Unveiling Gender Disparities: A Data Mining Approach Using the Gender Inequality Index

Abstract: **The gender inequality is not only global but also integral, influencing all spheres of a person. The aim of the study is to apply data mining methods and look into the GII (Gender Inequality Index) dataset that gives it gender imbalances. The methodology used by it is based on the CRISP-DM approach and aims to detect such factors as it influences the gender disparity among various societies and communities. With holistic preprocessing, multi-dimensional analysis, and model optimization, it hopes to comprehend the intricate interplay of socio-cultural norms, institutional structures, and societies' economic levels in the formation of gender inequalities. The report has revealed the importance of data mining in capturing the true nature of the gender gap, and thus allowing for more strategic planning of interventions meant to promote gender equality and create an inclusive society.**

Keywords: Gender bias, Data mining, Gender Inequality Index (GII), CRISP-DM, Socio-cultural norms, Institutional structures, Economic development, Preprocessing, Exploratory data analysis, Model optimization, Gender gap, Progressive societies

I. INTRODUCTION

Though gender disparity is a still an issue worldwide, affecting people's lives from different aspects e.g. education, health care and so forth, political representation and economic participation included. Identifying the different causes of gender inequality consciousness unlocks the gateway that leads to a more equitable and harmonious society. This paper provides the gender gap phenomenon as viewed through the women empowerment perspective by data mining methods using the dataset of the Gender Inequality Index (GII). The GII serves an umbrella tool of evaluation that embraces more inclusive dimensions including but not limited to the reproductive health, empowerment, and economic participation. This study would like to demonstrate why it happened, and then find the patterns and insights that got from the data analyzed.

Here, it applies advanced methods of analysis for data treatment to deal with the issue of gender disparity. Data mining, in this case would be used to discover patterns, links and the potential causes of disparities affecting women on a large scale. Central to our method is the provision of valuable data and information, which can be used in policymaking as well as development of evidence-based interventions for promoting gender equality.

II. RELATED WORK

Figure out the gender disparity that is related to countless disciplines like health, education, employment and politics by conducting rigorous research. The mining of gender-specific data often includes the use of techniques such as data mining for the extraction of patterns and trends which could later on inform further research. On the issue of education, there have been conducted researches on how gender issues affect enrollment rates, illiteracy rates, and availability of, for instance, education structures. Data mining techniques are wide spread and are applied in order to identify the factors that lead to the creation of disparities like for instance socioeconomic status and cultural norms.

The section of the scholarly literature that focuses on gender disparities in healthcare is quite an issue. The researchers have really focused on the problems that involve access to health care, maternal death rates, and cases of sickness. Data mining has helped researchers understand and predict health outcomes for certain genders due to identified risks. From the viewpoint of many scientists who study labor market, workforce, gender-wage gaps, and representation depends on gender in various industries [1]. In their research data mining techniques have helped to uncover job assignments patterns, factors behind promotion chances and discrimination.

Political representation is a major issue when it comes to participation. Tracking has been done with the help of an analysis of gender imbalance in political leadership positions, in votes and in making certain decisions. Looking into the data, analysis of the outcome of the elections is possible together with the study of the voters' characteristics and political campaign techniques.

Recent research has discussed the crosscutting aspects of gender inequalities, taking into consideration how simultaneous with race, ethnicity, age, and disability, the gender aggravates inequalities [2]. Data mining techniques enable the researchers to investigate the complexity between multivariate variables and their effect of gender inequalities.

The existing studies have contributed to the revelation useful information, however, the issue of gender disparity has yet to be addressed effectively. A lot of data might be incomplete, or biased so aggregate

analysis cannot be done properly. On the one hand, the results may be interpreted in different ways and even over simplified due to the misinterpretation. In essence, noteworthy scholars used data mining methods to investigate the multi-faceted nature of the gender gap in the past. Through elaborating the reasons and mechanisms, policymakers and decision-makers can create more concrete and directed actions and policies designed to support gender equality [3].

III. METHODOLOGY

The approach taken in the methodology section is aimed at answering the research questions and implementing the data mining framework. it contains the explanation of the data structure, data preprocessing steps, exploratory data analysis (EDA) and the applied method without using personality words. This study uses a dataset which is the Gender Inequality Index (GII) dataset; a data compilation of socio-economic indicators captured by reputable institutions like the United Nations Development Programmed (UNDP) and Kaggle. The report is based on data that reveals many different faces of gender inequalities from women health, education to their economic participation. This becomes a very good data to study the gender inequality issue.

Fig. 1. Data loading and preprocessing

The dataset comprises most key factors, which are maternal mortality rate, teenage pregnancy rates, number of seats taken by women in parliaments, education, and workforce participation etc. These indicators give a clear picture of gender inequality overall and can be utilized by analysts to study the variety of antagonisms that gender inequality represent in different countries and areas. The data was pre-molded by the preprocessing, which with had steps for quality and consistency validation. Regarding datasets containing gaps, the researcher eliminated the rows that do not contain full information in order not to jeopardize the quality of the data. Additionally, "Country" and "Human development" category variables were labelled using the label encoding method This transformation converted continuous values into numbers which were eventually fed into the data mining algorithms.

