Predictive Modelling of Population Dynamics: Analysing Demographic Shifts and Growth Rates across U.S. States with Machine Learning

Reetu Sheoran
Department of Data Science and Analytics
Sharda University, Greater Noida, Uttar Pradesh, India

Neetu Dabas
Department of Statistics
Amity University, Noida, Uttar Pradesh, India

Abstract- This paper presents comprehensive study utilizing machine learning techniques to predict and analyse population dynamics across U.S. states, focusing on demographic shifts and growth rates. Firstly, it delves into population analysis, examining trends and growth rates over time, thus identifying significant demographic shifts within various states. Secondly, the research investigates the educational landscape, specifically targeting bachelor's degree attainment among different racial groups. Through this analysis, the study aims to uncover patterns and disparities in educational achievements across states. Lastly, the research explores diversity insights by examining the intersection of racial demographics and educational attainment. Ordinary Least Squares (OLS) regression analysis is used to indicate the relationship between racial demography and bachelor's population. The model's R-squared values indicates that approximately 24.6%, 33.6% and 86.2% of the variability in the percentage of individuals with bachelor's degrees can be explained by the percentage of the Asian population, Black population and White population respectively.

Keywords – Machine learning, Population dynamic, Educational landscape, Racial demographic, Ordinary least squares, Migration, Economy.

I. INTRODUCTION

Population dynamics, encompassing the complex interplay of births, deaths, and migration, serve as fundamental indicators of societal change and development. Understanding these dynamics is crucial for policymakers, planners, and researchers alike to formulate informed strategies and interventions. The size of population as well as its age and sex composition affect the economy and budget of any country [1] . In the United States, the heterogeneous nature of its population, coupled with evolving societal trends, necessitates a nuanced approach to analyze demographic shifts and growth rates across its diverse states [2]. This paper presents a predictive modelling framework leveraging machine learning techniques to explore population dynamics, with a particular focus on demographic shifts and educational attainment across U.S. states.

The Population trends and growth rates provide essential insights into the evolving demographic landscape of a region. Identifying shifts in population demographics over time is crucial for understanding societal changes and planning future resource allocation strategies. Previous research has extensively explored population dynamics using various statistical and computational methods [3]. The U.S. becoming more diverse in terms of race and ethnicity. Immigration is the most influencing factor behind this diversity [4]. The effect of migration takes many forms on a host country especially on college going outcomes such as test scores and college application patterns [5].

Educational attainment serves as a key determinant of socioeconomic status and contributes significantly to individual and societal well-being. Analysing the educational landscape, particularly the prevalence of bachelor's degrees, allows for the assessment of disparities and opportunities across different racial groups. Existing literature has examined educational attainment patterns and disparities, highlighting the role of race, socioeconomic status, and educational policies [6].

Understanding the intersectionality of racial demographics and educational achievements provides critical insights into the diversity of educational attainment across states. Examining how demographic factors influence educational outcomes can inform efforts to promote equity and inclusivity in education. Previous studies have explored the intersection of race, ethnicity, and education, shedding light on disparities and opportunities for improvement [7] (Fuller-Thomson et al., 2009). The racial demographics of the United States are changing at a rapid pace. The past study demonstrates an effect of demographic change on race ratings and categorization [8] (Krosch et.al., 2022). By integrating population analysis, educational attainment, and diversity insights within a predictive modelling framework, this research aims to contribute to a deeper understanding of population dynamics and societal trends in the United States [9] (Choi & Lee, 2021). A report was prepared at the Urban Institute for U.S. Department of Labor, this report focused on demographic trends, work and family issues, health and pension patterns, technical change, adjustment to low unemployment, globalization, and the plight of low-skilled workers [10]. Demographic processes such as women's fertility and life expectancy are also undergoing extraordinary changes. The trends in fertility and mortality leads to a huge demographics transitions [11]. Demography certainly play a powerful role in shaping the social context. Demographic forces are significantly interact with other forces like economic, social, political, technical and environmental. Some of these forces affecting the U.S. in positive direction and some of them in negative direction and some are neural in their impact. But their net direction add up to the major challenges in the countries [12]. Ongoing demographic changes will bring about a substantial shift in the size and the age composition of the population, which will influences macroeconomic variables such as economic growth, inflation, savings, investment etc.[13-14]. Leveraging machine learning approaches offers the potential to uncover complex patterns and relationships within demographic data [15] , thus informing evidence-based decision-making and policy formulation.

