International Journal of Multidisciplinary Studies and Innovative Technologies

e-ISSN: 2602-4888
dergipark.org.tr/en/pub/ijmsit
Research Article

2022, 6 (2), 162-168
DOI: 10.36287/ijmsit.6.2.162
Received: June 15, 2022; Accepted: Nov. 10, 2022

Predicting the Income Groups and Number of Immigrants by Using Machine Learning (ML)

Belgin AYDEMİR [1], Hakan AYDIN [2], Ali ÇETİNKAYA [3*] and Doğan Şafak POLAT [4]

[1] *Department of Computer Engineering/Faculty of Engineering and Architecture, Istanbul Gelisim University, Istanbul, Turkey*
[2] *Department of Computer Engineering/Faculty of Engineering, Istanbul Topkapı University, Istanbul, Turkey*
[3*] *Department of Electronics Technology/Istanbul Gelisim Vocational School, Istanbul Gelisim University, 34310, Istanbul, Turkey*
[4] *Ankara, Turkey*
[*] *Corresponding Author: alcetinkaya@gelisim.edu.tr*

Abstract – Migration is one of the biggest problems in the history of mankind. It is important to predict human migration as accurately as possible in terms of many aspects such as urban planning, trade, pandemics, the spread of diseases, and public policy development. With the help of Artificial Intelligence (AI), which is now used in almost all areas of life, it is possible to make predictions about migration. The purpose of this study is to predict the income groups and the number of immigrants by using ML algorithms. Two different applications were carried out in the study. The first one was about predicting the income groups of immigrants and the second one was about predicting the number of immigrants. Data used in the study was obtained from the World Bank. In the first application of the study, Support Vector Machines (SVM), Naive Bayes (NB), Logistic Regression (LR), K-Nearest Neighbors (KNN) were used. In the second application of the study, Random Forest (RF), and Xgboost algorithms were used. As a result of the experiments conducted in the study, 98.37% success rates were obtained with Xgboost, 96.42% with RF, 86.04% with LR, 83.72% with SVM, 83.72% with KNN, and 69.76% with NB. The results of the study reveal that the highest success in the applications was achieved with the LR and Xgboost algorithms. In general, the predictive machine learning models of human migration used in this study will provide a flexible base with which to model human migration under different what-if conditions.

Keywords *– Artificial Intelligence (AI), Machine Learning (ML), Migration, Data Science*

I. INTRODUCTION

Migration is a population movement that results in the displacement of individuals or groups of people, regardless of its duration, nature, or cause. Throughout history, people have migrated for many reasons, such as wars, climatic conditions, displacement, natural disasters, or political, policy, and economic reasons. It is a fact that individuals and communities migrate by leaving the regions where they live for various reasons. The phenomenon of migration includes the migration of refugees, internally displaced persons, economic migrants, and people who move for other reasons, including family reunification [1-2]. Human migration is a form of human mobility in which a person moves by intending to change his place of residence. Shortly, the migration phenomenon is a population movement that includes all human movements, regardless of their duration, composition, and causes.

One of the subfields of Artificial Intelligence (AI) is machine learning (ML). ML can be defined as an automated process for discovering previously unknown structures (patterns) from a data set [3]. An existing problem can be modeled using ML algorithms, and predictions for the future can be made using computer algorithms and a data set. In ML, the goal of the models created using algorithms and datasets is to achieve the highest performance. By using models created by ML algorithms, the exact migration numbers of potential countries can be estimated and analyzed [4]. In [5], it was described how ML can be used in modeling migration processes, and they used the decision tree algorithm in their study. The accuracy of their estimation was 67%.

In the information age we live in, we see that it is possible to use AI to make various predictions about migration, which happens in almost all areas of life. For example, for countries that want to take precautions against the migration problem, it can be useful to predict the number of immigrants, income groups, and other information by using AI techniques and methods. Predicting or knowing immigration-related information can help countries to prepare better for this challenge in many ways, including social, cultural, economic, and security. The prediction of human migration; related to population size and growth are important for various policy on strategy, planning, and industry [6]. One of the advantages of the ML models is that countries and global or regional organizations that know the relationship between the number of immigrants and immigrant income groups can estimate the influx of immigrants and be prepared better for it.

The purpose of this study, whose motivation is to apply AI techniques and methods to the migration issue, is to predict the income groups of immigrants and their numbers by classification with different ML algorithms. In this study, two different applications were conducted. The first application was about predicting immigrants' income groups. In this application, Support Vector Machines (SVM), Naive Bayes

(NB), Logistic Regression (LR), and K-Nearest Neighbors (KNN) algorithms were used. The second application of the study, on the other hand, was about predicting the number of immigrants. For this application, the Random Forest (RF) and Xgboost algorithms were used. In both applications, it aimed to examine the comparative performances of different ML algorithms in human migration prediction. In the study, two separate datasets, created for the first time in this study and including the World Bank 2010 data, were used. For the study, a total of 21 different experiments were conducted.

