

Unemployment Rate Forecasting in India

A PROJECT REPORT



SRM UNIVERSITY, AP

submitted by

Name: T. Narasimha Naidu

Registration No. AP22111260039

Degree: BSc. (Hons.) Physics with research

Course Name: Mathematical Modelling Of Physical Data

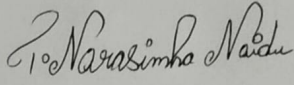
Course Code: SEC107

Course Instructor: Dr. Pankaj Bhalla

Certificate

It is certified that **T. Narasimha Naidu**(AP23111260039) has carried out the project work presented in this Report titled **Unemployment Rate Forecasting in India – State Wise Data in India** for the award of Bachelor of Science (Hons.) CS from SRM University AP, Amaravati under my supervision. The report embodies the result of the original work and studies carried out by the student themselves, and the contents of the report do not form the basis for the award of any other degree to the candidate or to anybody else.

Dr. Pankaj Bhalla
Physics Department
Assistant Professor
SRM University ,AP


T. Narasimha Naidu
AP23111260039
BSC (Hons.) CS
SRM University ,AP

ACKNOWLEDGMENT

We would like to express our sincere gratitude to our course instructor, Dr.Pankaj Bhalla, for their guidance, support, and valuable feedback throughout this project. Their insights have been instrumental in shaping our understanding of trend analysis and linear regression in the context of satellite growth.

We would also like to thank all the members involved in our project for their hard work, dedication, and collaboration.

T.Narasimha Naidu

(Reg.No. AP23111260039)

B.Sc (Hons.) BSC Computer Science

SRM University, A.P

ABSTRACT

This project uses a dataset that is broken down by year, gender, age group, and state to examine unemployment trends across Indian states. With a focus on the top 15 states with the most data entries, the analysis looks at unemployment trends over time and estimates future trends using exponential regression models. Visual comparisons between states, age groups, and gender are carried out, with a focus on gender disparities and inter-state differences. To understand potential future changes in unemployment rates, a two-year prediction is made for each category. Additionally, the model's performance is evaluated and the predictions' level of reliability is indicated by the coefficient of determination (R^2). *The results reveal clear demographic and geogra*

CONTENTS

ACKNOWLEDGMENT	i
ABSTRACT	ii
Chapter 1. INTRODUCTION	1
1.1 Introduction:	1
Chapter 2. METHODOLOGY	2
2.1 Linear Regression	2
2.2 Implementation	2
2.3 Tools Used	3
Chapter 3. DATA ANALYSIS	5
3.1 Actual Data Analysis	5
3.2 Best-Fit Line Analysis	5
3.3 Model Performance Evaluation	8
REFERENCES	12

LIST OF FIGURES

3.1 Male Unemployment Comparison for All states with Linear Forecast (2025-2026)	7
3.2 Female Unemployment Comparison for All states with Linear Forecast (2025-2026)	7

Chapter 1

INTRODUCTION

1.1 INTRODUCTION:

Forecasting India's unemployment rate is essential for comprehending and resolving the ever-changing labor market issues in one of the fastest-growing economies in the world. With a GDP of over \$3.94 trillion in 2024, India is currently the fifth-largest economy in the world and is expected to rise to the fourth position by 2025. Unemployment is still a problem despite this remarkable growth trajectory, reflecting differences between economic expansion and labor market absorption. The unemployment situation in India varies greatly between states and union territories due to a number of factors, including demographic trends, industrial growth, and educational attainment. For example, Madhya Pradesh maintains one of the lowest rates of youth unemployment at 2.6%, while Kerala and other states report high rates of 29.9%. Furthermore, there are differences between urban and rural areas; in April 2024, urban unemployment was 8.7% while rural unemployment was 7.8%. With the use of historical data and forecasting models, this report attempts to present a thorough analysis of India's state-by-state unemployment trends. This study aims to educate stakeholders and policymakers on practical methods for promoting inclusive economic growth and lowering unemployment nationwide by looking at regional disparities and demographic-specific issues.