The United States are the third most populous country in globally. According to the Population Reference Bureau, The U.S. is getting bigger, older, and more diverse. In this study, we undertake a comprehensive examination utilizing machine learning methodologies to forecast and scrutinize population dynamics among U.S. states, with a particular focus on discerning demographic shifts and growth rates. Initially, we conduct an in-depth analysis of population trends, scrutinizing growth rates over time to pinpoint significant demographic transitions occurring within diverse states. Subsequently, our investigation extends to the educational landscape, specifically targeting the attainment of bachelor's degrees among various racial groups. Through this lens, our study endeavours to uncover patterns and discrepancies in educational accomplishments across state boundaries. Lastly, we delve into insights regarding diversity by probing the intersection of racial demographics and educational achievements. By harnessing the power of machine learning techniques, this research provides a nuanced comprehension of population dynamics, educational attainment, and their intricate interplay. These findings furnish valuable insights for policymakers, researchers, and stakeholders invested in understanding demographic trends and diversity dynamics across U.S. states.

II. Emerging Trends in Predictive Modelling of Population Dynamics

In Currently, research in the area of Demographic Shifts and Growth Rates with Machine Learning is witnessing significant momentum and innovation. With the increasing availability of large-scale demographic datasets and advancements in machine learning techniques, researchers are exploring novel approaches to understand and forecast population dynamics at the state level. One prevalent trend involves the integration of diverse data sources, including census data, satellite imagery, social media data, and administrative records, to capture the multidimensional nature of demographic trends.

Moreover, there is a growing emphasis on developing interpretable machine learning models that not only provide accurate predictions but also offer insights into the underlying drivers of demographic shifts. Furthermore, researchers are actively exploring the application of advanced Machine Learning algorithms, such as deep learning and ensemble methods, to improve the predictive performance and robustness of population forecasting models. Interdisciplinary collaborations between demographers, data scientists, and policymakers are also gaining prominence, facilitating the translation of research findings into actionable insights for urban planning, resource allocation, and policy formulation. Overall, the current research landscape in this domain is characterized by a convergence of cutting-edge methodologies, rich datasets, and collaborative efforts aimed at addressing complex challenges in population dynamics analysis.

III. Approach and Methodology

2.1 Data Collection –

Data is the information in the digitized form. "The process of gathering and analyzing accurate data from various sources to find answers to research problems, trends and probabilities etc. to evaluate possible outcomes is known as Data Collection". It is an essential phase in all types of research, analysis, and decision-making. The first step

involves gathering comprehensive data on population dynamics and educational attainment across U.S. states. The source of the data is secondary data collection. This includes demographic information such as age, gender, race, and educational attainment levels, as well as historical population counts and growth rates. The Datasets have 21 features like 'state', 'densityMi', 'growth', 'BachelorsPercent', 'AsianPercent', 'BlackPercent', 'WhitePercent', 'AsianBachelors', 'AsianTotal', 'BlackBachelors', 'BlackTotal', 'WhiteBachelors', 'WhiteTotal' etc.

2.2 Data Pre-processing and Feature Engineering –

Data Pre-processing involves transforming raw data into a format that is suitable for analysis and modelling. It improve the efficiency of the machine learning model. So, before proceeding with the analysis, the collected data undergoes pre-processing to handle missing values, outliers, and inconsistencies [16]. Initially, done the feature selection part by removing some unnecessary columns from data set, after that checking the missing values and fortunately it is found that there are no missing values in data set [Fig. 1].

