. This helps to reduce the risk of overfitting and generates a smaller and more compact decision table (Rajalakshmi et al., 2016).

M5Rules

The M5Rules algorithm is designed to extract rules from model trees and is commonly used for classification and prediction problems. It operates through a straightforward algorithm that involves training a pruned tree using a tree learner on a set of training samples. The algorithm then selects the elite leaf as the rule and discards the tree. This approach has proven to be effective in producing accurate and interpretable models (Gao et al., 2019.

M5P model tree

The M5P model tree algorithm is a hybrid approach that combines a traditional decision tree with the probabilistic nature of linear regression. This allows the algorithm to efficiently handle various tasks and predict the class value of samples that arrive at the leaf nodes. Compared to regression trees, model trees are much smaller in size, have clear decision-making capabilities, and use regression functions with fewer variables. As a result, model trees offer significant advantages over regression trees. Dang and Singh (2021) highlight the efficiency and compactness of model trees as their biggest advantage.

Reduced Error Pruning Tree (REPTree)

The REPTree algorithm is a speedy and efficient approach for constructing decision trees. It uses information variance to create decision or regression trees and reduces overfitting by employing reduced error pruning. Additionally, the algorithm sorts values for numeric attributes only once, thereby saving computational resources. To handle missing values, the algorithm segments the corresponding samples. Overall, the REPTree algorithm is a useful tool for building decision trees quickly and effectively (George-Nektarios, 2013).

Locally weighted learning (LWL)

Locally weighted learning (LWL) is a type of lazy learning algorithm known for its optimal convergence speed and high minimum performance, which outperforms all possible linear regressions. LWL is capable of handling diverse data distribution types and can avoid cluster and boundary effects. The algorithm relies on a distance function to identify the nearest neighbors of a given query instance, as explained by Khosravi (2021).

Estimating of Tax Evasion in Turkey for the 1999-2020 Period

There are many factors that cause tax loss and evasion. Some factors that affect the tax loss and evasion are the inflation rate, unemployment, tax burden, economic growth (GDP), government size and trade openness. These factors are the most commonly used variables in related literature studies. Therefore; in this study, tax loss and evasion rates were estimated in Turkey in 1999-2020 depending on these factors. One of the most popular determinants of tax evasion is the tax burden. In the literature, it is accepted that the tax burden has a direct positive effect on the tax loss and evasion rate. In most studies, it has been seen that the increase in tax evasion is significantly related to the increase in the total tax burden. Tax burden was determined as ratio of tax revenue to GDP. Economic growth (GDP) is also significantly associated with tax evasion. There are two opposing forces that determine the relationship between GDP and tax evasion rate. First, tax evasion rate may be positively related to GDP and secondly, it may be negatively related to GDP. Therefore, it must be resolved by empirical analysis in each country to determine whether the expected sign of this variable is positive or negative. The size of the government is another important factor that has an impact on tax evasion rate. The size of the government was determined as the ratio of government spending to GDP (Schneider and Savasan, 2007; Savasan, 2003; Dell'Anno 2007; Dell'Anno et al., 2004). Other popular variables affecting the tax evasion rate are the unemployment and inflation rate. In general, the increase in unemployment and inflation rate positively affects the tax evasion rate (Tabandeh and Tamadonnejad, 2015; Caballé and Panadés, 2004). The relationship between tax evasion and trade openness has also been examined in the literature. Studies have generally shown that there is a negative relationship between trade openness and tax evasion rate (Sameti et al., 2009). In Table 2, dependent and independent variables used in the study is seen. These are variables used in the existing literature.

