Detecting Discrimination Using Selected Data Mining Methods

Otkrivanje diskriminacije pomoću odabranih postupaka dubinske analize podataka

Petra Kuljiš, Lucia Načinović Prskalo, Marija Brkić Bakarić
University of Rijeka
Faculty of Informatics and Digital Technologies
Radmile Matejčić 2, 51000 Rijeka, Croatia
petrakuljis@gmail.com, {lnacinovic, mbrkic}@inf.uniri.com

Abstract. *The study uses advanced data analysis techniques to better understand patterns and manifestations of gender discrimination and to provide a basis for the development of effective policies aimed at reducing it. A publicly available dataset is analysed using selected data mining methods. Descriptive analysis provided deep insights into the basic characteristics of the data, and clustering was used to identify countries with similar patterns of gender inequality. Regression analysis was used to identify the factors contributing to the increase or decrease in gender inequality. The Isolation Forest model identified countries that represent extreme cases, while the comparative analysis revealed Croatia's advantages compared to the global average as well as areas for improvement.*

Keywords. data mining, gender discrimination, gender inequality, descriptive analysis, regression analysis, clustering, anomaly detection

Sažetak. *U radu su primijenjene napredne tehnike analize podataka s ciljem boljeg razumijevanja uzoraka i manifestacija rodne diskriminacije te stvaranja temelja za oblikovanje učinkovitih politika usmjerenih k smanjenju iste. Primjenom odabranih tehnika dubinske analize podataka analizira se javno dostupan podatkovni skup. Deskriptivnom analizom dobiveni su duboki uvidi u osnovne karakteristike podataka, a grupiranjem su identificirane zemlje sa sličnim obrascima rodne nejednakosti. Regresijska analiza korištena je za utvrđivanje čimbenika koji doprinose povećanju ili smanjenju rodne nejednakosti. Modelom izolacijskih šuma prepoznate su zemlje koje predstavljaju ekstreme, dok su komparativnom analizom utvrđene prednosti Hrvatske u odnosu na svjetski prosjek, ali i identificirana područja za poboljšanje.*

Ključne riječi. dubinska analiza podataka, rodna diskriminacija, rodna nejednakost, deskriptivna analiza, regresijska analiza, grupiranje, otkrivanje anomalija

1 Introduction

The primary objective of this study is to show how advanced data analysis techniques can be used to detect patterns of discrimination and to conduct a comprehensive analysis of gender inequality at the global level and compare these findings with the Croatian context. The analysis focuses on specific domains in which gender inequality is most pronounced, as well as on the key factors influencing it. By identifying countries that exhibit similar patterns of gender inequality, alongside outliers that deviate substantially from prevailing trends, the study positions Croatia within a broader global

1 Uvod

Primarni cilj ovog istraživanja je prikazati kako se napredne tehnike analize podataka mogu koristiti za otkrivanje obrazaca diskriminacije te provesti sveobuhvatnu analizu rodne nejednakosti na globalnoj razini kako bi se mogla izvršiti usporedba s Hrvatskom. Analiza se usredotočuje na specifična područja u kojima je rodna nejednakost najizraženija, kao i na ključne čimbenike koji na nju utječu. Identificiranjem zemalja koje pokazuju

framework. This approach enables researchers and policy makers to gain a better understanding of the causes and manifestations of discrimination and to develop informed strategies to overcome it. However, as Kabeer (2016) notes, even when policy recommendations are present, there is often shortage of concrete outcomes.

Gender equality "refers to the equal rights, responsibilities and opportunities of women and men and girls and boys" (UN Women, 2025). This includes equal opportunities for education, employment, participation in the political process and access to healthcare (World Economic Forum, 2025).

Gender discrimination is a particular form of discrimination based on sex or sexual orientation. The term "discrimination" comes from the Latin term *dis-criminare*, which means "to distinguish between". In a social context, discrimination refers to the behaviour based on prejudice in which people are treated unfairly because they belong to a particular group, regardless of their individual characteristics. Such practices seriously hinder the realization of equality, social development and peace. Gender discrimination manifests itself in the systematic preferential treatment of one gender at the expense of another. Despite progress, it is still a serious social problem with global consequences (United Nations, 2025, World Economic Forum, 2025, Grant Thornton International, 2025).

Gender discrimination can manifest itself in different domains. In the area of education, gender discrimination can lead to unequal opportunities for girls and boys, often resulting in lower literacy rates and educational attainment for women. In the area of healthcare, women often have limited access to certain medical services and poorer health outcomes. In developing countries, women often face a higher risk of death due to complications during pregnancy and childbirth (United Nations, 2025). Even though women's workforce participation globally has risen, gender-based industry segregation persists, with women still concentrated in lower-paying, people-centric industries (United Nations, 2025). Unequal pay for equal work among male and female workers is one of the possible causes of the unadjusted gender pay gap (Leythienne & Pérez-Julián, 2022, Grissom et al., 2021, Poli & Maier, 2022). This inequality is particularly pronounced in certain sectors such as the financial industry and managerial roles (Grant Thornton International, 2025). Political discrimination is another major challenge as women remain significantly underrepresented in the political sphere (United Nations, 2025). This inequality is exacerbated by social norms and legal systems that often do not offer equal protection and opportunities to all genders.

The knowledge discovery in databases (KDD) process encompasses steps ranging from the data collection and preparation, application of analytical

slične obrasce rodne nejednakosti, uz izdvajanje ekstremnih slučajeva koji značajno odstupaju od prevladavajućih trendova, Hrvatska se smješta u širi globalni okvir. Ovakav pristup omogućuje istraživačima i donositeljima odluka bolje razumijevanje uzroka i manifestacija diskriminacije te razvoj informiranih strategija za njeno prevladavanje. Međutim, čak i kada su donesene odgovarajuće politike, često izostaju konkretni rezultati (Kabeer, 2016).

Rodna jednakost odnosi se na ista prava, odgovornosti i mogućnosti žena i muškaraca te djevojčica i dječaka (UN Women, 2025). To uključuje jednake mogućnosti za obrazovanje, zapošljavanje, sudjelovanje u političkom procesu i zdravstvenu skrb (World Economic Forum, 2025).

Rodna diskriminacija predstavlja specifičan oblik diskriminacije koji se temelji na rodu ili rodnoj orijentaciji. Pojam „diskriminacija" potječe od latinskog izraza *dis-criminare*, što znači „razlikovati između". U društvenom kontekstu, diskriminacija podrazumijeva postupanje utemeljeno na predrasudama, pri čemu se osobe nepravedno tretiraju zbog pripadnosti određenoj skupini, neovisno o njihovim individualnim kvalitetama. Takva praksa ozbiljno ugrožava postizanje jednakosti, društvenog razvoja i mira. Rodna diskriminacija se očituje u sustavnom favoriziranju jednog roda nauštrb drugoga. Unatoč napretku, rodna diskriminacija i dalje predstavlja ozbiljan društveni problem s globalnim posljedicama (United Nations, 2025, World Economic Forum, 2025, Grant Thornton International, 2025).