The main contributions of this study are as follows:

- The prediction of immigrants' income groups and their number by using AI can be shown as an important contribution to this study. Several experiments were conducted using SVM, NB, LR, KNN, Xgboost, and RF ML algorithms. In short, we wanted to show that AI techniques and methods can be applied to the migration issue, which is one of the biggest problems in human history.
- Another contribution of this study is that the used two separate data sets were created for the first time in this study.

In the remaining of this article, a literature review on the related studies is presented in the 2nd section. Migration and ML algorithms are described in the 3rd section. The experimental studies that were performed in this study are explained in the 4th section. Finally, in Section 6, the study is concluded and some ideas for future studies are explained.

II. Related Work

In the literature, by using AI and ML approaches, many studies have been conducted on different topics, such as big data, autonomous systems, handwritten character recognition, natural language processing, image processing, convolutional neural network, classification studies, and other topics. There are dozens of studies and researches on AI and ML in the literature, and the number of these types of studies is increasing very rapidly. One of the reasons for this is the increasing popularity of AI and its use in important issues that have entered all areas of life. However, for this study, mainly the studies focusing on migration and AI were reviewed. In [7], the impact of increasing digitization and AI on migration and mobility systems was critically examined in a post-COVID transnational context. The authors emphasized that the advancement of AI through the migration cycle goes beyond its original focus on the pre-departure and entry period. Different ML techniques and a set of pre-immigration variables for more than 6,281 immigrants from Mexico to estimate their legal status in the United States in [8]. As a result of the study, it was found that the legal status of 80% of Mexicans who immigrated to the United States could be accurately predicted. The central hypothesis of the study conducted by Beduschi in [9] was that AI technology could influence international migration management in three different dimensions: (1) by deepening existing asymmetries among states at the international level; (2) by modernizing traditional practices of states and international organizations; and (3) by reinforcing current calls for more evidence-based migration management and border management. This study examined each of these three hypotheses and discussed the key challenges in deploying AI solutions for international migration management. In [10], the authors addressed the following research questions: 1. How is AI used by immigrants, their employees, and stakeholders involved in labor integration? 2. How does AI improve immigrants' job skills? and 3. What are the pathways to the success of AI interventions for integration and inclusion? In [11] ML models that can combine any number of exogenous characteristics to predict the origin-arrival flows of human migration were proposed. Compared to traditional human mobility models based on a variety of evaluation criteria, their models outperformed in predicting migrations between the U.S. and international migrations. The model presented in [12] was based on the ML methods, which are necessary to construct a valid self-learning system for describing the social and economic behavior of a rational individual who decides to migrate in search of a job. In [13], a tree-based ML approach was proposed to analyze the role of air shocks on an individual's migration intention in six agriculture-dependent economies such as Burkina Faso, Côte d'Ivoire, Mali, Mauritania, Niger, and Senegal. The authors found that (i) although socioeconomic characteristics had a greater impact on migration intentions, weather characteristics improved prediction performance (ii) a country-specific model was required, and (iii) international moves were more influenced by the longer time scales of SPEIs, while general moves (which include internal moves) were influenced by the shorter time scales. The findings of Harrison presented in [14] suggested that Internally Displaced Persons (IDPs) are rational actors whose post-displacement migration can be predicted by some quantifiable factors. And according to these findings, ML methods can play a key role in developing innovative solutions for migration routes, targeting aid, and meeting the needs of these vulnerable populations.

The literature review shows that since it has become increasingly interesting today, AI is also being used in the migration issue, as in many other areas. The literature review also shows that applications of AI in the migration issue are on an upward trend.

III. Migration as a Global Phenomenon

Migration has been an important problem in the world for centuries [15]. Today, migration and displacement have reached unprecedented levels [16-17]. Throughout history and today, it has been observed that people and human communities have migrated for social, cultural, political, economic, and security reasons. Social events such as war and famine are among the best-known causes of migration. The economy is the most important reason for migration, whether permanent or seasonal [18]. Migration can be linked to labor market insecurity and unemployment, which are directly related to crime and lawlessness [19]. High incomes, labor demand, economic opportunities, and political freedom are the factors that attract migration [20]. It is believed that economic migration brings people from poorer developing areas to more prosperous areas where wages are higher and more jobs are available [21]. Economic (voluntary) migrants hope to improve their livelihoods and send money home [22]. Most refugees are located in developing countries and regions currently characterized by suboptimal economic productivity and widespread poverty. Migration is a dynamic phenomenon that is studied by various disciplines in terms of social changes other than displacement [23]. The complexity of migration processes can be described by the interaction of economic, demographic, and social factors that determine the value, skill level, and age structure of migrants from the country of origin to the host country [12]. In the migration phenomenon, it is

also common for people to move from rural areas to more competitive urban areas to find more opportunities [22].