Tax loss and evasion can be influenced by various factors, such as inflation rate, unemployment, tax burden, economic growth (GDP), size of government, and trade openness. These variables are commonly studied in the literature and are used to estimate tax loss and evasion rates in Turkey for the period of 1999-2020. Among these factors, tax burden is one of the most popular determinants of tax evasion. Previous research has shown that the increase in tax evasion is significantly related to the rise in the total tax burden, which is calculated as the ratio of tax revenue to GDP. Another important factor is economic growth, which has two opposing forces affecting the tax evasion rate. Empirical analysis is necessary to determine the expected sign of this variable in each country. The size of government, measured as the ratio of government spending to GDP, is also a significant factor that influences tax evasion rate (Schneider and Savasan, 2007; Savasan, 2003; Dell'Anno 2007; Dell'Anno et al., 2004). Different methods are used to measure the size of government. In the literature, the ratio of government spending to GDP is preferred in the measurement of the size of government, as it generally gives more accurate results (Kirmanoğlu, 2013; Sandalci and Sandalci, 2016). Additionally, unemployment and inflation rate are commonly associated with higher tax evasion rates (Tabandeh and Tamadonnejad, 2015; Caballé and Panadés, 2004). Finally, the relationship between tax evasion and trade openness has been examined in the literature, with studies generally showing a negative correlation between these two variables (Sameti et al., 2009). In Table 2, dependent and independent variables used in the study is seen. These are variables used in the existing literature.

Table 2: Dependent and independent variables

Independent variables	**Dependent variable**
Tax burden (%)	Tax evasion rate (%)
Size of government (%)	
GDP ($)	
Inflation rate(%)	
Trade openness (thousand $)	
Unemployment(%)	

Different modeling techniques (Gaussian processes, MLP, SLR, SMOreg, KStar, AR, Random Committee, DT, M5 Rules, M5P model tree, RepTree, LWL) were used for estimating tax evasion rate. In addition, comparison of these different modeling techniques was given out. The required dataset in data mining analysis were taken from official sources as Turkish Statistical Institute (2022) and T.R. Presidential Strategy and Budget Department (2022). The data set includes data for the period 1999-2020. Since there are no tax inspection results for 2010, the data set for the tax evasion rate included 21 data patterns. The data on GDP, inflation rate, trade openness and unemployment were sourced directly from the Turkish Statistical Institute. Tax burden, size of government and tax evasion rate data were created by us based on Presidential Strategy and Budget Department, Ministry of Treasury and Finance Activity Reports and Revenue Administration data. The data set has been divided into two groups for training (70 % of the data), for testing (30 % of the data). Three frequently used parameters for assessing models performance were applied: coefficient of determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE). These parameters can be calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{e,i}-t_{a,i}\right)^2}{\sum_{i=1}^{n}\left(t_{a,m}-\bar{t}_{a,m}\right)^2} \qquad (1)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{ei} - t_{a,i}\right| \qquad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{e,i}-t_{a,i}\right)^2}{n}} \qquad (3)$$

Specified in the equations, $y_{e,i}$ refers to the estimated value, $t_{a,i}$ to the actual value, $\bar{t}_{a,m}$ to the mean of the actual value, and n to the number of data.

Sensitivity Analysis

Sensitivity analysis is used to examine the effect of variations in one of its constituent independent variables on the dependent variable of a financial model. Sensitivity analysis is also used to reduce models to the most important variables and to eliminate or ignore less significant ones. In this study, sensitivity analysis was made using the partial rank correlation coefficient (PRCC). PRCC is the corresponding measure when input-output relationships are built using the ranks of the variables to linearize the relation. Detailed information about the analysis can be found on Mishra (2004).

