A DATA WAREHOUSE SYSTEM FOR AN ANALYSIS OF UNEMPLOYMENT RATE IN THE REPUBLIC OF SERBIA

Marija Đukić[1], Ivan Luković[2]
[1]University of Belgrade-Faculty of Organizational Sciences,
marija.djukic@fon.bg.ac.rs
[2]University of Belgrade-Faculty of Organizational Sciences,
ivan.lukovic@fon.bg.ac.rs

Abstract: *In the paper, we present a data warehouse system to analyze the unemployment rate in the Republic of Serbia. The goal of our research is to improve the analytical capabilities of the unemployment rate in Serbia by creating a new business intelligence tool and predictive machine learning models. First, we discuss research motives and the unemployment problem, and then we present the development process of the proposed data warehouse system. The Data Warehouse Quality methodology has been deployed to assess the quality of the data. Machine learning algorithms have been utilized to build predictive models and gain insights into the differences in unemployment rates between young and experienced workers. Finally, we have created several reports to visually present the results of the proposed data analyses.*
Keywords: *unemployment, data warehouse, data quality, machine learning.*

Apstrakt: *U radu predstavljamo sistem skladišta podataka za analizu stope nezaposlenosti u Republici Srbiji. Cilj našeg istraživanja je da unapredimo analitičke mogućnosti stope nezaposlenosti u Srbiji kreiranjem novog alata poslovne inteligencije i prediktivnih modela mašinskog učenja. Prvo razmatramo motive istraživanja i problem nezaposlenosti, a zatim predstavljamo proces razvoja predloženog sistema skladišta podataka. Metodologija kvaliteta skladišta podataka je primenjena za procenu kvaliteta podataka. Algoritmi mašinskog učenja su korišćeni za izgradnju prediktivnih modela i sticanje uvida u razlike u stopama nezaposlenosti između mladih i iskusnih radnika. Konačno, napravili smo nekoliko izveštaja kako bismo vizuelno predstavili rezultate predloženih analiza podataka.*
Ključne reči: *nezaposlenost, skladište podataka, kvalitet podataka, mašinsko učenje.*

1. INTRODUCTION

The Republic Institute of Statistics performs analyses combining data from different government and business sources. Currently, data is retrieved from individual institutions as needed, usually during survey periods. The establishment of a centralized system that would integrate data from administrative institutions, providing

a single data source for analysis has not been realized to the date. Similarly, the International Labor Office relies on labor market data for setting employment objectives and evaluating employment policy effectiveness over time. Data analysis often involves creating models or applying machine learning to pre-prepared data. Therefore, there is a need to develop a solution that can consolidate data from diverse sources and format it appropriately. The development of a data warehouse can contribute to obtaining the necessary data for labor market decision-making and facilitate the seamless integration and centralization of this data.

The scope of our research is confined to the labor market in the Republic of Serbia, examining registered unemployment from 2008 to 2020. Structural changes in the labor market, driven by globalization, rapid technological advancement, economic crises, and an aging population, have led to persistently high unemployment rates. In 2022, the unemployment rate in Serbia stood at 9.4%, with youth unemployment at a notably higher rate of 24.4%. Such high unemployment levels signify imbalances between labor supply and demand, reduced economic activity, and a slow job creation process. This research area is chosen as the unemployment rate is an important indicator for assessing society's economic well-being, and the published statistical data may not accurately reflect the true state.

The goal of our research is to enhance the analytical capabilities of the unemployment rate in Serbia by developing a data warehouse and predictive machine learning models. We use data from various sources consolidated within a data warehouse to gain insights into unemployment issues in the country. Leveraging the DW capabilities, we conduct comprehensive analyses to identify factors influencing the country's unemployment rate and predict increases through ML models. The research is expected to offer guidance for enhancing labor market analysis, providing an empirical foundation for addressing unemployment challenges. These findings will underpin the formulation of measures within the National Employment Service and the broader International Labor Office. Also, it will facilitate a better grasp of labor market dynamics and the identification of actions needed to improve national employment policies.