Rodna diskriminacija se može manifestirati u raznim sferama života. U području obrazovanja, rodna diskriminacija može rezultirati nejednakim mogućnostima za djevojčice i dječake, često uzrokujući niže stope pismenosti i obrazovnih uspjeha među ženama. U području zdravstva, žene se često suočavaju s ograničenim pristupom određenim zdravstvenim uslugama te lošijim zdravstvenim ishodima. U zemljama u razvoju, žene su nerijetko izložene većem riziku od smrtnosti zbog komplikacija tijekom trudnoće i poroda (United Nations, 2025). Iako je sudjelovanje žena u radnoj snazi globalno poraslo, rodna segregacija po industrijama i dalje postoji, pri čemu su žene još uvijek koncentrirane u slabije plaćenim, na ljude usmjerenim industrijama (United Nations, 2025). Nejednaka plaća za jednak rad među muškim i ženskim radnicima jedan je od mogućih uzroka neprilagođenog rodnog jaza u plaćama (Leythienne & Pérez-Julián, 2022, Grissom et al., 2021, Poli & Maier, 2022). To je posebno izraženo u određenim sektorima, primjerice u financijskoj industriji te rukovodećim pozicijama (Grant Thornton International, 2025). Politička diskriminacija predstavlja još jedan značajan izazov budući da su žene još uvijek podzastupljene na političkoj sceni (United Nations, 2025). Ova nejednakost dodatno se

techniques, to the interpretation and presentation of results (Tan et al., 2019). The process that was applied in this study was structured using CRISP-DM phases (Cross-Industry Standard Process for Data Mining) (Shearer, 2000). CRISP-DM is designed as a universal framework, independent of specific tools, industries or the problems to be analysed. It defines the key steps in the knowledge discovery process, such as understanding the problem, understanding the data, data preparation, modelling, model evaluation and deployment.

In order to achieve a comprehensive understanding of gender inequality at the global level and compare these findings with the Croatian context, selected data mining techniques are applied. Analysis is conducted using the Kaggle platform's World Data of Gender Inequality Index" dataset.

Descriptive analysis examines existing data to uncover patterns, relationships, and specific domains in which gender inequality is most pronounced. Linear regression estimated by OLS (ordinary least squares) is used to understand and quantify the relationship between variables in order to discover key discrimination factors. While clustering enables the identification of countries that exhibit similar patterns of gender inequality, anomaly detection enables the identification of extreme cases. Finally, comparative analysis positioned Croatia within a broader global framework.

The next section briefly presents related work, followed by an overview of the methodology and the dataset in Section 3. Section 4 reports the results of the selected data mining techniques, while the final section provides a discussion, the main conclusions, and directions for future research.

2 Related Work

Even though gender discrimination can manifest itself in various domains, the literature review reveals that existing studies mostly focus on analysing labour market discrimination. While some researchers focus on uncovering discriminatory behaviour of employers (Hangartner et al., 2021), others such as Edelsztein & Waisbrot (2023), Bonaccolto-Töpfer et al. (2023) and (Alatrista-Salas et al. (2017) focus on wage gaps, on geographical distribution of land ownership and wealth (Fisher & Naidoo, 2016), and on health benefit indicators (Heinz et al., 2020).

Even though gender bias and favoritism can also be detected using natural language processing techniques and sentiment analysis, as demonstrated by Torres-Berrú et al. (2023), the dataset used in this study contains only structured attributes. A detailed overview of text analysis methods used to identify and detect gender bias, inequality, and violence in language can be found in (Nacinovic Prskalo, 2024).

produbljuje uslijed društvenih normi i pravnih sustava koji često ne osiguravaju ravnopravnu zaštitu i mogućnosti.

Proces otkrivanja znanja u bazama podataka (KDD) uključuje korake od prikupljanja i pripreme podataka, preko primjene analitičkih tehnika, pa sve do interpretacije i prezentacije rezultata (Tan et al., 2019). Proces koji je primijenjen u ovom istraživanju je strukturiran po uzoru na CRISP-DM (engl. *Cross-Industry Standard Process for Data Mining*) (Shearer, 2000). CRISP-DM je osmišljen kao univerzalni okvir, neovisan o konkretnim alatima, industrijskim sektorima ili specifičnim problemima koji se analiziraju. On definira ključne korake u procesu otkrivanja znanja poput razumijevanja problema, razumijevanja podataka, pripreme podataka, modeliranja, evaluacije modela i primjene. Analiza je provedena na skupu podataka „World Data of Gender Inequality Indeks" preuzetom s platforme Kaggle.

Deskriptivna analiza ispituje postojeće podatke kako bi se otkrili obrasci, odnosi i specifična područja u kojima je rodna nejednakost najizraženija. Linearna regresija metodom najmanjih kvadrata (OLS) koristi se za razumijevanje i kvantificiranje odnosa između varijabli s ciljem otkrivanja ključnih čimbenika diskriminacije. Grupiranje omogućuje identifikaciju zemalja koje pokazuju slične obrasce rodne nejednakosti, dok detekcija anomalija omogućuje prepoznavanje ekstremnih slučajeva. Na kraju, komparativna analiza smješta Hrvatsku u širi globalni okvir.

U sljedećem poglavlju dan je kratki prikaz povezanih istraživanja, nakon čega u trećem poglavlju slijedi prikaz metodologije i korištenog skupa podataka. Četvrto poglavlje sadrži prikaz rezultata odabranih tehnika dubinske analize podataka. Posljednje poglavlje sadrži interpretaciju rezultata, glavne zaključke i smjernice za buduća istraživanja.

2 Srodna istraživanja

Iako se rodna diskriminacija može očitovati u raznim područjima, pregled literature pokazuje da se većina istraživanja fokusira na analizu diskriminacije na tržištu rada. Dok se neki istraživači bave otkrivanjem diskriminatornog ponašanja poslodavaca (Hangartner et al., 2021), drugi poput Edelsztein & Waisbrot (2023), Bonaccolto-Töpfer et al. (2023) i Alatrista-Salas et al. (2017) proučavaju razlike u plaćama, geografsku raspodjelu vlasništva nad zemljištem i bogatstvom (Fisher & Naidoo, 2016), te zdravstvene indikatore (Heinz et al., 2020).

Iako se rodna pristranost i favoriziranje mogu otkriti korištenjem tehnika obrade prirodnog jezika i

Since this research aims at conducting a comprehensive analysis of gender inequality at the global level in order to compare the findings with the Croatian context, effort was put into finding similar studies. In the research covering the countries of the European Union (Vinska & Tokar, 2021), an agglomerative hierarchical clustering with full association based on the Euclidean distance obtained five groups based on gross domestic product and the Gender Equality Index (GEI) as indicators of economic development and gender equality. The research has shown that there are no distinct "laggards" among EU member states in terms of gender equality; on the contrary, more developed countries almost all belong to the group of "leaders" in gender equality (GEI>90% of the EU average), while less developed countries are equally divided into those that are leaders and those who are gradually adopting the principles of gender equality (GEI between 75-90% of the EU average). It can be concluded that countries with a higher level of development have better GEI scores. Alatrista-Salas et al. (2017) used clustering to investigate gender wage gaps. The cluster analysis conducted by Heinz et al. (2020) challenged the popular belief that greater gender equality automatically means a smaller difference in social well-being among girls and boys and suggested that higher levels of gender equality are associated with increased pressure, dissatisfaction, and related challenges among girls. A global study covering 47 countries suggested that different dimensions of gender inequality may have different causes and occur independently of each other, but also that measures such as income potentially underestimate the extent of gender inequality (Fisher & Naidoo, 2016).

Regression methods help to determine not only whether discrimination exists (e.g., a statistically significant gender effect on wages or another outcome variable), but also how much of the difference remains unexplained after accounting for relevant factors – and it is this unexplained part that is often interpreted as an indicator of discrimination. Research conducted by Bonaccolto-Töpfer et al. (2023) found that the wage gap narrowed significantly among lower-paid workers but widened among the highest-paid, suggesting a persistent "glass ceiling" for women at the top of the income ladder. The authors applied an adapted Oaxaca-Blinder decomposition to estimate and decompose changes between two time points. The approach is based on the method of least squares to estimate the parameters of the regression model.

