

HOLISTIC ANALYSIS OF UNEMPLOYMENT IN INDIA

Ankit Saha
Vellore Institute of Technology
School of Information & Technology
Department of Computer Application
Vellore, India
ankit.saha2022@vitstudent.ac.in

Ishita Maurya
Vellore Institute of Technology
School of Information & Technology
Department of Computer Application
Vellore, India
ishita.maurya2022@vitstudent.ac.in

Tushar Mulwani
Vellore Institute of Technology
School of Information & Technology
Department of Computer Application
Vellore, India
tushar.mulwani2022@vitstudent.ac.in

Arpit Shourya
Vellore Institute of Technology
School of Information & Technology
Department of Computer Application
Vellore, India
arpit.shourya2022@vitstudent.ac.in

Dr. Harshita Patel
Sr. Assistant Professor
Vellore Institute of Technology
School of Information & Technology
Department of Computer Application
Vellore, India
harshita.patel@vit.ac.in

Deepak Singh
Vellore Institute of Technology
School of Information & Technology
Department of Computer Application
Vellore, India
deepak.singh2022@vitstudent.ac.in

1) ABSTRACT:

For many nations, including India, unemployment is still a major issue. Along with slowing economic growth, it also has an impact on social welfare and people's quality of life. In this project, we aim to perform a holistic analysis of unemployment in India by leveraging data mining and machine learning techniques. The objective is to predict the unemployment rate accurately using various algorithms, such as linear regression, multiple linear regression, support vector regression, k-nearest neighbors (KNN), random forest and voting regressor. Through this study, we find that the random forest and votingregressor algorithm provides the highest accuracy among the tested techniques. This project illustrates the value of using data-driven methodologies to gain understanding of the factors causing unemployment, ultimately assisting stakeholders and policymakers in developing successful strategies for lowering unemployment rates and promoting economic prosperity in India.

Keywords: Unemployment Analysis; Holistic Approach; Predictive Modelling; Accuracy Comparison; Economic Growth; Simple Linear Regression; Multiple Regression; Support Vector Regression; K-Nearest Neighbor; Random Forest; Voting Regressor.

2) INTRODUCTION:

The ongoing and complicated problem of unemployment presents serious difficulties for the Indian economy and for society at large. In addition to slowing economic growth, it has an effect on people's livelihoods and general well-being. Recognizing the gravity of this problem, we embark on a project titled "Holistic Analysis of Unemployment in India," aimed at understanding and addressing this pressing issue. This project has two goals: first, it wants to understand what causes unemployment in India, and second, it wants to properly anticipate unemployment using data mining and machine learning methods. To accomplish our goals, we make use of a wide range of attributes that are extremely important in determining the unemployment situation in India. These attributes include cities, regions, real GDP, NSDP, estimated unemployment rate, estimated employed, estimated labour participation rate, date, latitude and longitude. These attributes are taken into account in an effort to depict the complex nature of unemployment and the various social and geographic factors that affect it. In our study, we use a variety of methods, including simple linear regression, multiple linear regression, support vector regression, k-nearest neighbours (KNN), random forest and voting regressor. These methods allow us to extract patterns, connections, and forecast insights from big datasets, enabling a more complex understanding of the unemployment problem in India.

We seek to produce actionable insights that guide policy choices and promote change by investigating numerous attributes and using a range of approaches. With this all-encompassing strategy, we intend to help in the effort to lower unemployment and create a bright future for the Indian worker.

3) LITERATURE RIVEW:

3.1) " Advanced statistics: linear regression, part II: multiple linear regression":

Year of publication: 2003

Authors: Keith A Marill

This paper delves into the topic of advanced statistics, specifically focusing on multiple linear regression. The author provides an in-depth exploration of this statistical technique, building upon the foundations of linear regression to analyze multiple predictors and their impact on the dependent variable. The paper offers valuable insights and practical applications of multiple linear regression in various fields.

3.2) " Support Vector Regression"

Year of publication: 2007

Authors: Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis

This paper introduces support vector regression, a machine learning technique that combines support vector machines with regression analysis. The authors provide an overview of its concepts, mathematical formulation, and applications in regression problems. The paper emphasizes the strengths and advantages of support vector regression for prediction and modeling tasks.

3.3) "A Data Mining Approach to Construct Graduates Employability Model in Malaysia":

Year of publication: 2011

Authors: Myzatul Akmam Sapaat, Aida Mustapha, Johanna Ahmad, Khadijah Chamili, Rahamirzam Muhamad

This study aims to forecast whether a graduate would find employment, be unemployed, or be in an uncertain situation six months after graduation. In order to do so, they draw information for 2009 from the Tracer Study, a web- based survey system run by the Ministry of Higher Education of Malaysia (MOHE). The classification of a graduate profile as employed, jobless, or other has been accomplished through a series of classification tests utilizing several algorithms under Bayes and decision approaches.

3.4) " Forecasting the Unemployment Rate by Neural Networks Using Search Engine Query Data":

Year of publication: 2012

Authors: Wei Xu, Ziang Li and Qing Chen

In this paper, a novel neural network-based forecasting method is proposed for the unemployment rate prediction using search engine query data. The empirical results show that the proposed method outperforms other forecasting methods, which have been used for the unemployment rate prediction. These findings imply that web information, especially web search behavior, can improve the efficiency and effectiveness of the unemployment rate prediction.

3.5) "Data mining for unemployment rate prediction using search engine query data":

Year of publication: 2012

Authors: Wei Xu, Ziang Li, Cheng Cheng and Tingting Zheng

This paper proposes a data mining framework using search engine query data for unemployment rate prediction. The framework includes neural networks (NNs) and support vector regressions (SVRs) for unemployment rate prediction. The framework includes neural networks (NNs) and support vector regressions (SVRs) for forecasting unemployment trends. It involves extracting query data, feature selection, modelling with NNs and SVRs, optimizing parameters using genetic algorithm, selecting the best predictor via cross-validation, and forecasting unemployment trends with the chosen method.

3.6) " A Study on Multiple Linear Regression Analysis"

Year of publication: 2013

Authors: Nese Guler and Gulden Kaya Uyan1k

This paper's main goal is to fully explain the concept of multiple linear regression by using data from university students to predict test results.

3.7) "Labor Market Forecasting by Using Data Mining":

Year of publication: 2013

Authors: Yas A. Alsultanny

In this study, several data mining classification approaches, including Naive Bayes Classifiers, Decision Trees, and Decision Rules procedures, are used to analyse and forecast the labour force requirement from the databases of human resources. The Decision Tree technique achieved the highest accuracy when the three techniques were compared.

3.8) " An Ontology-based Web Mining Method for Unemployment Rate Prediction"

Year of publication: 2014

Authors: Ziang Li, Wei Xu, Likuan Zhang and Raymond Y.K. Lau

The proposed framework is underpinned by a domain ontology which captures unemployment related concepts and their semantic relationships to facilitate the extraction of useful prediction features from relevant search engine queries. In addition, state-of-the-art feature selection methods and data mining models such as neural networks and support vector regressions are exploited to enhance the effectiveness of unemployment rate prediction.

3.9) "Analyzing the Performance of MGNREGA Scheme using Data Mining Technique":

Year of publication: 2015

Authors: G. Sugapriyan and S. Prakasam

The success of MGNREGA in Kanchipuram District is analyzed using data mining techniques and comparisons to past government statistics. This work aims to assess the effectiveness of the program and its performance in two ways: through data mining and by comparing historical statistical data.

3.10) "A Review on Predicting Student's Performance using Data Mining Techniques "

Year of publication: 2015

Authors: Amirah Mohamed Shahiria, Wahidah Husaina, Nur'aini Abdul Rashid

This paper applies data mining methods to Malaysian student data to predict performance, benefiting teachers and students seeking improvement in teaching and learning methods. By considering attributes like external/internal assessments, extracurricular activities, and soft skills, valuable predictions are made for forecasting student outcomes.

3.11) "Students' Employability Prediction Model through Data Mining "

Year of publication: 2016

Authors: Tripti Mishra, Dharminder Kumar and Sangeeta Gupta

This paper uses a sample of 1400 Master of Computer Applications (MCA) students from various colleges in India to predict the employability of MCA students using a variety of classification techniques of data mining, such as Bayesian methods, Multilayer Perceptrons and Sequential Minimal Optimization (SMO), Ensemble Methods and Decision Trees, and finds the algorithm that is most suitable for this problem.

3.12) " Forecasting Unemployment Rates using Machine Learning Techniques":

Year of publication: 2017

Authors: V. Vijayalakshmi and R. Ravi

The authors' study reveals that both Artificial Neural Networks (ANN) and Support Vector Machine (SVM) models effectively forecasted unemployment rates. The ANN model displayed slightly higher accuracy compared to the SVR model. Notably, both models surpassed traditional statistical approaches like ARIMA and ES models.

