ADVANCED ANALYTICAL APPROACHES FOR ECONOMIC DEVELOPMENT OF TRIBAL COMMUNITIES IN JAMTARA DISTRICT OF JHARKHAND.

BY

SUBHAJEET KHAWAS

(23MB0063)

THESIS

SUBMITTED TO

INDIAN INSTITUTE OF TECHNOLOGY

(INDIAN SCHOOL OF MINES), DHANBAD

For the award of the degree of

MASTER OF BUSINESS ADMINISTRATION (BA)

DEPARTMENT OF MANAGEMENT STUDIES AND
INDUSTRIAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY
(INDIAN SCHOOL OF MINES),
DHANBAD – 826004

CERTIFICATE OF THESIS COMPLETION

This is to certify that **Mr. Subhajeet Khawas, Admission Number – 23MB0063** of the Department of Management Studies and Industrial Engineering, IIT (ISM), Dhanbad has done Thesis work titled "**Advanced Analytical Approaches for Economic Development of Tribal Communities in Jamtara district of Jharkhand.**" in his 3rd semester during the period of Aug-Nov 2024, during academic year 2024-25 under the guidance of **Dr. Rashmi Singh** towards partial fulfilment of the award of **Master of Business Administration (MBA) in Business Analytics**.

Dr. Rashmi Singh
(PROJECT GUIDE)

DEPARTMENT OF MANAGEMENT STUDIES AND
INDUSTRIAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY
(INDIAN SCHOOL OF MINES),
DHANBAD – 826004

DECLARATION

I hereby confirm that the present Thesis report, which is being submitted in partial fulfilment for attaining the Master of Business Administration-Business Analytics degree, is the result of my independent research and effort. The content presented below has not been included in any other academic submission to obtain a degree or diploma, either within the confines of this institution or any other accredited university.

Subhajeet Khawas

Mr. Subhajeet Khawas

(23MB0063)

Form No: PG11

INDIAN INSTITUTE OF TECHNOLOGY (INDIAN SCHOOL OF MINES) DHANBAD

DECLARATION BY THE STUDENT

(To be submitted at the time of Final Dissertation Submission)

I hereby declare that the work which is being presented in this dissertation entitled "ADVANCED ANALYTICAL APPROACHES FOR ECONOMIC DEVELOPMENT OF TRIBAL COMMUNITIES IN JAMTARA DISTRICT OF JHARKHAND" in partial fulfilment of the requirements for the award of the degree of Master of BUSINESS ADMINISTRATION in BUSINESS ANALYTICS is an authentic record of my own work carried out during the period from AUG 24 to NOV 24 under the supervision of PROF. RASHMI SINGH Department of MANAGEMENT STUDIES AND INDUSTRIAL ENGINEERING Indian Institute of Technology (ISM) Dhanbad, Jharkhand, India.

I acknowledge that I have read and understood the UGC (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations, 2018. These Regulations were published in the Indian Official Gazette on 31st July, 2018.

I confirm that this Dissertation has been checked for plagiarism using the online plagiarism checking software provided by the Institute. At the end of the Dissertation, a copy of the summary report demonstrating similarities in content and its potential source (if any) generated online using plagiarism checking software is enclosed. I herewith confirm that the Dissertation has less than 10% similarity according to the plagiarism checking software's report and meets the MoE/UGC Regulations as well as the Institute's rules for plagiarism.

I further declare that no portion of the dissertation or its data will be published without the Institute's or Guide's permission. I have not previously applied for any other degree or award using the topics and findings described in my dissertation.

Subhajeet Khawas

(Signature of the Student)

Name of the Student: SUBHAJEET KHAWAS

Admission No.: 23MBD063

Department : DMSIE

Form No: PG12

INDIAN INSTITUTE OF TECHNOLOGY (INDIAN SCHOOL OF MINES) DHANBAD

CERTIFICATE FOR CLASSIFIED DATA

(To be submitted at the time of Final Dissertation Submission)

This is to certify that the Dissertation entitled "ADVANCED ANALYTICAL APPROACHES FOR ECONOMIC DEVELOPMENT OF TRIBAL COMMUNITIES IN JAMTARA DISTRICT OF JHARKHAND" being submitted to the Indian Institute of Technology (Indian School of Mines), Dhanbad by Mr/Ms SUBHAJEET KHAWAS for award of Master Degree in BUSINESS ADMINISTRATION (BA) does not contain any classified information.

This work is original and yet not been submitted to any institution or university for the award of any degree.

18/11/24.

Signature of Supervisor (s)

Subhajeet Khawas

Signature of Student

INDIAN INSTITUTE OF TECHNOLOGY (INDIAN SCHOOL OF MINES) DHANBAD

COPYRIGHT AND CONSENT FORM

(To be submitted at the time of Final Dissertation Submission)

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IIT (ISM), Dhanbad and must accompany any such material in order to be published by the IIT (ISM). Please read the form carefully and keep a copy for your files.

_____

_____

COPYRIGHT TRANSFER

1. The undersigned hereby assigns to Indian Institute of Technology (Indian School of Mines), Dhanbad all rights under copyright that may exist in and to: (a) the above Work, including any revised or expanded derivative works submitted to the IIT (ISM) by the undersigned based on the work; and (b) any associated written or multimedia components or other enhancements accompanying the work.

CONSENT AND RELEASE

2. In the event the undersigned makes a presentation based upon the work at a conference hosted or sponsored in whole or in part by the IIT (ISM) Dhanbad, the undersigned, in consideration for his/her participation in the conference, hereby grants the IIT (ISM) the unlimited, worldwide, irrevocable permission to use, distribute, publish, license, exhibit, record, digitize, broadcast, reproduce and archive; in any format or medium, whether now known or hereafter developed: (a) his/her presentation and comments at the conference; (b) any written materials or multimedia files used in connection with his/her presentation; and (c) any recorded interviews of him/her (collectively, the "Presentation"). The permission granted includes the transcription and reproduction of the Presentation for inclusion in products sold or distributed by IIT(ISM) Dhanbad and live or recorded broadcast of the Presentation during or after the conference.
3. In connection with the permission granted in Section 2, the undersigned hereby grants IIT (ISM) Dhanbad the unlimited, worldwide, irrevocable right to use his/her name, picture, likeness, voice and biographical information as part of the advertisement, distribution and sale of products incorporating the Work or Presentation, and releases IIT (ISM) Dhanbad from any claim based on right of privacy or publicity.
4. The undersigned hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the undersigned has obtained any necessary permissions. Where necessary, the undersigned has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IIT (ISM) Dhanbad.

GENERAL TERMS

* The undersigned represents that he/she has the power and authority to make and execute this assignment.
* The undersigned agrees to indemnify and hold harmless the IIT (ISM) Dhanbad from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
* In the event the above work is not accepted and published by the IIT (ISM) Dhanbad or is withdrawn by the author(s) before acceptance by the IIT(ISM) Dhanbad, the foregoing copyright transfer shall become null and void and all materials embodying the Work submitted to the IIT(ISM) Dhanbad will be destroyed.
* For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.

Subhajit Ukanoq

Signature of the Author

Form No: PG10

INDIAN INSTITUTE OF TECHNOLOGY (INDIAN SCHOOL OF MINES) DHANBAD

CERTIFICATE FOR THE FINAL VERSION OF DISSERTATION

(To be submitted at the time of Final Dissertation Submission)

This is to certify that the Dissertation entitled "ADVANCED ANALYTICAL APPROACHES FOR ECONOMIC DEVELOPMENT OF TRIBAL COMMUNITIES IN JAMTARA DISTRICT OF JHARKHAND"

being submitted to the Indian Institute of Technology (Indian School of Mines), Dhanbad, by Mr/Ms ✓ SUBHAJEET KHAWAS, Admission No 23MB00063 for the award of the Degree of Master OF BUSINESS ADMINISTRATION from IIT (ISM), Dhanbad,is a bonafide work carried out by him/her, in the Department of MANAGEMENT STUDIES AND INDUSTRIAL ENGINEERING, IIT (ISM), Dhanbad, under my/our supervision and guidance. The dissertation has fulfilled all the requirements as per the regulations of this Institute and, in my/our opinion, has reached the standard needed for submission. The results embodied in this dissertation have not been submitted to any other university or institute for the award of any degree or diploma.

18/11/24.

Signature of Supervisor (s)

Name: PROF. RASHMI SINGH

Date: 18-11-24

Form No: PG13

INDIAN INSTITUTE OF TECHNOLOGY (INDIAN SCHOOL OF MINES) DHANBAD

CERTIFICATE REGARDING ENGLISH CHECKING

(To be submitted at the time of Final Dissertation Submission)

This is to certify that the Dissertation entitled "ADVANCED ANALYTICAL APPROACHES FOR ECONOMIC DEVELOPMENT OF TRIBAL COMMUNITIES IN JAMTARA DISTRICT OF JHARKHAND

_____ being submitted to the Indian Institute of Technology (Indian School of Mines), Dhanbad by ✓ Mr/Ms SUBHAJEET KHAWAS

_____,

Admission No 23MB0063, for the award of Master of BUSINESS ADMINISTRATION (BA)

has been thoroughly checked for quality of English and logical sequencing of topics.

It is hereby certified that the standard of English is good and that grammar and typos have been thoroughly checked.

Signature of Supervisor (s)

Name: PROF. RASHMI SINGH

Date: 18-11-24

Signature of Student

Name: SUBHAJEET KHAWAS

Date: 18-11-24

DEPARTMENT OF MANAGEMENT STUDIES AND
INDUSTRIAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY
(INDIAN SCHOOL OF MINES),
DHANBAD – 826004

ACKNOWLEDGEMENT

This project report has been submitted to the Department of Management Studies and Industrial Engineering at IIT (ISM) Dhanbad. I am deeply grateful to my project guide, Dr. Rashmi Singh, for her invaluable guidance and support throughout the study. Her insights and recommendations have been instrumental in shaping and refining this work, which was developed under her expert mentorship and encouragement.

I also extend my sincere appreciation to Dr. Sandeep Mondal, Head of the Department of Management Studies and Industrial Engineering, for his essential support in providing the resources and facilities that enabled the successful completion of this project.

I am thankful to IIT (ISM) for offering me the opportunity to apply my skills within the MBA Business Analytics program. My heartfelt gratitude goes to my parents, faculty members, and friends, whose unwavering support and motivation have driven me to reach this stage. Finally, I would like to acknowledge everyone who has shown confidence in my abilities and provided support, contributing significantly to the successful completion of this thesis.

Contents

List of Figures

Abstract

This thesis investigates advanced analytical approaches to enhance the economic development of tribal communities in the Jamtara District of Jharkhand, India. The study focuses on identifying key determinants influencing critical socio-economic variables, including annual income, access to public distribution system (PDS) cards, bank accounts, expenditure on food and labour, satisfaction with existing educational facilities in their respective areas, and access to LPG connections. Utilizing a comprehensive dataset with multiple demographic and socio-economic features, we employ regression and classification models, specifically ensemble methods, to address the complexity and interdependencies inherent in the data.

The analysis reveals the top five independent features that significantly impact the selected dependent variables. These findings are crucial for informing government policy aimed at fostering sustainable economic growth and improving living standards within these communities. By understanding which factors most influence economic outcomes, policymakers can design targeted interventions that address specific needs and challenges faced by tribal populations in Jamtara.

The results underscore the importance of financial inclusion, access to essential services, and educational satisfaction as pivotal elements in enhancing economic conditions. This research contributes to the broader discourse on economic development in marginalized communities and provides actionable insights for stakeholders aiming to implement effective development strategies tailored to the unique context of tribal societies in India.

Introduction

The economic development of tribal communities in India, particularly in regions like Jamtara District of Jharkhand, remains a critical area of concern for policymakers and researchers alike. These communities often face multifaceted challenges, including poverty, limited access to essential services, and socio-cultural barriers that hinder their overall development. Despite the rich cultural heritage and natural resources available to these populations, their socio-economic conditions are frequently characterized by low-income levels, inadequate access to education and healthcare, and insufficient infrastructure. This thesis aims to explore advanced analytical approaches to identify the key determinants influencing the economic development of these tribal communities.

In recent years, there has been a growing recognition of the importance of data-driven decision-making in addressing the complex issues faced by marginalized populations. By leveraging advanced analytical techniques such as regression and classification models, particularly ensemble methods that can handle high-dimensional data, this research seeks to uncover the underlying factors that significantly impact the economic variables of interest. The selected dependent variables for this study—annual income, access to public distribution system (PDS) cards, bank accounts, expenditure patterns on food and labour, satisfaction with educational facilities, and access to LPG connections—are crucial indicators of economic well-being and quality of life.

Understanding these interdependencies is essential for crafting effective policies aimed at improving the living standards of tribal communities. The analysis focuses on identifying the top independent features that influence these dependent variables. This knowledge will empower government agencies and local stakeholders to design targeted interventions that address specific needs within these communities. For instance, enhancing financial inclusion through improved access to banking services or increasing investment in educational infrastructure could significantly uplift the socio-economic status of these populations.

Moreover, this research contributes to the broader discourse on sustainable development by highlighting how tailored policies can foster economic resilience among tribal

communities. By focusing on the unique challenges faced by these groups and employing sophisticated analytical methodologies, this thesis not only aims to provide empirical insights but also advocates for a more inclusive approach to economic development that recognizes the diverse needs and potentials of tribal populations.