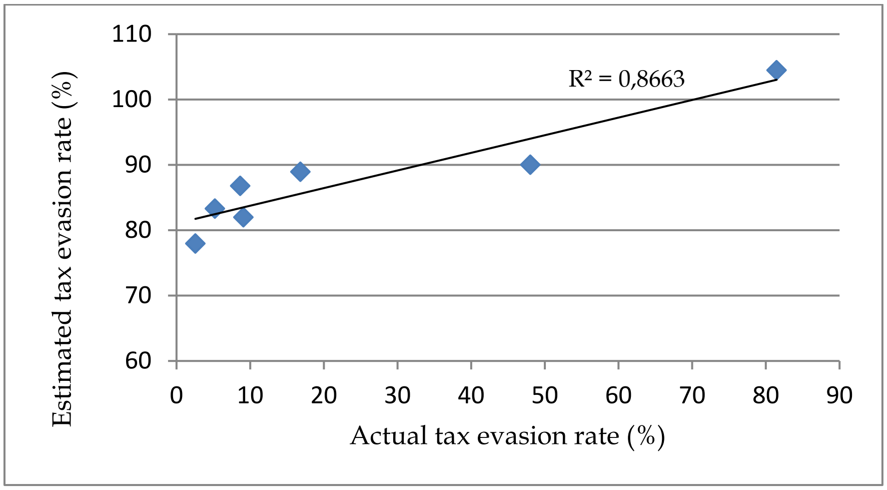
Results and Discussions

In this study, data mining method were used for estimating tax evasion depending on the inflation rate, unemployment, tax burden, economic growth (GDP), size of government and trade openness. Different data mining techniques were compared. Comparison parameters are R^2, RMSE and MAE values. Table 3 shows that the best technique in estimating tax evasion is the Gaussian processes. R^2, RMSE and MAE values were found as 0.931, 0.2473, and 0.2356, respectively. In addition, the regression curve of the dependent variable (tax evasion) for the test data set is given in Fig. 2. The correlation coefficient obtained is 0.8663, which is satisfactory.

Table 3: Comparison of different data mining techniques for tax evasion

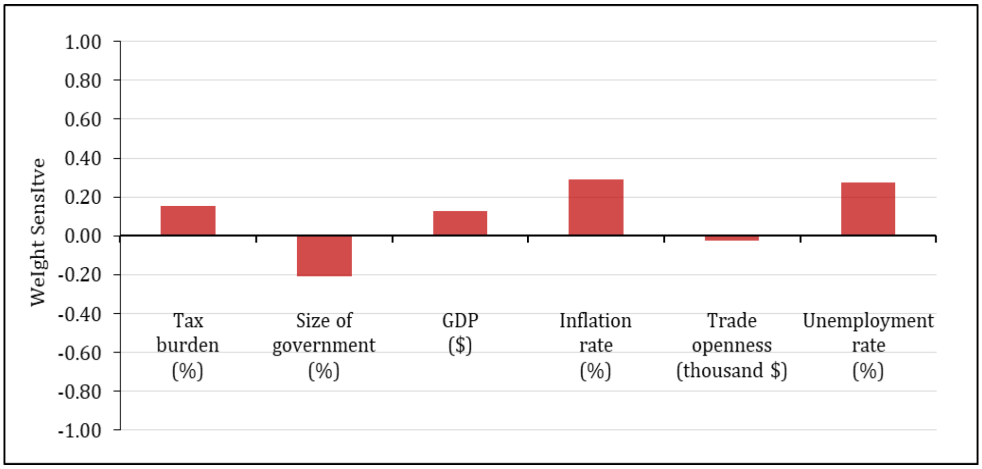
Method	**R^2**	**RMSE**	**MAE**
Gaussian processes	0.931	0.2473	0.2356
MLP	0.6759	0.606	0.5708
SLR	0.8597	0.2565	0.2472
SMOreg	0.8547	0.1567	0.1453
KStar	0.3478	0.4823	0.3737
AR	0.6393	0.3406	0.3341
Random Committee	0.3587	0.2123	0.1945
DT	0.1711	0.2642	0.2271
M5 Rules	0.8978	0.3162	0.312
M5P model tree	0.8978	0.3162	0.312
REPTree	0.3294	0.3168	0.2826
LWL	0.5427	0.2077	0.1741

Fig. 2: Comparison of actual and estimated values of tax evasion rate for the test data set



In this study, the relative contribution of factors causing tax evasion was also determined. Fig. 3 shows the importance of the each independent variable on tax evasion rate in modeling the data mining method.

