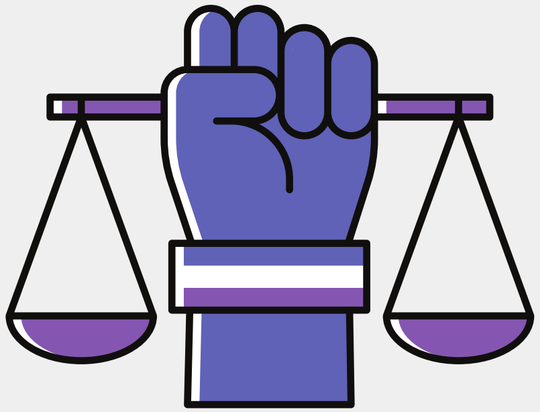
PREDICTING GENDER GAP INDEX SCORES USING ML AND DL MODELS



Bacha Saketh- 10722546005
S.Jayavardhan-107222546017
T.Shravya-107222546042
V.Jai kumar-107222546046

Overview

Introduction

The Global Gender Gap Index is a framework that measures the gender gap between men and women in four areas:

Health: Access to health care and survival

Education: Access to education and attainment

Economy: Access to economic opportunities and participation

Politics: Access to political empowerment

Abstract

Purpose:

The main aim is to track the relative gaps between women and men in health,education, the economy and politics it progress towards gender equality.The dataset contains the last 5 years data of Gender gap index of 146 countries .

Objective:

Predict the Score of the countries by using Machine learning and Deep learning models.

Methodology:

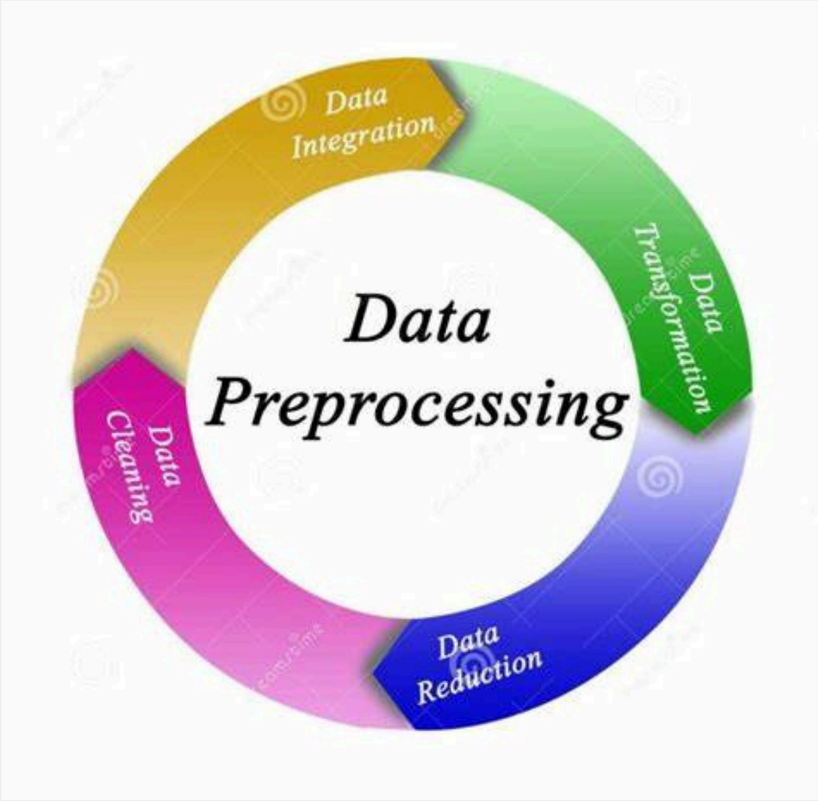
- Utilized Random forest ,XGBoost models to predict the score of the countries using label encoders .
- Splitting the dataset into training and testing sets.