In this study Croatia is positioned within a broader global framework by discovering key discrimination factors, and by identifying countries that exhibit similar patterns of gender inequality, alongside outliers that deviate substantially from the prevailing trends.

analize sentimenta, npr. Torres-Berrú i sur. (2023), u ovom istraživanju analizira se skup podataka koji sadrži samo strukturirane atribute. Međutim, detaljan pregled metoda analize teksta koje se koriste za prepoznavanje i otkrivanje rodne pristranosti, nejednakosti i nasilja u jeziku može se pronaći u radu (Nacinovic Prskalo, 2024).

Budući da je cilj ovog istraživanja provesti sveobuhvatnu analizu rodne nejednakosti na globalnoj razini kako bi se rezultati usporedili s hrvatskim kontekstom, nastojala su se pronaći slična istraživanja. U istraživanju koje obuhvaća zemlje Europske Unije (Vinska & Tokar, 2021) na temelju Euklidske udaljenosti je aglomerativnim hijerarhijskim grupiranjem s potpunim povezivanjem dobiveno pet grupa na temelju bruto domaćeg proizvoda i indeksa rodne jednakosti (GEI) kao pokazatelja ekonomskog razvoja i rodne jednakosti. Istraživanje je pokazalo da među članicama EU nema izrazito „zaostalih" po pitanju rodne jednakosti. Štoviše, razvijenije države gotovo sve spadaju u skupinu "lidera" u rodnoj jednakosti (GEI>90% EU prosjeka), dok su manje razvijene države podjednako podijeljene na one koje su lideri i one koje su "pristupnici" koji postupno usvajaju načela rodne jednakosti (GEI između 75-90% EU prosjeka). Može se zaključiti da zemlje s višim stupnjem razvoja imaju bolje GEI rezultate. Alatrista-Salas i sur. (2017) koriste grupiranje kako bi istražili rodne razlike u plaćama. Rezultati grupiranja koje su proveli Heinz i sur. (2020) osporili su uvriježeno mišljenje da veća rodna ravnopravnost automatski znači i manju razliku u zdravstvenoj dobrobiti između djevojaka i mladića i ukazuju da je kod većeg stupnja rodne jednakosti izgledniji veći pritisak i nezadovoljstvo kod djevojaka. Globalno istraživanje provedeno u 47 zemalja sugerira da različite dimenzije rodne nejednakosti mogu imati različite uzroke i javljati se neovisno jedna o drugoj, ali i da mjere poput dohotka mogu potencijalno podcijeniti razmjere rodne nejednakosti (Fisher & Naidoo, 2016).

Metode regresije mogu ne samo utvrditi postoji li diskriminacija (npr. statistički značajan učinak na plaće ili neku drugu zavisnu varijablu), nego i koliki dio razlike ostaje neobjašnjen nakon što se uzmu u obzir relevantni čimbenici, a upravo se taj neobjašnjeni dio često tumači kao pokazatelj diskriminacije. Istraživanje koje su proveli Bonaccolto-Töpfer i sur. (2023) pokazalo je da se jaz u plaćama značajno smanjio među niže plaćenim radnicima, ali je porastao među najviše plaćenima, što upućuje na postojan "stakleni strop" za žene na vrhu ljestvice dohodaka. Autori su primijenili prilagođenu Oaxaca-Blinder dekompoziciju za procjenu i dekompoziciju promjena između dvije vremenske točke. Pristup se temelji na metodi najmanjih kvadrata za procjenu parametara regresijskog modela.

3 Methodology and Data

Advances in technology and the proliferation of big data have facilitated the application of in-depth analytical methods capable of uncovering latent patterns within complex datasets. These insights contribute to the development of more effective business strategies, the enhancement of healthcare systems, and a deeper understanding of social phenomena. By applying advanced analytical techniques and machine learning algorithms, an in-depth analysis enables the discovery of hidden relationships that can serve as the basis for developing effective strategies and interventions to reduce gender discrimination and achieve gender balance.

3.1 Methodology

This research follows the CRISP-DM methodology. Data processing and modelling were performed using the R programming language. The workflow is shown in Fig. 1.

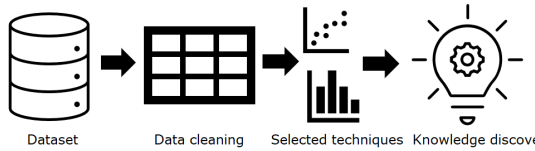


Figure 1. Phases of the research process

First, descriptive analysis was applied to the existing data to uncover patterns, relationships, and specific domains in which gender inequality is most pronounced. The descriptive analysis provided a deep insight into the basic characteristics of the data and revealed important patterns and trends that indicate the presence of discrimination.

Regression analysis is a classic statistical method for quantifying discrimination, and it is the second technique selected for this in-depth analysis. To understand and quantify the relationship between the predictors and the dependent variable, linear regression model estimated via OLS is used.

Clustering algorithms allow visualization of discrimination patterns that follow geographical, cultural, or economic lines, which is valuable for targeted policymaking. We used it to identify countries that exhibit similar patterns of gender inequality. The K-Means algorithm was selected due to its efficiency and simplicity. To determine the optimal number of clusters, the elbow method was used. This method visually illustrates the changes in total variance (sum of squared distances within groups) for different numbers of clusters. The optimal number corresponds to the point at which the decrease in variance begins to slow significantly – a

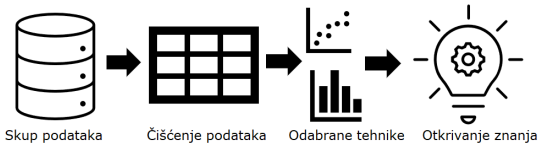
U ovom se istraživanju Hrvatska smješta u širi globalni okvir otkrivanjem ključnih čimbenika diskriminacije i prepoznavanjem zemalja koje pokazuju slične obrasce rodne nejednakosti, uz izdvajanje onih koje znatno odstupaju od prevladavajućih trendova.

3 Metodologija i skup podataka

Zahvaljujući razvoju tehnologije i dostupnosti velike količine podataka, dubinska analiza omogućuje prepoznavanje skrivenih obrazaca unutar kompleksnih skupova podataka, pružajući uvide koji mogu unaprijediti poslovne strategije, poboljšati zdravstvenu skrb te produbiti razumijevanje društvenih pojava. Primjenom naprednih analitičkih tehnika i algoritama strojnog učenja, dubinska analiza omogućuje otkrivanje skrivenih povezanosti koje mogu poslužiti kao temelj za oblikovanje učinkovitih politika i intervencija usmjerenih na smanjenje rodne diskriminacije i postizanje rodne ravnoteže.

3.1 Metodologija

Ovo istraživanje slijedi CRISP-DM metodologiju. Obrada podataka i modeliranje provedeni su korištenjem programskog jezika R. Proces je prikazan na slici 1.



Slika 1. Faze istraživačkog procesa

Najprije je provedena deskriptivna analiza kojom se analiziraju postojeći podaci kako bi se otkrili obrasci, odnosi i specifična područja u kojima je rodna nejednakost najizraženija. Deskriptivna analiza je pružila dubinski uvid u osnovne karakteristike podataka i otkrila važne obrasce i trendove koji mogu ukazivati na diskriminaciju.

Regresijska analiza klasična je statistička metoda za kvantificiranje diskriminacije, a odabrana je kao sljedeća tehnika dubinske analize podataka koja će se primijeniti. Za razumijevanje i kvantificiranje odnosa između prediktora i zavisne varijable, korišten je model linearne regresije i procjena metodom najmanjih kvadrata.