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51 entries, 0 to 50
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   fips              51 non-null     int64
 1   state             51 non-null     object
 2   densityMi         51 non-null     float64
 3   pop2024           51 non-null     int64
 4   growth            51 non-null     int64
 5   BachelorsPercent  51 non-null     float64
 6   AsianPercent      51 non-null     float64
 7   BlackPercent      51 non-null     float64
 8   WhitePercent      51 non-null     float64
 9   AsianBachelors    51 non-null     int64
 10  AsianTotal        51 non-null     int64
 11  BlackBachelors    51 non-null     int64
 12  BlackTotal        51 non-null     int64
 13  WhiteBachelors    51 non-null     int64
 14  WhiteTotal        51 non-null     int64
dtypes: float64(5), int64(9), object(1)
memory usage: 6.1+ KB
```

Figure 1.

IV. Population Analysis and Result Discussion

Machine learning models provide a robust toolkit for conducting descriptive analyses on individual variables within a dataset, offering insights into key statistical parameters such as central tendencies, dispersion, and data distributions (refer to Fig. 2). Furthermore, in order to facilitate a comparative examination of Bachelor's degree attainment across different states, we utilized multiple bar diagrams (depicted in Fig. 3) and line graphs (illustrated in Fig. 4) to delineate population trends.

	fips	densityMi	pop2024	growth	BachelorsPercent	AsianPercent	BlackPercent	WhitePercent	AsianBachelors	AsianTotal	BlackBachelors	BlackTotal	WhiteBachelors	WhiteTotal
count	51.000000	51.000000	5.100000e+01	51.000000	51.000000	51.000000	51.000000	51.000000	5.100000e+01	5.100000e+01	51.000000	5.100000e+01	5.100000e+01	5.100000e+01
mean	28.960784	421.657073	6.584305e+06	24627.509804	32.545098	0.537786	0.230887	0.348902	1.404355e+05	2.553980e+05	117757.313725	5.216675e+05	1.100074e+06	3.193804e+06
std	15.832828	1550.787453	7.485528e+06	99214.888042	6.492190	0.094218	0.048098	0.096483	3.347528e+05	6.210791e+05	153570.889807	6.348027e+05	1.147956e+06	3.255199e+06
min	1.000000	1.283440	5.851770e+05	-180341.000000	21.300000	0.261394	0.149247	0.210533	2.009000e+03	3.493000e+03	994.000000	3.021000e+03	1.034440e+05	2.221000e+05
25%	16.500000	51.170395	1.865538e+06	-4826.500000	28.800000	0.483274	0.191119	0.296185	1.419800e+04	2.582400e+04	9238.500000	4.268850e+04	3.125615e+05	1.028562e+06
50%	29.000000	109.262070	4.516527e+06	5713.000000	32.000000	0.537312	0.227423	0.333248	4.677700e+04	8.272800e+04	47703.000000	2.127680e+05	6.781150e+05	2.350417e+06
75%	41.500000	224.200155	7.711852e+06	27414.500000	35.250000	0.594649	0.261246	0.371203	1.420655e+05	2.419265e+05	155548.500000	8.755790e+05	1.471568e+06	3.913954e+06