Review of Literature

S.NO	AUTHOR'S	TITLE	MODELS USED
1	Radha R. Sharma, Sonam Chawla, and Charlotte M. Karam	Global Gender Gap Index: World Economic Forum perspective	ratio-based model, aggregate indexing model
2	Faustine Perrin	On the Construction of a Historical Gender Gap Index	regression analysis
3	Kavya Gupta; Ritu Rani; Arun Sharma; Poonam Bansal; Amita Dev; Rashmi Gandhi	Predicting gender gap index using machine learning	Linear Regression, Ridge Regression, and K-Nearest Neighbors
4	Rachel Forshaw, Vsevolod Iakovlev, Mark E. Schaffer & Cristina Tealdi	Using Machine Learning Methods to Estimate the Gender Wage Gap	stacking regression and Double-Debiased Machine Learning
5	Werner Kristjanpoller, Kevin Michell & Josephine E. Olson	Determining the gender wage gap through causal inference and machine learning models: evidence from Chile	causal inference and Metalearners,

Data Pre Processing

In data pre processing the data which includes data cleaning ,data transformation data reduction etc.Checking the null values .which we are replacing missing values (e.g, threshold 50%) and filling missing values using median and mode.



Data:

The dataset contains the last 5 years of gender gap index data,which has 695 rows and 32 columns.

Source:

https://in.docworkspace.com/d/sILyGioCIAvn1ur8G

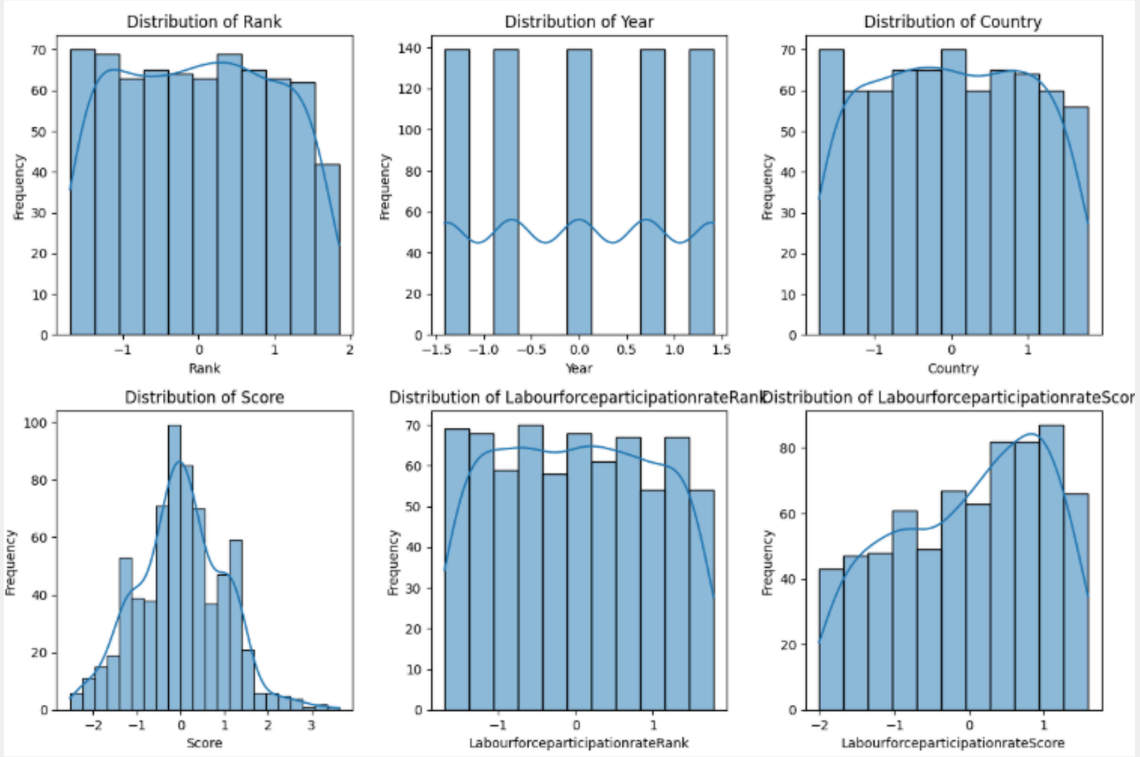
Dataset:

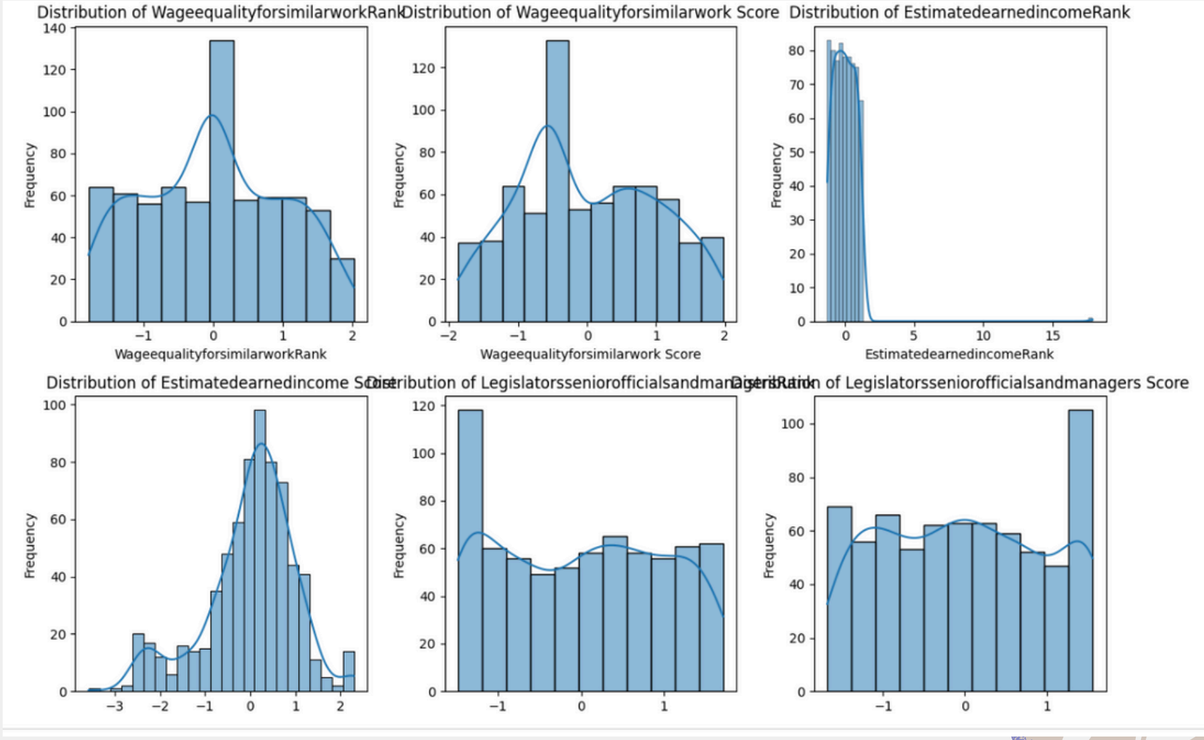
Rank	Year	Country	Score	Labourfo	Labourfo	Wageeq	Wageeq	Estimate	Estimate	Legislat	Legislat	Professi	Professi	Literacy	Literacy	Enrolme	Enrolme	Enrolme	Enrolme	Enrolme	Enrolm
1	2020	Iceland	0.877	73	0.825	131	0.502	73	0.621	34	0.628	1	1	51	0.999	117	0.984	1	1	1	1
2	2020	Norway	0.842	16	0.949	19	0.745	12	0.791	54	0.553	1	1	1	1	75	0.999	1	1	1	1
3	2020	Finland	0.832	13	0.959	9	0.798	33	0.72	77	0.467	1	1	1	1	1	1	1	1	1	1
4	2020	Sweden	0.82	14	0.955	50	0.694	15	0.769	35	0.628	1	1	1	1	103	0.994	106	0.997	1	1
5	2020	Nicaragua	0.804	120	0.627	113	0.56	14	0.774	57	0.543	1	1	1	1	1	1	1	1	1	1
6	2020	New Zealand	0.799	46	0.89	38	0.71	77	0.612	26	0.665	1	1	1	1	1	1	1	1	1	1
7	2020	Ireland	0.798	65	0.845	56	0.686	56	0.659	51	0.563	1	1	1	1	98	0.996	1	1	1	1
8	2020	Spain	0.795	54	0.872	115	0.558	55	0.661	73	0.473	74	0.978	69	0.99	1	1	1	1	1	1
9	2020	Rwanda	0.791	1	1	13	0.763	80	0.611	135	0.164	115	0.632	111	0.895	1	1	1	1	118	0.807
10	2020	Germany	0.787	38	0.898	68	0.671	41	0.695	89	0.416	1	1	1	1	1	1	138	0.878	1	1
11	2020	Latvia	0.785	20	0.934	34	0.717	31	0.721	11	0.814	1	1	1	1	1	1	1	1	1	1
12	2020	Namibia	0.784	55	0.871	69	0.662	13	0.787	15	0.772	1	1	61	0.998	1	1	1	1	1	1
13	2020	Costa Rica	0.782	118	0.637	111	0.573	105	0.551	66	0.513	96	0.862	1	1	77	0.999	1	1	1	1