3.13) " Using Machine Learning to Advance Personality Assessment and Theory":

Year of publication: 2018

Authors: Wiebke Bleidorn and Christopher James Hopwood

Machine learning has transformed several fields, including psychology science, by providing powerful predictions of human behaviour and personality traits. While previous machine learning approaches have mostly focused on integrating social media and digital records to recognised personality measures, this study presents a more thorough construct validation methodology. Researchers can improve the potential of personality assessment and get deeper insights into personality by implementing machine learning in this framework. The paper examines recent applications of machine learning in personality assessment, emphasises the relevance of construct validation principles, and offers ideas for harnessing machine learning to further our understanding of personality.

3.14) " Predicting the Unemployment Rate Using Social Media Analysis ":

Year of publication: 2018

Authors: Pum-Mo Ryu

This paper introduces a method for predicting the unemployment rate using social media analysis and statistical modeling. It collects and analyzes Korean social media content, including news articles, blogs, and tweets, using natural language processing techniques. The study fits ARIMAX and ARX models to the analyzed data, considering social moods expressed in the content, and highlights the method's advantage over existing approaches in capturing social tendencies.

3.15) " Linear regression analysis study":

Year of publication: 2018

Authors: Suniti Yadav, Khushbu Kumari

This article demonstrates that Regression expresses the relationship as an equation, whereas Correlation assesses the strength of the linear link between two variables.

3.16) "Case Study of UPNM Students Performance Classification Algorithm":

Year of publication: 2018

Authors: Sayarifah B.Rahayu, Nur D.Kamarudin and Zuraini Zainol

In this paper, authors analyze student performance using three classification algorithms (Naive Bayes, J48 Decision tree, and k-nearest neighbor). Among them, Naive Bayes shows the highest accuracy, aiding students and lecturers in improving instructional methods amidst lecturers' heavy teaching load and administrative duties.

3.17) " A Comparative Analysis of Machine Learning Techniques for Unemployment Rate Prediction":

Year of publication: 2019

Authors: B. K. Nayak and R. K. Behera

The authors utilize US and Indian monthly data on unemployment rates, inflation rates, and GDP to train and evaluate their models. Four machine learning algorithms—decision trees, random forests, support vector regression, and artificial neural networks—are compared. The results show that all four techniques effectively predict unemployment rates, with random forests performing best for US data and support vector regression for India data. Incorporating additional economic variables, such as GDP and inflation rates, improves predictive accuracy. Comparisons with traditional statistical models reveal the superiority of machine learning methods, except for ARIMA in short-term predictions.

3.18) " Machine Learning for Survival Analysis: A Survey"

Year of publication: 2019

Authors: Ping Wang, Yan Li, Chandan K. Reddy

This research paper applies survival analysis techniques to study the duration of unemployment spells. The author analyzes longitudinal data on individuals' unemployment spells, considering various demographic, socioeconomic, and labor market factors. Survival analysis models, such as Cox proportional hazards models, are used to examine the determinants of unemployment duration and estimate the probabilities of finding employment over time. The study provides valuable insights into the factors that influence the duration of unemployment and can inform policies aimed at reducing unemployment spells and supporting individuals in their job search efforts.

3.19) " The impact of artificial intelligence on employment before and during pandemic: A comparative analysis":

Year of publication: 2021

Authors: G Abuselidze, Lela Mamaladze

This comparative analysis explores the impact of technological advancements on employment across industries and regions. It examines the adoption of automation, artificial intelligence, and robotics and analyzes the resulting changes in employment patterns, considering factors like job displacement, skill requirements, and productivity gains. The study provides valuable insights into the complex relationship between technology and employment, aiding policymakers and workforce planners in making informed decisions.

3.20) " Prediction of Unemployment Rates in Turkey by k-Nearest Neighbor Regression Analysis":

Year of publication: 2021

Authors: Ş. Hatipoğlu, M. A. Belgrat, A. Degirmenci and Ö. Karal

The authors employ data on unemployment rates, GDP, inflation rates, and economic variables in Turkey from 2000 to 2017 for model training and testing. They utilize k-NN regression analysis to develop predictive models for Turkish unemployment rates. The findings indicate that k-NN regression models effectively forecast unemployment rates in Turkey, with enhanced accuracy when including GDP and inflation rates. In a comparison with linear regression, polynomial regression, and support vector regression, the k-NN regression models generally exhibit superior performance.

3.21) " Predicting Future Unemployment Rates Using Time Series Analysis and Machine Learning Techniques":

Year of publication: 2021

Authors: Christos Katris

The authors utilize US and Indian data from 1948 to 2014 and 1983 to 2014, respectively, to evaluate their models. Regression and neural network models exhibit effectiveness in predicting future unemployment rates, with the neural network model displaying slightly higher accuracy. The study reveals that GDP, inflation, and interest rates significantly influence unemployment rates, and incorporating these indicators enhances predictive accuracy in the models.

3.22) " Unemployment Rate Forecasting using Supervised Machine Learning Model"

Year of publication: 2022

Authors: Sidhari Manasa, M.Kalidas

Using a supervised machine learning model, this project forecasts unemployment for the next month or series. SVM, Random Forest, Gradient Boosting, and Extreme Machine Learning algorithms used in this project to estimate state unemployment rates. This application extracts all data from the selected state and then builds a training model using the aforementioned machine learning algorithms, which can subsequently be used to predict unemployment.

3.23) " Analysis of longitudinal causal relationships between gender role attitudes and labor market participation of young women in Korea":

Year of publication: 2022

Authors: Hanryeo Lim and Sungpyo Hong

This study looked at the association between gender role attitudes and labour market involvement among young women in Korea. The findings demonstrated that women's labor-force participation had a positive impact on establishing equal gender role attitudes, which maintained over time. However, having an equitable gender role attitude had no effect on eventual labour market participation. Additionally, childbearing was found to lower women's labor-force involvement, implying career disruptions. These findings highlight the importance of active employment strategies in assisting women's entry into the labour market and preventing career interruptions.

3.24) " Nowcasting Unemployment Using Neural Networks and Multi-Dimensional Google Trends Data":

Year of publication: 2023

Authors: Andrius Grybauskas, Vaida Pilinkiene, Mantas Lukauskas, Alina Stundžiene and Jurgita Bruneckiene

This article aimed to enhance predictive power by incorporating various dimensions of Google Trends into the analysis of initial jobless claims in the United States. The study successfully expanded the scope beyond traditional approaches by considering keywords related to job search, mental health, violence, leisure, consumption, and disasters. The utilization of keyword optimization, dimension reduction techniques, and long-short memory neural networks led to improved forecasting accuracy. However, it was noted that the relationship between jobless claims and specific Google keywords exhibited temporal instability. These findings underscore the potential of artificial intelligence and Google Trends for nowcasting unemployment trends, while emphasizing the need for ongoing exploration and refinement in this field.

3.25) " Unemployment Prediction using Machine Learning Techniques: A Comparative Study"

Authors: M. R. Islam, M. R. Khan and M. A. Rahman

Authors used monthly unemployment data (1983-2015) for Bangladesh, a high-unemployment developing country. They compared machine learning models (Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression, Gradient Boosting Regression, Support Vector Regression) based on MSE and MAE. Results: Random Forest Regression, Gradient Boosting Regression outperformed others with lower MSE/MAE. Including GDP, inflation, education improved predictive accuracy. Model performance varied based on training/testing periods. finds the algorithm that is most suitable for this problem.

4) ANALYSIS:

In our analysis, we will conduct exploratory data analysis (EDA) on three different datasets to gain insights into the unemployment situation. The first dataset is the national dataset, which provides a broader perspective on unemployment trends and patterns across the entire country. The second dataset focuses specifically on unemployment in Delhi, a major metropolitan area in India.

By analyzing this dataset, we can understand the unique dynamics of unemployment in the city, considering factors such as local industries, population density, and urban development policies. This localized analysis provides a more detailed understanding of the unemployment situation in Delhi and helps identify specific challenges and opportunities for employment growth in the region.

The third dataset is sourced from Kaggle and provides additional insights into unemployment in India.

By integrating this dataset with the national and Delhi datasets, we can gain a comprehensive understanding of the broader trends in India's unemployment landscape and how they relate to specific regional dynamics.

Through this three-fold analysis, we aim to identify commonalities, differences, and correlations among the datasets. We can uncover patterns, outliers, and potential causal relationships between various factors and unemployment rates. This integrated approach allows us to paint a more holistic picture of the unemployment situation, enabling policymakers, researchers, and stakeholders to make informed decisions and implement targeted interventions to address unemployment challenges effectively.