Algoritmi grupiranja omogućuju da se vizualiziraju obrasci diskriminacije koji prate geografske, kulturne ili ekonomske linije, što je iznimno važno za ciljano kreiranje politika. U ovom istraživanju se koristi za identifikaciju skupina zemalja sa sličnim obrascima rodne nejednakosti. K-Means algoritam je odabran zbog svoje učinkovitosti i jednostavnosti. Za određivanje

balance between the complexity of the model and its effectiveness in clustering the data. The parameter *random_state* is set to 42.

The Isolation Forest model was used to detect anomalies, with the parameter *contamination* set to 0.05 and *random_state* to 42 to ensure the reproducibility of the results. The parameter *contamination* sets the expected proportion of anomalies in the dataset to about 5%. Anomaly detection enabled the identification of countries that deviate significantly from the usual trends and patterns.

Finally, comparative analysis enabled positioning Croatia within a broader global framework.

3.1 Dataset

The dataset "WORLD DATA OF Gender Inequality Index"[1] was downloaded from the Kaggle website. The dataset contains a total of 195 instances. After pre-processing, i.e. removing instances with missing feature values, 170 instances remained in the dataset. The dataset consists of the features shown in Table 1. The following is a description of individual features and their significance for the conducted analysis. The Human Development Index (*HDI rank*) is a composite measure of performance in three key dimensions of human development: a long and healthy life, access to knowledge, and an adequate standard of living. HDI represents the geometric mean of the normalised indices for each of these three dimensions (UNDP, 2021). Lower numerical ranks indicate higher human development, while higher ranks indicate lower human development. *Maternal_mortality* indicates access to health services for women. *Adolescent_birth_rate* reflects education and health conditions for young girls. *Seats_parliament* (% held by women) measures the political participation and empowerment of women in political structures. *F_secondary_educ* and *M_secondary_educ* indicate the percentage of women and men, respectively, aged 25 and over, who have completed secondary education. *F_secondary_educ* and *M_secondary_educ* are used to assess the gender balance in education. While *F_Labour_force* reflects the economic engagement of women in the country, *M_Labour_force* reflects the engagement of men. These two features are used to assess gender balance in the labour market. The Gender Inequality Index (*GII rank*) reflects gender-specific disadvantages in three dimensions: reproductive health, empowerment and the labour market. The index scores range from 0, indicating full gender equality, to 1, indicating gender inequality in all dimensions measured (UNDP, 2021).

optimalnog broja klastera korištena je metoda „lakat". Ona vizualno prikazuje promjene u ukupnoj varijanci (zbroju kvadrata udaljenosti unutar grupa) u odnosu na broj grupa. Optimalni broj grupa odgovara točki u kojoj dolazi do naglog usporavanja smanjenja varijance i ta točka predstavlja kompromis između složenosti modela i njegove učinkovitosti u grupiranju podataka. Parametar *random_state* je postavljen na 42.

Otkrivanje anomalija pomaže u prepoznavanju zemalja koje značajno odstupaju od uobičajenih trendova i struktura. Za otkrivanje anomalija primijenjen je model izolacijskih šuma. Parametar *contamination* odnosi se na očekivani udio anomalija u skupu podataka, a u ovom istraživanju iznosi 0,05 čime se očekivani udio anomalija postavlja na oko 5%. Parametar *random_state* postavljen je na 42 kako bi se osigurala reproducibilnost rezultata.

Na kraju, komparativna analiza omogućila je smještanje Hrvatske u širi globalni okvir.

3.1 Podatkovni skup

Skup podataka "WORLD DATA OF Gender Inequality Indeks"[1] preuzet je u s web stranice Kaggle. Skup se sastoji od ukupno 195 instanci. Nakon pretprocesiranja, tj. uklanjanja instanci s nedostajućim vrijednostima značajki, u skupu je ostalo 170 instanci. Skup se sastoji od značajki, tj. atributa prikazanih u Tablici 1. U nastavku slijedi opis pojedinih atributa i njihov značaj za provedenu analizu.

Indeks ljudskog razvoja (*HDI rank*) je sažeta mjera dostignuća u tri ključne dimenzije ljudskog razvoja: dug i zdrav život, pristup znanju i životni standard. HDI predstavlja geometrijsku sredinu normaliziranih indeksa za svaku od ove tri dimenzije (UNDP, 2021). Niži brojčani rangovi označavaju visoki ljudski razvoj, a viši brojčani rangovi označavaju niži ljudski razvoj. *Maternal_mortality* pokazuje pristup zdravstvenim uslugama za žene. *Adolescent_birth_rate* odražava obrazovne i zdravstvene uvjete adolescenata. *Seats_parliament (% held by women)* mjeri političku participaciju i osnaživanje žena u političkim strukturama. *F_secondary_educ* i *M_secondary_educ* predstavljaju postotak žena, odnosno muškaraca, u dobi od 25 i više godina koji su stekli srednjoškolsko obrazovanje. Koriste se za procjenu rodne ravnoteže u obrazovanju. Analogno, dok *F_Labour_force* pokazuje ekonomski angažman žena u zemlji, *M_Labour_force* odražava angažman muškaraca. Te dvije značajke koriste se za procjenu rodne ravnoteže na tržištu rada. Indeks rodne nejednakosti (*GII rank*) odražava rodno zasnovane nepovoljnosti u tri dimenzije: reproduktivno zdravlje, osnaženost i tržište rada. Vrijednosti indeksa se kreću od 0, gdje

[1] https://www.kaggle.com/datasets/rajkumarpandey02/world-data-of-gender-inequality-index

Table 1. Features of the dataset

Feature	Feature description
HDI rank	the overall level of human development in a country based on factors such as health, education and standard of living
Country	name of the country
HDI value	level of human development in the country
GII value	GII value for the country
GII rank	ranking of countries based on their GII score
Maternal_mortality	number of women who die from pregnancy-related causes per 100,000 live births
Adolescent_birth_rate	number of births to women aged 15–19 per 1,000 women in this age group
Seats_parliamentt(% held by women)	percentage of seats held by women in the national parliament
F_secondary_educ	percentage of women aged 25 and over who have completed secondary education
M_secondary_educ	percentage of males aged 25 and over who have completed secondary education
F_Labour_force	percentage of women participating in the labour force
M_Labour_force	percentage of men in the labour force

4 Results

The following subsections present the results of the descriptive analysis, the regression analysis, the clustering, the detection of anomalies and, finally, the comparative analysis.

4.1 Descriptive Analysis

The boxplot analysis shows that most GII values are between 0.15 and 0.50 and that in most countries the rate of secondary school completion for both genders is between 40% and 90%. Female labour force participation shows a wider range, with several

postoji potpuna rodna jednakost, do 1, gdje je rodna nejednakost uočena u svim mjerenim dimenzijama (UNDP, 2021).

Tablica 1. Značajke skupa podataka

Značajka	Opis značajke
HDI rank	opća razina ljudskog razvoja u zemlji na temelju faktora kao što su zdravlje, obrazovanje i standard života
Country	naziv zemlje
HDI value	razina ljudskog razvoja u zemlji
GII value	vrijednost GII-ja za zemlju
GII rank	rangiranje zemalja prema njihovom GII-u
Maternal_mortality	broj žena umrlih od uzroka povezanih s trudnoćom na 100 000 živorođene djece
Adolescent_birth_rate	broj poroda u žena u dobi od 15 do 19 godina na 1000 žena u dobi od 15 do 19 godina
Seats_parliamentt(% held by women)	udio žena u parlamentu u ukupnom broju zastupnika/zastupnica
F_secondary_educ	postotak žena u dobi od 25 i više godina koje je steklo srednjoškolsko obrazovanje
M_secondary_educ	postotak muškaraca u dobi od 25 i više godina koje je steklo srednjoškolsko obrazovanje
F_Labour_force	postotak žena koje sudjeluju u radnoj snazi
M_Labour_force	postotak muškaraca koji sudjeluju u radnoj snazi

4 Rezultati

U sljedećim podpoglavljima prikazani su rezultati deskriptivne analize, regresijske analize, grupiranja, otkrivanja anomalija, i zaključno, komparativne analize.