14	2020	Denmark	0.782	19	0.935	52	0.693	32	0.721	102	0.364	1	1	1	1	1	1	1	1	1	1
15	2020	France	0.781	45	0.891	127	0.528	47	0.679	59	0.526	1	1	1	1	1	1	1	1	1	1
16	2020	Philippines	0.781	121	0.626	5	0.812	58	0.658	1	1	1	1	1	1	83	0.998	1	1	1	1
17	2020	South Africa	0.78	82	0.809	121	0.537	81	0.611	84	0.438	1	1	77	0.986	106	0.993	1	1	1	1
18	2020	Switzerland	0.779	37	0.899	40	0.707	37	0.706	58	0.526	85	0.934	1	1	86	0.997	119	0.969	1	1
19	2020	Canada	0.772	29	0.917	49	0.695	46	0.68	55	0.551	1	1	1	1	NA	NA	1	1	1	1
20	2020	Albania	0.769	94	0.747	3	0.823	49	0.67	76	0.469	1	1	65	0.992	1	1	1	1	1	1
21	2020	United Kingdom	0.767	49	0.886	76	0.642	102	0.562	47	0.569	71	0.99	1	1	88	0.997	1	1	1	1
22	2020	Colombia	0.758	97	0.742	122	0.535	44	0.684	1	1	1	1	1	1	1	1	1	1	1	1
23	2020	Moldova	0.757	0.903	39	22	0.749	23	0.68	1	1	63	0.995	87	0.997	110	0.991	1	1	1	1
25	2020	Mexico	0.754	128	0.57	129	0.503	127	0.463	48	0.565	76	0.973	79	0.983	1	1	1	1	1	1
26	2020	Estonia	0.751	32	0.913	36	0.713	92	0.582	50	0.564	1	1	1	1	1	1	1	1	1	1
27	2020	Belgium	0.75	56	0.871	80	0.638	59	0.657	64	0.513	1	1	1	1	1	1	1	1	1	1
28	2020	Barbados	0.749	21	0.933	72	0.655	39	0.701	8	0.963	1	1	1	1	127	0.981	1	1	1	1
29	2020	Belarus	0.746	23	0.929	NA	NA	68	0.626	10	0.897	1	1	52	0.999	85	0.997	1	1	1	1