4.1) **National Dataset:**

4.1.1) ***DATASET DESCRIPTION:***

```
[6] df.sample(5)
```

	Region	Date	Estimated Unemployment Rate (%)	Estimated Employed	Estimated Labour Participation Rate (%)
782	Jammu & Kashmir	31-12-2016	21.34	3865037	51.24
788	Jammu & Kashmir	30-06-2017	11.86	3503726	40.94
338	Chhattisgarh	31-10-2022	0.90	9407168	40.13
1582	Punjab	31-03-2021	7.35	9590215	40.73
987	Karnataka	31-10-2019	5.83	21602775	41.68

Fig 1. Sample of National Dataset

Our dataset contains 2314 unique rows and 7 columns. The columns namely [Region, Date, Frequency, Estimated Unemployment Rate (%), Estimated Employed, Estimated Labor Participation Rate (%)].

4.1.2) *DESCRIPTION OF COLUMNS:*

- Region: Contains the name of state.
- Date: Contains date when record is entered.
- Frequency: Contains the frequency accordingly which data was recorded.
- Estimated Unemployment Rate (%): Contains percentage of rate of unemployment recorded on particular date.
- Estimated Employed: Contains percentage of people estimated to gain employment on the particular date.
- Estimated Labor Participation Rate (%): Contains percentage of people estimated to participate for labor on the particular date.

The column 'Frequency' contained only one unique value which is 'M' which denotes that the data is recorded on Monthly basis. Hence, we dropped the column to avoid overfitting of model.

The column 'Date' contained data in the format dd-mm-yyyy. We transformed the data in the form of yyyy to get date year between data on yearly basis instead of monthly.

Code for transformation of dates:

```
j=0
for i in df[' Date']:
    df[' Date'][j]=i[7:11]
    j=j+1
```

Sample of data:

	Region	Year	Estimated Unemployment Rate (%)	Estimated Employed	Estimated Labour Participation Rate (%)
607	Haryana	2016	6.67	8100227	42.41
803	Jammu & Kashmir	2018	13.21	3551156	40.82

4.1.3) *PREPROCESSING:*

4.1.3.1) Outlier Removal:

For the preprocessing we have tried to remove the outliers from the data from the independent variable.

Outliers removal in Estimated Unemployment Rate (%):

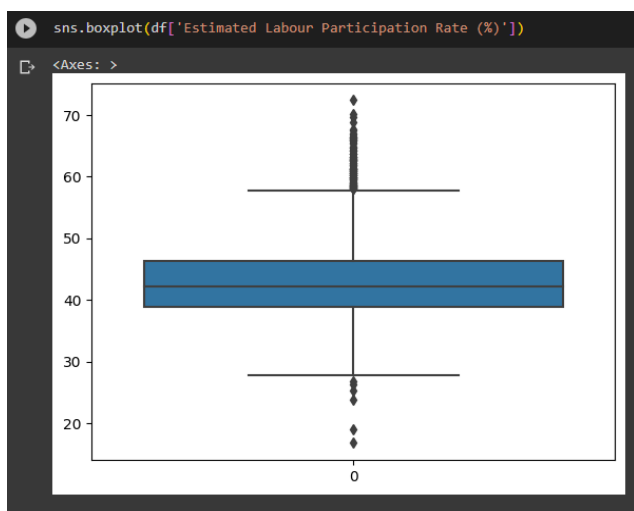


Fig.2: Outliers in Estimated Unemployment Rate (%) before removal

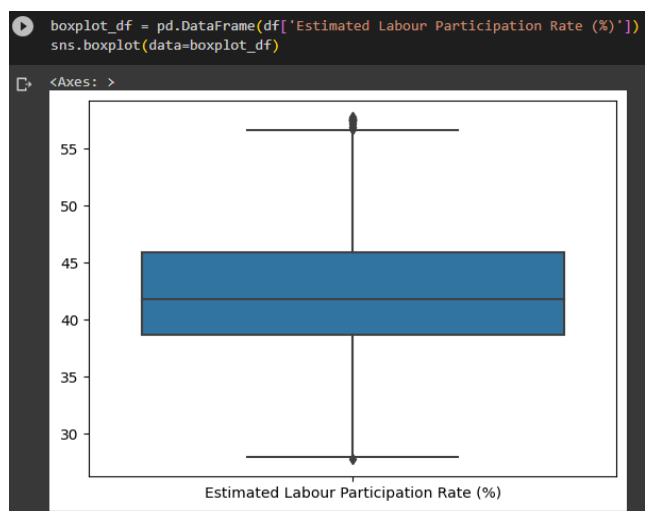


Fig.3: Outliers in Estimated Unemployment Rate (%) after removal

In Fig.2, we can see there are multiple outliers present in the column named Estimated Unemployment Rate (%). In order to remove the outliers from the columns we executed the following code:


```
[ ] Q1 = df['Estimated Labour Participation Rate (%)'].quantile(0.25)
    Q3 = df['Estimated Labour Participation Rate (%)'].quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5*IQR
    upper = Q3 + 1.5*IQR

# Create arrays of Boolean values indicating the outlier rows
upper_array = np.where(df['Estimated Labour Participation Rate (%)']>=upper)[0]
lower_array = np.where(df['Estimated Labour Participation Rate (%)']<=lower)[0]

# Removing the outliers
df = df[~df.index.isin(upper_array)]
df = df[~df.index.isin(lower_array)]
df = df.reset_index(drop=True)
```

In the code, we are calculating 1st and 3rd quantile of the column and getting the interquantile range. Then we are calculating the lower and upper bounds of the distribution upto which the distribution of data is allowed. Then we created two arrays namely upper and lower which contain all the values which are outliers. Then we are locating the index and removing those rows numbers mapped with outlier's index. The result of the outlier removal can be seen in **Fig.3**.

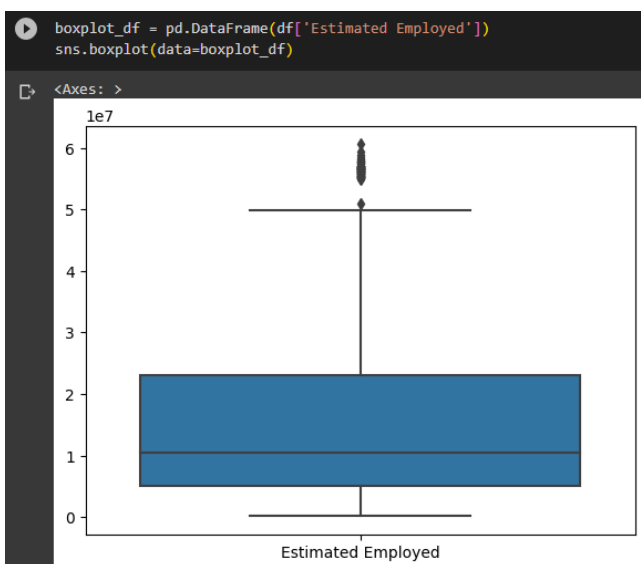


Fig.4: Outliers in Estimated employed before removal

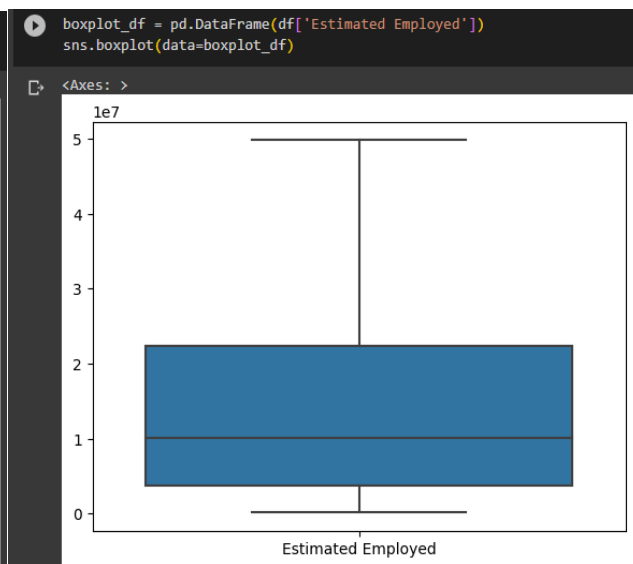


Fig.5: Outliers in Estimated employed after removal

In **Fig.4**, we can see there are multiple outliers present in the column named Estimated Employed. In order to remove the outliers from the columns we executed the following code:

```
Q1 = df['Estimated Employed'].quantile(0.25)
Q3 = df['Estimated Employed'].quantile(0.75)
IQR = Q3 - Q1
lower = Q1 - 1.5*IQR
upper = Q3 + 1.5*IQR

# Create arrays of Boolean values indicating the outlier rows
upper_array = np.where(df['Estimated Employed']>=upper)[0]
lower_array = np.where(df['Estimated Employed']<=lower)[0]

# Removing the outliers
df = df[~df.index.isin(upper_array)]
df = df[~df.index.isin(lower_array)]
df = df.reset_index(drop=True)
```

In the code, we are calculating 1st and 3rd quantile of the column and getting the interquantile range. Then we are calculating the lower and upper bounds of the distribution upto which the distribution of data is allowed. Then we created two arrays namely upper and lower which contain all the values which are outliers. Then we are locating the index and removing those rows numbers mapped with outlier's index. The result of the outlier removal can be seen in **Fig.5**.