max	56.000000	11111.918030	3.880204e+07	470708.000000	59.800000	0.834433	0.341737	0.897641	2.288312e+06	4.267150e+06	563381.000000	2.200749e+06	5.651032e+06	1.555447e+07

Figure 2. Insights into key statistical parameters

Through our graphical exploration, we uncovered notable patterns. For instance, California exhibited the highest proportions of White and Asian individuals with Bachelor's degrees, while Texas emerged as a leader in Black Bachelor's degree attainment. These findings highlight the pivotal role of visual data analysis in elucidating regional variations in educational accomplishments.

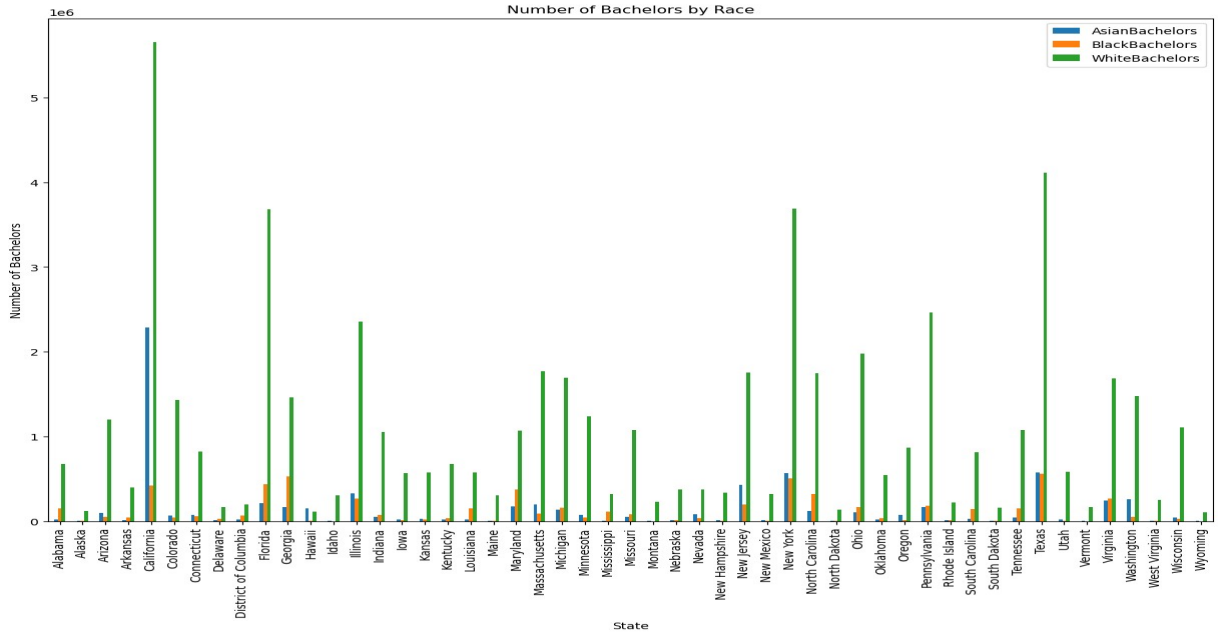

Figure 3. Comparative Examination of Bachelor's degree Attainment across different States

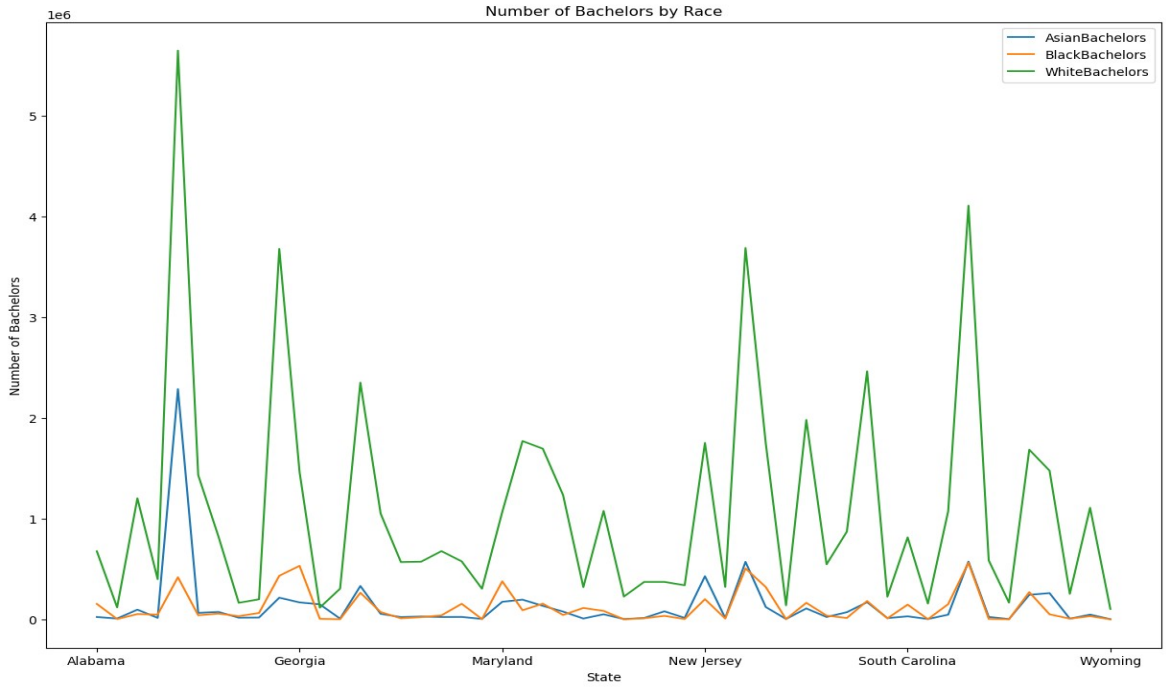

Figure 4. Population Trends

In case Upon analyzing the bar graphs represented in [Figure 5],[Figure 6], and [Figure 7], which outline the top 5 states concerning Black Bachelors, Asian Bachelors, and White Bachelors populations respectively, several notable observations surface. As depicted in [Figure 5], Texas emerges as the primary state in the Black Bachelors

demographic, showcasing the highest count. Georgia follows closely, also displaying a substantial population of Black Bachelors. Additionally, New York, Florida, and California present noteworthy figures, albeit slightly lower compared to Texas and Georgia. These insights illuminate the distribution of educational achievements among different racial groups across various states.

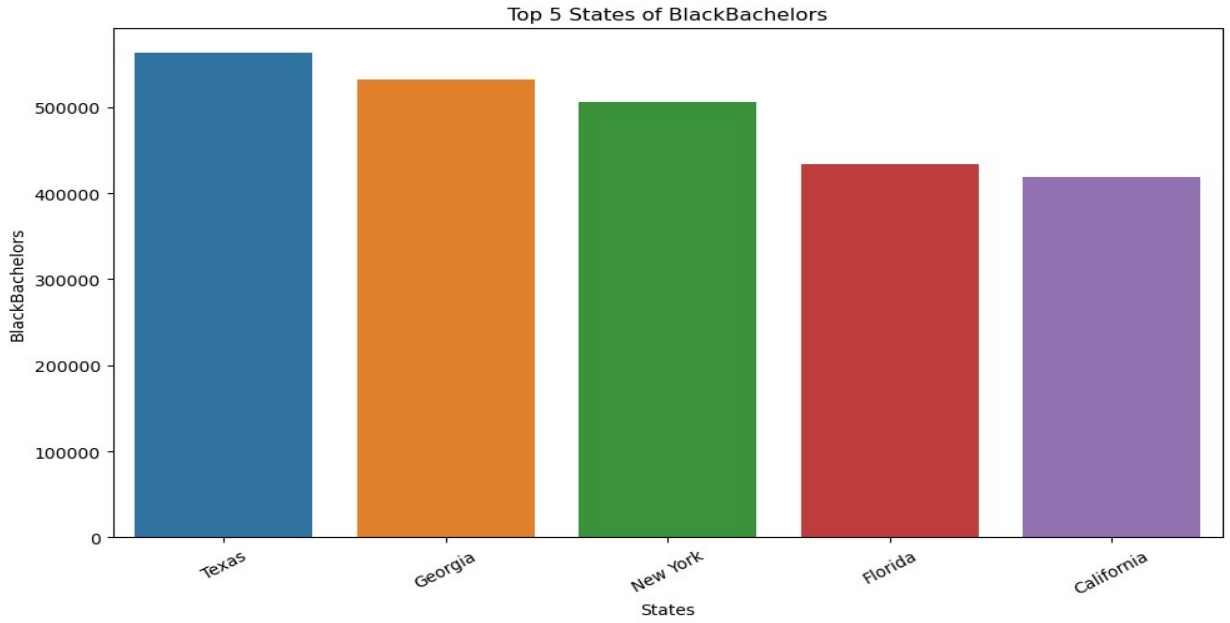

Figure 5. Top 5 States of Black Bachelors

After analysing the [Fig. 6] the top 5 states in terms of Asian Bachelors population, we can say that California emerges as the dominant state in this demographic category, boasting a substantially higher Asian Bachelors population compared to other states. Following California, Texas and New York also show significant figures, albeit noticeably lower than California's count. New Jersey and Illinois, while presenting Asian Bachelors populations, demonstrate comparatively smaller numbers.

In conclusion of [Fig. 7] the top 5 states highlighted in the bar graph offer valuable insights into the distribution of White Bachelors populations, California emerges as the clear frontrunner with a significantly higher White Bachelors population compared to the other states. Texas follows closely behind, showing a substantial White bachelors population as well. New York, Florida, and Pennsylvania also demonstrate notable figures, albeit trailing behind California and Texas.

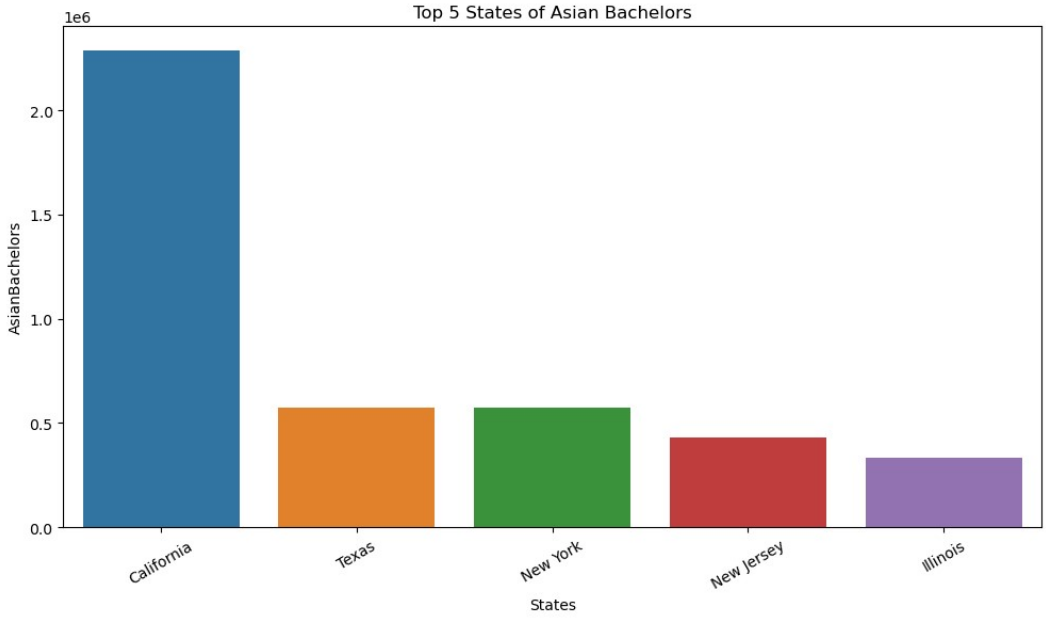

Figure 6. Top 5 States of Asian Bachelors

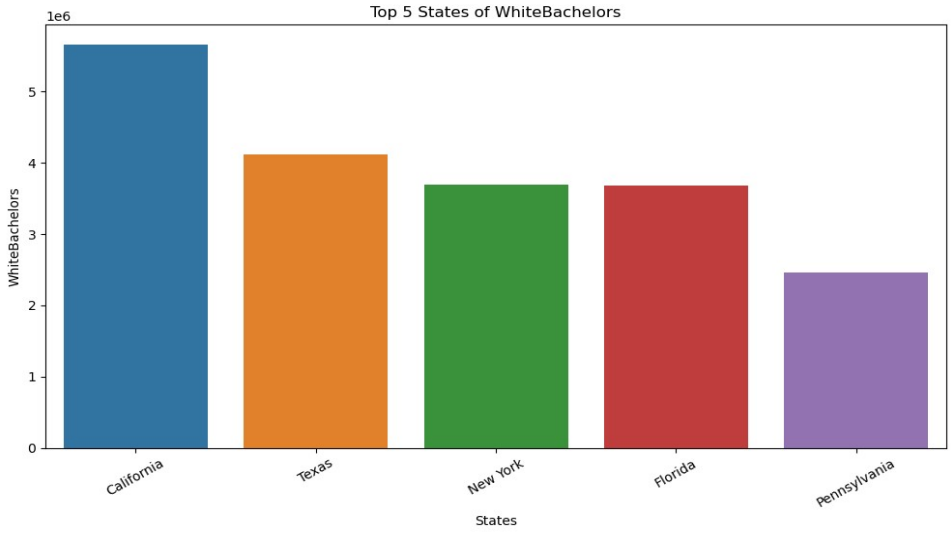

Figure 7. Top 5 States of white Bachelors