Chapter 2

METHODOLOGY

2.1 LINEAR REGRESSION

The relationship between time (year) and gender (independent variables) and the unemployment rate (dependent variable) will be modeled using linear regression. The general form of the linear regression equation is as follows:

The equation is:

$$y = \varepsilon + \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

iswherey = textUnemploymentRate

x_1 = Year

x_2 = Gender (0 for Male, 1 for Female)

β_0 = Intercept

β_1, β_2 = Coefficients for Year and Gender, respectively

ε = Error term

The impact of gender and time on unemployment rates is captured by this equation. By estimating the coefficients (β_0 , β_1 , and β_2) using historical data, it enables us to forecast future unemployment rates.

2.2 IMPLEMENTATION

1.Loading Data:

- The supplied CSV file ("my-pro.csv") should be loaded into a Pandas DataFrame.

2.Preprocessing Data:

- Create a numerical representation of the 'Year' column. For instance, divide the range of years (for instance, "2019-20") and calculate the average (for instance, 2019.5).
- Use one-hot encoding to convert the 'Gender' column into numerical values (0 for Male, 1 for Female).
- If there are any missing values, deal with them by deleting the rows or by applying imputation techniques.

3.Training Models:

- Create training and testing sets from the dataset.
- Fit the linear regression model using the training set. For this, the Python sklearn library will be utilized.

4.Assessment of the Model:

- To assess the model's performance, use the testing set.
- To evaluate the model's accuracy, compute metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

5.Predicting:

- Forecast future years' unemployment rates using the trained model.

2.3 TOOLS USED

Python is the main programming language used for implementing models and analyzing data.

- **Pandas:** A tool for working with DataFrames and manipulating data.
- **Scikit-learn:** Building and assessing the linear regression model.
- **NumPy:** For numerical calculations.
- **Seaborn and Matplotlib:** Data visualization tools.

Chapter 3

DATA ANALYSIS

3.1 ACTUAL DATA ANALYSIS

Data on unemployment for Indian states and union territories from 2019–20 to 2023–24 were used in the analysis. The yearly unemployment rates in this dataset are broken down by:

- Total Unemployment (%): The state of the labor market as a whole.
- Male Unemployment (%): Employment trends unique to men (Male (%)).
- Female Unemployment (%): Employment trends unique to women (Female (%)).

Category	Male (%)	Female (%)	Total Unemployment (%)
mean	5.593333	7.889333	13.496
median	5.8	8	13.6
std	1.833571	1.848416	3.534633

Table 3.1: Statistical summary of unemployment rates by gender and overall.

Pandas in Python was used to prepare the data. In order to guarantee data quality for the creation of summary statistics, this required verifying the proper data types and number formatting. The mean, median, and standard deviation of the unemployment rates for men, women, and the overall population were then determined using the prepared data.

3.2 BEST-FIT LINE ANALYSIS

Using data from 2019–2023 and projections through 2026, linear regression was utilized to model the relationship between year (x) and unemployment rate (y)

for every Indian state. To find gender-specific trends across states, distinct models were developed for unemployment rates for men and women.

The best-fit line $y = mx + b$ was computed using the least squares method, where: These formulas represent the equations used to calculate the slope (m) and y-intercept (b) of the best-fit line in simple linear regression:

1. Slope (m):

$$m = \frac{n \sum(xy) - \sum x \sum y}{n \sum(x^2) - (\sum x)^2}$$

Here, n is the number of data points, $\sum(xy)$ is the sum of the product of corresponding x and y values, $\sum x$ is the sum of x values, $\sum y$ is the sum of y values, and $\sum(x^2)$ is the sum of the squares of x values.

2. Y-intercept (b):

$$b = \bar{y} - m\bar{x}$$

\bar{y} is the mean of y values and \bar{x} is the mean of x values.

These equations are used to determine the line of best fit for a set of bivariate data in linear regression analysis.