As India strives towards achieving its Sustainable Development Goals (SDGs), understanding the dynamics of economic development in tribal areas like Jamtara is imperative. This thesis endeavors to bridge the gap between data analytics and policy formulation, ultimately aiming to promote sustainable growth and improve the quality of life for tribal communities in Jharkhand. Through this research, I hope to illuminate pathways for effective intervention that can lead to meaningful change in the socio-economic landscape of these marginalized groups.

Research Objectives

This thesis aims to systematically analyse the socio-economic factors influencing the economic development of tribal communities in the Jamtara District of Jharkhand. The following objectives have been formulated to guide the research:

1. Development of Predictive Regression Models:

The primary objective is to construct robust predictive regression models that quantitatively assess the relationships between selected independent variables and key dependent variables, such as annual income and expenditure patterns. By employing advanced regression techniques, this analysis will elucidate how variations in factors like access to public distribution system (PDS) cards, bank accounts, and LPG connections influence economic outcomes. This model will provide a statistical foundation for understanding the dynamics of economic development within tribal communities.

2. Implementation of Classification Algorithms for Socio-Economic Indicators:

This objective centers on deploying advanced classification algorithms to categorize households based on discrete socio-economic indicators, such as bank account possession, LPG connection, PDS card possession, and satisfaction with educational facilities in nearby areas. These models provide insights into patterns within the dataset, helping to reveal how specific factors contribute to socio-economic conditions. By examining classifications related to these essential services, we can identify disparities and develop data-driven recommendations for targeted interventions.

3. Assessment of Feature Importance through Advanced Variable Selection Techniques:

The third objective is to conduct a comprehensive assessment of feature importance to identify the top five independent variables that significantly impact the chosen dependent variables. Utilizing advanced variable selection techniques supported Machine Learning models such as Random Forest or Gradient Boosting, I will rank these features based on their predictive contributions. This analysis will be instrumental in highlighting critical areas for policy intervention, enabling stakeholders to focus on the most impactful factors driving economic development.

4. Formulation of Evidence-Based Policy Recommendations:

The final objective is to synthesize findings from the regression and classification analyses into actionable policy recommendations. By understanding which factors are most influential in determining economic outcomes, this research aims to inform government initiatives and community programs designed to enhance the quality of life for tribal populations in Jamtara. These recommendations will be grounded in empirical evidence, ensuring that they address specific needs and challenges faced by these communities.

Literature Review

Research on tribal communities in Jharkhand highlights the unique socio-economic challenges faced by Indigenous groups, often worsened by structural inequalities. In "Socio-Economic Issues of the Tribals in Jharkhand," Shekhar (Gargi College, University of Delhi) discusses the hardships these communities endure, particularly displacement from industrial expansion, low income, limited healthcare, and inadequate education. The study advocates for sustainable development policies that prioritize tribal needs.

Kumari's work, "Land and Property Rights Among Tribal Communities in Jharkhand" (Central University of Jharkhand), addresses land rights issues crucial for tribal livelihoods. It emphasizes integrating statutory and customary land relations and suggests that equitable land governance through inclusive strategies can reduce socio-economic disparities.

In "Enhancing the Socio-Economic Empowerment of Tribal Communities with Entrepreneurship and Skill Training Program in Jharkhand," Kumari (Vinoba Bhave University) explores how entrepreneurship and skill development programs can improve income and social inclusion. She argues that initiatives like the Grameen Udyami Yojana and BIRSA are vital for fostering self-reliance and gender equality among tribal women.

Raul, Majumdar, and Chatterjee (RKMVERI, Ranchi Campus) investigate the resilience of the Sauria-Paharia tribe in "Indigenous Knowledge and Endogenous Development**:** Exploring Survival Strategies of a Tribal Community of Jharkhand, India" highlighting the importance of traditional practices like shifting cultivation for survival amid globalization pressures.

Nandru and Rentala (2019) examine financial inclusion among Particularly Vulnerable Tribal Groups (PTGs) in "Demand-side analysis of measuring financial inclusion: Impact on socio-economic status of Primitive Tribal Groups (PTGs) in India". They identify five dimensions of financial inclusion and use Structural Equation Modelling (SEM) to show that financial access is a strong predictor of socio-economic improvement.

Bhattacharya (2023) evaluates the Public Distribution System (PDS) and the National Rural Employment Guarantee Act (NREGA) in "Effectiveness of public distribution system (PDS) in quality of life improvement – a study of the tribal population of Purulia district in West Bengal" identifying challenges that hinder PDS effectiveness and advocating for universal access to enhance nutritional security for tribal populations.

Majhi (2018) discusses the economic challenges of tribal fishing communities in Purulia in "A Comparative Study of Traditional Fishing Practice Among Tribal People at Some Selected Regions of Purulia District", concluding that targeted resource management and technical support could improve socio-economic conditions. Similarly, Chakrabarty and Bharati (2012) examine the Shabar tribe's household economy in "Household Economy and Nutritional Status in the Shabar Tribe of Orissa", noting adverse impacts of conservation policies and recommending community involvement in conservation efforts.

Ray, Rout, and Ray (2020) analyse banking service utilization in rural and tribal areas in "How do households utilize banking services and what are the determinants of it? An empirical analysis from the rural and tribal areas of an eastern Indian state", emphasizing that accessibility and actual utilization of banking services are vital for socio-economic growth.

Saxena and Bhattacharya (2017) highlight systemic biases affecting marginalized groups' access to LPG and electricity in "Inequalities in LPG and electricity consumption in India: The role of caste, tribe, and religion", advocating for improvements in supply infrastructure. Rashmi and Paul (2024) address the educational wellbeing gap between Scheduled Tribe and Non-Scheduled Tribe children in "Early childhood circumstances and educational wellbeing inequality among tribal and non-tribal children in India: Evidence from a panel study", calling for early childhood interventions to mitigate inequalities.

Xing (2024) explores happiness prediction using advanced regression models in “Global Happiness Ranking: An Ensemble Regression Vs. Traditional Approach Investigation”, finding GDP per capita as a significant predictor with XgBoost being the most apt Machine Learning model. Potdar et al. (2017) conduct a study on categorical variable encoding techniques in “A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers”, advocating for Sum Coding, Backward Difference Coding, One Hot Encoding and Ordinal Encoding for better accuracy. Lastly, Tsai and Hu (2022) investigate missing value imputation methods in “Empirical comparison of supervised learning techniques for missing value imputation”, finding CART superior for categorical data.

Research Gaps

1. Insufficient Focus on Tribal Economies:

    - While many studies have addressed economic development in various contexts, there is a lack of specific focus on the economic conditions of tribal communities, particularly in regions like Jharkhand. Existing literature often generalizes findings without delving into the unique socio-economic challenges faced by these populations.

2. Limited Use of Advanced Analytical Techniques:

    - Most previous research has utilized basic statistical methods for analyzing economic development in tribal areas. There is a gap in employing advanced analytical approaches, such as regression models and feature importance analysis, to derive actionable insights that can lead to effective policy formulations.

3. Neglect of Feature Importance in Policy Development:

    - Although several studies have identified socio-economic factors influencing development, they frequently overlook the systematic identification of critical features that contribute to economic progress. There is an opportunity to conduct feature importance analysis that can directly inform policymakers about the most impactful variables for intervention.

4. Lack of Empirical Evidence for Policy Formulation:

    - The existing body of literature often lacks empirical studies that demonstrate the practical applications of identified features in shaping policies aimed at improving the economic status of tribal communities. This research will fill that gap by providing empirical evidence derived from the analysis of relevant data.

5.Need for Contextual Understanding:

- Many studies fail to account for the specific cultural and economic contexts of tribal populations when proposing solutions. This research aims to bridge this gap by ensuring that the selected features for analysis are deeply rooted in the realities of the tribal communities in Jharkhand.

Motivation for the Research

1. Supporting Sustainable Development:

    - The motivation behind this research stems from the critical need for sustainable economic development in tribal areas of Jharkhand. By focusing on identifying the most influential features affecting economic status, the research aims to contribute to the formulation of policies that foster sustainable growth and improve the living conditions of these communities.

2. Enhancing Government Efficiency:

    - By distilling a comprehensive set of features into the top five most critical variables, this research seeks to enhance governmental efficiency in policy formulation. Policymakers often face information overload; providing them with targeted insights will allow for more strategic decision-making and resource allocation.

3. Empowering Tribal Communities:

    - Empowerment of tribal communities through informed policy interventions is a primary motivation. By identifying the key factors influencing their economic development, the research aims to provide actionable insights that can lead to tangible improvements in their socio-economic conditions.

4. Filling an Academic Void:

    - There is a personal and academic motivation to contribute to the literature on tribal economics and advanced analytical methods. This research aspires to add valuable insights that can be referenced by future scholars and practitioners working in similar fields.

5.Interdisciplinary Approach:

- The research takes an interdisciplinary approach by merging insights from economics, sociology, and data science. This blend not only enriches the analysis but also provides a holistic understanding of the factors at play in tribal economic development, thereby appealing to a broader audience of researchers and practitioners.

6. Potential for Real-World Impact:

    - Ultimately, the research is motivated by the potential for real-world impact. By translating analytical findings into policy recommendations, the study aims to bridge the gap between research and practical application, ensuring that the insights generated can lead to improved livelihoods for tribal populations in Jharkhand.

Research Methodology

This study employs a structured, quantitative research methodology to explore and analyse the socio-economic development of tribal communities in Jamtara District, Jharkhand. Given the focus on evaluating economic conditions, disparities, and development outcomes within marginalized communities, a quantitative approach allows for a rigorous statistical examination of relationships among various socio-economic indicators. By relying on a data-driven framework, this methodology aims to uncover key patterns and causal impacts that inform the socio-economic realities faced by tribal populations in Jamtara.

The research draws primarily on secondary data sources to ensure the comprehensiveness of socio-economic factors under investigation. This approach not only enhances the scope of analysis but also allows for the systematic examination of variables such as income levels, educational attainment, access to healthcare, infrastructure availability, and employment trends within these communities. By structuring the data collection and analysis process, this methodology facilitates robust findings that align with theoretical frameworks relevant to development studies and social inequality. Furthermore, secondary data sources provide a broad foundation, enabling the study to situate its findings within a wider regional and national context for comparative insight.

1. Nature of Research

The nature of this research is inherently quantitative, focused on numerical analysis to identify socio-economic patterns and correlations across various indicators affecting tribal communities. This approach is particularly suitable given the study's objective of assessing development disparities and socio-economic conditions among marginalized groups. A quantitative framework facilitates the use of statistical tools, enabling objective, data-driven conclusions that enhance the reliability and validity of the findings. Statistical methods, such as regression analysis, correlation coefficients, and descriptive statistics, are applied to quantify relationships among variables—such as income disparities, access to resources, literacy rates, and healthcare access—highlighting the

depth of socio-economic inequality and identifying key development needs within Jamtara's tribal population.

Moreover, this quantitative analysis offers insights into systemic barriers and enablers of socio-economic growth, allowing for a more focused understanding of development challenges faced by tribal communities in Jamtara. Given that quantitative methodologies provide robust evidence, they are invaluable for assessing causal relationships and establishing statistically significant trends, which are essential in addressing complex social issues and in shaping future policy recommendations for sustainable tribal development.

This rigorous quantitative approach, rooted in secondary data analysis, thus serves as the backbone of the study, providing a structured pathway to evaluate and interpret the socio-economic landscape of Jamtara's tribal communities with statistical precision and empirical depth.

2. Research Design

This study adopts a cross-sectional research design, selected for its effectiveness in examining and comparing multiple socio-economic variables at a single point in time. This design is particularly advantageous when assessing conditions and identifying correlations across diverse socio-economic indicators within tribal communities. By capturing data on various socio-economic factors—such as income, education, healthcare access, and infrastructure—within a specified timeframe, this approach provides a comprehensive snapshot of the socio-economic landscape in Jamtara District, Jharkhand.

The cross-sectional design is ideal for this study as it enables the identification of patterns and relationships without requiring a prolonged data collection period, which is often challenging in resource-limited settings. With this approach, we can analyse how specific factors, like education levels and employment opportunities, correlate with economic outcomes and living standards among tribal populations. It also allows for comparisons across different demographic groups within the community, such as age, gender, and

household size, to understand better the socio-economic disparities and inequalities present within the population.

3. Population and Sample

The study population includes all tribal households within the Jamtara District of Jharkhand, selected for its diverse socio-economic and demographic composition. To ensure that the findings accurately represent this varied population, a representative sample was drawn to allow for generalizable conclusions across the district's tribal communities.

3.1 Sample Size and Sampling Technique

A stratified sampling technique was chosen to account for the diversity within the tribal population, allowing subgroups—defined by characteristics such as age, gender, and socio-economic status—to be proportionately represented. This approach enhances the accuracy and relevance of the findings, ensuring that key demographic and economic variations within the community are captured. The sample size was determined through statistical power analysis, designed to secure a robust number of respondents for reliable statistical inference. A total of approximately 500 households were surveyed, providing a solid foundation for rigorous analysis and enabling the study to draw meaningful insights into the socio-economic conditions affecting tribal households across Jamtara District.

4. Data Collection

This study relies on secondary data, provided directly by supervising professor, who was collaborating with the Government of Jharkhand on a socio-economic development project centered on the tribal households in Jamtara District.

4.1 Data Sources

The dataset originates from socio-economic surveys conducted under the direction of professor in collaboration with the Government of Jharkhand. These surveys were designed to capture a wide range of socio-economic indicators, covering areas such as demographic characteristics, income levels, employment, healthcare access, and education. This comprehensive dataset serves as a reliable foundation for examining correlations and patterns essential to the study's objectives.