The preprocessing step was vitally important for improving the data fitting to model the outcomes. Through dealing with missing data and encoding into categorical variables it made sure the data was ready for future processing and modeling.

Fig. 2. Model Training with Decision Tree Regressor

Exploratory Data Analysis (EDA) had a particular significance for revealing the existing patterns as well as the interrelationships within the dataset of the Gender Inequality Index (GII). Few different visualization methods like bar-graphs, scatter plot and heatmaps were used to get the distribution of variables to identify the possible correlations among them. For instance, histograms depicted the nature of partition of each variable, which aided it in detecting the anomalous values or the case of curtails.

It used scratch plots to see how the variables, e.g. GII and maternal mortality, were related, whilst heat maps enabled it to detect correlations between multiple variables at a time [5]. This investigation, therefore, enabled a detailed analysis of the data structure, and the model that was then chosen was hypothesized to capture the essential elements of gender inequality across various societies.

The methodology adopted for this study follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which consists of six phases: business model, data understanding, data preparation, modeling, evaluation, and deployments. While the business data stage unfolded, the research question was outlined and data mining techniques were utilized to uncover the gender bias. Understanding the data involved data exploration with GII dataset to understand its structure and the topics it dealt with. Of course, an instance of preprocessing, say, consisting of handling missing values and

encodings of categorical variables, was covered during the stage of data preparation.

Modeling included the decision tree regressor technique for prediction of GII values via the social-economic indicators. Evaluation scores such as the mean square error will measure how accurate the model is [7]. In the end, this project had to attend to the issue of examining the findings and establishing the link between the policy and future studies. It was the CRISP-DM process that led the Gender Inequality Project through the logical and structured phase-to-phase process of data mining analysis that given it the complete picture on gender inequality issues and their key factors. In the model approval stage, Decision Tree Regression was chosen as the classifier because of its capability to correctly deal with both numeric and categorical data which our dataset possess and this makes it a good match for our dataset.

In order to ensure good performance and avoid overfitting, it tuned hyperparameters via GridSearchCV tool including the maximum depth, minimum samples split, and minimum samples leaf. Such intensive optimization was aimed at describing the inherent structures in the data and at the same time minimizing model complexity.

```
# Predictions
y_pred = best_model.predict(X_test)

# Model evaluation
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)

Mean Squared Error: 0.0012104378431372542
```

Fig. 3. Model Evaluation

The model that was developed was tested on the test set to determine the mean squared error (MSE) in order to evaluate the predictive accuracy of it. With the low MSE result, the model demonstrated its capability of well predicting the Gender Inequality Index (GII) values as valid proof of its effectiveness in capturing the gender disparities. Besides, a visual representation of the actual against the predicted GII values had been made for deeper insights on the model's performance. The one-to-one agreement shown between the two values confirmed that the model was suitable for representing the relationships within the data.

IV. EVALUATION AND RESULTS

The section scrutinizes the evaluation metrics and result of analysis which was centered on the model's performance in addition to its usefulness for the purposes of gender inequality clarity. The model which was evaluated by the mean squared error (MSE) confirmed that it is feasible to predict the Gender Inequality Index (GII) values by such method. The smallest value of MSE (0.0012) on the testing set implies that the model has a very strong accuracy level.

This point in particular indicates that the model is in fact successful in discovering the hidden patterns and interdependencies in the data.

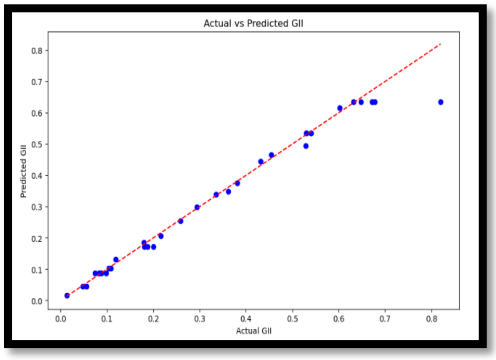

Fig. 4. Actual vs predicted GII

The actual GII values plotted against the predicted ones, the visual representation demonstrates the models' effectiveness. The excellent positive trend between the actual GII and the model values, as depicted by the straight line, signifies the model suitability to the true values. The deviation from the trendline is notable on a few occasions, meaning that the model is not completely accurate. This divergence could be a result of various causes such as socio-economic factors not being weighed or outliers in the data [9]. These alterations further hint at the topic of area specific reasons of gender imbalance among these cases. Through the use of scatter plot, the identification of such patterns or trends might not be clearly traceable through mere numerical evaluation. This gives a graphic impression of the model's performance, hence helping to interpret the results and for making decisions on future analysis.