Fig.3: The effect of independent variables on tax evasion



As can be seen Fig. 3, inflation rate and unemployment are the two most important factors that positively affect tax evasion. The results show that the increase in the inflation rate leads to higher tax evasion rate. When inflation rate is high, taxpayers prefer to save their money to protect their purchasing power and avoid paying taxes; thus increasing tax evasion rate. Similar results were found by Caballe and Panades (2004). There is a positive relationship between unemployment and tax evasion rate. High unemployment rates can lead to increased motivation for individuals to engage in shadow activities and seek employment in the informal economy. In such cases, individuals may choose to hide their income to avoid paying taxes, resulting in higher tax evasion rates. The relationship between unemployment and tax evasion is significant and has been studied in various literature. When people are unable to find formal employment, they may resort to working in the informal sector where they can avoid paying taxes. This can lead to significant revenue loss for the government and reduce the overall tax compliance of the population. Therefore, efforts to reduce unemployment rates and promote formal employment can play a critical role in reducing tax evasion rates. Similar results were found by Tabandeh and Tamadonnejad (2015). Tabandeh and Tamadonnejad used artificial neural network model for analysis.

The positive factors that contribute to tax evasion include GDP and tax burden. According to our research findings and Tabandeh et al.'s (2012) study using artificial neural networks model, there is a direct relationship between GDP and tax evasion rate. This implies that a surge in GDP leads to an increase in tax evasion. However, the effect of GDP on tax evasion is not uniform across countries and must be evaluated through empirical studies. The key determinants of this relationship are the penalty rate and the probability of detecting tax evaders. While a high penalty rate motivates taxpayers to declare their income and reduce tax evasion, it is insufficient to curb tax evasion without an effective tax audit system. In addition, when the tax burden is high, taxpayers look for ways to avoid paying taxes, leading to an increase in tax evasion rates, as suggested by Schneider and Savaşan (2007) and Dell'Anno (2007). Our study demonstrates that the variables of tax burden and GDP have lower significance compared to the unemployment and inflation rates, which are other factors that positively impact tax evasion.

The size of government, which is a measure of government spending, can theoretically have either a positive or negative impact on tax evasion. On one hand, an increase in government spending may result in increased administrative spending and pressure on the fiscal budget, which may lead to higher tax rates and tax evasion by taxpayers, as noted by Li and Ma (2015) and Sritharan and Salawati (2019). On the other hand, the size of government can refer to government capacity, and a government with strong capacity can enforce the rule of law and reduce tax evasion, as argued by Besley and Persson (2009). Our study found that the size of government, defined as the ratio of government expenditures to GDP, had a negative impact on the tax evasion rate. This finding is consistent with the results of D'Agostino et al. (2021), who examined the effect of government size on tax evasion at the provincial level in Italy between 2001 and 2015. Their analysis revealed heterogeneity between the provinces in the north and south of Italy, and that government spending had a negative impact on tax evasion.

There is a negative relationship between trade openness and tax evasion rate. The significance of this variable is relatively low among other independent variables. Similar results were found with the study of Tabandeh and Tamadonnejad (2015) using artificial neural networks method.

Conclusions

Tax evasion is very difficult to observe. Determining the tax evasion rate is also a very complex process. In this study to simplify this complex process, tax evasion rate were estimated using data mining techniques, depending on the basic causes such as inflation rate, unemployment, tax burden, economic growth (GDP), size of government and trade openness. The best technique for estimating tax evasion rate is Gaussian processes approach. The Gaussian process model is a probabilistic model that allows for nonlinear regression by limiting the prior distribution to fit the available training data. The non-parametric feature of the Gaussian processes model has been effective in finding this model as the best model. The estimated tax evasion rate with the Gaussian processes model was compared with the actual tax evasion rate. As a result of statistical analysis; R^2, RMSE and MAE values were found as 0.931, 0.2473, and 0.2356, respectively. These values were found to be at acceptable levels. The results of this study show that data mining method can use for estimating tax evasion rate. In addition, the relative contribution of factors causing tax evasion was also determined. While tax evasion rate is significantly and positively affected by variables such as unemployment, inflation and tax burden, it is negatively affected by size of the government.