The LinearRegression class in Scikit-learn was used to implement this technique. The analysis extracted the beginning year for computation purposes and converted year values (such as "2019-20") to numeric form.

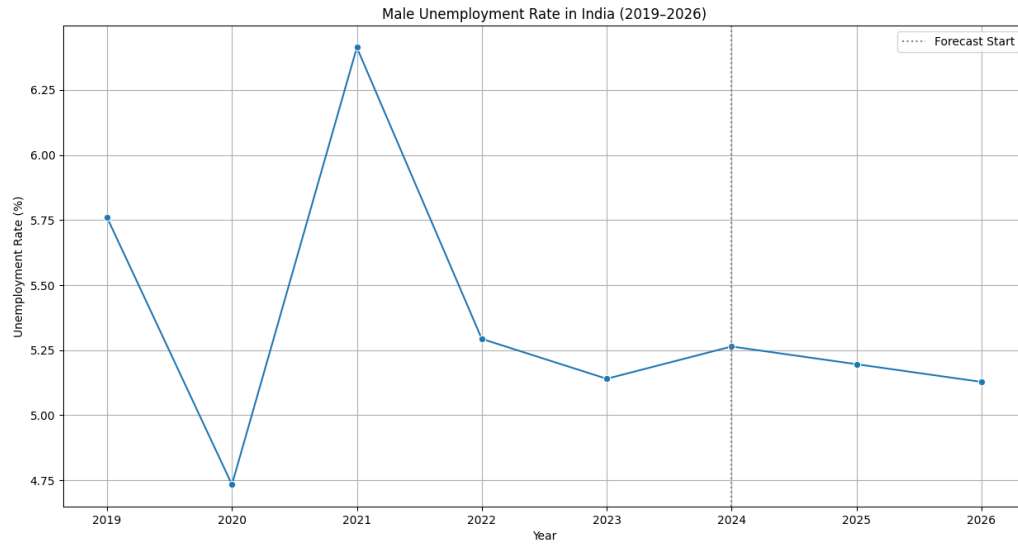


Figure 3.1: Male Unemployment Comparison for All states with Linear Forecast (2025 2026)

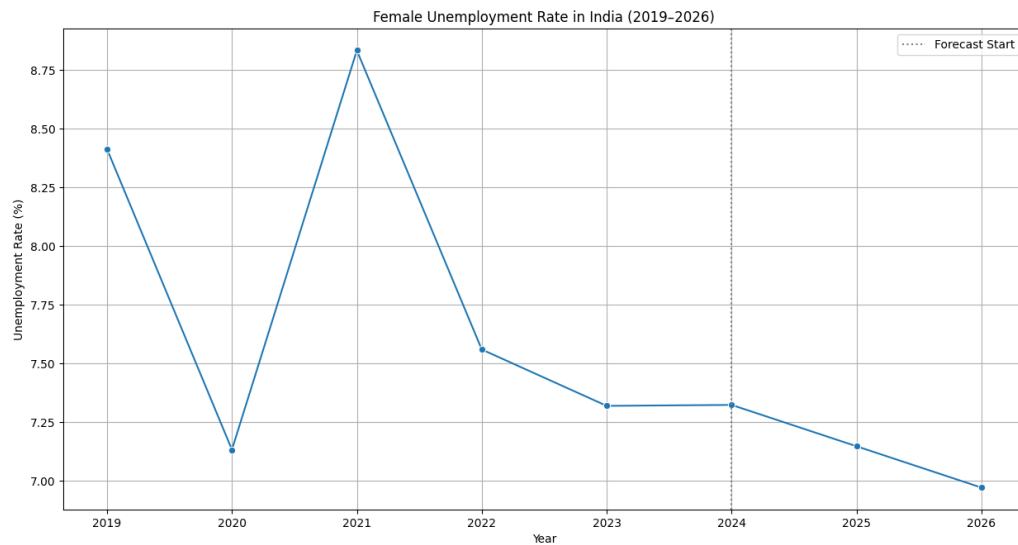


Figure 3.2: Female Unemployment Comparison for All states with Linear Forecast (2025 2026)