4.2 Overview of Variables

Upon receipt, the data was organized in an Excel file, which facilitated efficient data handling and analysis in Python. This structured format supported the application of a range of statistical techniques, ensuring that the data could be systematically explored and manipulated within the analytical framework of the study. The data encompassed a wide array of variables, each targeting specific socio-economic indicators relevant to the study. Below are detailed explanations of each variable:

1. Age: The age of the respondent, which is a critical demographic factor influencing socio-economic status.
2. Gender: Gender of the respondent, encoded using One Hot Encoding (OHE) to facilitate analysis of gender-specific impacts.
3. Religion: Respondents' religious affiliation, also encoded using OHE to allow for comparative analysis across different religious groups.
4. Caste Category: Caste classification of the respondent, encoded for analysis of caste-related socio-economic disparities.
5. If ST then: A variable indicating whether the respondent belongs to a Scheduled Tribe (ST), encoded using OHE.
6. Marital Status: The marital status of the respondent, used to assess its impact on economic outcomes, encoded using OHE.
7. Occupation: The primary occupation of the respondent, encoded to analyse the economic contributions of different occupational groups.

8. Origin (migrants/non-migrants): A categorical variable indicating if the respondent is a migrant or non-migrant, which can influence social dynamics and economic opportunities.
9. Information of Birth during Last 5 Years (Male/Female): Counts of births within the household, categorized by gender, to assess growth and health trends.
10. Information of Death during Last 5 Years (Male/Female): Counts of deaths within the household, categorized by gender, to evaluate mortality trends.
11. School Dropout (Male/Female): Indicates whether male or female children in the household have dropped out of school, providing insight into educational challenges.
12. PDS Card (Yes/No): A binary variable indicating access to the Public Distribution System, critical for understanding food security.
13. Type of House (Kaccha/Pakka): A categorical variable indicating the type of housing, relevant for assessing living conditions.
14. Bank Account (Yes/No): Indicates whether the household has a bank account, an important factor for financial inclusion.
15. Electricity Connection (Yes/No): Indicates access to electricity, a key determinant of quality of life.
16. LPG Connection (Yes/No): Indicates access to liquefied petroleum gas for cooking, which impacts health and time efficiency.
17. Deity Worshipped: The primary deity worshipped by the household, encoded to analyse cultural influences on economic behaviour.
18. Dialect Spoken: The dialect spoken at home, which could influence communication and access to resources.
19. Traditional Folk Dance: Indicates participation in traditional folk dances, useful for understanding cultural engagement.
20. Traditional Festival: Participation in traditional festivals, another measure of cultural involvement.
21. Activity Done in Free Time: Types of activities engaged in during leisure time, relevant for assessing community engagement.
22. Go to Community Library (Yes/No): Indicates access to educational resources, important for gauging literacy and learning opportunities.

23. Knowledge about Elderly Club at Panchayat (Yes/No): Awareness of community support structures for the elderly, relevant for assessing social support systems.
24. Decision Maker in Family: Indicates who in the household makes critical decisions, providing insight into gender roles and power dynamics.
25. Personality Type: Encodes personality traits, useful for understanding individual behaviour in economic decision-making.
26. Participation in Gram Sabha or Social Meetings (Yes/No): Indicates civic engagement and participation in local governance.
27. Approached Elected Representatives for Problem Resolution (Yes/No): Indicates civic engagement and responsiveness of local governance.
28. Problem Resolution Satisfactory: A measure of satisfaction with the outcomes of problem resolution efforts.
29. Lifestyle Description: Encodes various lifestyle choices, relevant for understanding consumption patterns.
30. Food Intake Habits: Encodes dietary habits, important for health and nutrition analysis.
31. Diet Choice (Best Possible Option): A measure of dietary preferences, useful for assessing nutritional quality.
32. Participation in Tribal Cultural Programs (Yes/No): Indicates engagement in cultural preservation activities.
33. Experience of Violence or Illegal Activities (Yes/No): A measure of safety and security within the community.
34. Life Satisfaction Measures: Multiple items assessing life satisfaction over various time frames, providing insight into overall well-being.
35. Happiness Measures: Similar to life satisfaction, measures of happiness at different time intervals.
36. Anxiety Measures: Questions assessing anxiety levels, relevant for mental health analysis.
37. Work Life Satisfaction Measures: Questions regarding satisfaction with work life, providing insight into economic engagement.
38. Women's Safety Rating: A measure of perceived safety for women in the community.

39. Positive and Negative Experience Measures: Questions assessing emotional well-being over the past four weeks.
40. Self-Perception of Life Quality: Measures reflecting self-assessment of life quality and future optimism.
41. Health Satisfaction Measures: Questions assessing satisfaction with health services.
42. Income and Expenditure Evaluations: Measures assessing perceptions of income and expenditure, relevant for economic analysis.
43. Caste Inequality in Village: A binary measure assessing perceived caste disparities.
44. Land Area and Agricultural Measures: Various quantitative measures related to agricultural land, crop area, and irrigation sources, crucial for understanding agricultural economics.
45. Livestock and Agricultural Inputs (Yes/No): Measures regarding livestock ownership and agricultural inputs used, important for assessing economic activity.
46. Participation in Skill Development Programs (Yes/No): Indicates community engagement in skill enhancement initiatives.
47. Water Quality and Accessibility Measures (Yes/No): Various questions assessing access to safe drinking water and sanitation facilities.
48. Health Risks and Awareness (Yes/No): Measures assessing awareness of health risks and participation in health programs.
49. Child Mortality and Health Facility Measures: Questions related to maternal health, child mortality, and perceptions of health services.
50. School Attendance and Education Measures (Yes/No): Questions assessing school attendance and educational attainment within households.

6. Summary of Respondents

A demographic analysis was conducted to capture the diversity within Jamtara District's tribal communities, examining key variables such as age, gender, education, and household size. These factors provide essential context for interpreting socio-economic conditions and access to resources across demographic groups.

- Age: Respondents were grouped by age to assess differences in socio-economic experiences, such as educational aspirations among youth versus household priorities among older adults.
- Gender: Gender analysis highlighted disparities in access to services, employment, and education, revealing potential differences in income levels and quality of life between men and women.
- Education Level: Educational attainment was categorized to examine its correlation with income, employment status, and overall quality of life, helping to identify its role in economic outcomes.

7. Statistical Analysis Techniques

The study employed both descriptive and inferential statistical methods to analyse the socio-economic data of tribal households in Jamtara District, allowing for a thorough understanding of variable distributions and interrelationships.

- Descriptive Statistics: Key metrics like means, medians, and standard deviations were calculated for variables such as income, expenditure, education, and employment status. Frequency distributions and cross-tabulations highlighted demographic-specific conditions, such as literacy rates and service access, forming a foundational view of the socio-economic landscape.
- Inferential Statistics: Inferential methods, particularly regression analysis, were used to examine the impact of factors like education and household size on outcomes such as income and healthcare access. Classification analyses helped identify high-risk groups facing socio-economic disadvantages, facilitating targeted insights.
- Correlation Analysis: Correlation techniques explored associations between variables, such as the relationship between income and household size or education and employment, revealing significant connections that influence overall well-being.

8. Regression and Classification Models

The model development phase involved building various statistical models to analyse relationships between independent and dependent variables. This process included both regression and classification techniques tailored to the nature of the data and research objectives.

8.1 Regression Model Development

In this study, regression analysis was utilized to quantify the relationships between independent variables and continuous dependent variables, such as annual income and expenditure patterns. This method provides a framework for understanding how socio-economic factors influence economic outcomes within tribal communities in Jamtara District. The following regression models were implemented:

1. Random Forest Regressor: Implemented using `RandomForestRegressor(n_estimators=100, random_state=42)`, this ensemble learning method builds multiple decision trees and averages their predictions. It effectively captures complex relationships in the data and is robust against overfitting, making it suitable for this analysis.

2. Decision Tree Regressor: A fundamental approach that uses `DecisionTreeRegressor(random_state=42)` to partition the data into branches based on feature values. This model is easy to interpret and can capture non-linear relationships, although it may be prone to overfitting if not properly tuned.

3. CatBoost Regressor: This model, implemented via `CatBoostRegressor(verbose=0, random_state=42)`, is designed specifically to handle categorical features efficiently. CatBoost is known for its high performance with minimal parameter tuning and is particularly effective in structured data scenarios.

4. XGBoost Regressor: Using `XGBRegressor(n_estimators=100, random_state=42)`, this model leverages gradient boosting techniques to optimize the prediction accuracy. XGBoost is recognized for its speed and efficiency, particularly in handling large datasets.

5. Gradient Boosting Regressor: Implemented with `GradientBoostingRegressor(n_estimators=100, random_state=42)`, this method builds models sequentially, where each new model attempts to correct the errors made by the previous ones. It is particularly useful for improving model performance through iterative refinement.

6. AdaBoost Regressor: The `AdaBoostRegressor(n_estimators=100, random_state=42)` combines multiple weak learners to create a strong predictive model. This approach is effective in enhancing the performance of simple models.

7. Bagging Regressor: Implemented using `BaggingRegressor(n_estimators=100, random_state=42)`, this model reduces variance by averaging predictions from multiple models trained on different subsets of the data.

The performance of each regression model was evaluated using several key metrics to ensure robustness and accuracy:
- Mean Squared Error (MSE): This metric calculates the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.
- Root Mean Squared Error (RMSE): The square root of MSE, RMSE provides an interpretable metric in the same units as the dependent variable, offering insight into the average error in predictions.
- Mean Absolute Error (MAE): This metric measures the average magnitude of the errors in a set of predictions, treating all errors equally without considering their direction.

- R^2 Score: This statistic indicates the proportion of variance in the dependent variable that can be explained by the independent variables in the model. A higher R^2 score suggests a better fit of the model.
- Adjusted R^2 Score: This adjusts the R^2 score based on the number of predictors in the model, providing a more accurate measure for models with multiple independent variables.

8.2 Classification Model Development

In addition to regression models, classification algorithms were employed to categorize households based on their socio-economic status or access to resources. These classification techniques play a crucial role in understanding the factors that classify households into specific socio-economic groups. The following classification models were implemented in this study:

1. Random Forest Classifier: Implemented with `RandomForestClassifier(n_estimators=100, random_state=42)`, this ensemble learning method constructs multiple decision trees during training. By combining the predictions from these trees, Random Forest enhances predictive accuracy while significantly reducing the risk of overfitting. It is particularly effective in handling diverse feature sets and capturing complex interactions among variables.

2. Decision Tree Classifier: Utilizing `DecisionTreeClassifier(random_state=42)`, this model partitions the data into branches based on feature values. Decision trees are intuitive and easy to interpret, allowing for straightforward visualization of decision paths. However, they can be prone to overfitting if not properly pruned.

3. CatBoost Classifier: The `CatBoostClassifier(verbose=0, random_state=42)` is designed to efficiently handle categorical features and complex datasets. Its capacity to work with categorical data without extensive pre-processing makes it a valuable tool for this analysis, while its gradient boosting framework enhances model performance.

4. XGBoost Classifier: Implemented via `XGBClassifier(n_estimators=100, random_state=42)`, XGBoost is recognized for its speed and performance in classification tasks. It employs gradient boosting techniques to optimize accuracy and is particularly effective for large datasets.

5. Gradient Boosting Classifier: Using `GradientBoostingClassifier(n_estimators=100, random_state=42)`, this model builds classifiers sequentially, where each new model attempts to correct errors made by the previous ones. This iterative approach helps improve model accuracy significantly.

6. AdaBoost Classifier: The `AdaBoostClassifier(n_estimators=100, random_state=42)` combines multiple weak learners to create a strong predictive model. The focus on misclassified instances from previous learners enhances overall model performance.

7. Bagging Classifier: Implemented with `BaggingClassifier(n_estimators=100, random_state=42)`, this model reduces variance by averaging the predictions from multiple classifiers trained on different subsets of the data.

The performance of each classification model was rigorously evaluated using several key metrics, which are essential for assessing the effectiveness of the models:

- Accuracy: This metric indicates the proportion of correctly classified instances over the total instances, providing a straightforward measure of overall performance.
- F1-Score: The F1-score is the harmonic mean of precision and recall. It is particularly useful for imbalanced datasets as it provides a balance between false positives and false negatives.
- Precision: This metric measures the ratio of true positive predictions to the total predicted positives, indicating the accuracy of the positive predictions made by the model.
- Recall: Also known as sensitivity, recall measures the ratio of true positives to the actual positives, reflecting the model's ability to identify relevant instances.

9. Feature Importance Analysis

Feature importance analysis is essential for understanding which factors play the most significant roles in predicting socio-economic outcomes within the tribal communities of Jamtara District. By identifying key features that influence dependent variables in both regression and classification models, this analysis highlights critical areas that may benefit from targeted policy interventions, helping prioritize efforts to improve economic and social well-being within these communities.

9.1 Feature Importance Calculation Using model.feature_importances_

In tree-based machine learning models such as Random Forests and Gradient Boosting Machines, feature importance can be assessed directly through the model.feature_importances_ attribute in Scikit-learn. This attribute provides a quantitative score for each variable, indicating its contribution to the model's predictive accuracy by measuring how much it reduces impurity (e.g., Gini impurity or entropy) across all trees in the ensemble.