When tax evasion rates in Turkey, which is one of the developing countries, are examined in the 1999-2020 period, it is seen that the highest tax evasion was during the economic crisis in 2001 and the global economic crisis in 2008. In times of economic crisis, inflation in the country was also high. Therefore, controlling inflation, which is one of the most important factors causing tax evasion, is one of the most important suggestions for reducing tax evasion. Another important factor in the tax evasion rate is unemployment. In order to reduce tax evasion, it can be suggested that policy makers take into account that high unemployment in the economy may lead people to find jobs in the informal economy and increase tax evasion. Therefore, policy makers should try to reduce unemployment. In addition, lowering the tax rates (tax burden) on individual and corporate incomes will greatly help reduce tax evasion rates.

In addition; factors such as inadequacy of tax inspection, tax awareness and morality, education and culture level, frequent tax amnesties, presence of reconciliation institutions, the abundance of exceptions and exemptions, frequent legislative changes and very little deterrent effect of penalties cause an increase in tax loss and evasion rates in our country. For this reason, governments need to take social and legal measures as well as economic and financial measures to reduce tax loss and evasion rates.

References

Albarea, A., Bernasconi, M., Marenzi, A., & Rizzi, D. (2020). "Income underreporting and tax evasion in Italy: Estimates and distributional effects". *Review of Income and Wealth*, *66*(4), 904-930.

Amoh, J. K., & Adafula, B. (2019). "An estimation of the underground economy and tax evasion: Empirical analysis from an emerging economy". *Journal of Money Laundering Control*.

Angour, N., & Nmili, M. (2019). "Estimating shadow economy and tax evasion: Evidence from Morocco". *International Journal of Economics and Finance*, *11*(5), 1-7.

Armağan, A. (2016). *Yargı Kararları Işığında Türkiye'de Vergi Kayıp ve Kaçakları ile Mücadele ve Alternatif Çözüm Arayışları*. Doktora Tezi, Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü, Isparta.

Athanasios, A., Eleni, K., & Charalampos, K. (2020). "Estimation of the size of tax evasion in Greece". *Bulletin of Applied Economics*, *7*(2), 97.

Ay, A. , Sugözü, İ. H. & Erdoğan, S. (2014). "Türkiye'de Vergi Yükünün, Enflasyonun ve Vergi Affı Beklentisinin Kayıt Dışı Ekonomiye Etkisi Üzerine Ampirik Bir Uygulama 1985-2012". *Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, (31.1) , 23-32.

Besley, T., & Persson, T. (2009). "The origins of state capacity: Property rights, taxation, and politics". *American Economic Review*, *99*(4), 1218-44.

Caballé, J., & Panadés, J. (2004). "Inflation, tax evasion, and the distribution of consumption". *Journal of Macroeconomics*, *26*(4), 567-595.

Çomakli, Ş. E. (2008). "AB İlerleme Raporlari Çerçevesinde Türkiye'deki Vergi Kayip ve Kaçaklarinin Önlenmesine Yönelik Uygulamalar". *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 22(1), 51-82.

D'Agostino, E., De Benedetto, M. A., & Sobbrio, G. (2021). "Tax evasion and government size: evidence from Italian provinces". *Economia Politica*, *38*(3), 1149-1187.

Dang, S. K., & Singh, K. (2021). "Predicting tensile-shear strength of nugget using M5P model tree and random forest: An analysis". *Computers in Industry*, *124*, 103345.

Dell'Anno, R., Gómez, M., & Pardo, Á. A. (2004). "Shadow Economy in Three Very Different Mediterranean Countries: France, Spain and Greece". A Mimic Approach. *CRISS, http://www. unisi. it/criss/download/meeting2004/papers/dellanno. pdf*.

Dell'Anno, R. (2007). "The shadow economy in Portugal: An analysis with the MIMIC approach". *Journal of Applied Economics*, *10*(2), 253-277.

Dell'Anno, R., & Davidescu, A. A. (2019). "Estimating shadow economy and tax evasion in Romania. A comparison by different estimation approaches". *Economic Analysis and Policy*, *63*, 130-149.