The study (Axelrad et al., 2018), investigating the factors influencing the employment chances of young and older workers, was taken as a starting point for our research. The findings indicate that youth and adult unemployment are influenced by different characteristics. Additionally, (O'Higgins, 2001) highlights a relationship between youth and adult unemployment, with youth unemployment rates typically exceeding adult unemployment rates across countries, irrespective of the overall unemployment rate. To verify these claims in the context of the labor market in the Republic of Serbia and validate the analytical capabilities of the newly developed business intelligence tool, a research hypothesis was formulated as follows: "The unemployment rate among young workers is higher compared to the unemployment rate of older workers."

2. RELATED WORK

Over the past decades, the labor market and unemployment have garnered significant attention. Interest in the topic notably surged after the economic crisis of 2008, as unemployment emerged as a critical political concern.

In recent research, (Gogas et al., 2022) forecast the unemployment rate in Eurozone member states, employing various machine learning algorithms. The analysis revealed that the Random Forest algorithm yielded the most accurate results. (Devashish et al., 2019) developed a model to analyze the causes of youth unemployment in India, applying machine learning algorithms to collected data. However, the model lacks a data warehouse, assuming that it operates on cleaned and prepared data. (Dahliah & Nur, 2021) examined the economic consequences of unemployment, particularly its impact on the poverty rate. Through linear regression, it is determined that high unemployment has a noticeable effect on the poverty rate. In (Zhang & Shi, 2019), a data warehouse was developed for youth employment data in China. The research analyzed employment rates based on factors such as education level, and majors studied at universities. A decision tree classification algorithm is used to identify in-demand majors and determine youth employment rates relative to education level, degree, and city of residence. (Dieni et al., 2021) conducted research on the unemployment rate while simultaneously developing a data warehouse using the Business Intelligence roadmap approach. The K-Nearest Neighbor algorithm was applied to identify patterns in the unemployment rate. (Mahringer, 2004) described the development of a data warehouse by the Austrian Ministry of Labor and Economy. The objective was to establish a system capable of integrating data and translating administrative data into relevant labor market information.

Our research incorporates a data warehouse, integrating various data sources, similar to (Mahringer, 2004), and machine learning algorithms for analyzing labor market data similar to (Devashish et al., 2019). However, unlike previous research, our study addresses data quality by applying the Data Warehouse Quality methodology to evaluate the suitability of retrieved data for data warehouse development and subsequent analyses.

3. UNDERSTANDING OF THE UNEMPLOYMENT NOTION

Unemployment is an important indicator of the economic state of society and affects various aspects of the lives of individuals and communities. However, the definition of unemployment may vary among authors, indicating the lack of a unique and precise definition and resulting in different unemployment data figures. In our study, unemployment is defined as „the presence of skilled workers who are out of work, out of paid work, or out of work within a recognized profession“ (Giddens & Griffiths, 2006). Unemployment can impact not just a country's economy but also the social well-being and health of its population. Some obvious causes include low productivity, inadequate education levels, and underdeveloped infrastructure. It was believed that

technological progress contributed to unemployment. As technological advancements have led to the creation of new jobs, unemployment has become a significant subject of research.

Per definition in (International Labour Office, 2015), an unemployed person is defined as "a person over the age of 15 who has been out of work in a given week, is available for work, and is actively looking for paid employment or income from self-employment". The indicator used to measure unemployment is the unemployment rate. This rate is defined as the percentage of unemployed individuals in the labor force. The unemployment rate provides insight into the economy's capacity to provide employment opportunities for individuals willing to work, and actively seeking employment.

The research spans from 2008 to 2020, excluding data for 2021 and 2022. This omission is due to the implementation of a new data collection methodology. The changes introduced refer to altered principles for categorizing individuals as employed, which subsequently impacts the „unemployed“ class.

4. DATA QUALITY ASSESSMENT

The use of poor-quality data is a primary cause of the failure of business intelligence projects (Singh & Singh, 2010). Inadequate data quality impacts not only data warehouses but also affects decision-making processes and day-to-day operations. Data quality is measured through several quality dimensions. When data meets these quality dimensions, it is considered to be of „high quality“.

Data Warehouse Quality methodology is used to evaluate and improve data quality within a data warehouse. It is data-driven methodology, meaning that improvements in data quality result from alterations in the actual data values (Batini et al., 2009). Following the example of other methodologies, the quality of the data is evaluated through quality dimensions such as precision, completeness, consistency, accuracy, adaptability, and data integrity.