- The feature_importances_ attribute assigns a score to each feature, reflecting its overall contribution to improving model accuracy. Higher scores indicate that the feature is more critical for the model's predictions. For example, if the feature "access to healthcare facilities" has a high importance score, it suggests a substantial impact on outcomes, such as income level or quality of life.
- Calculation Methodology: Feature importance scores are calculated by evaluating the impact of each feature on impurity reduction at each split across all decision trees in the model. Specifically:
    - Each time a feature is used to split a node within a tree, it contributes to reducing the weighted impurity at that point in the model.

  - This reduction in impurity is aggregated for each feature across all trees and then normalized by the total number of trees, resulting in an average importance score for each feature.

  - For instance, in a model predicting economic outcomes, features like household size, access to PDS (Public Distribution System) cards, or education level may appear frequently in splits, thus gaining high importance scores if they consistently reduce impurity.

- Interpretation: Features with higher scores are considered more influential in determining the model's predictions. For instance, if "household income" has a higher feature importance score than "marital status" or "occupation," it suggests that household income is a stronger predictor of economic outcomes in the model. This insight helps direct resources toward interventions that could have a substantial impact, such as improving income-generating activities or expanding access to resources that directly affect economic well-being.

9.2 Visualization of Feature Importance

To enhance the interpretability of feature importance scores, visualization techniques are employed, which allow for a clearer understanding of which variables are most impactful.

- Bar Plots: Bar plots are used to display feature importance scores for each variable in a visually accessible format. By representing importance scores on a scale, stakeholders can quickly see the relative influence of different features. For example, a bar plot showing high scores for "access to education" and "access to healthcare" might visually emphasize the need for interventions in these areas.

- Feature Importance Ranking: Ranking features based on their importance scores provides a prioritized list of variables, facilitating targeted decision-making. By focusing on top-ranked features, policymakers and researchers can prioritize efforts on the most impactful variables, such as access to essential services, household economic status, or educational attainment.

10. Mitigation of Research Biases

To enhance the validity and reliability of this study's findings on Jamtara's tribal communities, several strategies were implemented to mitigate potential biases in secondary data analysis.

- Data Cleaning: A thorough cleaning process addressed inaccuracies, outliers, and missing values through methods like imputation and categorical encoding. This minimized the impact of data quality issues and improved result accuracy.
- Transparent Methodology: Each analytical step, from pre-processing to model selection, was meticulously documented. This transparency promotes reproducibility, allowing others to replicate the study and verify its findings.
- Cross-Validation: K-fold cross-validation was used to train models, reducing the risk of overfitting and ensuring consistent model performance across data subsets. This approach enhances the generalizability of results to new data.
- Control of Confounding Variables: Potential confounders, such as age, gender, and household size, were accounted for in the statistical models to prevent misleading associations. This control improved the accuracy of identified socio-economic relationships.
- Expert Consultation: Regular input from subject matter experts and the supervising professor helped validate interpretations, adding context-specific insights and reducing overlooked biases.

11. Rationale for Chosen Methods

The methodologies employed in this research were selected to rigorously analyse the complex socio-economic conditions of tribal communities in Jamtara District, Jharkhand. Given the study's objective of generating data-driven insights to inform policy, a quantitative approach was deemed most suitable. This approach allows for the objective measurement of relationships among multiple socio-economic variables, enabling a

precise evaluation of factors that impact economic and social well-being within marginalized populations.

The chosen methods include descriptive and inferential statistical techniques, regression analysis, and feature importance analysis. Each of these methods brings specific advantages:

- Descriptive Statistics: Descriptive methods were selected to provide a clear overview of the data and identify basic patterns across demographic and socio-economic variables. By summarizing factors such as household income, education levels, and healthcare access, descriptive analysis offers a foundational understanding of the socio-economic landscape in Jamtara, setting the stage for more detailed analysis.

- Inferential Statistics: Inferential techniques, including regression and correlation analyses, allow for the examination of relationships between independent and dependent variables. These methods are particularly valuable in identifying the impact of specific socio-economic factors—such as education level, household size, and access to services—on outcomes like income and quality of life. By establishing statistically significant associations, inferential methods enable the research to draw evidence-based conclusions that inform targeted policy recommendations.

- Feature Importance Analysis: Feature importance analysis was chosen to pinpoint which variables most strongly influence the outcomes being studied. This is particularly useful for policy development, as it highlights the key areas that have the highest potential for impact. Tree-based machine learning models, such as Random Forests, provide robust feature importance scores, helping prioritize interventions that address the most pressing needs within tribal communities.

Data Analysis

The analysis phase of this research involved a systematic approach to data preparation and model implementation using Python, ensuring that the dataset was suitable for drawing meaningful insights regarding the socio-economic development of tribal communities in Jamtara District. The analysis consisted of several key steps, as outlined below.

1. Importing Necessary Libraries

The first step in the coding process involved importing essential Python libraries that facilitate data manipulation, statistical analysis, and visualization. The primary libraries used include:

- Pandas (2.2.2): For data manipulation and analysis, particularly for handling DataFrames.
- NumPy (1.26.4): For numerical computing and array operations.
- Scikit-learn (1.4.2): For implementing machine learning algorithms and model evaluation.
- Matplotlib (3.8.4) and Seaborn (0.13.2): For data visualization to illustrate key findings.

2. Reading the Master Excel File

The dataset was read into a Pandas DataFrame using the `pd.read_excel()` function, allowing for efficient handling of the data. The initial inspection included checking the shape of the DataFrame ‘df.shape()’ and displaying the first few records ‘df.head()’ to gain an understanding of the structure and content of the data.

3. Data Preprocessing

Data preprocessing was a critical phase that addressed various issues within the dataset to ensure its quality and relevance for analysis:

3.1 Duplicate Rows Removal

Duplicate entries were identified using `df.duplicated()` and eliminated with `df.drop_duplicates()` to maintain data integrity.

3.2 Null Value Check

A comprehensive assessment for null values was performed using `df.isnull().sum()`, which provided a count of missing values per column.

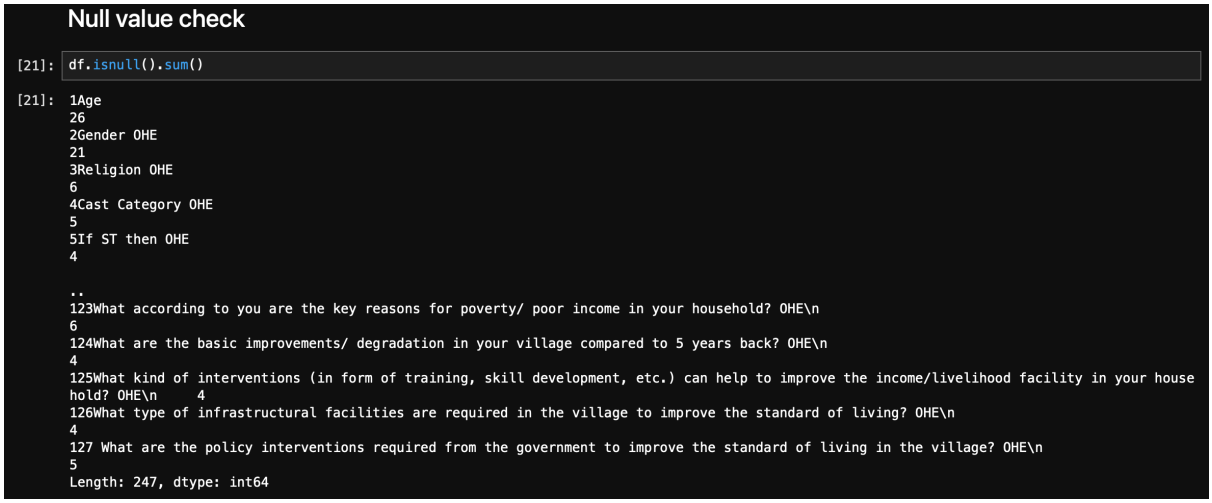


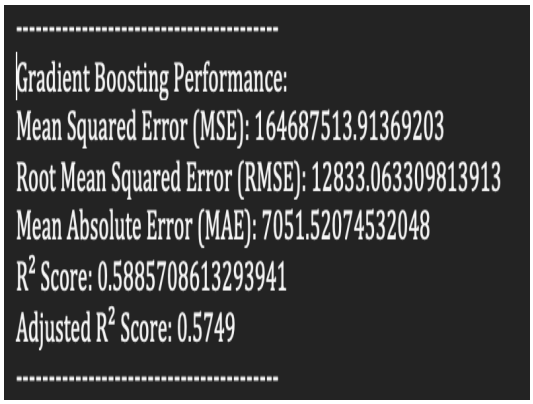
Figure 1

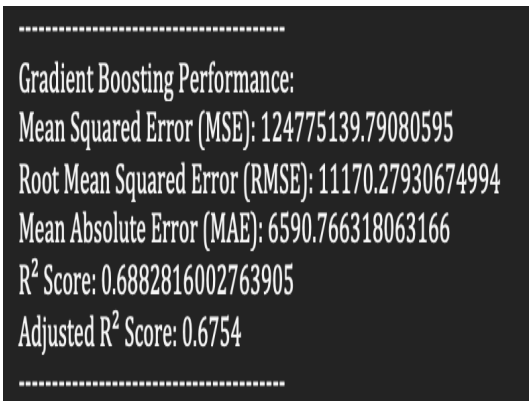
3.3 Independent Columns Removal (based on percentage of null values)

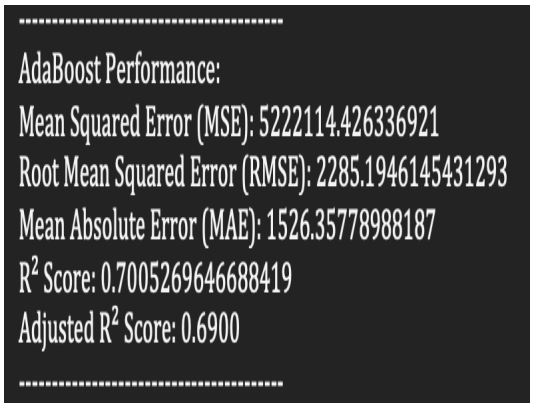
To enhance the reliability of the dataset, columns exhibiting an excessive proportion of null values—specifically, those with over 40% missing data—were systematically removed. This step is pivotal in safeguarding the integrity of the data, as retaining features with a high prevalence of missing values can result in biased analyses and diminish the reliability of model predictions. By meticulously evaluating the extent of null values across each column, only those features that offer meaningful insights and positively contribute to the overall quality of the dataset were preserved. This process not only

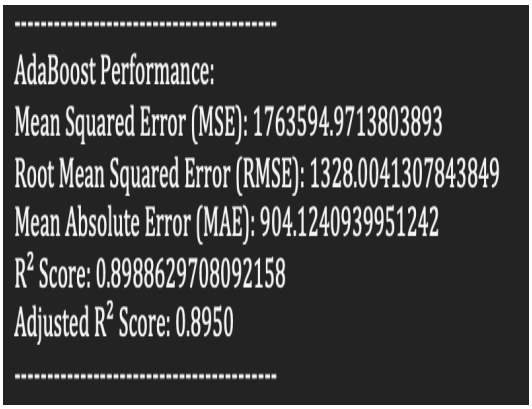
streamlines the dataset for more efficient analysis but also strengthens the robustness of subsequent modelling efforts.

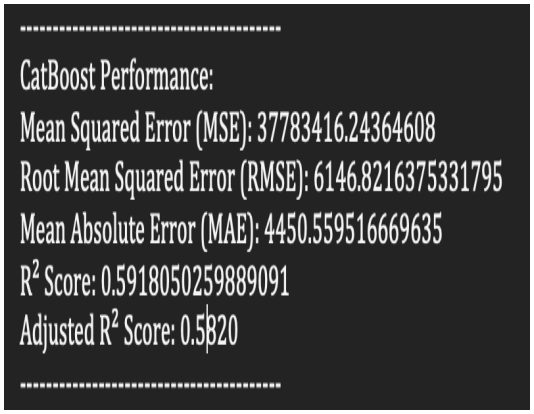
Given the lack of supporting literature on this topic, two separate models were developed for each of the seven identified dependent variables. The first model for each variable included the columns with greater than 40% null values, while the second model excluded these columns. The performance of both models was evaluated based on the Adjusted R^2 Score and R^2 Score for the regression models, and the Accuracy Score for the classification models. Notably, for all seven variables examined, the second models consistently outperformed the first, indicating that the exclusion of columns with excessive null values contributed to improved predictive performance.