This data underscores the importance of understanding regional demographics, particularly concerning educational attainment among individuals. It highlights the need for tailored initiatives and support systems to address the specific educational and professional needs of Bachelors populations in respective states. Additionally, further investigation into the factors driving the disparities in Bachelors populations across different regions could provide valuable insights for targeted interventions and resource allocation.

The Ordinary Least Squares (OLS) regression analysis [Fig. 8] reveals that there is a statistically significant relationship between the percentage of Asian population and the percentage of individuals with bachelor's degrees. The model suggests that for every one-unit increase in the percentage of the Asian population, there is a corresponding increase of

OLS Regression Results

Dep. Variable:	BachelorsPercent	R-squared:	0.246
Model:	OLS	Adj. R-squared:	0.230
Method:	Least Squares	F-statistic:	15.96
Prob(F-statistic):	0.000217	Log-Likelihood:	-160.07
No. Observations:	51	AIC:	324.1
Df Residuals:	49	BIC:	328.0
Df Model:	1	Covariance Type:	nonrobust

	coef	std err	t	P>|t|	[0.025	0.975]
const	14.1762	4.666	3.038	0.004	4.799	23.554
AsianPercent	34.1564	8.549	3.995	0.000	16.976	51.337

Omnibus:	3.149	Durbin-Watson:	2.144
Prob(Omnibus):	0.207	Jarque-Bera (JB):	2.148
Skew:	0.366	Prob(JB):	0.342
Kurtosis:	3.690	Cond. No.	13.8

Figure 8.

approximately 34.16 units in the percentage of individuals with bachelor's degrees, holding other variables constant. The model's R-squared value of 0.246 indicates that approximately 24.6% of the variability in the percentage of individuals with bachelor's degrees can be explained by the percentage of the Asian population alone.

In summary of [Fig. 9] the Ordinary Least Squares (OLS) regression analysis indicates a significant relationship between the percentage of the Black population and the percentage of individuals with bachelor's degrees. The model suggests that for every one-unit increase in the percentage of the Black population, there is an estimated

increase of approximately 78.27 units in the percentage of individuals with bachelor's degrees, holding other variables constant.

OLS Regression Results

Dep. Variable:	BachelorsPercent	R-squared:	0.336
Model:	OLS	Adj. R-squared:	0.323
Method:	Least Squares	F-statistic:	24.82
Prob (F-statistic):	8.22e-06	Log-Likelihood:	-156.81
No. Observations:	51	AIC:	317.6
Df Residuals:	49	BIC:	321.5
Df Model:	1	Covariance Type:	nonrobust

	coef	std err	t	P>|t|	[0.025	0.975]
const	14.4735	3.703	3.908	0.000	7.031	21.916
BlackPercent	78.2704	15.709	4.982	0.000	46.701	109.840

Omnibus:	25.499	Durbin-Watson:	2.033
Prob(Omnibus):	0.000	Jarque-Bera (JB):	69.343