4.1 Deskriptivna analiza

Analizom kutijastih dijagrama zaključeno je da se većina vrijednosti GII-ja nalazi između 0.15 i 0.50 i da većina zemalja ima udio srednjoškolskog obrazovanja između 40% i 90% za oba roda. Sudjelovanje žena u radnoj snazi ima veći raspon, uz

outliers at the lower end, while male labour force participation is higher and more constant (65–75%). In most countries, women hold between 10% and 35% of parliamentary seats. Although most countries have a low maternal mortality rate, there are some exceptions. The situation is similar with the adolescent birth rate, which is below 75 in most countries, but far higher in some. With both features, extreme values are observed at the upper end.

The correlation matrix is important to understand how different dimensions of gender inequality influence each other. There is a strong positive correlation between GII, maternal mortality and adolescent birth rate (Fig. 2) and a significant negative correlation between these variables and the completion of secondary school for both females and males (Fig. 3).

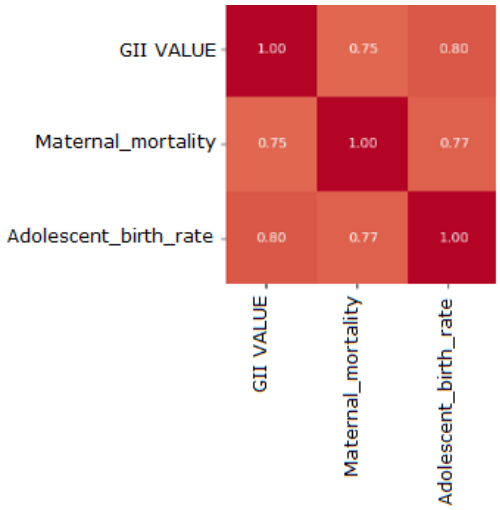


Figure 2. Correlation between GII, maternal mortality and adolescent birth rate

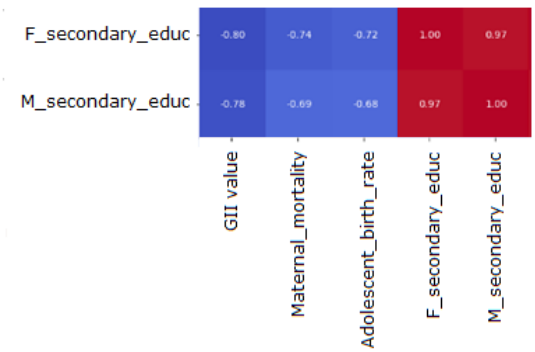


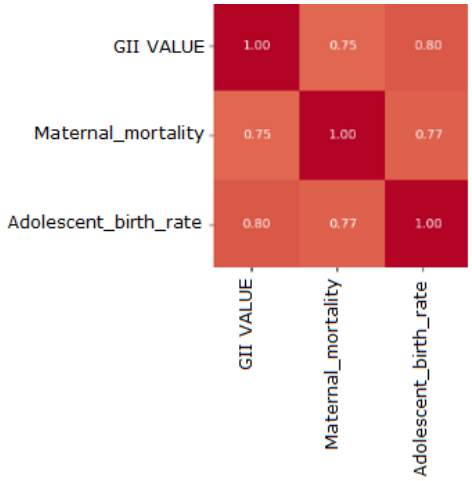
Figure 3. Correlation with secondary education

4.2 Regression Analysis

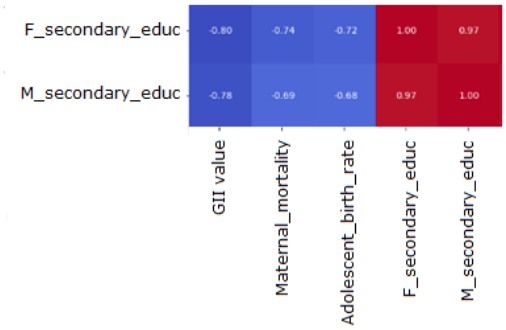
Regression analysis is used to identify key factors that contribute significantly to the variability of GII values in different countries. Scatterplots are generated to provide better understanding of the

nekoliko stršećih podataka prema nižim vrijednostima, dok je kod muškaraca sudjelovanje veće i ujednačenije (65-75%). Većina zemalja imala je između 10% i 35% zastupnica. Uz nekoliko iznimaka, većina zemalja je imala nisku stopu mortaliteta rodilja. Također, iako je većina zemalja imala stopu nataliteta kod adolescenata ispod 75, neke su znatno premašile tu vrijednost. Dakle, u obje značajke uočene su ekstremne vrijednosti na višem kraju.

Korelacijska matrica ključna je za razumijevanje kako pojedine dimenzije rodne nejednakosti utječu jedna na drugu. Utvrđena je visoka pozitivna korelacija između GII, stope mortaliteta rodilja smrtnosti i stope nataliteta kod adolescenata (Slika 2) i značajna negativna korelacija navedenih varijabli sa srednjoškolskim obrazovanjem oba roda (Slika 3).



Slika 2. Korelacija između GII, stope mortaliteta rodilja smrtnosti i stope nataliteta kod adolescenata



Slika 3. Korelacija sa srednjoškolskim obrazovanjem

4.2 Regresijska analiza

Regresijska analiza korištena je za prepoznavanje ključnih čimbenika koji značajno doprinose varijabilnosti GII vrijednosti među različitim zemljama. Za bolje razumijevanje odnosa između

relationships between the independent variables and GII.

The results of the least squares method show that the linear regression model effectively explains the variation in the GII, with a coefficient of determination (R^2) of 0.919, which means that the model explains 91.9% of the total variation. The F-statistic is 258.1, and the extremely low p-value (2.17e-83) confirms the high statistical significance of the model.

The variables HDI rank and adolescent birth rate have a positive and statistically significant impact on GII, indicating that countries with lower levels of human development (i.e. higher HDI rank) and higher adolescent birth rates contribute to greater gender inequality (Table 2). On the other hand, both, female representation in parliament and female labour force participation, have negative and statistically significant impact on GII, suggesting that a higher proportion of women in political decision-making and the labour market contributes to a decrease in gender inequality (Table 2).

Table 2. Coefficients

	Coefficient	p>|t|
Constant	0.2448	0.000
HDI rank	0.0019	0.000
Adolescent_birth_rate	0.0015	0.000
Seats_parliamentt(% held by women)	-0.0032	0.000
F_secondary_educ	-0.0009	0.210
M_secondary_educ	0.0003	0.728
F_Labour_force	-0.0018	0.000
M_Labour_force	0.0009	0.141

4.3 Clustering

Clustering enabled the identification of groups of countries with similar patterns of gender inequality. The K-Means algorithm was selected due to its efficiency and simplicity. The optimal number of clusters is set to 3, as determined by the elbow method (Fig. 4).

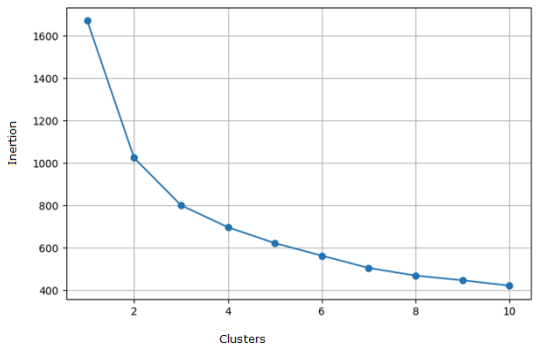


Figure 4. Elbow method

nezavisnih varijabli i GII-ja kreirani su dijagrami raspršenosti.