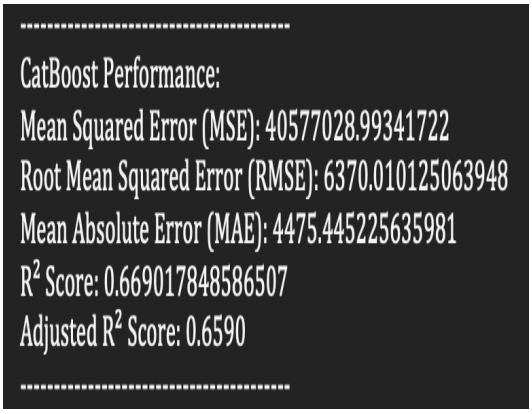


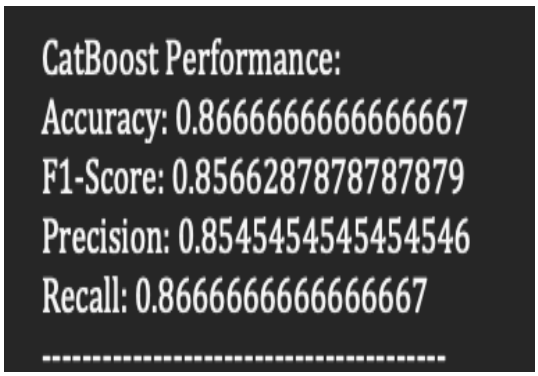


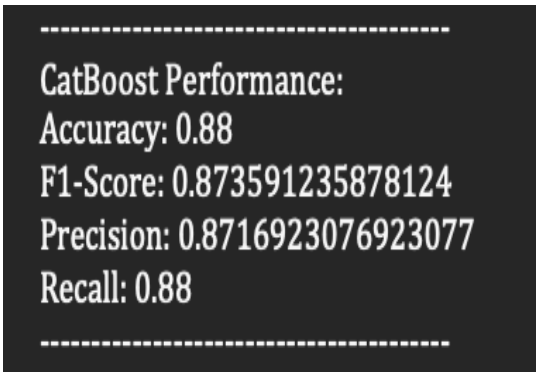


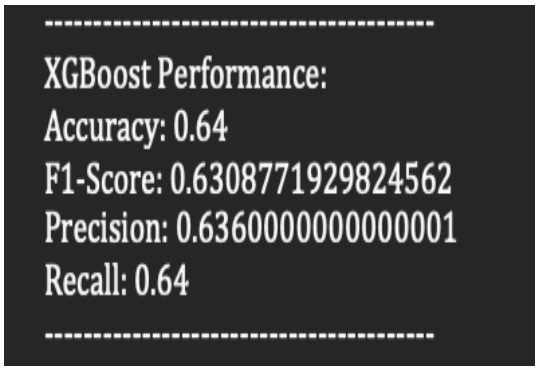


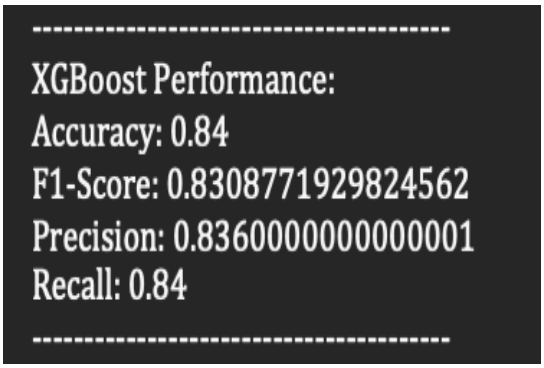

CatBoost Performance:
Mean Squared Error (MSE): 37783416.24364608
Root Mean Squared Error (RMSE): 6146.8216375331795
Mean Absolute Error (MAE): 4450.559516669635
R^2 Score: 0.5918050259889091
Adjusted R^2 Score: 0.5820


CatBoost Performance:
Mean Squared Error (MSE): 40577028.99341722
Root Mean Squared Error (RMSE): 6370.010125063948
Mean Absolute Error (MAE): 4475.445225635981
R^2 Score: 0.669017848586507
Adjusted R^2 Score: 0.6590

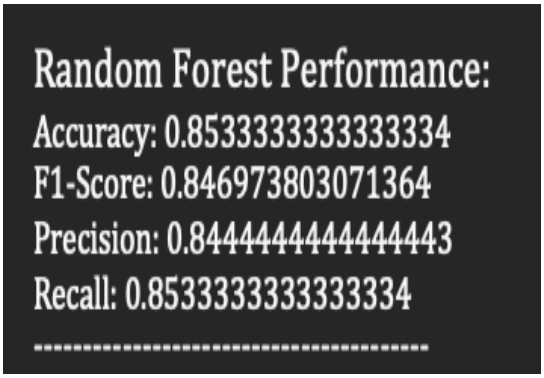

CatBoost Performance:
Accuracy: 0.8666666666666667
F1-Score: 0.8566287878787879
Precision: 0.8545454545454546
Recall: 0.8666666666666667


CatBoost Performance:
Accuracy: 0.88
F1-Score: 0.873591235878124
Precision: 0.8716923076923077
Recall: 0.88


XGBoost Performance:
Accuracy: 0.64
F1-Score: 0.6308771929824562
Precision: 0.6360000000000001
Recall: 0.64


XGBoost Performance:
Accuracy: 0.84
F1-Score: 0.8308771929824562
Precision: 0.8360000000000001
Recall: 0.84


Random Forest Performance:
Accuracy: 0.8266666666666667
F1-Score: 0.8191508581752484
Precision: 0.8150793650793651
Recall: 0.8266666666666667


Random Forest Performance:
Accuracy: 0.8533333333333334
F1-Score: 0.846973803071364
Precision: 0.8444444444444443
Recall: 0.8533333333333334

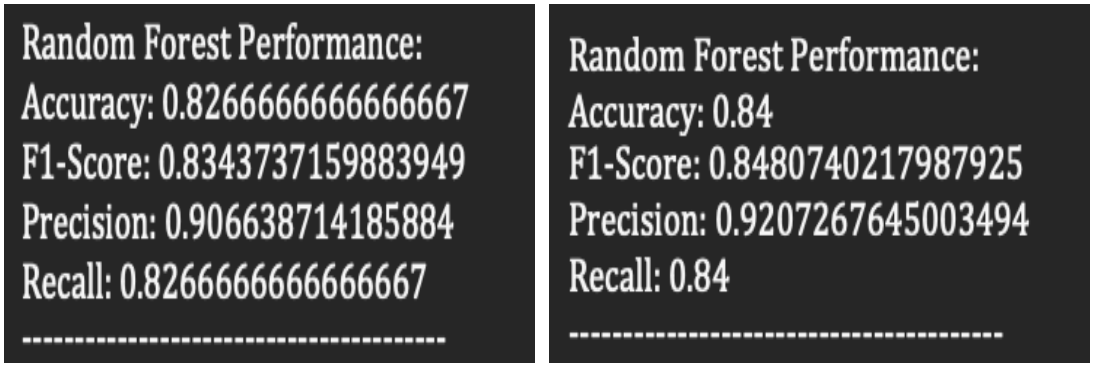


Figure 2

In all the figures above, the 2nd model shows better metrics in terms of Adjusted R^2 Score and R^2 Score for the regression models, and the Accuracy Score for the classification models.

Hence removing those columns which have greater than 40% null values.

3.4 Categorical and Numerical Columns Identification

To facilitate more accurate statistical analyses and model development, all columns within the dataset were categorized as either numerical or categorical. This classification enables the application of appropriate data preprocessing techniques tailored to the nature of each variable type, thus ensuring that each feature is handled in a way that aligns with its inherent characteristics.

Numerical columns, which represent quantifiable data, were prepared for scaling, transformation, and normalization techniques to improve model convergence and performance. On the other hand, categorical columns, which denote qualitative data, were preprocessed through encoding techniques suited for categorical variables, such as Ordinal Encoding or One-Hot Encoding, to transform them into a format usable by machine learning algorithms. This distinction between categorical and numerical columns not only streamlines the preprocessing workflow but also enhances the interpretability and efficacy of statistical analyses, as it allows for more precise Modelling based on the data type.

3.5 Handling Non-Numeric Data in Numerical Columns

In this study, the presence of non-numeric entries within columns designated as numerical posed a potential obstacle to accurate data analysis and model performance. These non-numeric entries, which often represented categorical responses or data entry anomalies, required careful handling to ensure consistency and validity across the dataset.

To address this, non-numeric entries within numerical columns were systematically identified and processed. Where feasible, categorical responses embedded in these columns were transformed into binary or ordinal formats, thus preserving the information while enabling numerical analysis. For example, responses indicating the presence or absence of a condition or characteristic (e.g., “Yes” or “No”) were converted into binary representations (1 for “Yes,” 0 for “No”). In cases where categorical values represented a range or level, such responses were encoded in a way that retained their ordinal relationships, allowing them to be analyzed quantitatively without introducing bias.

3.6 Imputing Missing Values

To ensure data completeness and enhance the reliability of the subsequent analysis, missing values in the dataset were imputed using established statistical methods. For numerical variables, the mean or median was used as the imputation measure, depending on the distribution of the data. Specifically, the mean was applied to normally distributed variables, while the median was chosen for skewed distributions to prevent the influence of outliers. For categorical variables, the mode—representing the most frequently occurring category—was utilized to fill in missing entries, thereby preserving the dataset’s categorical structure.

The decision to use mean and mode imputation methods aligns with established practices in statistical data imputation. Tsai and Hu (2022) highlight the efficacy and simplicity of these techniques, stating, “The mean and mode imputation methods are simple to implement; all missing values are replaced by the mean of the observed variables where the mean and mode methods are used for continuous (or numerical) and discrete (or

categorical) data types, respectively" (Tsai & Hu, 2022, p. 3). This approach not only maintains the integrity of the dataset but also supports the assumptions of many statistical and machine learning models by ensuring that no data point remains incomplete.

Table 8 Average percentage of correct predictions and average RMSE for 10–50% missing rates

	Mean + mode	MLP	SVM	KNN	CART
10%	96.97/0.1705	97.29/0.1949	**97.69**/0.1338	96.84/0.226	97.68/**0.131**9
20%	94.59/0.1716	94.51/0.1831	**95.12**/0.1439	93.79/0.2138	95/**0.1427**
30%	92.24/0.1675	92.06/0.1818	**92.63**/0.1442	91.15/0.2163	92.33/**0.1435**
40%	90.04/0.1643	89.72/0.1798	**90.24**/0.1464	88.48/0.2148	89.9/**0.1461**
50%	**88.05**/0.163	87.52/0.1739	87.9/0.1437	86.25/0.2195	87.55/**0.1467**
Avg	92.38/0.1658	92.22/0.1821	**92.71**/0.1437	91.3/0.2181	92.49/**0.1431**

Bold values indicate the highest performance for each dataset

Figure 3

Through the application of these imputation techniques, the dataset was rendered fully populated, enabling the use of robust statistical analyses and enhancing the validity of model outcomes.

3.7 Identifying Ordinal and Nominal Variables

In the data preparation phase, a crucial step involved the classification of categorical variables into ordinal and nominal types. This classification is essential for informing subsequent encoding decisions and ensuring that the data is appropriately prepared for analysis in machine learning models.

3.7.1 Nominal Variables

Nominal variables refer to categorical data that do not have a specific order or ranking. These variables can be divided into distinct categories, where each category is mutually exclusive. In the context of this study, several nominal variables were identified, including:

- Gender: Coded as "Male" or "Female," this variable provides essential demographic information without implying any order.
- Religion: Represented as categories without a hierarchical structure, allowing for analysis of socio-economic factors across different religious affiliations.
- Caste Category: Similar to religion, the caste variable classifies individuals into distinct groups based on social stratification, with no inherent ranking.

3.7.2 Ordinal Variables

Ordinal variables, on the other hand, are categorical data that possess a clear order or ranking among the categories. This characteristic allows for the assessment of relationships based on the relative positioning of these variables. Examples of ordinal variables identified in this study include:

- Satisfaction Levels: Questions regarding satisfaction with life, work life, and educational facilities were measured on a scale (e.g., "Very Dissatisfied" to "Very Satisfied"). These variables capture qualitative perceptions while allowing for quantitative analysis.
- Problem Resolution Satisfaction: This variable measures the satisfaction level regarding the effectiveness of problem resolution efforts, with responses ranging from "Very Unsatisfied" to "Very Satisfied."
- Health Satisfaction: Questions assessing satisfaction with health services were also ordinal, providing insights into the perceived quality of healthcare among respondents.

3.7.3 Importance of Classification

The identification and classification of ordinal and nominal variables are critical for effective model training and analysis. Proper encoding ensures that machine learning algorithms can interpret the data accurately, allowing them to capture the underlying patterns that influence economic outcomes in tribal communities. The distinction between these variable types not only informs the choice of encoding methods but also shapes the analytical strategies used in subsequent phases of the study.

4. Noise Reduction

Noise reduction is a crucial step in data preprocessing that aims to enhance the reliability and validity of the analysis by identifying and addressing outliers within the dataset. Outliers can skew results and lead to misleading interpretations, particularly in real-world data that may not conform to a normal distribution hence Z-score method was not used. In this study, the Interquartile Range (IQR) method was employed as the primary outlier detection technique, given its effectiveness in dealing with skewed data.

4.1 Outlier Detection Using the IQR Method

The IQR method is widely recognized for its robustness in identifying outliers in datasets with non-normal distributions. The IQR is calculated by determining the first quartile (Q1) and the third quartile (Q3) of the data, which represent the 25th and 75th percentiles, respectively. The IQR is then computed as follows:

$IQR=Q3-Q1$

Using the IQR, outliers can be defined as values that lie outside the range of:

$Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$

Any data points falling below the lower threshold or above the upper threshold are considered outliers.

4.2 Addressing Identified Outliers

Once outliers were identified using the IQR method, the following strategy was implemented to address them:

- Capping: For outliers that represented valid observations but were significantly deviated from the main data distribution, a capping approach was employed. This involved replacing extreme values with the nearest non-outlier value within the threshold. This method preserves the data while mitigating the influence of extreme values.

4.3 Importance of Noise Reduction

By employing the IQR method for outlier detection and addressing these outliers appropriately, the study aimed to enhance the overall quality of the dataset. Reducing noise allows for more accurate modelling and analysis, leading to more reliable insights into the socio-economic dynamics of tribal communities in Jamtara District. This step is critical in ensuring that the findings of the study are valid and can inform effective policy interventions aimed at improving the livelihoods of marginalized populations.