3.3 MODEL PERFORMANCE EVALUATION

The coefficient of determination, or R², value was used to evaluate the model's performance:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

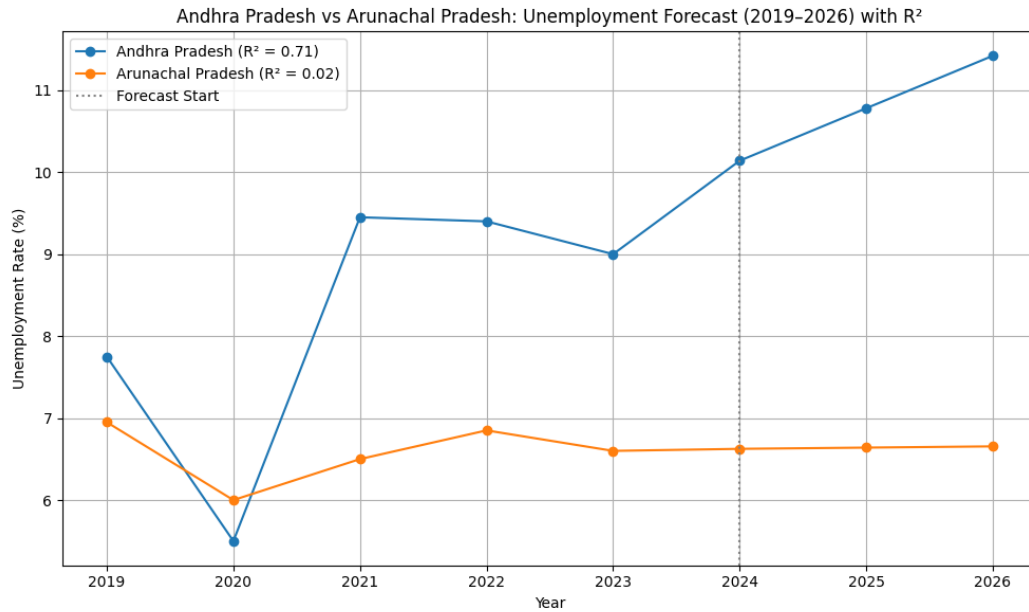


fig 3.3: Unemployment Rate Comparison with Linear Regression: Andhra Pradesh, Arunachal Pradesh(Including R²)

This shows the percentage of the 2019–2026 unemployment variance that the model can account for. Arunachal Pradesh displays a low R² of 0.02, suggesting that linear trends only capture a small portion of the real variations, while Andhra Pradesh has a relatively high R² of 0.71, suggesting a good fit, as seen in Figure. This disparity most likely stems from state-specific economic variables as well as the linear model's poor ability to forecast Arunachal Pradesh's unemployment patterns.

Total Unemployment Rate: Visualization and Forecast Data Solid lines represent Andhra Pradesh's and Arunachal Pradesh's historical unemployment rates from 2019 to 2024. The forecasts for 2025 and 2026 are shown by dashed lines. These were produced using a basic linear regression model based on the data from the previous years.