The migration problem has become increasingly evident, especially after the collapse of the Union of Soviet Socialist Republics (USSR) in 1990. In addition, the gap between the Global North and the Global South, unemployment, lack of adequate resources, and growing instability in some regions have forced people to seek better living conditions. The prosperity and stability in the Global North have attracted large numbers of people. The problems posed by international migration can affect all nations, depending on the individual and the need. Preventing or reducing internal and external migration plays an important role in the healthy development of a country [24]. The high level of capital, employment, education, culture, and urbanization in developed regions attracts immigrants [25].

Migration movements not only change places where people live but also cause many social, cultural, economic, and political consequences [27]. The migration problem affects many countries and has various dimensions [28]. It is a reality that many migrants today face extreme conditions, such as a lack of access to health and social services beyond jobs, income, and emergency humanitarian assistance. Migration can occur in the form of internal migration and international migration. In international migration, two or more states are affected by the movement in some way [20]. Developing economies encourage immigration by creating points of attraction thanks to the modernization process [29]. The social or economic impact of immigrants in the countries they go to creates the political and security agenda [30]. These include situations such as migration, war, political tension, disease, birth rate, and disruption of social order. One of the most important reasons why migration continues to increase in one direction is the development of industry and trade in the migrated region [31]. Voluntary migration is assumed to be driven by economic motives [32]. Involuntary or forced migration, whether it is natural or human in origin, refers to a migration movement where there is an element of coercion, including threats to livelihoods [1]. For example, COVID-19 is defined as a virus that is transmitted from person to person and spreads between countries around the world and has become a pandemic, and its progression has affected all sectors and also the issue of migration. This pandemic has brought many problems, especially for immigrants and it continues to significantly disrupt international migration and mobility. Due to COVID-19, people were stranded and helpless for months without being able to find food [7].

It is estimated that there were approximately 272 million international migrants worldwide in 2019, and this number represents 3.5 percent of the world's population [33]. Approximately two-thirds of international migrants (about 176 million) lived in high-income countries in 2019 [33]. More than 70 million forcibly displaced people, including more than 25 million refugees, nearly 3 million asylum seekers, and 41 million internally displaced persons, account for nearly 1%of the world's population [34]. Table 1 shows the number of immigrants and the ratio of this number to the global population by years. From this data, it appears that the world population has increased over time, but the ratio of immigrants to the world population has remained nearly constant. This is an indication that the phenomenon of migration will continue to be on the agenda as an important problem for humanity, as it has been in the past.

Table 1. The ratio of the number of immigrants to the world's population

Year	Number of Immigrants	Ratio
1970	84.460.125	2.3%
1975	90.368.010	2.2%
1980	101.983.149	2.3%
1985	113.206.691	2.3%
1990	153.011.473	2.9%
1995	161.316.895	2.8%
2000	173.588.441	2.8%
2005	191.615.574	2.9%
2010	220.781.909	3.2%
2015	248.861.296	3.4%
2019	271.642.105	3.5%

Source: UN DESA, 2008, 2019a, 2019b.

IV. Machine Learning (ML)

ML is an important component in the field of data science. ML can be defined as a branch of computer science that incrementally increases the accuracy of data by using algorithms that mimic the way humans learn in the field of AI. ML Algorithms are algorithms that accurately predict outcomes without programming any software. The goal is basically to take the data used as input in many algorithms and draw conclusions from the input data, analyze it with new output data, and produce results. This method focuses on how the data is used and how the program will proceed according to the flowchart. The algorithms used in ML are included in the training to make classifications or predictions, and part of the data is allocated for training, and part for testing. Accuracy values are calculated by comparing the predictions with the test data. The ML algorithms used in this study are explained below.