Rezultati metode najmanjih kvadrata pokazali su da linearni regresijski model vrlo učinkovito objašnjava varijacije u vrijednosti indeksa rodne nejednakosti (GII), s koeficijentom determinacije (R^2) od 0,919, što znači da model objašnjava 91,9 % ukupne varijabilnosti. F-statistika iznosi 258,1, a izuzetno niska p-vrijednost (2,17e-83) potvrđuje visoku statističku značajnost modela.

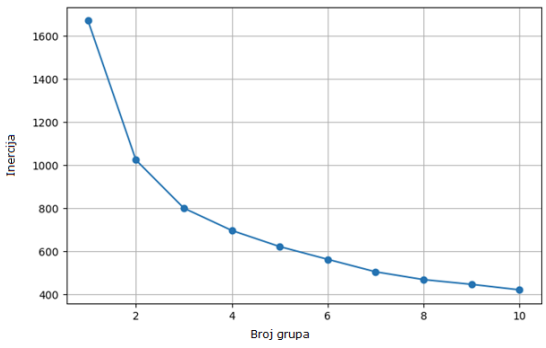
Varijable HDI rang i stopa nataliteta kod adolescenata pokazali su pozitivan i statistički značajan utjecaj na GII, upućujući na to da zemlje s nižim stupnjem ljudskog razvoja, tj. višim HDI rangom te većom stopom nataliteta kod adolescenata doprinose višoj razini rodne nejednakosti (Tablica 2). Varijable zastupljenost žena u parlamentu i udio žena u radnoj snazi su pokazali negativan i statistički značajan utjecaj na vrijednost GII-a, ukazujući na to da veći udio žena u političkom odlučivanju i na tržištu rada doprinosi smanjenju rodne nejednakosti (Tablica 2).

Tablica 2. Koeficijenti

	Koeficijent	p>|t|
Konstanta	0.2448	0.000
HDI rank	0.0019	0.000
Adolescent_birth_rate	0.0015	0.000
Seats_parliamentt(% held by women)	-0.0032	0.000
F_secondary_educ	-0.0009	0.210
M_secondary_educ	0.0003	0.728
F_Labour_force	-0.0018	0.000
M_Labour_force	0.0009	0.141

4.3 Grupiranje

Grupiranje je omogućilo identifikaciju skupina zemalja sa sličnim obrascima rodne nejednakosti. Odabran je K-Means algoritam s 3 grupe. Za određivanje optimalnog broja grupa korištena je metoda „lakat“ (Slika 4).



Slika 4. Metoda “lakat”

The cluster with high GII values represented countries with greater gender inequality, whereas the cluster with low GII values corresponded to countries with lower levels of gender inequality (Fig. 5).

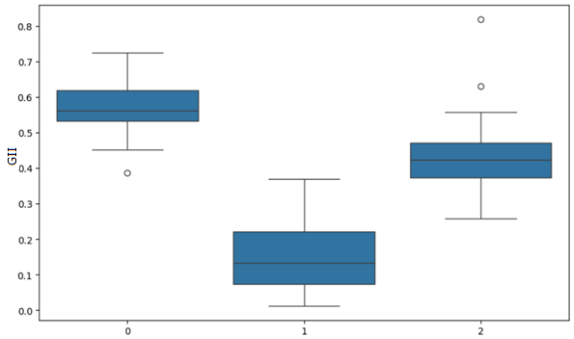


Figure 5. GII distribution per cluster

4.4 Anomaly Detection

The Isolation Forest model was used to detect anomalies, which enabled the identification of countries that represent extreme cases of gender inequality.

Countries such as Tonga, Papua New Guinea, Mauritania, Nigeria, Rwanda, Afghanistan, Yemen, Mozambique and the Central African Republic are characterised by unfavourable indicators and require special attention and targeted measures to reduce discrimination and promote gender equality.

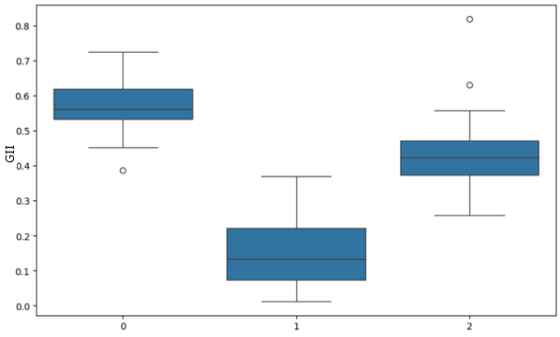
4.5 Comparative Analysis

In order to draw well-founded conclusions about the situation in Croatia, a comparative analysis was carried out. A comparison of Croatia's feature values with global average and the average for countries with very high HDI is given in Table 3.

Table 3. Comparison of Croatia's feature values with global averages

	CRO	World	Countries with high HDI
GII VALUE	0.09	0.3	0.15
Maternal_mortality	8	139.59	13.98
Adolescent_birth_rate	8.6	43.62	13.51
Seats_parliamentt (% held by women)	31.1	25.30	28.40
F_secondary_educ	97	62.78	85.82
M_secondary_educ	100	67.07	87.83
F_Labour_force	45.9	50.22	52.70
M_Labour_force	58.8	70.09	69.68

Grupa s visokim GII vrijednostima predstavlja zemlje s većom rodnom nejednakošću, dok grupa s niskim GII vrijednostima predstavlja one s nižom razinom rodne nejednakosti (Slika 5).



Slika 5. Raspodjela GII po grupama

4.4 Otkrivanje anomalija

Za otkrivanje anomalija u ovom istraživanju je primijenjen model izolacijskih šuma, čime su prepoznate zemlje koje predstavljaju ekstremne slučajeve rodne nejednakosti.

Zemlje poput Tonge, Papue Nove Gvineje, Mauritanije, Nigerije, Ruande, Afganistana, Jemena, Mozambika i Centralnoafričke Republike izdvojile su se po svojim nepovoljnim pokazateljima te su ukazale na potrebu ciljanih intervencija kako bi se smanjila diskriminacija i unaprijedila rodna ravnopravnost.

4.5 Komparativna analiza

Kako bi se donijeli utemeljeni zaključci o situaciji u Hrvatskoj, izvršena je komparativna analiza. Usporedba značajki Hrvatske s globalnim prosjekom i s prosjekom zemalja s vrlo visokim indeksom ljudskog razvoja dana je u Tablici 3.

Tablica 3. Usporedba značajki Hrvatske s globalnim prosjecima

	HR	Svijet	Države visokog HDI-ja
GII VALUE	0.09	0.3	0.15
Maternal_mortality	8	139.59	13.98
Adolescent_birth_rate	8.6	43.62	13.51
Seats_parliamentt (% held by women)	31.1	25.30	28.40
F_secondary_educ	97	62.78	85.82
M_secondary_educ	100	67.07	87.83
F_Labour_force	45.9	50.22	52.70
M_Labour_force	58.8	70.09	69.68

5 Discussion and Conclusion

In this study, the selected techniques were used to comprehensively analyse the causes and manifestations of gender inequality. This resulted in a detailed analysis of gender inequality at the global level and enabled comparing these findings with the Croatian context.

Descriptive analysis revealed the underlying patterns and characteristics of the data. A strong positive correlation was determined between GII, maternal mortality rate, and adolescent birth rate, and a negative correlation was determined between these indicators and secondary education.