4.1.3.2) Scaling:

Scaling is a method used to normalize the range of independent variables or features of data. Our dataset contains such columns which contains high values. The machine learning models assign weights to the independent variables according to their data points and conclusions for output. In that case, if the difference between the data points is high, the model will need to provide more significant weight to the farther points, and in the final results, the model with a large weight value assigned to undeserving features is often unstable. This means the model can produce poor results or can perform poorly during learning. Hence we need to scale the data.

For the purpose of scaling we have used min max scaler.

The formula for min max scaling is:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

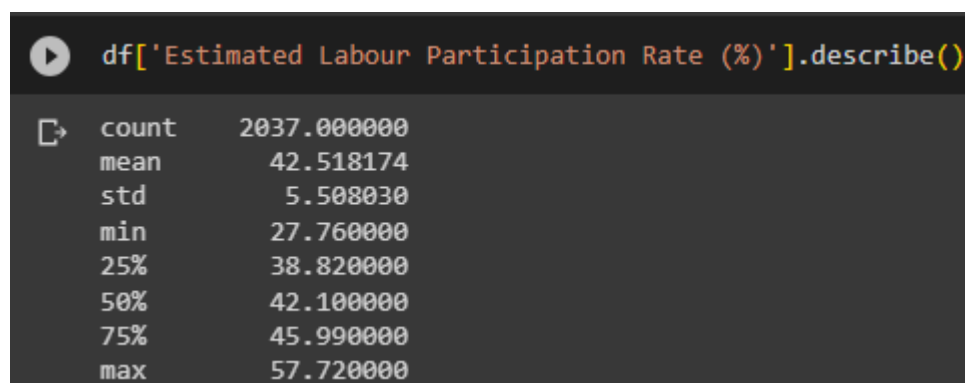


Fig.6: Estimated Labour Participation Rate (%) before scaling.

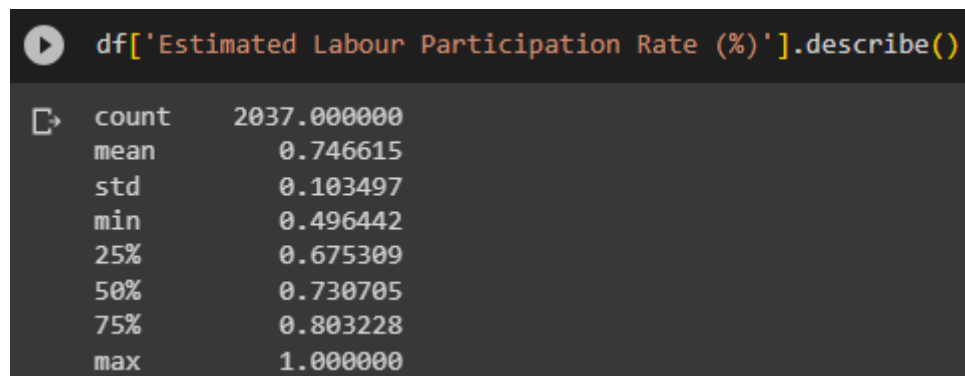
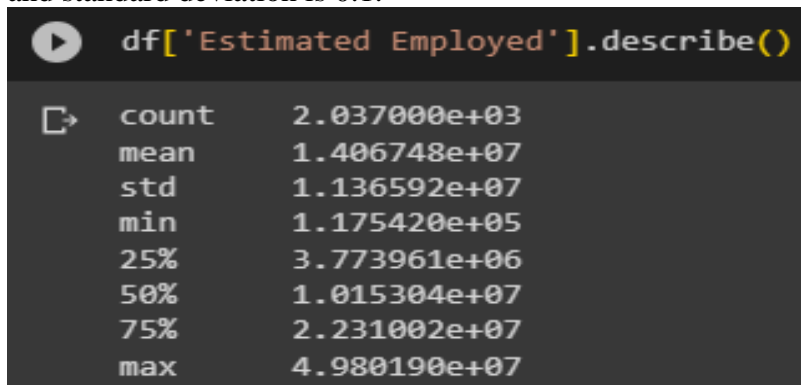


Fig.7: Estimated Labour Participation Rate (%) after scaling.

In fig.6 we can see the value of the column titled Estimated Labour Participation Rate (%) lies between 57.72 to 27.76. The mean of data is 42.5 with standard deviation of 5.5. To scale this data, we executed the following code:

```
[ ] i=0
while i!=2037:
    df['Estimated Labour Participation Rate (%)'][i]=(df['Estimated Labour Participation Rate (%)'][i]-
                                                    df['Estimated Labour Participation Rate (%)'].min())/
                                                    (df['Estimated Labour Participation Rate (%)'].max()-
                                                    df['Estimated Labour Participation Rate (%)'].min())
    i=i+1
```

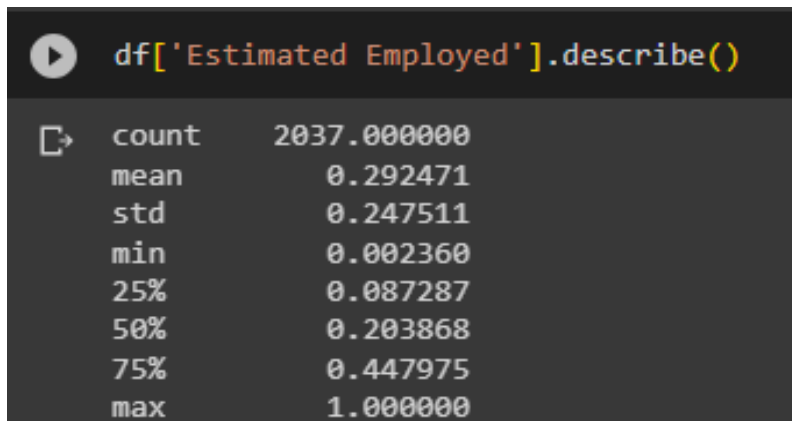
In fig 7 we can see the same column but the data values are now in range between 0.49 to 1.00. The mean is 0.7 and standard deviation is 0.1.



```
df['Estimated Employed'].describe()
```

count	2.037000e+03
mean	1.406748e+07
std	1.136592e+07
min	1.175420e+05
25%	3.773961e+06
50%	1.015304e+07
75%	2.231002e+07
max	4.980190e+07

Fig.8: Estimated Employed before scaling.

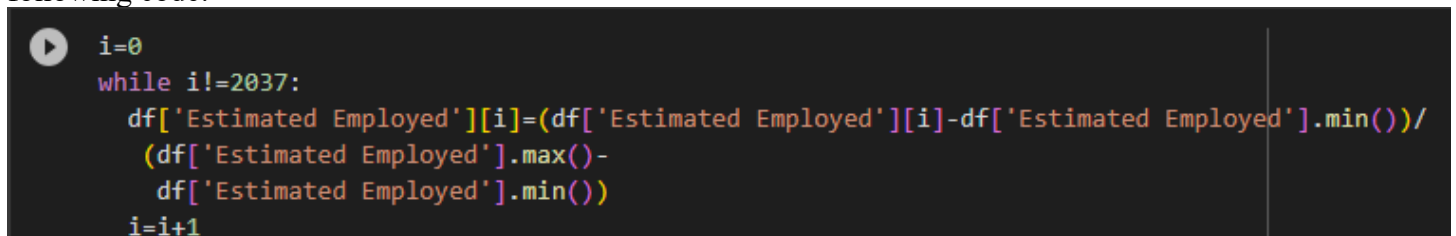


```
df['Estimated Employed'].describe()
```

count	2037.000000
mean	0.292471
std	0.247511
min	0.002360
25%	0.087287
50%	0.203868
75%	0.447975
max	1.000000

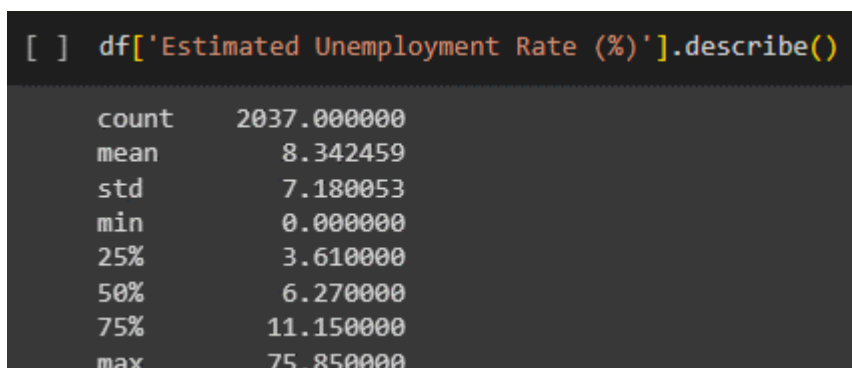
Fig.9: Estimated Employed after scaling.