5. Encoding of Categorical Variables: One Hot Encoding and Ordinal Encoding

To ensure compatibility with machine learning models, categorical variables were transformed into numerical formats using One Hot Encoding (OHE) and Ordinal Encoding, depending on the nature of each variable. Encoding categorical data is essential in preprocessing, as most machine learning models require numerical inputs and cannot process raw categorical data effectively.

One Hot Encoding (OHE) was applied to nominal categorical variables—those without an inherent order. OHE creates binary (0/1) columns for each unique category in the variable, where each column indicates the presence or absence of a category. This approach is widely adopted due to its ability to handle categorical data without imposing any artificial ordering. Potdar et al. (2017) describe One Hot Encoding as follows: “One Hot Coding is the most widely used coding scheme. It compares each level of the categorical variable to a fixed reference level. One hot encoding transforms a single variable with n observations and d distinct values, to d binary variables with n observations each. Each observation indicating the presence (1) or absence (0) of the dichotomous binary variable” (Potdar et al., 2017, p. 7). This transformation enabled seamless integration of nominal categorical features into the machine learning models without introducing any implicit rank or hierarchy.

Ordinal Encoding was employed for ordinal categorical variables, where an intrinsic order exists among categories. In Ordinal Encoding, an integer value is assigned to each category, preserving the original order. This method is computationally efficient as it does not add new columns, unlike OHE. However, it implies an order to the categories, which may not necessarily reflect real-world relationships. Potdar et al. (2017) caution that "In ordinal encoding, an integer is assigned to each category, provided the number of existing categories are known. It does not add any new columns to the data, but implies an order to the variable that may not actually exist" (Potdar et al., 2017, p. 7). For our dataset, Ordinal Encoding was carefully applied only to categorical variables with a meaningful ordinal structure, ensuring the encoding accurately represented the variable's nature.

According to Table 3 from the study by Potdar et al. (2017, p. 9), different encoding techniques can yield varied accuracy percentages when applied in machine learning models. One Hot Encoding demonstrated an accuracy of 90%, making it a widely preferred choice due to its effectiveness in creating binary representations for categorical data. Conversely, Ordinal Encoding achieved an accuracy of 81%, which, while still useful, may be less effective in certain applications due to its imposition of an arbitrary ordinal relationship among categories.

This evidence supports the careful selection of encoding techniques based on the nature of the data and the requirements of the model. For this analysis, the choice between One Hot and Ordinal Encoding was made with consideration to these observed accuracy rates, aiming to enhance model reliability and predictive performance.

Table 3. Encoding technique and the accuracy percent

Encoding Technique	Accuracy (Percentage)
One Hot Coding	90
Ordinal Coding	81

Figure 4

6. Final Data Integrity Check

A thorough check confirmed that all non-numeric data had been appropriately handled, and a summary of remaining null values was generated. This process is executed for all the 7 models.

Final check of any non numeric data in the dataframe

```
[93]: import pandas as pd

      # Check if all columns are numeric
      if df_final.select_dtypes(exclude=['number']).empty:
          print("All columns in df_final are numeric.")
      else:
          print("Some columns in df_final are not numeric.")
          # Optionally, print the non-numeric columns
          non_numeric_columns = df_final.select_dtypes(exclude=['number']).columns
          print("Non-numeric columns:", non_numeric_columns)
          print(len(non_numeric_columns))

      All columns in df_final are numeric.
```

Figure 5

7. Creating Independent and Dependent Variables

With the data preprocessed, independent and dependent variables were established based on the research objectives. Independent variables included socio-economic indicators, while the dependent variables focused on key outcomes like annual income and expenditure patterns.

Independent and dependant columns

```
[95]: X=df_final.drop(columns=["79Total annual income", "79Whats your household income"])
      y=df_final["79Total annual income"]
```

Figure 6

8. Train-Test Split

The dataset was divided into training and testing subsets using an 70-30 split ratio. This partitioning allowed for effective model training and subsequent evaluation of model performance on unseen data. This process is done for all the 7 models.

Figure 7

9. Dependent Variables Selection

To conduct an in-depth analysis and develop models focused on improving the socio-economic conditions of tribal communities, seven dependent variables were carefully selected. These variables have been substantiated by existing literature, which highlights their influence on enhancing the economic and social well-being of tribal populations in India.

9.1 Total Annual Income

Total annual income is an essential indicator of a household's financial health and overall socio-economic status. In the study titled "Demand-side analysis of measuring financial inclusion: Impact on socio-economic status of Primitive Tribal Groups (PTGs) in India" by Nandru and Rentala, it is emphasized that "financial inclusion has a significant positive impact on the socio-economic conditions of Primitive Tribal Groups" (Nandru and Rentala, p. 5). By measuring and understanding annual income levels, policies can be tailored to improve financial inclusion, thereby uplifting the socio-economic status of tribal households.

Hypothesis testing was done with 2 of the hypotheses from the study by (Nandru and Rentala p. 7):

H6. Financial inclusion has a significant impact on economic status.
H7. Financial inclusion has a significant impact on social status.

According to Figure 1 from the study by (Nandru and Rentala p. 7) the following hypotheses were drawn:

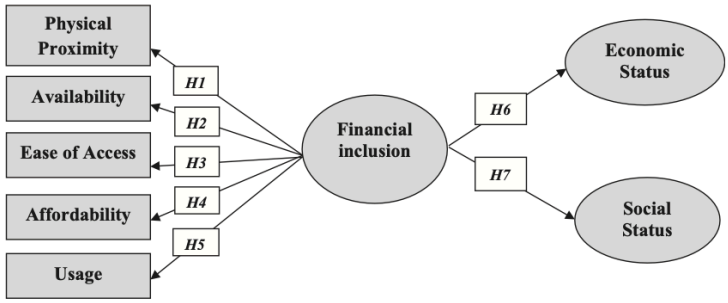


Figure 1.
Research framework

Figure 8

From the study by Nandru and Rentala (p. 7), "further, it was found that financial inclusion had a significant impact on the social status and economic status of the PTGs in India. The hypothesis H6 and H7 were accepted and found to be statically significant level at 0.001. It can, thus, be inferred that better financial inclusion status has a positive impact on the socio-economic status of the primitive tribal households in India."

Hence Total Annual Income was selected to be a dependent variable which can drive the socio-economic growth of the tribal people of Jamtara.

9.2 PDS Card (yes/no)

Access to Public Distribution System (PDS) cards is crucial for improving the nutritional and health status of tribal populations. In the research paper "Effectiveness of PUBLIC DISTRIBUTION SYSTEM (PDS) in quality-of-life improvement- A study of the Tribal population in Purulia District in West Bengal" it is mentioned that Bhattacharya (p. 279)

“Thus the above literature review has provided a number of interesting clues regarding the impact of PDS on meeting the nutritional requirements of poor in India”
PDS card possession is a key indicator of socio-economic stability for tribal communities, as it ensures access to subsidized food, enhances household savings, and supports better health, education, and productivity outcomes. By bridging resource gaps, especially in marginalized regions, it serves as a crucial factor in improving overall well-being and resilience.
The study of Bhattacharya (p. 279) confirms “The studies (Puri, 2012; Sadasivam and Senthamarai, 2012; Dreaze and Khera, 2010; Gundegowda and Nagraj, 2011; Singh et al, 2011) have confirmed the success of PDS in terms of benefitting the poor”.

9.3 Expenditure on Agricultural Activities in Family: Labour

Expenditure on agricultural activities, particularly labour, is indicative of a household’s engagement in agricultural productivity, which is a primary source of livelihood for many tribal families. According to the study “A Comparative Study of Traditional Fishing Practice Among Tribal People at Some Selected Regions of Purulia District”, it is quoted that “Sixthly, considering the correlation between labour charge with other variables, there exist a significant high positive correlation with harvesting cost, total input and total output.” (Majhi p. 5).

Also 2 regression models were calculated in the above-mentioned research paper considering Total Output and Total Input as the Dependent variables and both have a significant impact from the Labour Charge (-1.331 times and59.97 times respectively) (Majhi p. 6)

Total Output = 9724.046 + (-12.726 x Stocking) + (.041 x Raw Cow Dung) + (-2.835 x Liming) + (-.119 x feeding) + (-1.331 x Labour Charge) + (105.147 x harvesting cost.

Figure 9

Total Input = -71392.59 + (299.81 x Stocking) + (-60.97 x Raw Cow Dung) + (153.11 x Liming) + (8.80 x Feeding) + (59.97 x Labour charge) + (-1996.19 x Harvesting cost).

Figure 10

This variable highlights the investment in agriculture labour as a means of income generation.

9.4 Expenditure on Household: Food

Expenditure on food directly impacts the nutritional status and health of a household. The study "Household Economy and Nutritional Status in the Shabar Tribe of Orissa" (Chakrabarty and Bharati) underlines that focusing on this expenditure helps to assess food security and overall well-being in tribal communities.

The following Statistical Analysis was done in the above-mentioned research paper: "ANOVA was used to evaluate the difference of nutritional status between two economic sub-groups. Chi-square was used to understand the association between nutritional groups and economic sub-groups. Regression analyses were used to understand the association between economic condition and nutritional status." (Chakrabarty and Bharati p. 28)

The results are as follows (Chakrabarty and Bharati p. 31; p. 33):

Table 4: Degree of uncertainty in food supply by economic group

	Degree of uncertainty (monthly)							
	0-6		6-11		12 plus		Total	
Economic group	n	%	N	%	n	%	n	%
Lower	8	24.24	19	57.58	6	18.18	33	100.00
Higher	13	22.03	25	42.37	21	35.59	59	100.00
Total	21	22.83	44	47.83	27	29.35	92	100.00

Figure 11

Table 7: Mean weight and BMI among adult by economic group

Variable	Economic group				F- value
	Higher		Lower		
	Mean	SD	Mean	SD	
Male					
Height (cm)	160.06	6.03	159.99	6.55	0.04
Weight (kg)	48.61	6.95	47.24	5.51	1.11
BMI (kg/m^2)	18.95	2.36	18.41	1.31	1.68
Female					
Height (cm)	149.58	4.55	150.09	5.39	0.25
Weight (kg)	41.02	5.58	39.63	4.58	1.60
BMI (kg/m^2)	18.32	2.27	17.59	1.82	2.70

Table 8: Nutritional status among adult by economic group

Economic group	Nutritional status						
	Undernourished		Normal and above		Total		OR (95% CI)
Male	n	%	n	%	n	%	
Lower	19	50.00	19	50.00	38	100.00	1.19 (0.539-2.619)
Higher	32	47.06	36	52.94	68	100.00	
Total	51	48.11	55	51.89	106	100.00	
Female							
Lower	26	74.29	9	25.71	35	100.00	1.93 (0.778-4.765)
Higher	39	60.00	26	40.00	65	100.00	
Total	65	65.00	35	35.00	100	100.00	

Figure 12

These values clearly show a rift in the food consumption pattern between different socio-economic level groups, thereby indicating expenditure on food is a driving factor towards socio economic development of tribal people.

9.5 Bank Account (yes/no)

Access to banking services is a cornerstone of financial inclusion, especially for rural and tribal populations. In the paper “How do households utilize banking services and what

are the determinants of it? An empirical analysis from the rural and tribal areas of an eastern Indian state", it is asserted that "As we argue that simply having access will not fulfil the broader goal of financial inclusion, hence steps should be taken to motivate and make people aware about the various usage of formal financial services. The study result will be helpful to the policy makers and other stake holder of financial inclusion programme to understand the problems of utilization and accordingly could direct the policy measures to minimize the constraints of utilization." (Ray et al. p.14)

The research includes several models to show how tribal people are using the banking services from the following figure (Ray et al p.3)

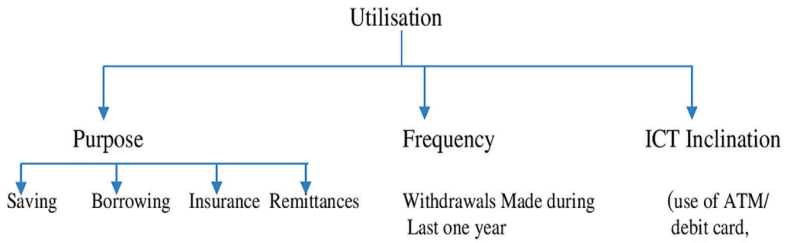


FIGURE 1 Indicators of measurement. Source: Compiled by the Authors

Figure 13

The following 3 models were created (Ray et al. p3; p4):

Model 1: Determinants of Utilization of Account to Save.

$$Pr(Y = 1) = \beta + \beta_i X_i + \epsilon,$$

where Y = Use of bank account for saving, the dependent variable
β = The Intercept
βi = Coefficients of the Explanatory Variables 1–20
Xi = The Explanatory Variables and i = 1–20
€ = The Error Term

Figure 14

Model 2: Determinants of Utilization of Account for Borrowing.

$$Pr(Y = 1) = \beta + \beta iXi + \epsilon, \qquad (2)$$

where Y = Use of bank account for Borrowing, the dependent variable
β = The Intercept
βi = Coefficients of the Explanatory Variables 1–9.
Xi = The Explanatory Variables and i = 1–9
€ = The Error Term

Figure 15

Model 3: Determinants of Frequent Utilization of Account.

$$Pr(Y = 1) = \beta + \beta_i X_i + \epsilon, \qquad (3)$$

where Y = Frequent utilization of Account, the dependent variable
β = The Intercept
βi = Coefficients of the Explanatory Variables 1–12.
Xi = The Explanatory Variables and i = 1–12
€ = The Error Term

Figure 16

The results of these model evaluations as stated “The use of bank accounts for financial transactions is mostly done to receive DBT and NREGS payments” (Ray et al. p.14) which states that financial inclusion is possible via bank accounts, better policy intervention is the need.