A. Logistic Regression (LR)

Logistic regression (LR) is a statistical method used to analyze a data set containing one or more independent variables that determine an outcome. The logistic regression model is the most commonly used regression model for analyzing data [35]. In LR, the outcome is measured with a binary variable. That is, LR is the appropriate regression analysis when the dependent variable is binary. The purpose of using the LR analysis is the same as other model-building techniques used in statistics. In LR, it is aimed to find the optimal model to describe the relationship between the bidirectional characteristic and a set of related independent variables. Mathematically, LR estimates the multiple linear regression function defined in Equation 1:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta 0 + \beta 1x1 + \cdots + \beta pxp \qquad (1)$$

The analysis LR is about estimating the daily rate of an event. As with all regression analyses, LR is a predictive analysis. The discriminant function approach was first popularized by Cornfield [36] for coefficient estimation in LR. In [37] it was focused on linear logistic models for cross-trial schedules. In [38], it was worked on the use and development of the LR model. In LR analysis, which is used to classify and study the relationships between dependent and independent variables, the dependent variable consists of categorical data and takes discrete values.

B. Support Vector Machine (SVM)

SVM has been successfully used in solving many classification problems, from face recognition systems to speech analysis, and has taken place in the literature as one of the efficient and effective ML algorithms with high generalization performance. SVMs have supervised learning models with associated learning algorithms that analyze data used for classification and regression analyses. The most important advantage of SVM is that it converts the classification problem into a quadratic optimization problem and solves it. Thus, at the stage of learning the solution to the problem, the number of operations is reduced compared to other techniques/algorithms, and a faster solution is provided. Due to this characteristic of the technique, it offers a great advantage, especially for large data sets. Moreover, since it is based on optimization, it is more successful than other techniques in terms of classification performance, computational cost, and usefulness.

C. K-Nearest Neighbors (KNN)

The KNN algorithm is one of the classification techniques performed based on the distance measure of the features in the data set. It is usually calculated by the Euclidean distance given in Equation (2):

$$(\mathbf{x}i\ ,\ \mathbf{x}j\ ) = \sqrt{\sum_{k=1}^{p}(xik\ -\ yjk)^2} \tag{2}$$

To determine which class a new sample will belong to, the distance from this point to all points in the training data set is calculated, and the calculated distances are ordered from smallest to largest. Considering the parameter of the predetermined number of neighbors (K), the K points with the closest distance are selected. To determine the class of the new observation, weighted voting methods or majority voting are generally applied within the known training data points in the K-selected classes. According to the majority voting, while assigning the class of the new sample, the most repeating class among K neighbors is selected. In the weighted voting method, it is aimed that the neighbors who are closer to the new sample point have a higher say in the classification. The vote of the nearest neighbor $m = 1, \ldots, K$ is given in Equation (3), where it is inversely proportional to the distance to the new sample point $\mathbf{x}Z$:

$$(\mathbf{x}m) = \begin{cases} \infty, \mathrm{d}(\mathrm{xm}, \mathrm{xZ}\ ) \ =\ 0 \ \mathrm{is}e \\ \frac{1}{\mathrm{d}(\mathrm{xm},\mathrm{xZ})} \end{cases} \tag{3}$$

D. Support Vector Machine (SVM)

The concept of situational probability within the scope of Bayes' theorem is the basis for the NB algorithm. The Bayesian formula is presented in Equation 4.

$$P(A \mid B)\ =\ \frac{P(A \cap B)}{P(B)}\ \ if\ P(B)\ \neq\ 0 \tag{4}$$

The NB algorithm makes it possible to predict new and unlabelled observations with the same logic by using the feature information of labeled observations based on this conditional probability function. This algorithm is successfully used not only on numeric data but also on non-numeric textual data.

E. Random Forest (RF)

The RF algorithm is a regression model that uses more than one decision tree to create more effective models and make accurate predictions. In this model, k nodes are randomly selected from their partitions. The training set consists of a random weight set. A subset that is used to grow each tree is selected. The kth tree is a random vector. k is generated independently of the past random vectors, … k − 1 is generated using the training set and k to grow a tree, resulting in a classifier. In the dataset D= {($\boldsymbol{x}$k, yk):, k = 1, ...} where {k} is independently distributed in the same way, $\boldsymbol{x}$k = (xk1, ⋯, $x$kp) is the input vector with p and the output corresponding to yk.

F. XGBoost (eXtreme Gradient Boosting)

XGBoost was first appeared in the literature with the article in [39]. The most important feature of the algorithm is that it can achieve high predictive power and prevent redundancy. Its working logic is quite similar to Gradient Boosting. The first step in XGBoost is to get the base score. This estimate can be any number (this number is 0.5 by default); when combined with the actions to be taken in the next steps, the correct result will be obtained. How good this prediction is decided by comparing it with the model's erroneous predictions. Errors are found by subtracting the estimated value from the observed value. XGBoost shows higher prediction success using different techniques and is optimized to work on large datasets.