Skew:	1.275	Prob(JB):	8.76e-16
Kurtosis:	8.111	Cond. No.	22.1

Figure 9.

The model's R-squared value of 0.336 implies that approximately 33.6% of the variability in the percentage of individuals with bachelor's degrees can be explained by the percentage of the Black population alone.

The Ordinary Least Squares (OLS) regression analysis [Fig. 10] reveals a highly significant relationship between the percentage of the White population and the percentage of individuals with bachelor's degrees. The model suggests that for every one-unit increase in the percentage

OLS Regression Results

Dep. Variable:	BachelorsPercent	R-squared:	0.862
Model:	OLS	Adj. R-squared:	0.859
Method:	Least Squares	F-statistic:	306.0
Prob (F-statistic):	1.03e-22	Log-Likelihood:	-116.76
No. Observations:	51	AIC:	237.5
Df Residuals:	49	BIC:	241.4
Df Model:	1	Covariance Type:	nonrobust

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.7483	1.292	8.320	0.000	8.152	13.344
WhitePercent	62.4724	3.571	17.493	0.000	55.296	69.649

Omnibus:	2.695	Durbin-Watson:	1.750
Prob(Omnibus):	0.260	Jarque-Bera (JB):	1.835
Skew:	-0.435	Prob(JB):	0.400
Kurtosis:	3.326	Cond. No.	11.8

Figure. 10

of the White population, there is an estimated increase of approximately 62.47 units in the percentage of individuals with bachelor's degrees, holding other variables constant.The remarkable R-squared value of 0.862 indicates that approximately 86.2% of the variability in the percentage of individuals with bachelor's degrees can be explained by the percentage of the White population alone, highlighting the substantial impact of demographic composition on educational attainment levels within this population. These findings underscore the importance of considering

demographic factors such as ethnicity when examining educational attainment levels within a given population. However, it's crucial to recognize that this analysis is based on a specific dataset and context, and further research may be needed to generalize these findings to other populations or time periods

V. CONCLUSION

In conclusion, our study demonstrates the power of machine learning models in conducting descriptive analyses and visualizing complex datasets to uncover insights into educational attainment across different racial groups and states in the U.S. By employing multiple graphical representations, we identified notable patterns and variations in Bachelor's degree attainment among Black, Asian, and White populations across various states. Specifically, our analysis revealed California's dominance in Asian Bachelor's degree attainment, Texas's leadership in Black Bachelor's degree attainment, and the substantial White Bachelor's degree population in both California and Texas. Furthermore, Ordinary Least