Thus, this variable is critical to understanding financial participation and economic empowerment in tribal areas.

9.6 LPG Connection (yes/no)

Access to LPG connections is an indicator of energy inclusion and improved living conditions. The study "Inequalities in LPG and electricity consumption in India: The role of caste, tribe, and religion" discusses how "The results of our empirical analysis suggest that, after controlling for the determinants which impinge on the households' microeconomic demand and regional supply characteristics, the households belonging to the scheduled tribe and scheduled caste communities do have significantly poorer access to LPG and electricity usage as compared to the upper caste households." (LPG Study, p. 1).

The following model is used in the above-mentioned research paper (LPG Study p.8):

$$Y_i^* = \beta + \sum_{m=1}^{q} \Omega_m X_{im} + \sum_{j=1}^{k} \alpha_j X_{ij} + \varepsilon_i \quad \text{(A)}$$

and

$$Y_i = \begin{cases} 1 \; if \; Y_i^* > 0 \\ 0 \; otherwise \end{cases} \quad \text{(B)}$$

Considering that the Generalised Linear Model in equation (A) is consistent for both the continuous (observable) and binary (unobservable) dependent variables (Liao, 1994), for the access to LPG, equation (B) is the measurement equation that connects the latent dependent variable in equation (A). In equation (B) the variable takes value 1 if the household has access to LPG, zero otherwise. Thus, the link function of Y_i^* in equation (A) is unity for the log of electricity usage and it changes to probit for the access to LPG.[6] $\sum_{m=1}^{q} \Omega_m X_{im}$ is the sum of dummy variables representing the scheduled caste, scheduled tribe and Muslim households and $\sum_{j=1}^{k} \alpha_j X_{ij}$ is the set of other control variables in the specification. Therefore, $\beta + \sum_{j=1}^{k} \alpha_j$ can be defined as the coefficients for the upper caste households. The specification of dependent variable and covariates in the equations above are discussed below. For the identification of the vector of coefficients we assume that the covariance between errors ε_i and covariates X_i is zero.

Figure 17

The results obtained from the research are as follows:
“Thus, without controlling for the determinants of access to LPG, the Muslim households have significantly better access to LPG than the Scheduled Tribe and Scheduled Caste households in this sample. However, the upper caste households, as a social group, have the best access to LPG in this sample.” (LPG Study p.10)

Table 1: Distribution of LPG and Electricity by Social Groups.

	LPG (per cent access)	Average Electricity consumption (in KWh)
Upper Caste Hindu Households (Others)	0.545 (0.497)	95.212 (96.158)
Scheduled Tribes Households	0.388 (0.488)	64.090 (55.218)
Scheduled Caste Households	0.337 (0.472)	72.715 (63.622)
Muslim Households	0.453 (0.497)	89.140 (92.010)
Total Sample Size	87753	87753

mean coefficients; sd in parentheses

Figure 18

It is stated “At the policy level, there are a number of large-scale interventions and policies by the government to overcome inequalities and poverty in general, but very few for dealing with the issue of accessing cleaner energy sources by socially marginalised groups in India” (LPG Study p.25)
Hence policy intervention is needed in order to secure clean fuel access to tribal people and lower marginalized section of the society.

This variable signifies the adoption of cleaner fuel and its impact on quality of life in tribal regions.

9.7 Satisfaction with Educational Facilities

Educational attainment is foundational to economic development. The paper “Early Childhood Education Among Tribal Communities in India” highlights that “Historically,

tribal children faced rejection and discrimination in terms of their backwardness. Such discrimination can be seen in inequality in their educational wellbeing due to their early life circumstances. Commendable progress has brought tribal children to schools in the past few years. Still, efforts should also be made towards reducing their discontinuation and improving their quality of education which can improve their educational wellbeing." (Rashmi and Paul p. 12).

The following statistical method is used in the above-mentioned paper:
"Further, we used the Blinder-Oaxaca twofold decomposition technique to identify the contribution of explanatory covariates in 2005 behind the caste differential in the educational wellbeing of children in - 2012. We show the overall and detailed decomposition of the caste differential in educational wellbeing." (Rashmi and Paul p.5)

Results of the analysis are present in the below Table 5 (Rashmi and Paul p.12):
"Table 5 shows the decomposition of the educational wellbeing gap among children in ST and Non-ST, ST and SC, ST and OBC, and ST and Others groups, respectively. We find that the direction of contribution is the same across all the statically significant contributors in the four decomposition estimates. The explained educational wellbeing difference between ST and Non-ST groups is similar for the ST and OBC, and ST and Others groups. The magnitude of the percentage contribution of each statistically significant contributor varies across the four decomposition estimates." (Rashmi and Paul p.6)

Characteristics	Caste differential in educational wellbeing among children in round-II							
	Explained difference of ST and Non-ST group		Explained difference of ST and SC group		Explained difference of ST and OBC group		Explained difference of ST and Others group	
	Coefficient	Percent	Coefficient	Percent	Coefficient	Percent	Coefficient	Percent
Age of children (in years)	−0.003	0.6	0.005	−1.7	−0.005	1.3	−0.005	0.7
Gender of the children	0.000	0.0	0.001	−0.5	0.000	0.0	0.000	0.0
Stunting satus of children	−0.012*	2.5	−0.006	2.0	−0.010*	2.4	−0.018*	2.7
Type of school attended in round-II	−0.028*	6.2	−0.011*	3.5	−0.039*	9.8	−0.025	3.7
Takes private tuition in round-II	−0.031*	6.9	−0.029*	9.2	−0.017*	4.3	−0.054*	8.2
Place of cooking in household	0.001	−0.3	−0.008	2.6	0.004	−1.1	0.005	−0.7
Type of cooking fuel	0.002	−0.4	0.001	−0.2	0.007*	−1.8	−0.009	1.4
Household sanitation condition	−0.026*	5.7	−0.011*	3.4	−0.022*	5.6	−0.049*	7.3
Water purification in household	0.000	−0.1	0.004	−1.2	0.000	−0.1	0.002	−0.3
Household wealth quintile	−0.111*	24.3	−0.109*	34.9	−0.092*	22.8	−0.100*	15.0
Household poverty status	−0.022*	4.9	−0.007	2.1	−0.022*	5.5	−0.069*	10.3
Highest educational level of male adults in household	−0.062*	13.7	−0.025*	7.9	−0.057*	14.2	−0.121*	18.1
Highest educational level of female adults in household	−0.060*	13.1	−0.018*	5.6	−0.057*	14.3	−0.092*	13.9
Gender of household head	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
Religion of household head	−0.010*	2.3	0.013	−4.1	−0.020*	5.0	−0.001	0.2
Types of mass media viewed by children	−0.004	0.8	0.002	−0.5	−0.005	1.4	0.002	−0.3
Women's autonomy in child health-care in household	−0.003*	0.8	−0.003	0.9	−0.002	0.4	−0.009*	1.3
Attack/threat on household	0.000	−0.1	0.000	−0.1	0.001	−0.3	−0.001	0.1
Solving community problem	−0.001	0.2	−0.006	1.8	0.001	−0.1	−0.002	0.4
Domestic violence in community	0.003*	−0.8	0.003	−1.1	0.002	−0.5	0.000	−0.1
Type of community	0.008	−1.7	0.011	−3.5	0.002	−0.6	−0.002	0.3
Country regions	−0.002	0.3	0.009	−3.0	0.001	−0.2	−0.004	0.6
Explained difference (E)	−0.360*	79.1	−0.181*	57.8	−0.330*	82.3	−0.552*	82.9
Unexplained difference (U)	−0.095*	20.9	−0.132*	42.2	−0.071	17.7	−0.114*	17.1
Total difference (T)	−0.454*		−0.314*		−0.401*		−0.666*	

Table 5. Blinder-Oaxaca decomposition of the caste differential in educational wellbeing score between ST and Non-ST, ST and SC, ST and OBC, and ST and Others children for sensitivity analysis. (a) Statistical significance denoted by asterisks where *p-value < 0.05.

Figure 19

Thus, understanding satisfaction with education services helps address gaps in educational support for tribal families.

10. Model Building

A total of seven different regression models were implemented to analyse the relationships within the data:

1. Random Forest Regressor/ Classifier
2. Decision Tree Regressor/ Classifier
3. CatBoost Regressor/ Classifier
4. XGBoost Regressor/ Classifier
5. Gradient Boosting Regressor/ Classifier
6. AdaBoost Regressor/ Classifier

7. Bagging Regressor/ Classifier

10.1 Assumptions of Linear Regression and Data Normality

In statistical modelling, particularly for linear regression, certain assumptions about the data are essential for accurate predictions and reliable interpretations. One of the key assumptions is normality, which states that the residuals (errors) of the model should be normally distributed. However, in real-world data, these assumptions are often violated due to factors such as outliers, skewed distributions, or complex underlying relationships within the data. In this study, since the data is real-world and collected from various sources, it inherently fails to meet the normality assumption required for linear regression.

10.1.1 Shapiro-Wilk Test for Normality

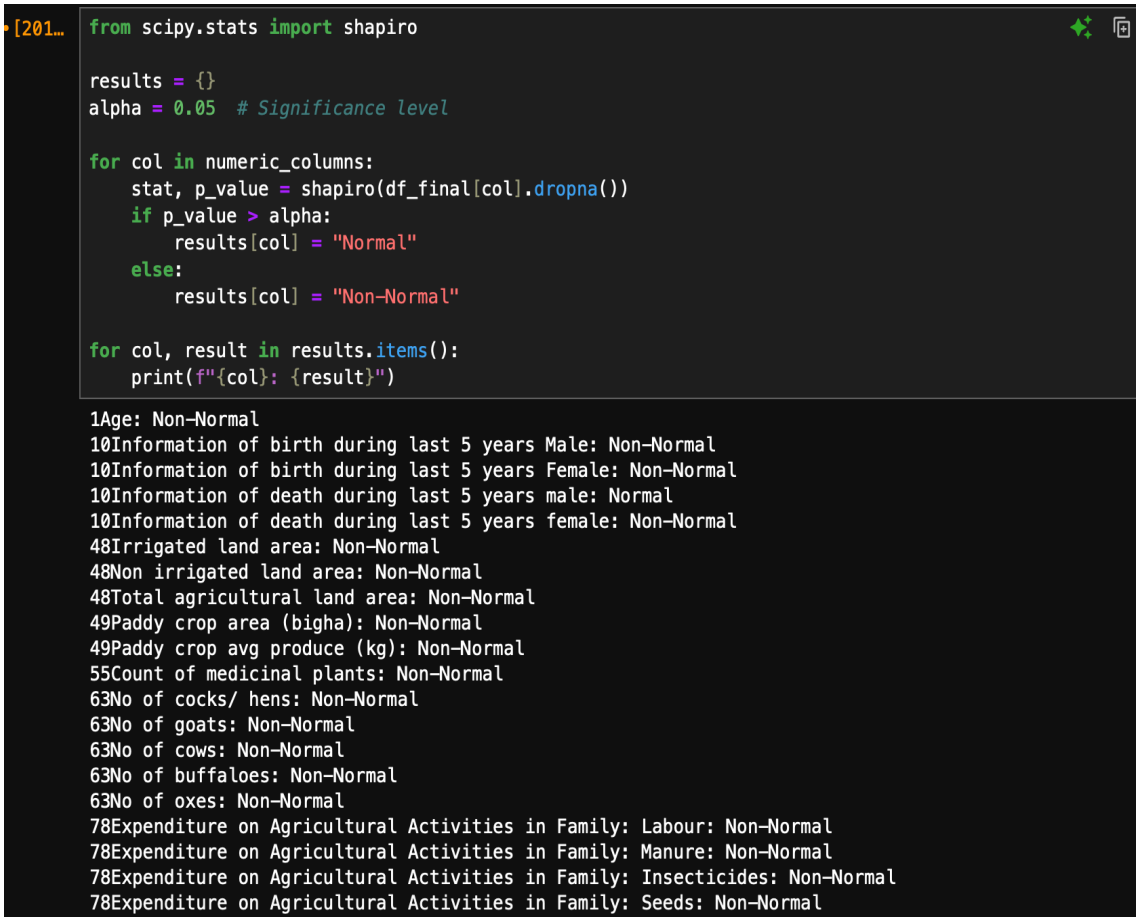
To assess if the dataset's numerical variables follow a normal distribution, the Shapiro-Wilk test was conducted, where:

- Null Hypothesis (H0): Data is normally distributed.
- Alternative Hypothesis (H1): Data is not normally distributed.

Using a significance level of $\alpha = 0.05$, results are interpreted as follows:

- p-value > 0.05: Fail to reject H0, suggesting the data is approximately normal.
- p-value ≤ 0.05: Reject H0, indicating significant deviation from normality.