V. EXPERIMENTS

A. Dataset

The dataset of this study consists of data obtained from the World Bank Economic Development Indicators [40]. In this dataset, there is a numerical representation of development indicators of countries by year. Based on this dataset, two separate datasets were created and used in this study: 1st dataset for income groups and 2nd dataset for immigration numbers. The first dataset named "Country.csv" was created using the "Country Migration Number" data and based on country regions and income groups by taking into account the countries' migration percentages, industry, population, female mortality rates, male mortality rates, under-14 age groups, 15-65 age groups, agricultural areas, and service indicators. This dataset was used to calculate the success rate in the experiments associated with our first application, which used the LR, SVM, KNN, and NB ML algorithms to estimate immigrant income groups. Some examples from the dataset are presented in Table 2.

Table 2. Income group dataset

Country	International migrant stock (% of population)	Industry, value added (% of GDP)	Mortality rate, adult female (per 1,000 female adults)	Mortality rate, adult male (per 1,000 male adults)	Population, ages 0-14 (% of total)	Population, ages 15-64 (% of total)	Agriculture, value added (% of GDP)	Population growth (annual %)	Number of under-five deaths	Services, etc., value added (% of GDP)	Region	Income Group
Afghanistan	0.325021	21.862341	253.259000	296.684000	47.589344	50.135406	27.091540	2.736886	114576.00	0.325030	South Asia	Low income
Albania	3.058886	28.690547	51.258800	97.999200	21.425123	67.620059	20.658189	-0.496462	540.0000	3.058895	Europe & Central Asia	Upper middle income
Algeria	0.672447	53.853213	93.879000	142.464000	27.191308	67.257704	9.029438	1.776047	24683.00	0.325031	Middle East & North Africa	Upper middle income
American Samoa	50.992163	28.385815	154.472289	217.866116	28.852485	63.430890	12.540303	-1.054881	36525.79	3.058896	East Asia & Pacific	Upper middle income
Andorra	66.154539	13.760281	154.472289	217.866116	28.852485	63.430890	0.463861	-1.241974	2.000000	0.325032	Europe & Central Asia	High income: non OECD

The second dataset named "Indictors.csv" was created using data obtained from "International Immigrant Quantity". Data on the number of immigrants between 1960 and 2010 was used in the creation of this dataset. This dataset was used to calculate the success rate in experiments related to our second application, which used the Xgboost and RF ML algorithms to estimate migration quantity. Regarding the dataset used in this study, the 1960-2010 data for Afghanistan are shown in Table 3 as an example.

Table 3. The ratio of the number of immigrants to the world's population

Country Name	Year	International migrant stock, total
Arab World	1960	3324685.0
Caribbean small states	1960	187880.0
Central Europe and the Baltics	1960	3294191.0
East Asia & Pacific (all income levels)	1960	8810259.0
East Asia & Pacific (developing only)	1960	3654693
Virgin Islands (U.S.)	2010	61798.0
West Bank and Gaza	2010	1923808.0
Yemen, Rep.	2010	517926.0
Zambia	2010	23314.0

In the study, data pre-processing steps were also applied to the datasets. Data pre-processing involves preparing the data for analysis by using methods of data cleaning, data merging, data transformation, and data reduction [36]. In these processes, missing data are added and deleted line by line. In the pre-processing of the data in our study, the columns "region" and "income group" with the value "nan" were deleted row by row in the datasets. One of the most successful methods to complete missing values in the dataset is to average missing and non-missing values [37]. In this context, the missing values in our dataset were completed by taking the average of the non-missing values. As an example of this situation, column-based row numbers of the missing data are shown in Figure 1.

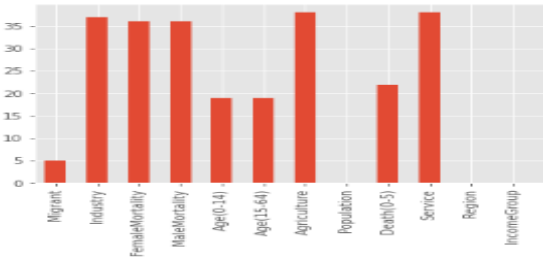

Fig. 1. Data preprocessing and missing data

B. Programming Language

In the experiments of this study, the Python programming language and various libraries were used. The library "Pandas" is one of them. This library takes data and makes it available for analysis. Another library used in this study is the "Numpy" library. This library is also a mathematical library that can be used to perform scientific calculations quickly and reliably. "Numpy" solves the problem of slowness by providing functions and operators that work efficiently with multidimensional arrays and fields. This library was used in our applications for list building and mathematical operations. The "Matplotlib" library is a cross-platform library for its numerical extension "Numpy"; that is, it is used to add visualizations such as data visualization and charts to the project. The "Sklearn" library, the "Pyecharts" package, and the "Bubbly package" were used in the study. The "StandardScaler" feature of the "Sklearn" library was used to scale our dataset. "Scikit-Learn (Sklearn)" is a widely used open source ML library.