Exploratory Data Analysis (EDA)

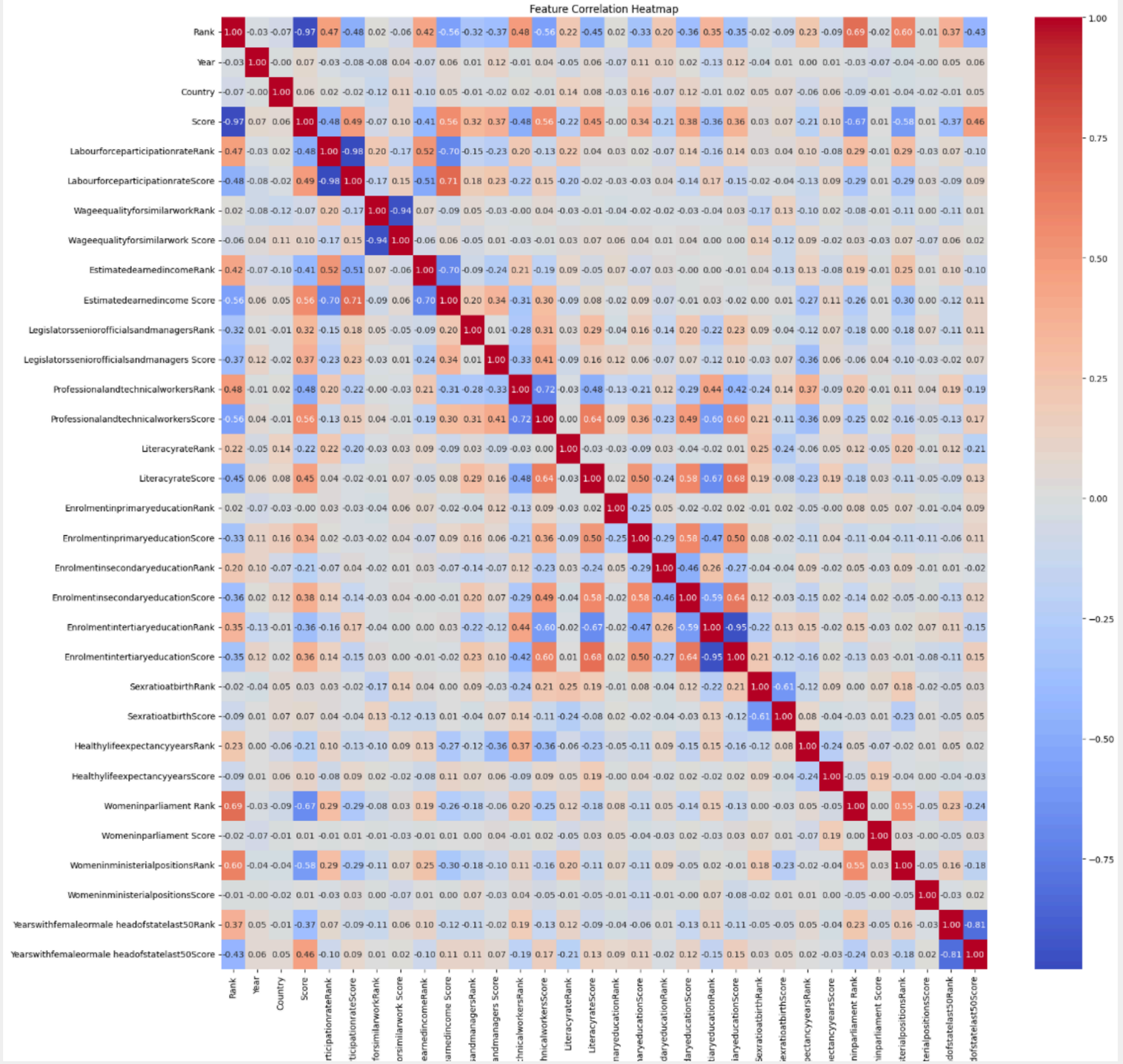
Body

Histogram

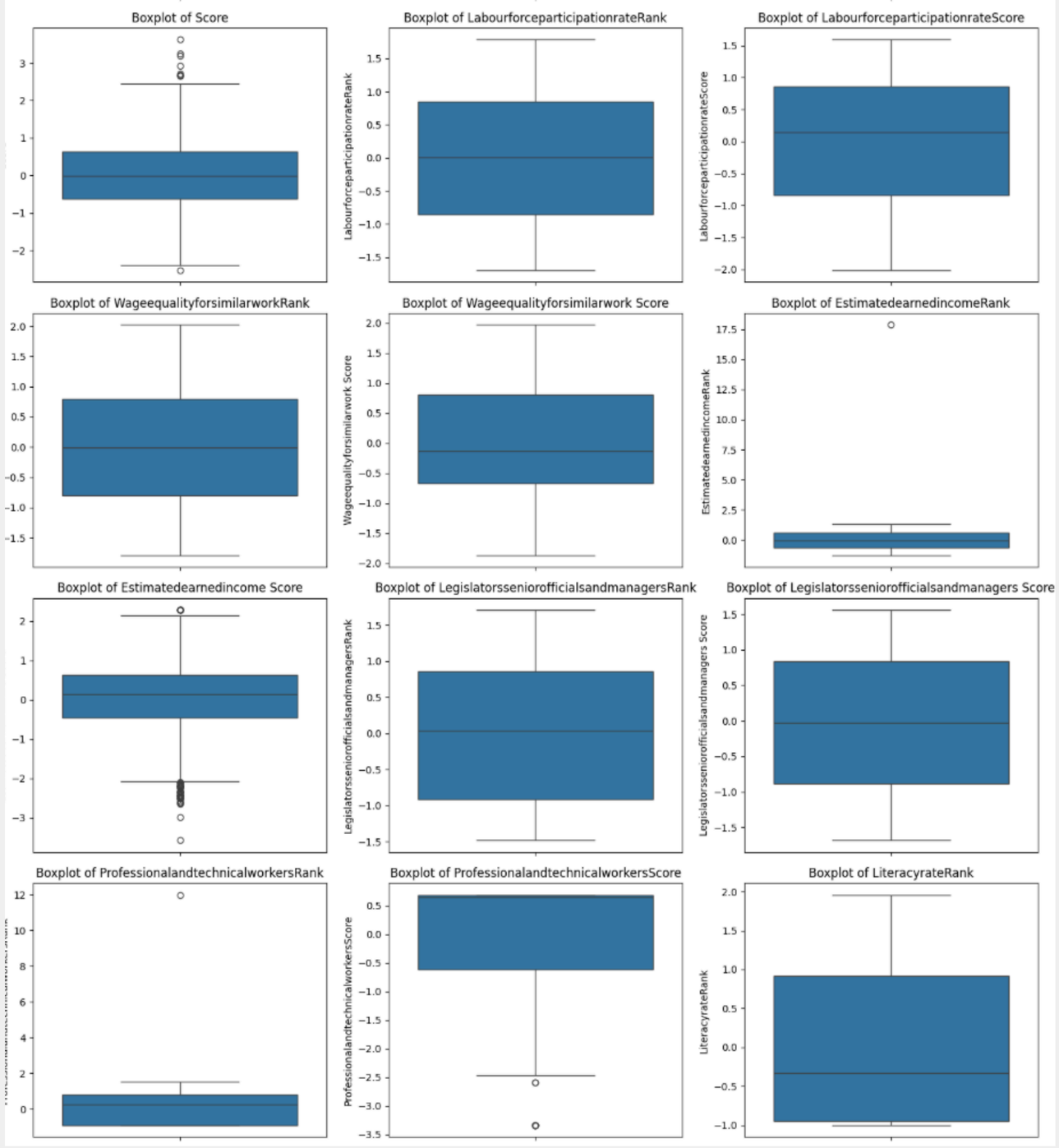

Distribution of Rank
Rank
Frequency
Distribution of Year
Year
Frequency
Distribution of Country
Country
Frequency
Distribution of Score
Score
Frequency
Distribution of LabourforceparticipationrateRank
LabourforceparticipationrateRank
Frequency
Distribution of LabourforceparticipationrateScor
LabourforceparticipationrateScore
Frequency


Distribution of WageequalityforsimilarworkRank
WageequalityforsimilarworkRank
Frequency
Distribution of Wageequalityforsimilarwork Score
Wageequalityforsimilarwork Score
Frequency
Distribution of EstimatedearnedincomeRank
EstimatedearnedincomeRank
Frequency
Distribution of Estimatedearnedincome Score
Frequency
Distribution of Legislatorsseniorofficialsandmanagers Rank
Frequency
Distribution of Legislatorsseniorofficialsandmanagers Score
Frequency