In fig.8 we can see the value of the column titled Estimated Employed lies between 1.17×10^5 to 2.23×10^7 . The mean of data is 1.40×10^7 with standard deviation of 1.13×10^7 . To scale this data, we executed the following code:



```
i=0
while i!=2037:
    df['Estimated Employed'][i]=(df['Estimated Employed'][i]-df['Estimated Employed'].min())/
    (df['Estimated Employed'].max()-
     df['Estimated Employed'].min())
    i=i+1
```

In fig 9 we can see the same column but the data values are now in range between 0.002 to 1.00. The mean is 0.29 and standard deviation is 0.24.



```
[ ] df['Estimated Unemployment Rate (%)'].describe()
```

count	2037.000000
mean	8.342459
std	7.180053
min	0.000000
25%	3.610000
50%	6.270000
75%	11.150000
max	75.850000

Fig.10: Estimated Unemployment Rate (%) before scaling.

```
[ ] df['Estimated Unemployment Rate (%)'].describe()

count    2037.000000
mean      0.151551
std       0.160438
min       0.000000
25%      0.052999
50%      0.103494
75%      0.186233
max       1.000000
```

Fig.11: Estimated Unemployment Rate (%) after scaling.

In fig.10 we can see the value of the column titled Estimated Unemployment Rate (%) lies between 0.00 to 75.85. The mean of data is 8.34, with standard deviation of 7.18. To scale this data, we executed the following code:

```
[ ] i=0
while i!=2037:
    df['Estimated Unemployment Rate (%)'][i]=(df['Estimated Unemployment Rate (%)'][i]-
                                                df['Estimated Unemployment Rate (%)'].min())
                                                /(df['Estimated Unemployment Rate (%)'].max()
                                                -df['Estimated Unemployment Rate (%)'].min())

    i=i+1
```

In fig 11 we can see the same column but the data values are now in range between 0.00 to 1.00. The mean is 0.15 and standard deviation is 0.16.

4.1.3.3) One-hot encoding:

For the categorical column 'Region' and 'Date (renamed to Year)' are made to go through one-hot encoding in which, for every unique value in the column a new column will form.

Code for one hot encoding:

```
[ ] #One-hot encoding
df = pd.get_dummies(df, columns = ['Region', 'Year'])
```

Sample of data after one hot encoding:

Region_Delhi	Region_Goa	Region_Gujarat	...	Region_Uttarakhand	Region_West_Bengal	Year_2016	Year_2017	Year_2018	Year_2019	Year_2020	Year_2021	Year_2022	Year_2023
0	0	0	...	0	0	0	1	0	0	0	0	0	0
0	0	0	...	0	0	0	1	0	0	0	0	0	0
0	0	0	...	0	0	0	0	0	0	1	0	0	0
0	0	0	...	0	0	0	0	0	0	1	0	0	0
0	0	0	...	0	0	0	0	0	0	1	0	0	0

After the whole pre-processing steps the numbers of unique tuples are: 2037 and numbers of columns are: 38.

4.1.4) *Model Training and Evaluation:*

For the training of models, we have split the data into 80:20 ratios between test and train.

Our target variable is “Estimated Unemployment Rate (%)”, rest of the columns are deciding variables.

```
[ ] X = df.drop('Estimated Unemployment Rate (%)', axis=1)
    y = df['Estimated Unemployment Rate (%)']
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

We have trained three basic existing models (Linear regression, Support Vector Regression and KNN) and also trained an ensemble voting classifier of the two best performing model.

1. Linear Regression:

Initialization and fitting of model:

```
[ ] Linear_reg_model=LinearRegression()
    Linear_reg_model.fit(X,y)
```

```
▼ LinearRegression
LinearRegression()
```

10-fold cross validation test for mean squared error:

```
[ ] mse=cross_val_score(Linear_reg_model,X_train,y_train, scoring='neg_mean_squared_error', cv=10)

[ ] np.mean(mse)

-1.774035423103319e+18
```

Prediction of test data using model:

```
[ ] y_pred=Linear_reg_model.predict(X_test)
```

Calculating R2 Score:

```
[ ] score_lr=r2_score(y_pred,y_test)

[ ] score_lr

0.5594141463728004
```

2. KNN Regression:

Initialization, fitting and prediction of model:

```
▶ knn_model = KNeighborsRegressor(n_neighbors=5)
  knn_model.fit(X_train, y_train)
  y_pred = knn_model.predict(X_test)
```

10-fold cross validation for mean squared error and calculation of R2 score

```
[ ] mse=cross_val_score(knn_model,X_train,y_train, scoring='neg_mean_squared_error', cv=10)

[ ] np.mean(mse)

-0.008087346880392502

[ ] score_knn=r2_score(y_pred,y_test)

[ ] score_knn

0.7871491538780709
```

3. Support Vector Regression:

Initialization, fitting and prediction of model:

```
[ ] svr_model = SVR(kernel='rbf')
svr_model.fit(X_train, y_train)
y_pred = svr_model.predict(X_test)
```

10-fold cross validation for mean squared error and calculation of R2 score:

```
[ ] mse=cross_val_score(svr_model,X_train,y_train, scoring='neg_mean_squared_error', cv=10)

[ ] np.mean(mse)

-0.009668312889950632

[ ] score_svr=r2_score(y_pred,y_test)

[ ] score_svr

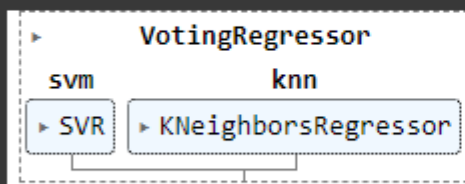
0.7007257265909861
```

4. Ensemble Voting Regressor:

Initializing the voting classifier with top two performing model, fitting model and predicting the test data.

```
[ ] voting_reg=VotingRegressor(estimators=[('svm',svr_model),('knn',knn_model)])

[ ] voting_reg.fit(X_train,y_train)
```



```
[ ] y_pred = voting_reg.predict(X_test)
```

10-fold cross validation for mean squared error and calculation of R2 score

```
[ ] mse=cross_val_score(voting_reg,X_train,y_train, scoring='neg_mean_squared_error', cv=10)

np.mean(mse)

-0.008164999163301039

[ ] score_voting=r2_score(y_pred,y_test)