Python was used as the programming language in the study. The reason for choosing this programming language for the study is that it is easy to learn, has a simple and flexible structure, and gives a command to write code in the program with ML algorithms. Some of the IDEs (Integrated Development Environment) used to develop Python are Eclipse, Pydev, Eric, and Anaconda.

C. Results

The first application of the study was about predicting income groups of immigrants. In this application, SVM, NB, LR, and KNN algorithms were used. In the study, four different ML algorithms selected were used to investigate the accuracy and effectiveness of different ML algorithms. The aim here is to test the success rates of different ML models within the scope of evaluation criteria. In the experiments, "Migration", "Industry", "Female Mortality", "Male Mortality", "0-14 Age Range", "15-64 Age Range", "Agriculture", "Population", "Deaths Under 5", "Service" and "Territory" attributes in the dataset were used. In each experiment, the number of attributes was changed and the success rates were recalculated. The results of the experiments are given in Table 4.

Table 4. The success rates of the experiments

Experiment Number	LR	KNN	SVM	NB
Experiment 1	86.05	83.72	83.72	69.77
Experiment 2	86.05	83.72	86.04	67.44
Experiment 3	83.72	83.72	86.04	67.44
Experiment 4	81.4	83.72	86.04	69.77
Experiment 5	83.72	83.72	86.04	67.44
Experiment 6	81.4	83.72	83.72	65.11
Experiment 7	76.74	83.72	76.74	67.44
Experiment 8	65.11	83.72	69.76	55.81
Experiment 9	67.44	83.72	67.44	55.81
Experiment 10	46.51	83.72	48.83	45.51
Experiment 11	44.18	83.72	39.53	34.88

In the experiments, a success rate of 86.04% was obtained for LR, 83.72% for SVM, 83.72% for KNN, and 69.76% for NB. It can be seen that the highest success in this application was achieved with the LR algorithm. When examining the success rates of the experiments conducted within the scope of this application, it was found that the success rates ranged from 34.88% to 86.06%. Figure 2 shows the success rates of the experiments LR, SVM, KNN, and NB. For the first implementation in this article, LR accuracy is higher than the other three algorithms.

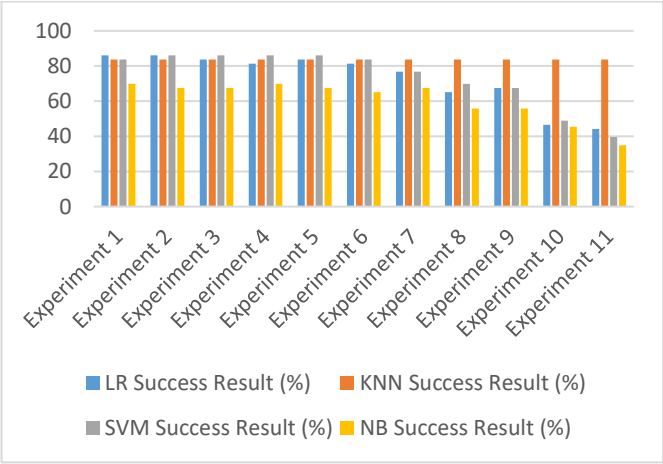

Fig. 2. The success rates of LR, SVM, KNN, and NB experiments

For the second application conducted to estimate the number of immigrants, the RF and Xgboost algorithms were used and 10 different experiments were carried out. In each experiment, the number of attributes was changed and the success rates were recalculated. Figure 3 shows the results of the experiments performed with the algorithm RF. Figure 4 shows the results of the experiments performed with the XGBoost ML algorithm.