Squares (OLS) regression analysis highlighted statistically significant relationships between demographic composition and educational attainment, with ethnicity playing a significant role. The findings underscore the importance of understanding regional demographics and tailoring educational initiatives to address the specific needs of diverse populations. Future research should delve deeper into the factors driving these disparities to inform targeted interventions and resource allocation strategies aimed at promoting educational equity and excellence across all demographic groups. It's essential to acknowledge that these findings are based on specific datasets and contexts, and further research is warranted to generalize these insights to broader populations and timeframes.

VI. FURTHER SCOPE OF STUDY

The future scope of study following the application of Ordinary Least Squares (OLS) regression analysis Demographic Shifts and Growth Rates across U.S. States with Machine Learning entails several avenues for further exploration. Firstly, integrating additional machine learning algorithms, such as ensemble methods or deep learning architectures, could enhance predictive accuracy and capture nonlinear relationships inherent in demographic data. Secondly, extending the analysis to incorporate spatial-temporal factors, such as geographic clustering and temporal trends, would provide a more comprehensive understanding of population dynamics at the state level. Furthermore, exploring the impact of exogenous variables, such as environmental factors or policy interventions, on demographic shifts could offer valuable insights for policymakers and urban planners. Additionally, incorporating uncertainty analysis techniques, such as bootstrapping or Monte Carlo simulations, would enable the assessment of prediction intervals and enhance the robustness of forecasting models. Lastly, fostering interdisciplinary collaboration between demographers, statisticians, and machine learning experts is essential for advancing methodological developments and addressing emerging challenges in predictive modelling of population dynamics.

REFERENCES

I. Cyn-Young Park, Kwanho Shin, and Aiko Kikkawa, “Demographic Change, Technological Advances, and Growth: A Cross-Country Analysis", ADB Economics Working Paper Series, (2020), 617.

II. Rashmi Mishra and S.P. Thakur, “A Study of Brics Nation Being the Real Growth Drivers of World Economy”, International Journal of Management, 8:2 (2017), 188–198.

III. Adrian Timpson, Ramiro Barberena , Mark G Thomas , César Méndez and Katie Manning, “Directly modelling population dynamics in the South American Arid Diagonal using 14C dates”, Phil. Trans. R. Soc. B, (2021), 376, 20190723.