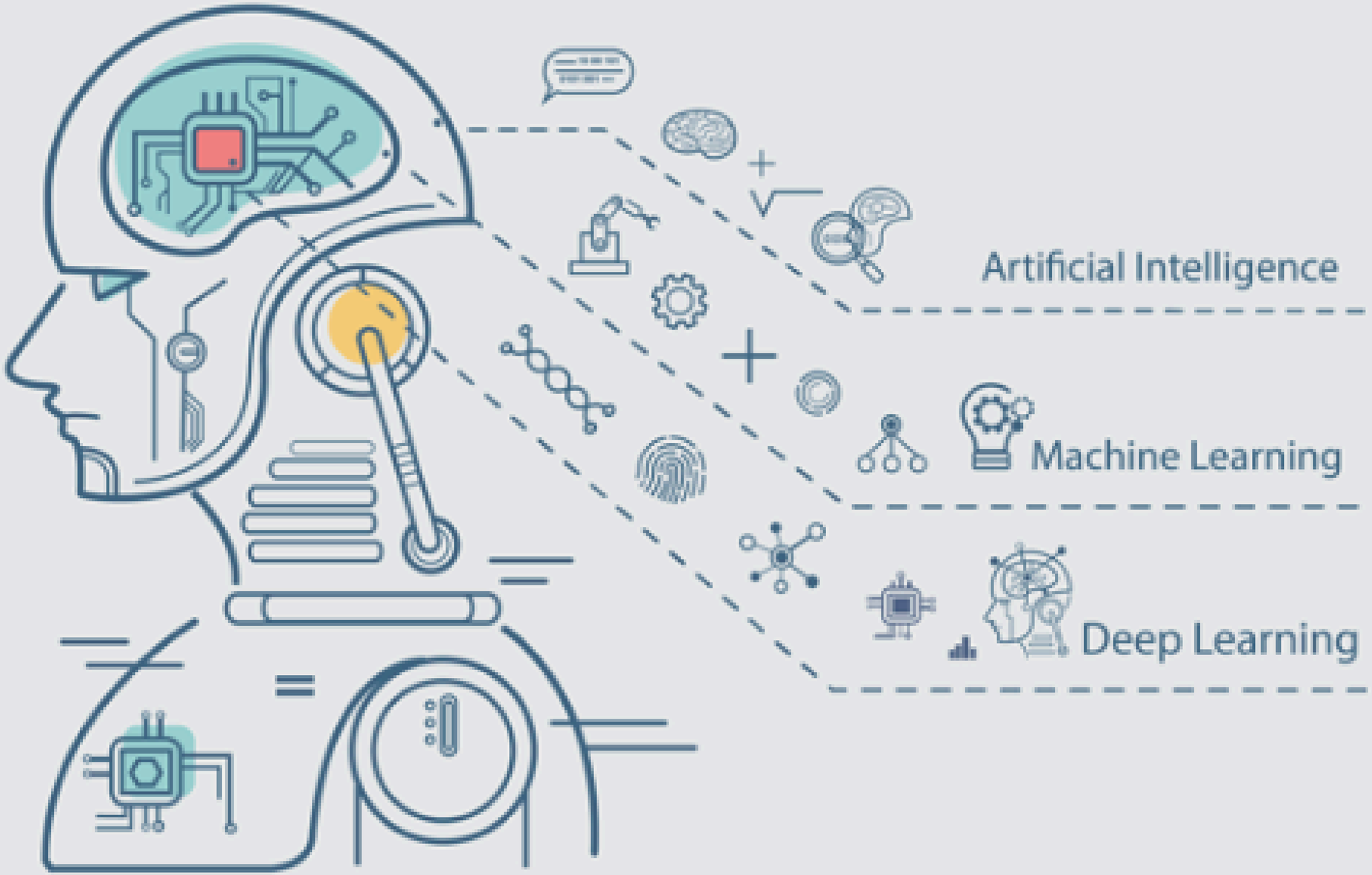
Correlation Matrix



Boxplot


Boxplot of Score
Boxplot of LabourforceparticipationrateRank
LabourforceparticipationrateRank
Boxplot of LabourforceparticipationrateScore
LabourforceparticipationrateScore
Boxplot of WageequalityforsimilarworkRank
Boxplot of Wageequalityforsimilarwork Score
Wageequalityforsimilarwork Score
Boxplot of EstimatedearnedincomeRank
EstimatedearnedincomeRank
Boxplot of Estimatedearnedincome Score
Boxplot of LegislatorsseniorofficialsandmanagersRank
LegislatorsseniorofficialsandmanagersRank
Boxplot of Legislatorsseniorofficialsandmanagers Score
Legislatorsseniorofficialsandmanagers Score
Boxplot of ProfessionalandtechnicalworkersRank
Boxplot of ProfessionalandtechnicalworkersScore
ProfessionalandtechnicalworkersScore
Boxplot of LiteracyrateRank
LiteracyrateRank