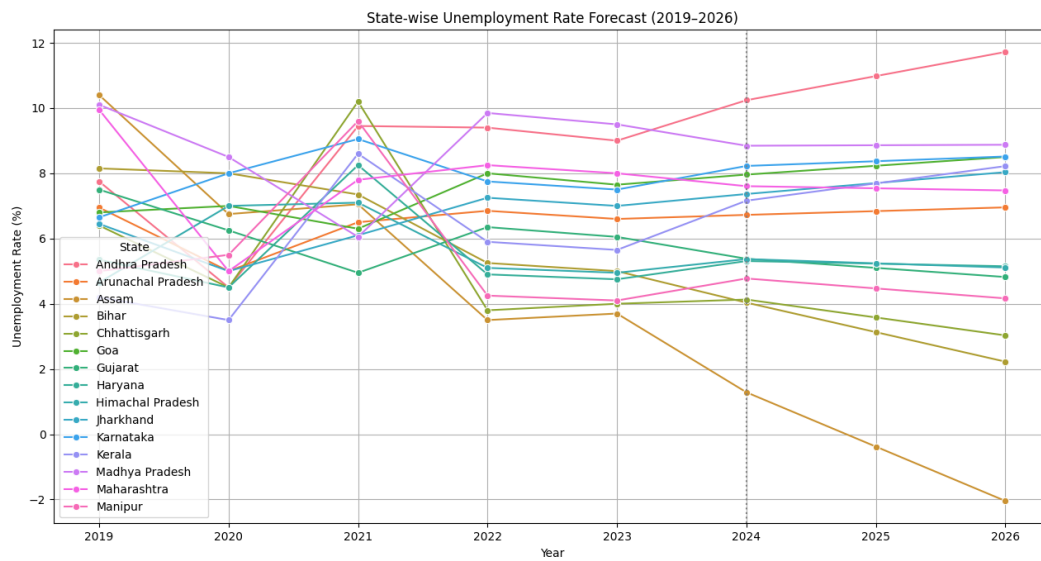


Figure 3.4: Total Unemployment Trends (2019-2024) and Linear Forecasts (2025-2026)

The estimated numbers for the 2025 and 2026 Total Unemployment Rate projections, as determined by the linear regression method. This table provides a quantitative evaluation of projected future rates derived from the linear extrapolation of the 2019–2024 trend. An increase to roughly 10.8% in 2025 and 11.6% in 2026 is projected for Andhra Pradesh. The unemployment rate in Arunachal Pradesh, on the other hand, is anticipated to stay largely unchanged in 2025 and 2026 at roughly 6.7%.

Forecasted Percentage for States/UTs for 2022 and 2026

State/UT	Forecast 2022 (%)	Forecast 2026 (%)
Andhra Pradesh	10.8	11.6
Arunachal Pradesh	6.7	6.7
Assam	1.5	0.2
Bihar	8.2	7.5
Chhattisgarh	8.2	8.2
Goa	15.5	15.6
Gujarat	11.8	11.5
Haryana	9.4	9.3
Himachal Pradesh	9.8	9.7
Jharkhand	14.2	14.3
Karnataka	14.7	14.4
Kerala	11	10.7
Madhya Pradesh	18.6	18.2
Maharashtra	15.8	15.6
Manipur	8	7.8

CONCLUSION

This study examined the overall, male, and female unemployment rates in several Indian states and union territories (UTs) from 2019 to 2024. It projected unemployment rates for 2025 and 2026 using basic linear regression.

Significant patterns emerged from the analysis, including shifts driven by changes in the economy, especially in 2021. The linear regression model provided a fundamental understanding of the overall trend in each state/UT and allowed for early projections of future unemployment rates. However, the model was unable to adequately capture the complexity of unemployment dynamics across all regions, as evidenced by the relatively low R^2 values in a number of states, including Arunachal Pradesh ($R^2=0.02$). Other states, like Andhra Pradesh, on the other hand, demonstrated a better alignment between the model and the data with higher R^2 values ($R^2=0.71$). These results show that while the model can detect general trends, its predictive accuracy is limited during periods of economic volatility and varies by region.

In essence, we found that while our simple mathematical method helped us to detect some general patterns, it was not enough to make accurate predictions in each state. In addition to taking into account additional variables impacting employment and the economy, we must employ increasingly sophisticated mathematical and computer methods if we are expected to forecast the future more accurately. This would increase the knowledge of those who make economic decisions.

REFERENCES

- 1.Periodic Labour Force Survey (PLFS):<https://mospi.gov.in/>
- 2.Centre for Monitoring Indian Economy
(CMIE):<https://www.adda247.com/upsc-exam/unemployment-rate-in-india/>