Based on the software-generated report, key quality dimensions were calculated according to the DWQ methodology, as shown in Table 1. The Accuracy and Interpretability dimensions for EmploymentRate and UnemploymentRate indicators reveal uninterpretable values, attributed to null values per the Completeness dimension. However, the Consistency dimension shows no integrity rules are violated for any indicator.

Table 1: Quality dimensions per DWQ methodology for unemployment data

Field name	Accuracy [num]	Completeness [%]	Consistency	Interpretability [num]
AgeCategory	624	100	/	624
Population	624	100	/	624
ActivityRate	624	100	/	624
EmploymentRate	598	96	/	598
UnemploymentRate	598	96	/	598
InactivityRate	624	100	/	624

The values of the quality dimensions show the data is of appropriate quality, affirming its suitability for the data warehouse. However, certain indicators have null values, and these need to be addressed during the ETL process to guarantee quality analysis.

5. DATA WAREHOUSE DEVELOPMENT

The data warehouse in our research was developed following Kimball's approach, dimensional modeling. The main conception was to establish a centralized system that would integrate data from various administrative institutions, providing a single data source for the analysis most often conducted by the Republic Institute of Statistics.

Dimensional modeling focuses on business processes and streamlines DW design by eliminating the need for model normalization (Kimball, 1996). Data is organized per dimensions representing key aspects of the business context, making it more comprehensible to specific business requirements. Additionally, the denormalized model enhances DW performance, enabling swift data access and efficient retrieval, simplifying query writing, and facilitating data analysis. The DW schema is shown in Figure 1.

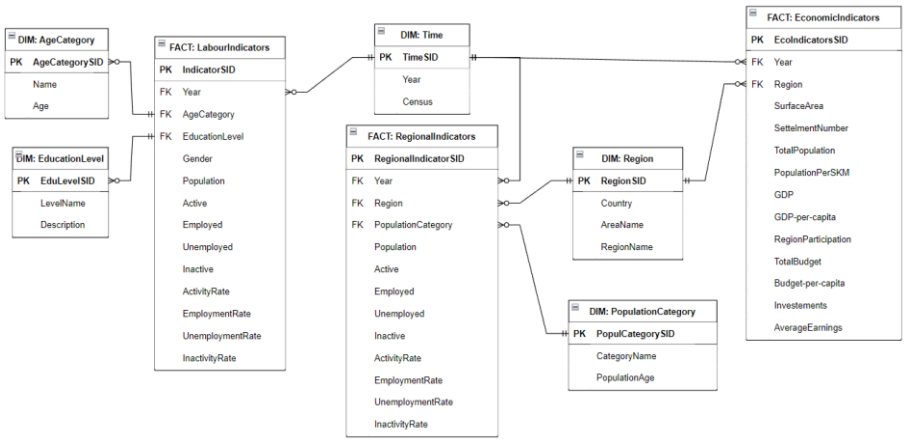

Figure 1: Data warehouse schema

In the ETL process, data is extracted from various data sources and then transformed into appropriate data formats. Subsequently, NULL values detected during data quality assessment are addressed. The data is loaded into data warehouse tables using an "incremental" loading method. The ETL model is designed to load dimension tables first, followed by the loading of fact tables in a sequential manner.

6. RESULTS

Machine learning algorithms are employed to analyze the unemployment rate and test the established research hypothesis. The initial model was developed for selecting indicators that hold significance in determining the unemployment rate, and the selected indicators are used in subsequent models. The objective was to enhance the performance of machine learning models by eliminating irrelevant or redundant indicators that could potentially undermine prediction accuracy. Due to its superior performance highlighted in the literature, the random forest algorithm was chosen to identify significant indicators. Among the 14 input variables, the model identified 7 variables as significant contributors to determining the unemployment rate: "AgeCategories", "Population", "ActivityRate", "EmploymentRate", "NumOfActive", "NumOfEmployed", and "NumOfUnemployed".