ML and DL models



80-20 Train Test Split

Models	R2 Score	MAE
Random Forest	0.8890	0.0437
XGBoost	0.9102	0.0353
ANN	0.7104	0.0532
CNN	0.8183	0.0448

75-25 Train Test Split

Models	R2 Score	MAE
Random Forest	0.9183	0.0360
XGBoost	0.9405	0.0262
ANN	0.7939	0.0400
CNN	0.8555	0.0380

70-30 Train Test Split

Models	R2 Score	MAE
Random Forest	0.9535	0.0249
XGBoost	0.9820	0.0097
ANN	0.8606	0.0330
CNN	0.8883	0.0280

65-35 Train Test Split

Models	R2 Score	MAE
Random Forest	0.9603	0.0244
XGBoost	0.9752	0.0152
ANN	0.7604	0.0312
CNN	0.8488	0.0280

60-40 Train Test Split

Models	R2 Score	MAE
Random Forest	0.9544	0.0262
XGBoost	0.9728	0.0152
ANN	0.8282	0.0298
CNN	0.8674	0.0278

Best Split 70-30 Train Test Split

Models	R2 Score	MAE
Random Forest	0.9535	0.0249
XGBoost	0.9820	0.0097
ANN	0.8606	0.0280
CNN	0.8883	0.0270

Forecasting:

Country	Year	Actual Score	Predicted Score	Rank
Japan	2020	0.652	0.648	121-124
India	2020	0.668	0.665	112-116
Spain	2021	0.788	0.791	12-15
Costa Rica	2021	0.786	0.784	15-17
Iceland	2022	0.908	0.910	1-2

Country	Year	Actual Score	Predicted Score	Rank
Norway	2022	0.845	0.847	2-4
USA	2023	0.748	0.748	43
Finland	2023	0.863	0.863	3
Rwanda	2024	0.757	0.756	37-40
Nambia	2024	0.805	0.804	8-10

Future Prediction:

Country	Year	Predicted Score	Rank
Iceland	2025	0.892	1-3
USA	2025	0.724	57-62
Norway	2025	0.849	3-7
India	2025	0.625	130-135
Japan	2025	0.756	105-115

Summary:

As we can see the best split is 70-30 Train Test Split and the best algorithm is XGBoost algorithm.The main aim is to predict the score of the countries.The Accuracy for XGBoost is 98%.Apart from predicting the score of the countries which already has its Actual score ,we are preddicting the Score of countries in future ,based on the past score of the countries we can predict the future score of countries by predicting the score we can determine the Rank of the country.

Future Scope:

- At the current rate, it will take 134 years to reach full gender parity globally
- Real-time data from worldwide reports and surveys can be integrated into future research to improve the predicted accuracy and applicability of the model.

Insights

- The global **gender gap** score in 2024 is 68.5%, meaning 31.5% of the gap remains unaddressed

- Progress has been extremely slow, with only a 0.1% point improvement from 2023

Reference

1.References: Sharma, R. R., Chawla, S., & Karam, C. M. (2021). Global gender gap index: world economic forum perspective. In Handbook on diversity and inclusion indices (pp. 150-163). Edward Elgar Publishing..

2.Reference: Perrin, F. (2014). On the construction of a historical gender gap index. An Implementation on French Data, 05-14.

3.Reference: Gupta, K., Rani, R., Sharma, A., Bansal, P., Dev, A., & Gandhi, R. (2023, January). Predicting Gender Development Index using Machine Learning. In 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 530-535). IEEE.

4.Reference: Forshaw, R., Iakovlev, V., Schaffer, M. E., & Tealdi, C. Using Machine Learning Methods to Estimate the Gender Wage Gap Check for updates. Machine Learning for Econometrics and Related Topics, 109.

5.Reference: Kristjanpoller, W., Michell, K., & Olson, J. E. (2023). Determining the gender wage gap through causal inference and machine learning models: evidence from Chile. Neural Computing and Applications, 35(13), 9841-9863.

Thank you

S.Jayavardhan

V.Jai kumar

T.Shravya

B.Saketh

Colab Notebook