Regression analysis identified the key socio-economic factors contributing to inequality. Countries with higher HDI rank (worse human development), more adolescent births, fewer women in parliament, and lower female labour force participation tend to have lower gender equality (higher GII). Education (both male and female secondary education) and male labour force participation did not show significant effects on GII.

Countries were categorised into groups based on the extent of gender discrimination. The visualisation of the partial plots clearly showed that variables such as the maternal mortality rate, the adolescent birth rate and HDI had a strong influence on clustering. In contrast, variables such as the representation of women in parliament and the labour force participation of women showed considerable overlaps between the clusters, which indicated that they are less effective in distinguishing countries based on the characteristics analysed. Cluster 0 had the highest rates of maternal mortality and adolescent pregnancies, indicating countries with greater health risks, cluster 2 included countries with the lowest rates, while cluster 1 fell in-between.

Anomaly detection highlighted countries that deviated significantly from the usual patterns, highlighting the most critical cases.

Finally, comparative analysis enabled the cross-country comparison of key indicators, providing a deeper insight into global and local trends in gender inequality. Significantly lower Croatia's GII indicated relatively higher level of gender equality in Croatia and suggested the effectiveness of measures to reduce gender inequality. Additionally, an exceptionally low maternal mortality rate and a low number of adolescent pregnancies indicated the efficiency of its health and education system. The representation of women in parliament above the global average and above the average of countries with a very high HDI indicated progress in the political participation of women. Croatia demonstrated exceptional strength in the area of education, as the secondary school completion rate for both men and women was well above the global average and the average for industrialised countries.

5 Diskusija i zaključak

U ovom istraživanju primijenjen je niz analitičkih tehnika s ciljem sveobuhvatnog sagledavanja uzroka i manifestacija rodne nejednakosti. To je rezultiralo detaljnom analizom rodne nejednakosti na globalnoj razini i omogućilo pozicioniranje Hrvatske po tom pitanju.

Deskriptivna analiza otkrila je temeljne obrasce i karakteristike podataka. Utvrđena je snažna pozitivna korelacija između GII-ja, stope mortaliteta rodilja i stope nataliteta kod adolescenata, a negativna između navedenih značajki i srednjoškolskog obrazovanja.

Regresijom su identificirani ključni socioekonomski čimbenici koji doprinose nejednakosti. Zemlje s višim rangom prema HDI-ju (što ukazuje na lošiji ljudski razvoj), većom stopom nataliteta kod adolescenata, manjim udjelom žena u parlamentu te nižom stopom sudjelovanja žena u radnoj snazi pokazuju nižu razinu rodne jednakosti (viši GII). Obrazovanje (srednjoškolsko, i žensko i muško) te sudjelovanje muškaraca u radnoj snazi nisu pokazali značajan učinak na GII.

Grupirane su zemlje prema razinama rodne diskriminacije. Vizualizacija parcijalnih dijagrama jasno je pokazala da varijable poput stope mortaliteta rodilja, stope nataliteta kod adolescenata i razine ljudskog razvoja imaju snažan utjecaj na formiranje grupa. Nasuprot tome, varijable poput zastupljenosti žena u parlamentu i udjela žena u radnoj snazi pokazale su značajno preklapanje među grupama, što upućuje na njihovu manju učinkovitost u razlikovanju zemalja prema analiziranim obilježjima. Klaster 0 se istaknuo s najvišim stopama mortaliteta rodilja i adolescentskih trudnoća ukazujući na zemlje s većim zdravstvenim rizicima, klaster 2 je obuhvatio zemlje s najnižim stopama, dok je klaster 1 obuhvatio zemlje sa srednjim rizicima.

Detekcija anomalija je izdvojila države koje značajno odstupaju od uobičajenih obrazaca, ističući najkritičnije slučajeve.

Konačno, komparativna analiza je omogućila usporedbu ključnih pokazatelja među državama, pružajući dublji uvid u globalne i lokalne trendove rodne nejednakosti. Za Hrvatsku je zabilježena značajno niža vrijednost indeksa rodne nejednakosti (GII) u usporedbi sa svjetskim prosjekom, ali i s prosjekom zemalja s vrlo visokim indeksom ljudskog razvoja (HDI). Ova razlika upućuje na relativno višu razinu rodne ravnoteže u Hrvatskoj te sugerira učinkovitost politika usmjerenih na smanjenje rodnih nejednakosti. Dodatno, za Hrvatsku je zabilježena izuzetno niska stopa mortaliteta rodilja i adolescentskih trudnoća, upućujući na učinkovitost zdravstvenog i obrazovnog sustava. Zastupljenost žena u parlamentu je premašila globalni prosjek i prosjek

However, the labour market participation of both genders was still below average, which indicated that further efforts are needed to promote greater integration of the population into the labour market.

Data mining is a powerful tool for uncovering hidden patterns and understanding complex phenomena such as gender discrimination, enabling informed and targeted decision-making. It should be used as the basis for developing effective policies and initiatives to reduce inequality at a global level.

A key limitation of the present study lies in the time-bound nature of the dataset. Access to more recent or feature-rich data would allow for deeper insights and broaden the scope of investigation.

Future research will focus on evaluating model diagnostics to assess the validity of the linear regression approach, as well as applying alternative data mining techniques for comparative analysis.

Acknowledgments

This work was fully supported by the University of Rijeka as part of the scientific project "LinguaGender: Primjena računalne obrade jezika u analizi rodne pristranosti i nasilja u jeziku", uniriiskusni-drustv-23-271.

References

Alatrista-Salas, H., Esposito, B., Nunez-del-Prado, M., & Valdivieso, M. (2017). Measuring the gender discrimination: A machine learning approach. *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 1–6.

Bonaccolto-Töpfer, M., Castagnetti, C., & Rosti, L. (2023). Changes in the gender pay gap over time: the case of West Germany. *Journal for Labour Market Research, 57(*1). https://doi.org/10.1186/s12651-023-00338-7

Edelsztein, V. C., & Waisbrot, S. A. (2023). Breaking down the Gender Pay Gap through a machine learning model. *Convergencia*, 30. https://doi.org/10.29101/crcs.v30i0.20656

Fisher, B., & Naidoo, R. (2016). The geography of gender inequality. *PLoS ONE, 11*(3). https://doi.org/10.1371/journal.pone.0145778

World Economic Forum. (2025). *Global Gender Gap Report*. https://reports.weforum.org/docs/WEF_GGGR_2025.pdf

Grissom, J. A., Timmer, J. D., Nelson, J. L., & Blissett, R. S. L. (2021). Unequal pay for equal work? Unpacking the gender gap in principal

zemalja s vrlo visokim HDI-jem, odražavajući napredak u političkoj participaciji žena. Hrvatska se posebno istaknula u području obrazovanja, s udjelom žena i muškaraca koji završavaju srednju školu znatno iznad svjetskog prosjeka i prosjeka razvijenih zemalja. Ipak, sudjelovanje u radnoj snazi bilo je ispodprosječno za jedne i druge, upućujući na potrebu za dodatnim naporima u poticanju radne aktivnosti stanovništva.

Dubinska analiza podataka predstavlja snažan alat za otkrivanje skrivenih obrazaca i razumijevanje složenih pojava poput rodne diskriminacije, omogućujući donošenje informiranih i ciljanih odluka. Ona treba činiti temelj za oblikovanje učinkovitih politika i inicijativa usmjerenih na smanjenje nejednakosti na globalnoj razini.

Ključno ograničenje ovog istraživanja odnosi se na vremenski ograničenu prirodu skupa podataka. Pristup novijim ili sadržajnijim skupovima podataka omogućio bi dublje uvide i proširio opseg istraživanja.