To gain insights into the impact of each variable on the unemployment rate, a second model was constructed. Linear regression was employed to derive a linear equation that effectively describes how changes in independent variables influence changes in the unemployment rate, which serves as the dependent variable. The equation (1) represents the outcome of the linear regression model.

$$\begin{aligned} UnemploymentRate &= -0{,}017 * Population + (-0{,}504) * NumOfActive \\ &+ 0{,}535 * NumOfEmployed + 0{,}50 * NumOfUnemployed + 0{,}95 \\ &* ActivityRate \\ &+(-1{,}231) * EmploymentRate + (-0{,}455) * AgeCategory + 41{,}754 \end{aligned} \quad (1)$$

The coefficients obtained from the linear regression model reveal that a 1% increase in the employment rate corresponds to a 1.231% decrease in the unemployment rate. Conversely, a 1% increase in the activity rate leads to a 0.950% increase in the unemployment rate. This suggests that individuals newly categorized as part of the "active population" may also fall into the "unemployed" category.

In the final model, classification algorithms are employed to analyze the unemployment rate and classify age categories based on whether there is an increase or decrease in the unemployment rate. The performance of this model is evaluated using performance metrics shown in Table 2.

Table 2: Classification model performance metrics

	Decision Tree	Random Forest	Boosting	Gradient Boosted Tree
Error rate	0.187	0.198	0.200	0.205
True positive	0.575	0.535	0.575	0.722
True negative	0.922	0.925	0.904	0.829
False positive	0.078	0.075	0.096	0.171
Accuracy	0.813	0.802	0.800	0.795
Precision	0.775	0.768	0.735	0.662

To effectively implement preventive measures for reducing unemployment, the accurate prediction of true positive cases is of great importance. The gradient-boosted tree algorithm demonstrated the best performance in correctly predicting the positive class, achieving a sensitivity of 0.722, which means that it correctly predicts the increase in the unemployment rate 72.2% of the time. The decision tree algorithm performed the best in predicting instances across the entire dataset, boasting an accuracy rate of 81.3%. Additionally, the decision tree achieved a precision rate of 77.5% in identifying true positive cases.

The research provided concrete answers and thus confirmed the hypothesis through the use of data visualization software. Figure 2 demonstrates that the unemployment rate among young workers is notably higher than the unemployment rate among older workers, irrespective of the overall unemployment rate level.

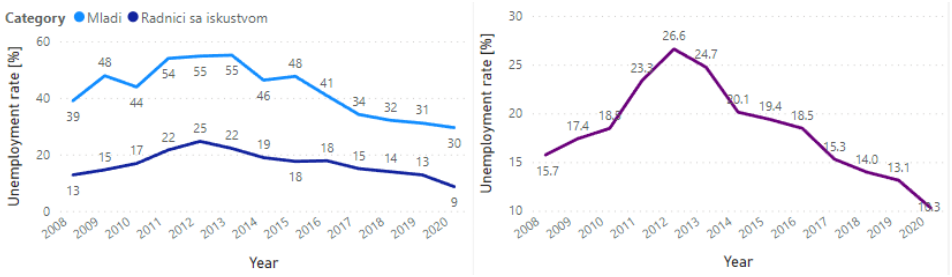

Figure 2: Category-based unemployment rate (left) and overall unemployment (right)

Figure 3 additionally confirms the unemployment rate is higher among young individuals compared to older workers. The data illustrates that as workers age, the unemployment rate decreases. Notably, the age group from 15 to 24 exhibits particularly high unemployment rates, which is understandable as this age range often overlaps with the period of education and early career entry. Additionally, the unemployment rate for the 25 to 34-year-old category is higher compared to other age groups, suggesting that individuals with more experience may find employment more readily or enjoy more stable job opportunities.