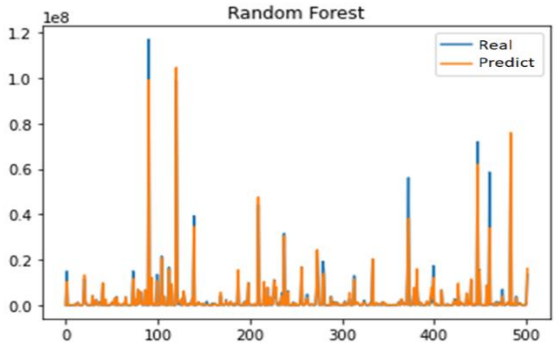

Fig. 3. shows the results of the experiments performed with the Xgboost ML algorithm

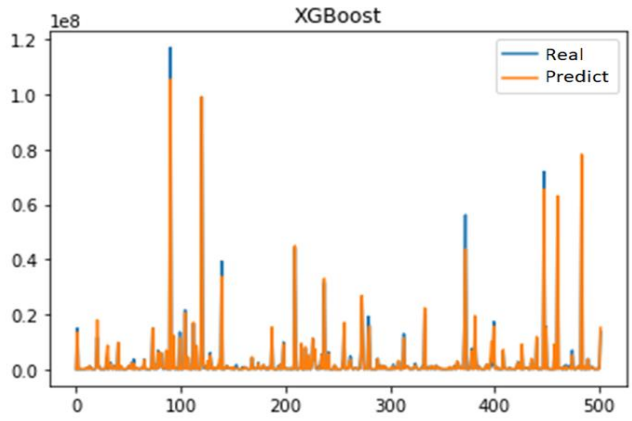

Fig. 4. Experimental results of Xgboost

In the experiments conducted in the second application of the study, success rates of 98.37% and 96.42% were achieved with XGBoost and RF, respectively. It can be seen here that the XGBoost model can provide a feature priority ranking. Essentially, the main factor behind this high success prediction rate is the power of the XGBoost when solving sequence prediction problems.

As seen from the results of this study, it is possible to use ML models to predict human migration. Regardless of the performances of the algorithms, these results also show the potential of ML algorithms in human migration prediction. ML models offer a high level of modelling flexibility. In addition, ML models can be customized to the human migration problems at hand. The use of ML algorithms in human migration prediction will make a significant contribution to countries. Thus, it will be easier to plan the national needs of the countries and more consistent social, economic, and environmental decisions will be made. When producing migration projections of countries, it may be necessary to estimate 50 years ahead, for example. In such a scenario, all variables beyond climatic events must be estimated with high success rates for each intervening year. After these input data such as temperature and drinking water status are estimated with ML for each year, the data obtained can be used in the human migration estimation method.

VI. CONCLUSION

In this work, the prediction of the income groups and the number of immigrants by using different ML algorithms are presented. The study is based on the implementation and evaluation of different ML algorithms. SVM, NB, LR, and KNN algorithms were used in the first application of the study, and RF and XGBoost algorithms were used in the second application. In this context, 21 different experiments were conducted. In the study, two separate datasets were created and used for the first time by using data obtained from the World Bank 2010. Experiments showed that the highest success rate was 86.04% with LR in the first application and 98.37% with XGBoost in the second application. It can be seen from the experiments done in the study that the highest success in the applications was achieved with the LR and XGBoost algorithms. The use of ML algorithms can make significant contributions to predicting the future of human migration. It is expected that this study will contribute to the literature by

showing that AI techniques and methods can be used in the field of migration, which is one of the most important issues in international relations. For future studies, we would like to extend this study using AI by investigating the relationship between climate change and migration

REFERENCES

[1] Richard, P., & Jillyanne, R. C. (2011). Glossary on migration. Book Glossary on Migration, 2.

[2] Zimmermann, K. F. (2014). Circular migration. IZA World of Labor.

[3] Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2020). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press.

[4] Micevska, M. (2021). Revisiting forced migration: A machine learning perspective. European Journal of Political Economy, 70, 102044.

[5] Iman, H. S., & Tarasyev, A. (2018). Machine learning methods in individual migration behavior. In Russian Regions in the Focus of Changes: Conference proceedings. Ekaterinburg, 2018 (pp. 72-81). LLC Publishing office EMC UPI.

[6] Hussain, N. H. M. (2021). Machine Learning of the Reverse Migration Models for Population Prediction: A Review. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(5), 1830-1838.

[7] McAuliffe, M., Blower, J., & Beduschi, A. (2021). Digitalization and Artificial Intelligence in Migration and Mobility: Transnational Implications of the COVID-19 Pandemic. Societies, 11(4), 135.

[8] Azizi, S., & Yektansani, K. (2020). Artificial intelligence and predicting illegal immigration to the USA. International Migration, 58(5), 183-193.

[9] Beduschi, A. (2021). International migration management in the age of artificial intelligence. Migration Studies, 9(3), 576-596.

[10] Lindström, N., Koutsikouri, D., Stier, J., & Arvidsson, M. (2020, June). Migrant Employment Integration and Artificial Intelligence (AI). The 32nd annual workshop of the Swedish Artificial Intelligence Society (SAIS) will be held as an online conference in June (pp. 16-17).