Demircan, E.S. (2004). "Türkiye'de Vergi Politikalarının Siyasi Analizi: Siyasi Değişimin Vergi Kayıp ve Kaçaklarına Etkisi Üzerine Bir İnceleme", *19. Türkiye Maliye Sempozyumu*, 10-14 Mayıs 2004, Belek/Antalya.

Faúndez-Ugalde, A., Mellado-Silva, R., & Aldunate-Lizana, E. (2020). "Use of artificial intelligence by tax administrations: An analysis regarding taxpayers' rights in Latin American countries". *Computer Law & Security Review*, *38*, 105441.

Gao, W., Alsarraf, J., Moayedi, H., Shahsavar, A., & Nguyen, H. (2019). "Comprehensive preference learning and feature validity for designing energy-efficient residential buildings using machine learning paradigms". *Applied Soft Computing*, *84*, 105748.

George-Nektarios, T. (2013). Weka classifiers summary. *Athens University of Economics and Bussiness Intracom-Telecom, Athens*.

Gerçek, A., Uygun, E. (2022). "Türkiye'de Vergi Kayıp ve Kaçakların Vergi Türlerine Göre Hesaplanması ve Değerlendirilmesi (2005-2020 Yılları)", *Vergi Raporu Dergisi*, 268, (163-177).

Hemberg, E., Rosen, J., Warner, G., Wijesinghe, S., & O'Reilly, U. M. (2016). "Detecting tax evasion: a co-evolutionary approach". *Artificial Intelligence and Law*, *24*(2), 149-182.

Khosravi, K., Khozani, Z. S., & Cooper, J. R. (2021). "Predicting stable gravel-bed river hydraulic geometry: A test of novel, advanced, hybrid data mining algorithms". *Environmental Modelling & Software*, *144*, 105165.

Kirmanoğlu, H. (2007). *Kamu Ekonomisi Analizi*, Beta Yayınevi, İstanbul.

Levin, J., & Widell, L. M. (2014). "Tax evasion in Kenya and Tanzania: Evidence from missing imports". *Economic Modelling*, *39*, 151-162.

Li, L., & Ma, G. (2015)." Government Size and Tax Evasion: Evidence from China". *Pacific Economic Review*, *20*(2), 346-364.

Madžarević-Šujster, S. (2002). "An estimate of tax evasion in Croatia". *Occasional paper series*, *6*(13), 1-23.

Mishra, S. 2004. Sensitivity analysis with correlated inputs—An environmental risk assessment example. In *Proceedings of the 2004 Crystal Ball User Conference*.

Niranjan, A., Nutan, D. H., Nitish, A., Shenoy, P. D., & Venugopal, K. R. (2018, April). ERCR TV: Ensemble of random committee and random tree for efficient anomaly classification using voting. In *2018 3rd international conference for convergence in technology (I2CT)* (pp. 1-5). IEEE.

Petanlar, S. K., Samimi, A. J., & Aminkhaki, A. (2011). "An Estimation of Tax Evasion in Iran". *Journal of Economics and Behavioral Studies*, *3*(1), 8-12.

Rahimikia, E., Mohammadi, S., Rahmani, T., & Ghazanfari, M. (2017). "Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran". *International Journal of Accounting Information Systems*, *25*, 1-17.

Rajalakshmi, A., Vinodhini, R., & Bibi, K. F. (2016). "Data Discretization Technique Using WEKA Tool". *International Journal of Science, Engineering and Computer Technology*, *6*(8), 293.

Raikov, A. (2021). "Decreasing tax evasion by artificial intelligence". *IFAC-PapersOnLine*, *54*(13), 172-177.

Sandalci, U., Sandalci, İ. (2016). "Kamu Kesimi Ekonomik Büyüklüğü ve Kamu Etkinlik Düzeyi İlişkisi". *Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, (Ek1), 413-429.

Sameti, M.A. &Sameti, M.O. & Dalaeemillan, A. (2009). "Underground Economy in Iran". International Economics Studies of Iran, 35 (2), 89-114.