[ ] score_voting

0.7717436942891198
```

Evaluation table:

MODEL/METRIC	R ² Score	MSE
Linear Regression	0.55	-1.77
K Nearest Neighbour	0.78	-0.008
Support Vector Regressor	0.70	-0.009
Voting Regressor	0.77	-0.008

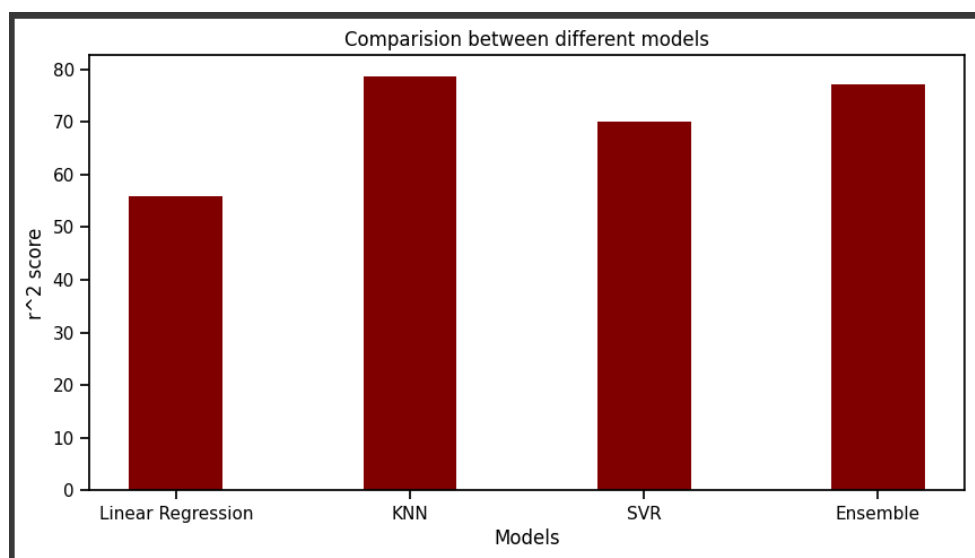


Fig.12: Comparison between different models.

Link to Colab Notebook:

https://colab.research.google.com/drive/1OUywBVbD2p0w_amyJBjC763sh3Ef2BeL?usp=sharing

4.2) Delhi Dataset:

4.2.1) DATASET DESCRIPTION:

	Estimated Unemployment Rate (%)	Estimated Employed	Estimated Labour Participation Rate (%)	NSDP	real gdp	Date of estimation	Month of estimation
0	10.61	5482637	44.00	270261.0	520159.0	31	1
1	10.84	4829996	38.78	270261.0	520159.0	29	2
2	11.05	5513829	44.26	270261.0	520159.0	31	3
3	6.30	5954845	45.28	295558.0	579653.0	30	4
4	17.30	5523948	47.48	295558.0	579653.0	31	5

Fig.13: Sample of Delhi Dataset

Our dataset contains 86 unique rows and 6 columns. The columns namely [Date, Estimated Unemployment Rate (%), Estimated Employed, Estimated Labor Participation Rate (%), NSDP, real gdp].

4.2.2) DESCRIPTION OF COLUMNS:

- Date of estimation: Contains date when record is entered.
- Estimated Unemployment Rate (%): Contains percentage of rate of unemployment recorded on particular date.
- Estimated Employed: Contains percentage of people estimated to gain employment on the particular date.
- Estimated Labor Participation Rate (%): Contains percentage of people estimated to participate for labor on the particular date.
- Month of Estimation: Contains month when record is entered.

The column 'Real GDP' is one of crucial factor for predicting future unemployment. As in most of project Nominal GDP is utilized where as in our project Real GDP has more correlation than Nominal GDP.

4.2.3) PREPROCESSING:

4.2.3.1) Outlier Removal:

For more precise and accurate prediction we have identified outlier is **above 40 percent** in Estimated Unemployment Rate (%) and handled it by replacing it with mean value of Estimated Unemployment Rate (%).

For Visualizing outliers we have used histogram and Box Plot

Histogram

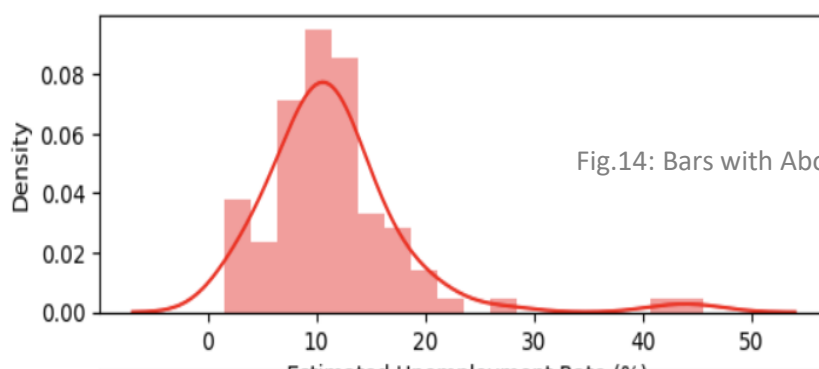


Fig.14: Bars with Above 30 are Outliers

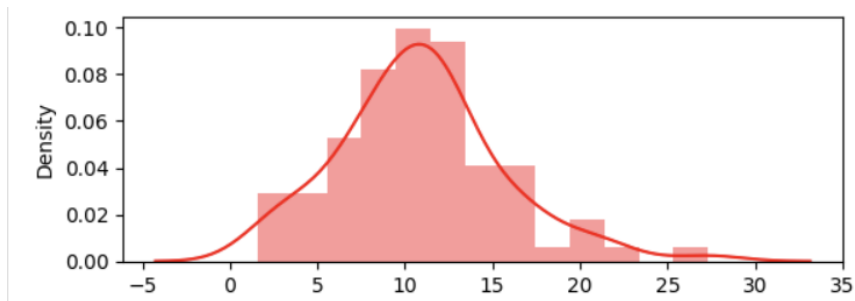


Fig.15: Outliers are adjusted by replacing values with 30

Box Plot:

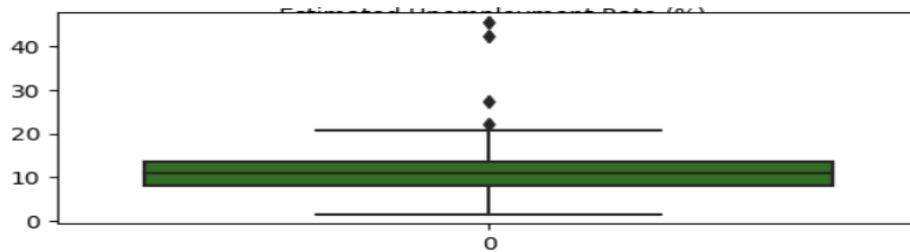


Fig.16: Above 30 is Outlier

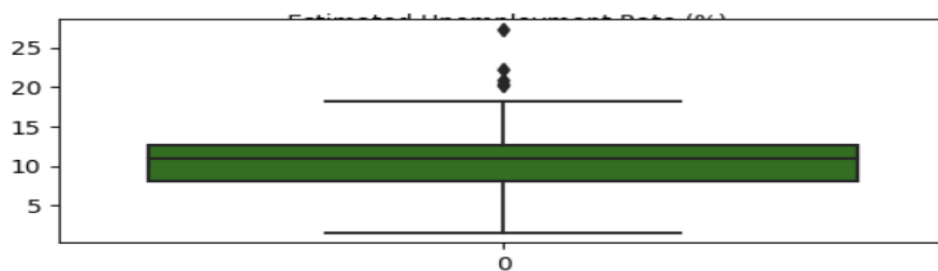


Fig.17: Adjusting the Outlier

```
[20] data1[' Estimated Unemployment Rate (%)'] = np.where(data1[' Estimated Unemployment Rate (%)'] > 30, data1[' Estimated Unemployment Rate (%)'].median(), data1[' Estimated Unemployment Rate (%)'])