[11] Robinson, C., & Dilkina, B. (2018, June). A machine learning approach to modeling human migration. In Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (pp. 1-8).

[12] Tarasyev, A. A., Agarkov, G. A., & Hosseini, S. I. (2018, July). Machine learning in labor migration prediction. In AIP Conference Proceedings (Vol. 1978, No. 1, p. 440004). AIP Publishing LLC.

[13] Aoga, J., Bae, J., Veljanoska, S., Nijssen, S., & Schaus, P. (2020). Impact of weather factors on migration intention using machine learning algorithms. arXiv preprint arXiv:2012.02794.

[14] Harrison, E. (2020). Modeling Movement: A machine-learning approach to track migration routes after displacement.

[15] Günay, E., Atılgan, D., & Serin, E. (2017). Migration management in the world and Turkey. Kahramanmaraş Sütçü İmam University Journal of the Faculty of Economics and Administrative Sciences, 7(2), 37-60.

[16] Hayakawa, T. (2020). Skill levels and inequality in migration: A case study of Filipino migrants in the UK. Asian and Pacific Migration Journal, 29(3), 333-357.

[17] Errichiello, G., & Nyhagen, L. (2021). "Dubai is a transit lounge": Migration, temporariness and belonging among Pakistani middle-class migrants. Asian and Pacific Migration Journal, 30(2), 119-142.

[18] Sinha, V. N. P., Ataullah, M. D., & Ataullah, M. (1987). Migration: an interdisciplinary approach. Seema Publications.

[19] Rubin, R. L., & Melnick, J. (2007). Immigration and American popular culture: An introduction (Vol. 4). NYU Press.

[20] Castles, S., & Miller, M. J. (1998). The Age of Migration (Mcmillan, London).

[21] Massey, D. S. (1989). Economic development and international migration in comparative perspective (No. 1). Commission for the Study of International Migration and Cooperative Economic Development.

[22] Taylor, J. E., & Fletcher, P. L. (2001). Remittances and development in Mexico: the new labour economics of migration: a critical review. Rural Mexico Research Project, 2.

[23] Memisoglu, F., & Yiğit, C. International Migration and Development: Theory and Current Issues. Yildiz Social Science Review, 5(1), 39-62.

[24] Yusuf, G. E. N. Ç., Gündüz, D. U., & Çöpoğlu, M. The Relationship of Migration and Development. Avrasya Uluslararası Araştırmalar Dergisi, 7(18), 479-498.

[25] Özdemir, D. (2018). Determinants of interregional internal migration movements in Turkey. Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 22(3), 1337-1349.

[26] Zlotnik, H. (1995). Migration and the family: The female perspective. Asian and Pacific Migration Journal, 4(2-3), 253-271.

[27] Çatalbaş, G. K., & Yarar, Ö. (2015). Determining the factors affecting interregional internal migration in Turkey with panel data analysis. Alphanumeric Journal, 3(1), 99-117.

[28] Schutte, S., Vestby, J., Carling, J., & Buhaug, H. (2021). Climatic conditions are weak predictors of asylum migration. Nature communications, 12(1), 1-10.

[29] Ravenstein, E. G. (1889). The laws of migration. Journal of the royal statistical society, 52(2), 241-305.

[30] Sirkeci, İ., Deniz, U. T. K. U., & YÜCEŞAHİN, M. M. (2019). An evaluation of the migration conflict model through participation, development and mass gaps. Journal of Economy Culture and Society, (59), 157-184.

[31] Nalbant, T. E. (2020). International Migration and Security. Avrasya İncelemeleri Dergisi , 9 (2) , 309-313 . DOI: 10.26650/jes.2020.020

[32] Betts, A. (2009). Forced migration and global politics. John Wiley & Sons.

[33] United Nations Department for Economic and Social Affairs. (2020). World economic situation and prospects 2020. UN. Available at: https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/WESP2020_FullReport.pdf

[34] Albu, D. (2019). UNHCR Global Trends Report: Forced displacement in 2018. Drepturile Omului, 114.

[35] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

[36] Cornfield, J. (1962, July). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. In Fed Proc (Vol. 21, No. 4, pp. 58-61).

[37] Le, C. T. (1984). Logistic models for cross-over designs. Biometrika, 71(1), 216-217.

[38] Bonney, G. E. (1987). Logistic regression for dependent binary observations. Biometrics, 951-973.

[39] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[40] World Development Indicators, 2018, Available at: https://www.kaggle.com/theworldbank/world-development-indicators. Date of access: 26.11.2021