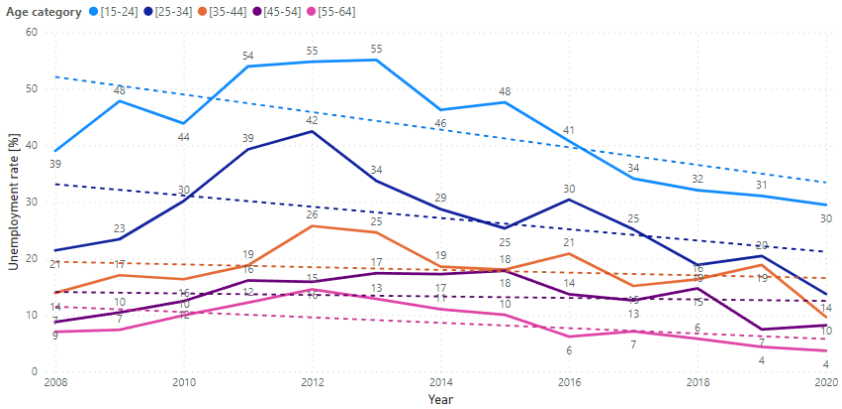

Figure 3: The unemployment rate in relation to age category

7. CONCLUSION

In this paper, we have proposed a new business intelligence solution to improve the analysis of the unemployment rate in the Republic of Serbia by centralizing data from various state institutions. Machine learning algorithms have been applied to select and analyze key indicators affecting the unemployment rate and predict its growth.

Our findings confirm a higher unemployment rate among young workers compared to older individuals. Notably, the 15 to 24 age group exhibited particularly high unemployment rates, while the 25 to 34 age group also had a relatively higher rate, possibly indicating the role of experience in job search and employment. Additionally, our research revealed a declining trend in the unemployment rate over time, suggesting the potential effectiveness of national measures in reducing unemployment. To obtain a more comprehensive understanding of unemployment, we consider it important to expand the dataset, including data that may indicate an "artificial" decrease in the unemployment rate such as migrations.

Machine learning models that we proposed in our research demonstrate potential for accurate predictions of changes in the unemployment rate. Such models can be valuable for state institutions, enabling proactive measures to address unemployment more effectively. Future research may involve a deeper analysis of unemployment causes, its consequences, and the effectiveness of measures aimed at reducing it.

Our DW solution has potential to enhance labor market data analysis for the Republic Institute of Statistics by eliminating the need for periodic data retrieval. Furthermore, the DW ensures that data are appropriately prepared, rendering it convenient for use in setting employment goals and policies by the National Employment Service.

REFERENCES

Axelrad, H., Malul, M., & Luski, I. (2018). Unemployment among younger and older individuals: does conventional data about unemployment tell us the whole story?. *Journal for labour market research*, *52*(1), 1-12.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, *41*(3), 1-52.

Dahliah, D., & Nur, A. N. (2021). The influence of unemployment, human development index and gross domestic product on poverty level. *Golden Ratio of Social Science and Education*, *1*(2), 95-108.

Devashish, D., Varun, S., & Vikram, D. (2019). Modeling Youth Unemployment. *Towards Data Science*. Retrieved from https://towardsdatascience.com/https-medium-com-vikramdevatha-modeling-youth-unemployment-d0f7cbcd078a, accessed 12.09.2022.

Dieni, T. O., Tania, K. D., Fathoni, F., Jambak, M. I., & Putra, P. (2021). Business Intelligence for Unemployment Rate Management System. *The IJICS (International Journal of Informatics and Computer Science)*, *5*(2), 111-117.

Giddens, A., & Griffiths, S. (2006). *Sociology*. Polity.

Gogas, P., Papadimitriou, T., & Sofianos, E. (2022). Forecasting unemployment in the euro area with machine learning. *Journal of Forecasting*, *41*(3), 551-566.

International Labour Office. (2015). *National Employment Policies - A guide for workers' organisations.* Geneva. International Labour Office.

Kimball, R. (1996). *The data warehouse toolkit: practical techniques for building dimensional data warehouses*. John Wiley & Sons, Inc..

Mahringer, H. (2004). Data Warehouse (DWH) Monitoring in the Public Employment Service (PES).

O'Higgins, N. (2001). Youth unemployment and employment policy: a global perspective. Geneva: International Labor Office.

Singh, R., & Singh, K. (2010). A descriptive classification of causes of data quality problems in data warehousing. *International Journal of Computer Science Issues (IJCSI)*, *7*(3), 41.

Zhang, S., & Shi, Y. (2019). Application Research of Data Warehouse Technology in College Student Enrollment and Employment Decision. In *1st International Symposium on Education, Culture and Social Sciences (ECSS 2019)* (pp. 577-580). Atlantis Press.