plot(data1, ' Estimated Unemployment Rate (%)')
```

4.2.4) Feature Selection:

We have performed **Feature Selection** to enhance model training and prediction efficiency, reduce overfitting, improve generalization, and provide insights into the important features for decision-making

	Attribute	Score
0	Estimated Employed	1.401419
1	Estimated Labour Participation Rate (%)	3.905683
2	NSDP	0.985549
3	real gdp	0.590293
4	Date of estimation	0.446393
5	Month of estimation	0.433494
6	Year of estimation	6.722430

Fig.18: Selecting influential Attributes

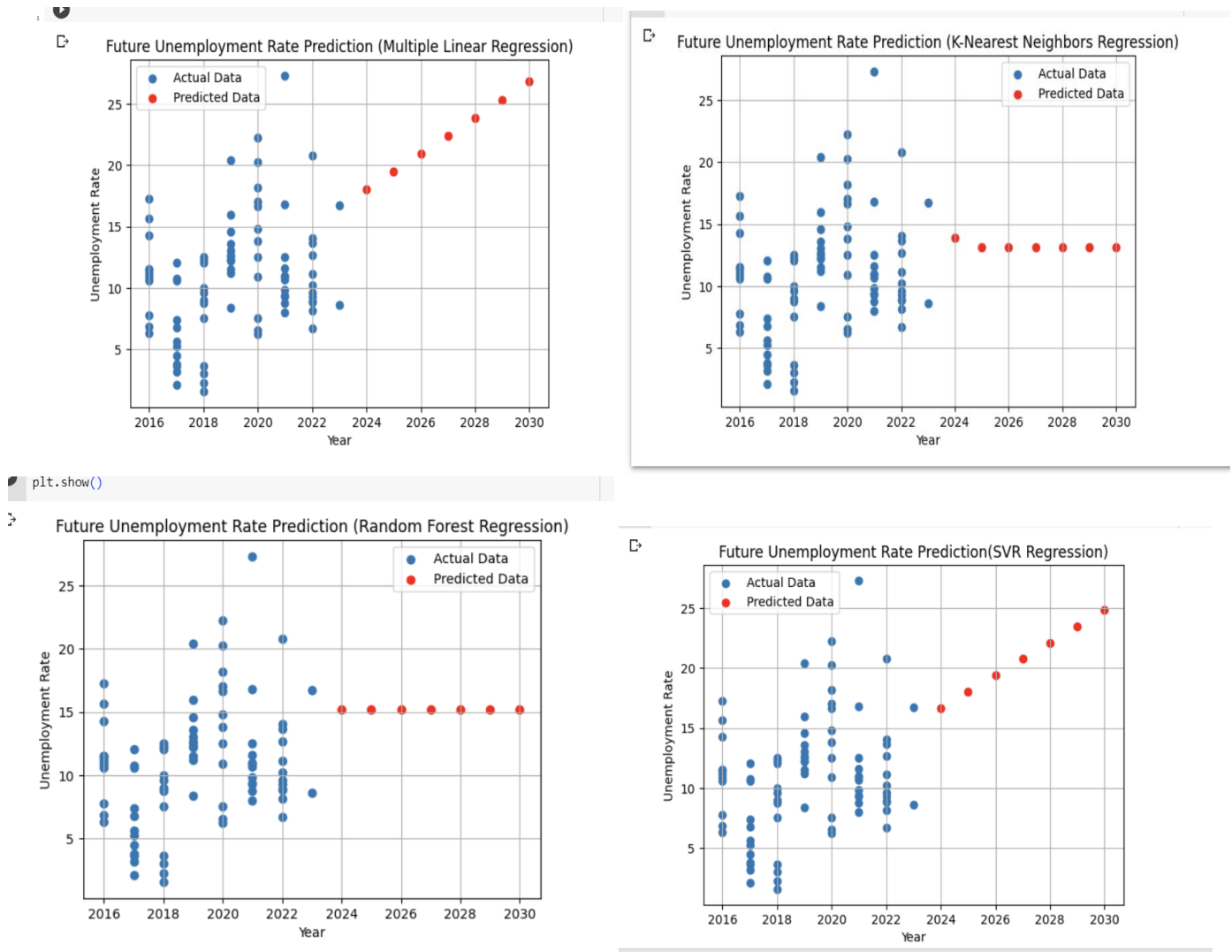
For Accurate Prediction We have taken 3 dependent variables which are: Year of estimation, Estimated Labour Participation Rate and Real GDP

4.2.5 Model Training and Evaluation:

For the training of models, we have split the data into 80:20 ratios between test and train.

We have trained five basic existing models (Linear regression, Multiple Linear regression, Support Vector Regression, Random Forest and KNN).Evaluation of models are done by R^2 score:

1. KNN Regression: 0.22273896314921648
2. Support vector Regression: 0.13837704341661905
3. Multiple Linear Regression: 0.236633751331482
4. Random Forest Regression: 0.30616481254230543



4.3) Kaggle India Dataset:

4.3.1) DATASET DESCRIPTION:

	Region	Date	Frequency	Estimated Unemployment Rate (%)	Estimated Employed	Estimated Labour Participation Rate (%)	Region.1	latitude	longitude	Level
0	Andhra Pradesh	31-01-2020	M	5.48	16635535	41.02	South	15.9129	79.74	0.6
1	Andhra Pradesh	29-02-2020	M	5.83	16545652	40.90	South	15.9129	79.74	0.6
2	Andhra Pradesh	31-03-2020	M	5.79	15881197	39.18	South	15.9129	79.74	0.6
3	Andhra Pradesh	30-04-2020	M	20.51	11336911	33.10	South	15.9129	79.74	0.8
4	Andhra Pradesh	31-05-2020	M	17.43	12988845	36.46	South	15.9129	79.74	0.7

Fig.19: Sample of Covid-19 Kaggle India Dataset

Our dataset contains 268 unique rows and 9 columns. The columns namely [Region, Date, Frequency, Estimated Unemployment Rate (%), Estimated Employed, Estimated Labor Participation Rate (%), Region.1, latitude, longitude, Level].

4.3.2) Visualization:

In the project's final analysis, a Kaggle unemployment dataset was used to evaluate the geographic distribution of unemployment rates in India. The project successfully highlighted the hotspot locations with various levels of unemployment by harnessing the power of data visualisation, notably through the usage of a folium heat map.

```
✓ [3] import folium
    map1 = folium.Map(location = [28.7041, 77.1025], zoom_start = 6)
    from folium.plugins import HeatMap
    import pandas as pd
    ab = pd.read_csv('india.csv')
```

Fig.20: Importing Libraries and setting HeatMap Locations

Code in Fig.20 imports the Folium library for creating interactive maps, creates a map centered at latitude 28.7041 and longitude 77.1025 with a zoom level of 6, and imports the HeatMap plugin from Folium. It also imports the Pandas library and reads a CSV file named 'india.csv' into a DataFrame variable named 'ab'.

```
✓ [11] da1 = ab[['latitude', 'longitude', 'Level']]
    da1.columns
    da1['latitude'] = da1['latitude'].astype(float)
    da1['longitude'] = da1['longitude'].astype(float)
    da1['Level'] = da1['Level'].astype(float)
```

Fig.21: Assigning and Converting Values

Code in Fig.21 selects three columns ('latitude', 'longitude', 'Level') from the DataFrame 'ab' and assigns them to a new DataFrame called 'da1'. It then converts the 'latitude', 'longitude', and 'Level' columns of 'da1' to float data type.