IV. Laura B. Shrestha and Elayne J. Heisler, “The Changing Demographic Profile of the United States”, Congressional Research Service 7-5700, (2011) www.crs.gov RL32701.

V. Neymotin Florence, “Immigration and its effect on the college-going outcomes of natives”, Economics of Education Review, 28:5 (2009), 538-550.

VI. Dominic Weinberg, Gonneke W. J. M. Stevens, Catrin Finkenauer, Bert Brunekreef, Henriëtte A. Smit and Alet H. Wijga, “The pathways from parental and neighbourhood socioeconomic status to adolescent educational attainment: An examination of the role of cognitive ability, teacher assessment, and educational expectations”, PLoS ONE, 14:5 (2019), 0216803.

VII. Esme Fuller-Thomson, A. Nuru-Jeter, Meredith Minkler and Jack M. Guralnik, “Black–White Disparities in Disability among Older Americans: Further Untangling the Role of Race and Socioeconomic Status”, J Aging Health, 21:5 (2009), 677–698.

VIII. Amy R. Krosch , Suzy J. Park , Jesse Walker and Ari R. Lisner, “The threat of a majority-minority U.S. alters white Americans' perception of race”, Journal of Experimental Social Psychology, 99 (2022), 104266.

IX. Choi Soobin and Lee Moosung, “Diversity as an opportunity or a challenge? A cross-national study of ethnic diversity and students generalized trust”, International Journal of Educational Development, 80 (2021), 102-301.

X. Robert I. Lerman and Stefanie R. Schmidt, “An Overview of Economic, Social, and Demographic Trends Affecting the U.S. Labor Market”, The Urban Institute Washington, (1999), D.C.

XI. Bongaarts John, “Human Population Growth and the Demographic Transition”, Philosophical Transactions of the Royal Society B: Biological Sciences, 364:1532 (2009), 2985–2990.

XII. R. Scott Fosler, William Alonso, Jack A. Meyer and Rosemary Kern, "Demographic Change and the American Future", University of Pittsburgh Press, (1990), e-Book.

XIII. Jong-Won Yoon, Jinill Kim and Jungjin Lee, "Impact of Demographic Changes on Inflation and the Macroeconomy", International Monetary Fund Working Paper, (2014), WP/14/210.

XIV. Ichiro Muto, Takemasa Oda and Nao Sudo, "Macroeconomic Impact of Population Ageing in Japan: A Perspective from an Overlapping Generations Model", Bank of Japan Working Paper Series, (2012), No. 12-E-9.

XV. Li Fe, Yigitcanlar Tan, Nepal Madhav, Nguyen Kien and Dur Fatih, "Machine learning and remote sensing integration for leveraging urban sustainability: A review and framework", Sustainable Cities and Society, 96 (2023), 104653.

XVI. Abdelrahman Elsharif Karrar, "The Effect of Using Data Pre-Processing by Imputations in Handling Missing Value", Indonesian Journal of Electrical Engineering and Informatics, 10:2 (2022), 375-384.

XVII. Sebastian Rausch, "Macroeconomic Consequences of Demographic Change", Springer, (2009).