Buduća istraživanja usmjerit će se na procjenu dijagnostičkih pokazatelja modela kako bi se ocijenila valjanost pristupa linearne regresije, kao i na primjenu alternativnih metoda dubinske analize podataka radi komparativne analize.

Zahvale

Ovaj rad je u potpunosti podržan od strane Sveučilišta u Rijeci kao dio znanstvenog projekta "LinguaGender: Primjena računalne obrade jezika u analizi rodne pristranosti i nasilja u jeziku", uniriiskusni-drustv-23-271.

Reference

Alatrista-Salas, H., Esposito, B., Nunez-del-Prado, M., & Valdivieso, M. (2017). Measuring the gender discrimination: A machine learning approach. *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 1–6.

Bonaccolto-Töpfer, M., Castagnetti, C., & Rosti, L. (2023). Changes in the gender pay gap over time: the case of West Germany. *Journal for Labour Market Research, 57(*1). https://doi.org/10.1186/s12651-023-00338-7

Edelsztein, V. C., & Waisbrot, S. A. (2023). Breaking down the Gender Pay Gap through a machine learning model. *Convergencia*, 30. https://doi.org/10.29101/crcs.v30i0.20656

Fisher, B., & Naidoo, R. (2016). The geography of gender inequality. *PLoS ONE, 11*(3). https://doi.org/10.1371/journal.pone.0145778

compensation. *Economics of Education Review, 82*, 102114.

Hangartner, D., Kopp, D., & Siegenthaler, M. (2021). Monitoring hiring discrimination through online recruitment platforms. *Nature, 589*(7843), 572–576. https://doi.org/10.1038/s41586-020-03136-0

Heinz, A., Catunda, C., van Duin, C., Torsheim, T., & Willems, H. (2020). Patterns of Health-Related Gender Inequalities—A Cluster Analysis of 45 Countries. *Journal of Adolescent Health, 66*(6), S29–S39. https://doi.org/10.1016/j.jadohealth.2020.02.011

Kabeer, N. (2016). Gender Equality, Economic Growth, and Women's Agency: the "Endless Variety" and "Monotonous Similarity" of Patriarchal Constraints. *Feminist Economics*, 22(1), 295–321. https://doi.org/10.1080/13545701.2015.1090009

Leythienne, D., & Pérez-Julián, M. (2022). *Gender pay gaps in the European Union-a statistical analysis*.

Nacinovic Prskalo, L. (2024). Text Analysis Methods for Identifying and Recognizing Gender Bias, Inequality, and Violence in Language: An Overview. *Proceedings of the Central European Conference on Information and Intelligent Systems*. https://www.researchgate.net/publication/389627282

Poli, D., & Maier, S. (2022). Enforcing`EqualEnforcing`Equal Pay for Equal Work' in the EU: what would it take? In *JRC Working Papers on Taxation and Structural Reforms* (11; Issue 11). https://joint-research-centre.ec.europa.eu

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing, 5(*4), 13–22. https://mineracaodedados.wordpress.com/wp-content/uploads/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining (2nd global ed.)*. Pearson, New York.

UNDP. (2021). *Technical Notes. Human Development Reports.* https://hdr.undp.org/sites/default/files/2021-22_HDR/hdr2021-22_technical_notes.pdf

United Nations. (2025). *The Gender Snapshot 2025*. https://reports.weforum.org/docs/WEF_GGGR_2025.pdf

World Economic Forum. (2025). *Global Gender Gap Report*. https://reports.weforum.org/docs/WEF_GGGR_2025.pdf

Grissom, J. A., Timmer, J. D., Nelson, J. L., & Blissett, R. S. L. (2021). Unequal pay for equal work? Unpacking the gender gap in principal compensation. *Economics of Education Review, 82*, 102114.

Hangartner, D., Kopp, D., & Siegenthaler, M. (2021). Monitoring hiring discrimination through online recruitment platforms. *Nature, 589*(7843), 572–576. https://doi.org/10.1038/s41586-020-03136-0

Heinz, A., Catunda, C., van Duin, C., Torsheim, T., & Willems, H. (2020). Patterns of Health-Related Gender Inequalities—A Cluster Analysis of 45 Countries. *Journal of Adolescent Health, 66*(6), S29–S39. https://doi.org/10.1016/j.jadohealth.2020.02.011

Kabeer, N. (2016). Gender Equality, Economic Growth, and Women's Agency: the "Endless Variety" and "Monotonous Similarity" of Patriarchal Constraints. *Feminist Economics*, 22(1), 295–321. https://doi.org/10.1080/13545701.2015.1090009

Leythienne, D., & Pérez-Julián, M. (2022). *Gender pay gaps in the European Union-a statistical analysis*.

Nacinovic Prskalo, L. (2024). Text Analysis Methods for Identifying and Recognizing Gender Bias, Inequality, and Violence in Language: An Overview. *Proceedings of the Central European Conference on Information and Intelligent Systems*. https://www.researchgate.net/publication/389627282

Poli, D., & Maier, S. (2022). Enforcing`EqualEnforcing`Equal Pay for Equal Work' in the EU: what would it take? In *JRC Working Papers on Taxation and Structural Reforms* (11; Issue 11). https://joint-research-centre.ec.europa.eu

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing, 5(*4), 13–22. https://mineracaodedados.wordpress.com/wp-content/uploads/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining (2nd global ed.)*. Pearson, New York.

Grant Thornton International. (2025). *Women in Business 2025 – Impacting the missed generation.*

Torres-Berru, Y., Lopez-Batista, V. F., & Zhingre, L. C. (2023). A Data Mining Approach to Detecting Bias and Favoritism in Public Procurement. *Intelligent Automation and Soft Computing, 36*(3), 3501–3516. https://doi.org/10.32604/iasc.2023.035367

UN Women. (2025). *Gender mainstreaming: strategy for promoting gender equality*. https://www.un.org/womenwatch/osagi/pdf/factsheet1.pdf

United Nations. (2025). *The Sustainable Development Goals Report*. https://unstats.un.org/sdgs/report/2025/The-Sustainable-Development-Goals-Report-2025.pdf

Vinska, O., & Tokar, V. (2021). Cluster analysis of the european union gender equality and economic development. *Business, Management and Economics Engineering, 19*(2), 373–388. https://doi.org/10.3846/bmee.2021.15382

UNDP. (2021). *Technical Notes. Human Development Reports.* https://hdr.undp.org/sites/default/files/2021-22_HDR/hdr2021-22_technical_notes.pdf

United Nations. (2025). *The Gender Snapshot 2025*. https://reports.weforum.org/docs/WEF_GGGR_2025.pdf

Grant Thornton International. (2025). *Women in Business 2025 – Impacting the missed generation.*

Torres-Berru, Y., Lopez-Batista, V. F., & Zhingre, L. C. (2023). A Data Mining Approach to Detecting Bias and Favoritism in Public Procurement. *Intelligent Automation and Soft Computing, 36*(3), 3501–3516. https://doi.org/10.32604/iasc.2023.035367

UN Women. (2025). *Gender mainstreaming: strategy for promoting gender equality*. https://www.un.org/womenwatch/osagi/pdf/factsheet1.pdf

United Nations. (2025). *The Sustainable Development Goals Report*. https://unstats.un.org/sdgs/report/2025/The-Sustainable-Development-Goals-Report-2025.pdf

Vinska, O., & Tokar, V. (2021). Cluster analysis of the european union gender equality and economic development. *Business, Management and Economics Engineering, 19*(2), 373–388. https://doi.org/10.3846/bmee.2021.15382