```
▶ mapdata = [[x[0], x[1], (x[2])] for x in da1]
    HeatMap(da1, gradient = {1.0: 'red', 0.8: 'orange', 0.7: 'yellow', 0.6: 'blue', 0.5: 'green'}).add_to(map1)
    map1.save("india_unemployment.html")
```

Fig.22: Plotting Points in Map

Code in Fig.22 creates a list of lists called 'mapdata' where each element contains the latitude, longitude, and level values extracted from the DataFrame 'da1'. It then creates a heatmap using the 'da1' DataFrame, specifying color gradients for different levels, and adds it to the 'map1' map. Finally, it saves the map as an HTML file named "india_unemployment.html".

The India country map visualisation used a colour gradient approach, with red dots denoting locations with highest unemployment rates (e.g. Delhi, Bihar, Tripura) and blue points reflecting places with the lowest unemployment rates (e.g. Assam, Meghalaya, Gujarat). This method allows for a quick and intuitive comprehension of the country's geographical inequalities in unemployment.

The latitude and longitude coordinates of each region were used to precisely map the unemployment levels. The heat map presented a comprehensive and visually attractive portrayal of the employment issues encountered by different regions in India by linking geographical information with corresponding unemployment rates.

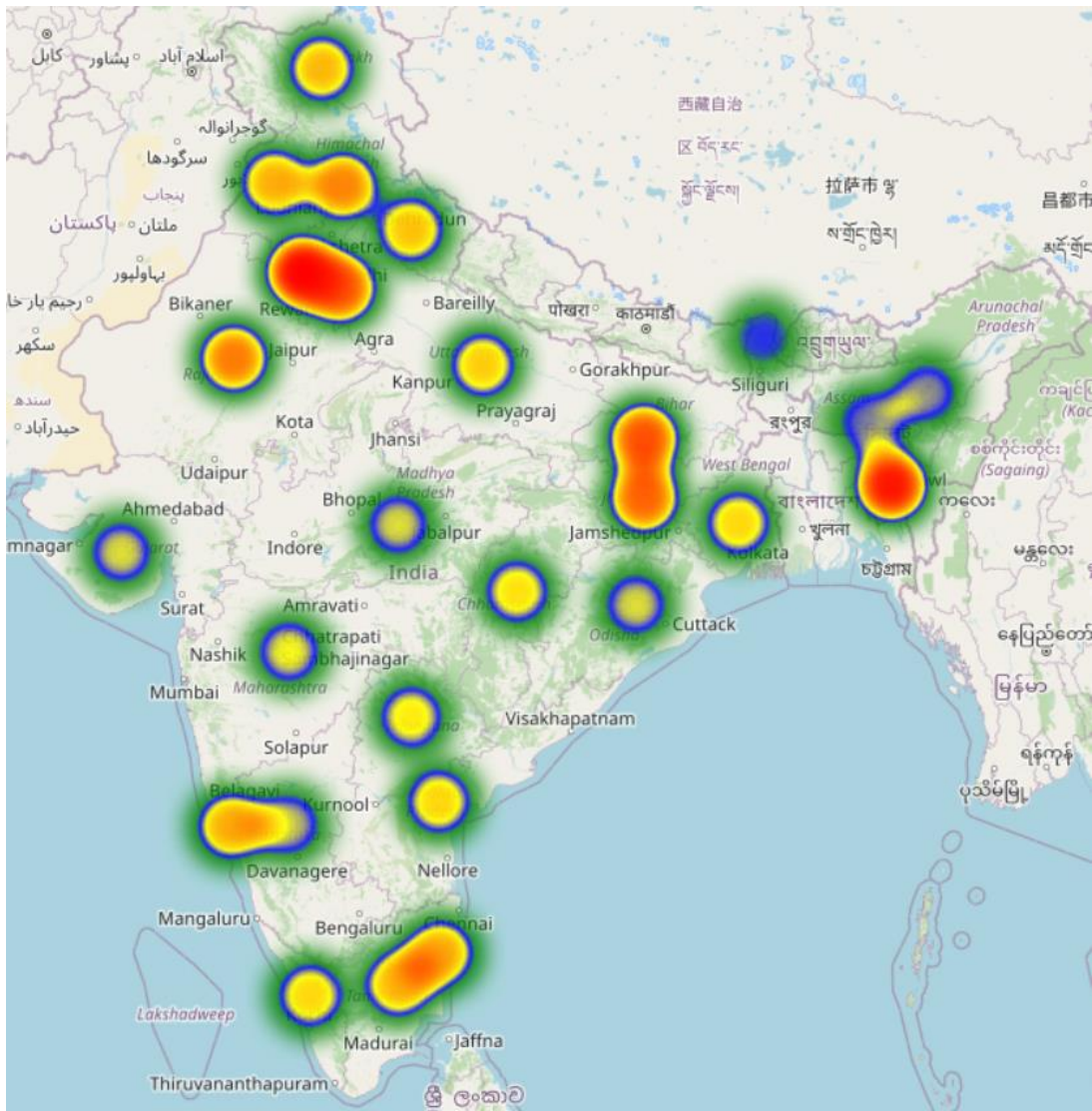


Fig 23. Hotspots Showing Unemployment

The use of the Kaggle unemployment dataset and the implementation of a folium heat map to visualise the variation in jobless rates across India adds an important layer of insight to this project. It highlights the importance of geographical context in employment research and provides a complete picture of the country's unemployment landscape.

Table of National Dataset

Model/Factors	R-squared	Mean absolute error	Mean squared error	Accuracy
Linear regression	0.559	32990381.849071693	1.774035423103319e+18	55.9%
Support vector regression	0.787	0.06330724265812257	0.009668312889950632	78.7%
KNN	0.700	0.04591719266633794	0.008087346880392502	70.0 %
VotingRegressor	0.7717	0.05116222844613375	0.008164999163301039	77.17%

Table of Delhi Dataset

Model/Factors	R-squared	Mean absolute error	Mean squared error	Accuracy
Simple linear regression	0.049	3.1626844166360883	18.186620120470533	4.9%
Multiple linear regression	0.237	-	14.592719993722628	23.7%
Support vector regression	0.138	2.8286406535272173	16.471022353314122	13.8%
Random forest	0.306	2.4665029411764654	13.263545028382325	30.6%
KNN	0.223	2.7609803921568634	14.858336601307192	22.3 %

5) Conclusion:

In this project, a three-fold analysis was conducted to understand employment patterns and make predictions using different datasets. The first analysis utilized a national dataset and employed various machine learning algorithms, including linear regression, support vector regression, KNN, and the voting regressor. The results revealed that the voting regressor outperformed the other models in terms of accuracy. Moving on to the second analysis, a dataset specific to Delhi was considered, with attributes such as NSD (National Skill Development) and real GDP. Machine learning algorithms like linear regression, multiple linear regression, support vector regression, KNN, and random forest were applied. Interestingly, the random forest model demonstrated superior performance compared to other algorithms. For the final analysis, an unemployment dataset from Kaggle was used to identify hotspot areas in India using a folium heat map. This visualization provided valuable insights into the regions facing significant unemployment challenges.

Comparing this three-fold analysis to previous studies, our project stands out in several ways. Firstly, we incorporated the use of the voting regressor, a powerful ensemble method, which effectively combined the strengths of multiple regression models. This approach yielded higher accuracy in predicting employment patterns. Additionally, the inclusion of unique attributes like NSD and real GDP in the analysis specific to Delhi provided a deeper understanding of the factors influencing employment in the region. The identification of the random forest model as the best-performing algorithm highlights its suitability for capturing complex relationships within the data. Lastly, our utilization of the Kaggle unemployment dataset and the visualization of hotspot areas using a heat map showcased the spatial distribution of unemployment across India, enhancing our understanding of regional disparities.

In conclusion, this three-fold analysis, comprising national data, Delhi-specific data, and the Kaggle unemployment dataset, along with the use of various machine learning algorithms and the voting regressor, has provided valuable insights into employment patterns and predictive capabilities. The uniqueness of our approach, including the incorporation of specific attributes, visualization techniques, and the superior performance of certain models, sets our study apart from existing research in the field.

6) References:

1. Keith A Marill. Advanced statistics: linear regression, part II: multiple linear regression. DOI: 10.1197/j.aem.2003.09.006.
2. Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. Support Vector Regression. Neural Information Processing – Letters and Reviews Vol. 11, No. 10, October 2007.
3. Myzatul Akmal Sapaat, Aida Mustapha, Johanna Ahmad, Khadijah Chamili, Rahamirzam Muhamad. A Data Mining Approach to Construct Graduates Employability Model in Malaysia. International Journal on New Computer Architectures and Their Applications (IJNCAA) 1(4): 1086-1098
4. The Society of Digital Information and Wireless Communications, 2011 (ISSN: 2220-9085).
5. Wei Xu, Ziang Li and Qing Chen. Forecasting the Unemployment Rate by Neural Networks Using Search Engine Query Data. DOI:10.1109/HICSS.2012.284.
6. Wei Xu, Ziang Li, Cheng Cheng and Tingting Zheng. Data mining for unemployment rate prediction using search engine query data. DOI:10.1007/s11761-012-0122-2.
7. Nese Guler and Gulden Kaya Uyanlk. A Study on Multiple Linear Regression Analysis. Doi: 10.1016/j.sbspro.2013.12.027.
8. Yas A. Alsultanny. Labor Market Forecasting by Using Data Mining. Procedia Computer Science 18 (2013) 1700 – 1709.
9. Ziang Li, Wei Xu, Likuan Zhang and Raymond Y.K. Lau. An Ontology-based Web Mining Method for Unemployment Rate Prediction. DOI:10.1016/j.dss.2014.06.007.
10. G. Sugapriyan and S. Prakasam. Analyzing the Performance of MGNREGA Scheme using Data Mining Technique. International Journal of Computer Applications (0975 – 8887) Volume 109 – No. 9, January 2015.
11. Amirah mohamed Shahiri, Wahidah Husain, Nur'Aini Abdul Rashid. A Review on Predicting Student's Performance Using Data Mining Techniques. DOI:10.1016/j.procs.2015.12.157.
12. Tripti Mishra, Dharminder Kumar and Sangeeta Gupta. Students' Employability Prediction Model through Data Mining . International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 4 (2016) pp 2275-2282.
13. V. Vijayalakshmi and R. Ravi. Forecasting Unemployment Rates using Machine Learning Techniques.
14. Wiebke Bleidorn and Christopher James Hopwood. Using Machine Learning to Advance Personality Assessment and Theory. <https://doi.org/10.1177/1088868318772990>.
15. Pum-Mo Ryu. Predicting the Unemployment Rate Using Social Media Analysis. DOI:10.3745/JIPS.04.0079.
16. Suniti Yadav, Khushbu Kumari. Linear regression analysis study. DOI: 10.4103/jpcs.jpcs_8_18.
17. Syarifah Bahiyah Rahayu, Nur Diyana Kamarudin, Zuraini Zainol. Case Study of UPNM Students Performance Classification Algorithms. DOI:10.14419/ijet.v7i4.31.23382.
18. B. K. Nayak and R. K. Behera. A Comparative Analysis of Machine Learning Techniques for Unemployment Rate Prediction.
19. Ping Wang, Yan Li, Chandan K. Reddy. Machine Learning for Survival Analysis: A Survey. <https://doi.org/10.1145/3214306>.
20. G Abuselidze, Lela Mamaladze. The impact of artificial intelligence on employment before and during pandemic: A comparative analysis. DOI:10.1088/1742-6596/1840/1/012040.

21. Ş. Hatipoğlu, M. A. Belgrat, A. Degirmenci and Ö. Karal. Prediction of Unemployment Rates in Turkey by k-Nearest Neighbor Regression Analysis. DOI:10.1109/asyu52992.2021.9598980.
22. Christos Katris. Predicting Future Unemployment Rates Using Time Series Analysis and Machine Learning Techniques. DOI:10.1007/s10614-019-09908-9.
23. Sidhari Manasa, M.Kalidas. UNEMPLOYMENT RATE FORECASTING USING SUPERVISED MACHINE LEARNING MODEL. 2022 IJCSPUB | Volume 12, Issue 4 October 2022 | ISSN: 2250-1770.
24. Hanryeo Lim and Sungpyo Hong. Analysis of longitudinal causal relationships between gender role attitudes and labor market participation of young women in Korea. <https://doi.org/10.1177/02685809221141010>.
25. Andrius Grybauskas, Vaida Pilinkiene, Mantas Lukauskas, Alina Stundžiene and Jurgita Bruneckiene. Nowcasting Unemployment Using Neural Networks and Multi-Dimensional Google Trends Data. *Economies* 2023, 11, 130. <https://doi.org/10.3390/economies11050130>.
26. M. R. Islam, M. R. Khan and M. A. Rahman. Unemployment Prediction using Machine Learning Techniques: A Comparative Study".