Savaşan, F. (2003). "Modeling the underground economy in Turkey: randomized response and MIMIC models". *The Journal of Economics*, *29*(1), 49-76.

Schneider, F., & Savasan, F. (2007). "Dymimic estimates of the size of shadow economies of Turkey and of her neighbouring countries". *International Research Journal of Finance and Economics*, *9*(5), 126-143.

Seeger, M. (2004). "Gaussian processes for machine learning". *International Journal of Neural Systems*, *14*(02), 69-106.

Shakil, M. H., & Tasnia, M. (2022). "Artificial Intelligence and Tax Administration in Asia and the Pacific". In *Taxation in the Digital Economy* (pp. 45-55). Routledge.

Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., & Murthy, K. R. K. (2000). "Improvements to the SMO algorithm for SVM regression". *IEEE transactions on neural networks*, *11*(5), 1188-1193.

Sritharan, N., & Salawati, S. (2019). "Economic factors impact on individual taxpayers' tax compliance behaviour in Malaysia". *Int. J. Acad. Res. Account. Finance. Manag. Sci*, *9*, 172-182.

Tabandeh, R., Jusoh, M., Nor, N. G. M., & Zaidi, M. A. S. (2012). "Estimating factors affecting tax evasion in Malaysia: A neural network method analysis". *Prosiding Persidangan Kebangsaan Ekonomi Malaysia Ke VII, 1525*.

Tabandeh, R., & Tamadonnejad, A. (2015). "The application of artificial neural network method to investigate the effect of unemployment on tax evasion". *Journal of Research in Business, Economics and Management*, *4*(3), 393-402.

Ture, M., Kurt, I., Kurum, A. T., & Ozdamar, K. (2005). "Comparing classification techniques for predicting essential hypertension". *Expert Systems with Applications*, *29*(3), 583-588.

Turkish Statistical Institute (TÜİK). (2022). https://www.tuik.gov.tr/

T.R. Presidential Strategy and Budget Department (2022). https://www.sbb.gov.tr

Uyar, A., Bani-Mustafa, A., Nimer, K., Schneider, F., & Hasnaoui, A. (2021). "Does innovation capacity reduce tax evasion? Moderating effect of intellectual property rights". *Technological Forecasting and Social Change*, *173*, 121125.

Van Dunem, J. E., & Arndt, C. (2009). "Estimating border tax evasion in Mozambique". *The Journal of Development Studies*, *45*(6), 1010-1025.

Warner, G., Wijesinghe, S., Marques, U., Badar, O., Rosen, J., Hemberg, E., & O'Reilly, U. M. (2015). "Modeling tax evasion with genetic algorithms". *Economics of Governance*, *16*(2), 165-178.

WEKA 3.9 software, https://waikato.github.io/weka-wiki/downloading_weka/

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., (2005, June). "Practical machine learning tools and techniques". In *Data Mining* (Vol. 2, No. 4).

Xiangyu, X., Youlin, Y., & Qicheng, X. (2018, July). Intelligent Identification of Corporate Tax Evasion Based on LM Neural Network. In *2018 37th Chinese Control Conference (CCC)* (pp. 4507-4511). IEEE.

Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). "Correlation and simple linear regression". *Radiology*, *227*(3), 617-628.

Zumaya, M., Guerrero, R., Islas, E., Pineda, O., Gershenson, C., Iñiguez, G., & Pineda, C. (2021). "Identifying tax evasion in Mexico with tools from network science and machine learning". In *Corruption Networks* (pp. 89-113). Springer, Cham.

Hakem Değerlendirmesi: *Dış bağımsız.*
Çıkar Çatışması: *Yazar, çıkar çatışması olmadığını beyan etmiştir.*
Finansal Destek: *Yazar, bu çalışma için finansal destek almadığını beyan etmiştir.*

Peer-review: *Externally peer-reviewed.*
Conflict of Interest: *The author have no conflicts of interest to declare.*
Financial Disclosure: *The author declared that this study has received no financial support.*