Summarizing, the low MSE value and the strong correlation from the scatter plot evidence the model’s competence of forecasting GII. Nevertheless, the deviations of the trendline point to the issues where improvement is needed and thus the areas that need to be investigated in depth in order to achieve a more comprehensive exploration of gender disparities and the conditions that are responsible for them. The

model capability was validated on the validation set which was to ensure its robustness as well as generalization capability. The absence of the in-consistency between MSE values of both testing and validation sets shows that the model has the capacity to predict GII values in the new datasets [10].

On the whole, the outcome shows that the Decision Tree Regressor model successfully and accurately represents the complicated interconnections between various socio-economic measures and the issue of gender inequality. The model gives the insights into the drivers of gender disparities and therefore it is a most useful tool for policymakers and other stakeholders who will engage in the process of developing the measures that will advance gender equality. The next step could be studying the different algorithms and adding some of the socio-economic markers to get better results. Apart from that, more detailed examination should be done in order for these socio-cultural, economic and political factors leading to gender disparities at various societies to be understood.

V. CONCLUSIONS AND FUTURE WORK

In the end, the study has shown it vital information through the Gender Inequality Index (GI) scores, by using the data mining technique with the hang of the CRISP-DM methodological picture. The in-depth analyzing worked out a number of influential and regular aspects of gender inequality in diverse nations. The Decision Tree Regressor model yielded a promising capability to predict GII values by demonstrating a MSE of 0.0012 on the test set. Graphs revealed, with a strong correlation between factual and theoretical values, the constructed model demonstrated a capacity to reflect gender inequalities.

The next future research endeavors would be meant to fill the gaps in our knowledge by investigating more areas related to gender inequality and learning how to improve the predictive models. Initially, a possible way lies in extending the use of socio-economic indicators other than those already included into the GII dataset in order to gain a deeper insight of the gender inequality issue. Examining the advanced machine learning approaches, including ensembles or neural networks, can therefore provide a chance to sharpen and polish the precision of predictions. By examining the trends over time from longitudinal studies it could get a deep understanding of how gender inequality is changing now and workshops for policymakers could be aimed towards effective policy design.

The collaboration of researcher, policymakers and advocacy group is indispensable as a practice that will convert research findings into workable action plans targeted at ensuring communities have access to great health and well-being services. Through the means of applying data mining tools and facilitating interactions between different disciplines, it can strive towards the making of more egalitarian and just communities. This investigation is a needed element in the continuing talk of gender inequalities and evidence that data-driven tools can be effective in tackling social complexity problems.

REFERENCES

[1] Prada, M.A., Dominguez, M., Vicario, J.L., Alves, P.A.V., Barbu, M., Podpora, M., Spagnolini, U., Pereira, M.J.V. and Vilanova, R., 2020. Educational data mining for tutoring support in higher education: a web-based tool case study in engineering degrees. *IEEE Access*, *8*, pp.212818-212836.

[2] Wu, K., 2022, October. Construction of ITIL Intelligent Platform for Management of Economics and Management Laboratory based on Circular Python Data Mining Algorithm. In *2022 International Conference on Edge Computing and Applications (ICECAA)* (pp. 225-228). IEEE.

[3] Li, Z., Zhao, Y., Botta, N., Ionescu, C. and Hu, X., 2020, November. COPOD: copula-based outlier detection. In *2020 IEEE international conference on data mining (ICDM)* (pp. 1118-1123). IEEE.

[4] Feng, G., Fan, M. and Chen, Y., 2022. Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access*, *10*, pp.19558-19571.

[5] Liu, Y., Xue, J. and Zhu, S., 2021, November. Knowledge map analysis of school-enterprise cooperation education in electronics and communications based on python. In *2021 2nd International Conference on Information Science and Education (ICISE-IE)* (pp. 1227-1231). IEEE.

[6] He, X., Xu, L., Zhang, X., Hao, R., Feng, Y. and Xu, B., 2021, May. Pyart: Python api recommendation in real-time. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)* (pp. 1634-1645). IEEE.

[7] Leng, S., Lin, J.R., Hu, Z.Z. and Shen, X., 2020. A hybrid data mining method for tunnel engineering based on real-time monitoring data from tunnel boring machines. *Ieee Access*, *8*, pp.90430-90449.

[8] Wang, Y., 2023, January. Research on Python Crawler Search System Based on Computer Big Data. In *2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA)* (pp. 1179-1183). IEEE.

[9] Ortigoza, G.M. and Castillo, W.A., 2023, October. Some Useful Data Science Techniques in Python to Analyze and Improve City Resilience Indexes: Case Study Mexico. In *2023 IEEE International Conference on Engineering Veracruz (ICEV)* (pp. 1-5). IEEE.

[10] Yeshchenko, A., Di Ciccio, C., Mendling, J. and Polyvyanyy, A., 2021. Visual drift detection for event sequence data of business processes. *IEEE Transactions on Visualization and Computer Graphics*, *28*(8), pp.3050-3068.

[11] Fernandez-Basso, C., Ruiz, M.D. and Martin-Bautista, M.J., 2020. A fuzzy mining approach for energy efficiency in a Big Data framework. *IEEE Transactions on Fuzzy Systems*, *28*(11), pp.2747-2758.