The test results showed several numerical columns with p- values below 0.05, confirming non- normality in the data. This deviation from normality suggested that linear regression may be unsuitable for this dataset due to potential compromises in model performance and interpretability.



```
[201…
from scipy.stats import shapiro

results = {}
alpha = 0.05  # Significance level

for col in numeric_columns:
    stat, p_value = shapiro(df_final[col].dropna())
    if p_value > alpha:
        results[col] = "Normal"
    else:
        results[col] = "Non-Normal"

for col, result in results.items():
    print(f"{col}: {result}")
```

```
1Age: Non-Normal
10Information of birth during last 5 years Male: Non-Normal
10Information of birth during last 5 years Female: Non-Normal
10Information of death during last 5 years male: Normal
10Information of death during last 5 years female: Non-Normal
48Irrigated land area: Non-Normal
48Non irrigated land area: Non-Normal
48Total agricultural land area: Non-Normal
49Paddy crop area (bigha): Non-Normal
49Paddy crop avg produce (kg): Non-Normal
55Count of medicinal plants: Non-Normal
63No of cocks/ hens: Non-Normal
63No of goats: Non-Normal
63No of cows: Non-Normal
63No of buffaloes: Non-Normal
63No of oxes: Non-Normal
78Expenditure on Agricultural Activities in Family: Labour: Non-Normal
78Expenditure on Agricultural Activities in Family: Manure: Non-Normal
78Expenditure on Agricultural Activities in Family: Insecticides: Non-Normal
78Expenditure on Agricultural Activities in Family: Seeds: Non-Normal
```

Figure 20

10.2 Transition to Ensemble Methods

Given the limitations of linear regression with non-normal data, an alternative modelling approach was necessary to improve the robustness and predictive accuracy of the model. Ensemble methods emerged as a suitable solution, as they do not require the strict assumptions of linear regression and are capable of handling complex, non-linear relationships in the data. Ensemble methods combine multiple models to achieve better predictive performance, greater stability, and increased resilience against noise and non-normal data.

10.2.1 Overview of Ensemble Methods

Ensemble methods combine predictions from multiple models, or base learners, to improve accuracy by reducing variance, bias, and enhancing generalization. Common techniques include:

1. Bagging (Bootstrap Aggregating):
    - Trains multiple models (e.g., decision trees) on different random subsets of data sampled with replacement, and combines predictions via averaging (regression) or majority voting (classification).
    - *Example*: Random Forest, a bagging method, builds decision trees on varied subsets and aggregates results, reducing bias and variance, thus being resilient to non-normal data and outliers.
2. Boosting:
    - An iterative approach where each new model focuses on correcting previous errors, enhancing accuracy by prioritizing misclassified points.
    - *Example*: Algorithms like Gradient Boosting, XGBoost, and AdaBoost excel with complex patterns, incrementally improving by targeting past mistakes.
3. Stacking:
    - Combines different models (e.g., decision trees, logistic regression) as base learners, using a meta-model to blend their predictions.
    - This strategy leverages the unique strengths of each model, creating an adaptive, high-performing ensemble suitable for diverse data.

10.2.2 Why Ensemble Methods Work Well with Non-Normal Data

Ensemble methods, especially decision tree-based approaches like Random Forests and Gradient Boosting, are robust against non-normality due to the following characteristics:

- Lack of Parametric Assumptions: Unlike linear regression, which assumes a specific linear relationship and normally distributed errors, ensemble methods (especially tree-based ones) do not make any assumptions about the distribution of the input data. They rely on data-driven splitting criteria, making them adaptable to non-normal, skewed, or multi-modal data.

- Resilience to Outliers: Outliers can severely impact linear regression models, but ensemble methods are often more resilient because they aggregate multiple models or focus only on errors in a gradual, corrective manner.

- Adaptability to Complex Patterns: Ensemble methods can capture non-linear relationships by integrating multiple models that, together, can approximate more intricate patterns in the data. This adaptability makes them well-suited for real-world data that deviates from the ideal conditions assumed by linear models.

To prove the efficacy of ensemble models on real world data, the study "Global happiness ranking: An ensemble regression vs traditional approach investigation" (Xing 2024) uses various traditional ML models and ensemble models on the same set of data from the World Happiness website.

The following questions are answered in the paper (Xing p.5):

> *How can models such as XGB Regressor, LGBM Regressor, CatBoost Regressor, Random Forest Regressor, Linear Regression, Polynomial Regression, and Ridge Regression be utilized to predict and evaluate the impact of future well-being related features on happiness ranking scores over the next decade years?*

The sub-questions can be listed separately, as such:

RQ1 *Which regressor or regression model fits the best for the small global happiness ranking dataset's predicted happiness-related feature values without overfitting?*

RQ2 *How can the analysis of a model's features and the investigation to determine which feature has the best-predicted accuracy score using SHAP be conducted?*

Figure 21

The results of applying these models on the dataset, the metrics are noted for every model from Table 11 (Xing p.21):

Models	RMSE
XGBoost	0.0084
LightGBM	0.0346
CatBoost	0.01222
Random Forest	0.0679
Linear Regression	0.5721
Polynomial Regression	0.5685
Ridge Regression	0.2386

Table 11: Comparison of prediction effects of seven types of models

Figure 22

Also, the research states “It reveals XGBoost’s robustness, known for handling large datasets efficiently, extends to smaller datasets as well, evident in its consistent performance and low RMSE. This finding is significant given the complexity of happiness prediction (Park & Ho, 2019).” (Xing p.23)

Hence ensemble methods are used for model building in this thesis work over traditional ML models.

10.3 Rationale for Not Using Deep Learning Models

While deep learning (DL) models, such as Artificial Neural Networks (ANNs), have shown impressive performance across various domains, they were not used in this study due to their inherent limitations in interpretability and feature importance assessment. Although DL models might potentially yield higher accuracy or R-squared scores compared to traditional ensemble methods, they operate as black-box models, meaning

that the internal workings and the contribution of individual features are largely opaque and difficult to interpret.

10.3.1 Limitations of Deep Learning in Interpretability and Feature Importance

Interpretability is key for this study, as understanding which features impact socio-economic outcomes is essential for guiding policy recommendations.

1. Black-Box Nature of DL Models:

- Deep learning (DL) models achieve high accuracy by capturing complex, non-linear relationships but lack interpretability due to their multi-layered architectures. This makes it challenging to trace individual feature impacts on predictions.
- Unlike ensemble methods (e.g., Random Forests), which offer feature importance metrics, DL models lack inherent transparency, limiting insights into feature contributions.

2. Difficulty in Identifying Key Features:

- Identifying influential features is crucial for targeted improvements in socio-economic areas, like banking services. Ensemble methods can indicate priority features, guiding actionable decisions.
- While DL models might offer marginally better accuracy, they do not clearly indicate feature importance without advanced interpretability methods, which remain complex and limited.

3. Limitations in Policy-Oriented Research:

- DL's lack of feature transparency limits its suitability for studies needing actionable insights. Although DL models can provide strong predictions, they lack the interpretability required to determine which socio-economic indicators should be prioritized for improvements in areas like financial inclusion, education, or healthcare.

10.4 Why Ensemble Methods Were Chosen Over Deep Learning Models

To balance predictive accuracy with interpretability, ensemble methods such as Random Forests, XGBoost, Gradient Boosting Machines and AdaBoost were selected. These methods provide a robust solution that not only handles complex, non-linear relationships but also enables the identification of feature importance, offering a clear advantage for interpretative and actionable insights. By utilizing ensemble techniques, we can achieve high model performance while also understanding which features most significantly affect the dependent variables, such as banking access or annual income.

This interpretability is vital for creating actionable recommendations to assist policymakers in improving socio-economic conditions. While DL models might offer slightly better accuracy, the trade-off with interpretability makes ensemble methods a more suitable choice for this study.

Finally, each of the 7 models was trained on the training dataset, and predictions were generated for the testing set.

11. Model Evaluation

The performance of both regression and classification models was thoroughly evaluated using a range of metrics to ensure robust assessment and selection of the best-performing models.

For regression models, the evaluation metrics included:

- Mean Squared Error (MSE): Measures the average of the squares of the errors, indicating how close the predictions are to the actual values.
- Root Mean Squared Error (RMSE): The square root of MSE, providing error in the same units as the target variable, making it easier to interpret.

- Mean Absolute Error (MAE): Captures the average magnitude of errors in predictions, offering insight into the overall prediction accuracy without penalizing large errors excessively.

- R^2 Score: Indicates the proportion of variance in the dependent variable explained by the model, showing how well the model fits the data.

- Adjusted R^2 Score: A modified version of R^2 that accounts for the number of predictors, helping to avoid overfitting by adjusting for model complexity.

For classification models, the following metrics were used:

- Accuracy: The percentage of correct predictions, providing a straightforward measure of model performance.

- Precision: The ratio of true positives to the sum of true positives and false positives, indicating the accuracy of positive predictions.

- Recall: The ratio of true positives to the sum of true positives and false negatives, showing the model's ability to identify all positive instances.

12. Feature Importance Extraction

Once the best model was identified, an in-depth feature importance analysis was conducted to pinpoint which independent variables had the most significant impact on the dependent variables. This step is essential for understanding the socio-economic factors that most strongly influence economic outcomes for tribal communities, providing insights into actionable areas for policy improvement.

In ensemble methods such as Random Forest, XGBoost, and Gradient Boosting, feature importance can be derived using the model.feature_importances_ attribute in scikit-learn. This attribute provides a numerical score that represents the relative importance of each feature in making predictions. The higher the score, the more significant the feature in contributing to the model's predictive accuracy.

The entire working of how model.feature_importances_ is explained in the Research Methodology section.

For this analysis, the model.feature_importances_ method was used to compute the feature importance scores for each model, and the top 5 features were identified for further investigation. These features offer targeted insights into the primary factors affecting economic conditions within the tribal communities, helping policymakers focus on the most impactful areas for intervention and development.

13. Visualization of Results

The top 5 most important features were visualized using a bar plot, effectively illustrating the significance of each predictor in the context of socio-economic development. This visual representation helps communicate findings clearly by highlighting the variables with the greatest impact on target outcomes. Additionally, the feature importance values were displayed as a DataFrame, providing an accessible summary of the relative contribution of each key feature. This combination of visual and tabular formats enhances the interpretability of the analysis, enabling stakeholders to easily identify and prioritize critical socio-economic factors.

Research Findings

A total of 7 dependent variables were selected and models were built based on the nature of the dependent variable selected (discrete or continuous).

Out of the 7 dependent variables selected, 3 were continuous in nature and the remaining 4 were discrete.

Analysis of the different models built:

1. Annual income

This variable is continuous in nature, so a regression model was developed depicted in Figure 23

The model with the best Adjusted R^2 Score and R^2 Score was selected and top 5 most im portant independent features for this dependent variable were calculated:

Gradient Boosting Performance:
Mean Squared Error (MSE): 124775139.79080595
Root Mean Squared Error (RMSE): 11170.27930674994
Mean Absolute Error (MAE): 6590.766318063166
R^2 Score: 0.6882816002763905
Adjusted R^2 Score: 0.6754

Model building

```
[10…
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error

# Importing ensemble models
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor, AdaBoostRegressor, Bagg
from sklearn.tree import DecisionTreeRegressor
from catboost import CatBoostRegressor
from xgboost import XGBRegressor


# Create a dictionary to store models and their names
models = {
    'Random Forest': RandomForestRegressor(n_estimators=100, random_state=42),
    'Decision Tree': DecisionTreeRegressor(random_state=42),
    'CatBoost': CatBoostRegressor(verbose=0, random_state=42),  # Verbose is set to 0 to suppress outp
    'XGBoost': XGBRegressor(n_estimators=100, random_state=42),
    'Gradient Boosting': GradientBoostingRegressor(n_estimators=100, random_state=42),
    'AdaBoost': AdaBoostRegressor(n_estimators=100, random_state=42),
    'Bagging': BaggingRegressor(n_estimators=100, random_state=42)
}

# Function to calculate Adjusted R²
def adjusted_r2(r2, n, p):
    #return 1 - (1 - r2) * (n - 1) / (n - p - 1)
    return 1 - ((1-r2)*(n-1)/(n-p-1))

# Iterate through models, train, predict, and evaluate each model
for model_name, model in models.items():
    # Train the model
    model.fit(X_train, y_train)

    # Make predictions
    y_pred = model.predict(X_test)

    # Calculate evaluation metrics
    mse = mean_squared_error(y_test, y_pred)
    rmse = np.sqrt(mse)
    mae = mean_absolute_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    # Calculate Adjusted R²
    n = X_test.shape[0]  # number of observations
    p = X_test.shape[1]  # number of features
    adj_r2 = adjusted_r2(r2, n, p)

    # Print performance metrics
    print(f"{model_name} Performance:")
    print(f"Mean Squared Error (MSE): {mse}")
    print(f"Root Mean Squared Error (RMSE): {rmse}")
    print(f"Mean Absolute Error (MAE): {mae}")
    print(f"R² Score: {r2}")
    print(f"Adjusted R² Score: {adj_r2}")
    print("-" * 40)
```

Figure 23

The results of the model are depicted in Figure 24

The top 5 features which effect Annual Income are as follows:

1. Total Monthly Income
2. Expenditure on Household: Food
3. Main sources of Income in Household: Others
4. No. of Cows
5. What are the different types of health risks existing in the area related to poor sanitati on?

```
Random Forest Performance:
Mean Squared Error (MSE): 127118756.42666666
Root Mean Squared Error (RMSE): 11274.695402833137
Mean Absolute Error (MAE): 6037.626666666667
R² Score: 0.6824266805502259
Adjusted R² Score: 0.6513
----------------------------------------
Decision Tree Performance:
Mean Squared Error (MSE): 474200000.0
Root Mean Squared Error (RMSE): 21776.13372479146
Mean Absolute Error (MAE): 9346.666666666666
R² Score: -0.18466599513942183
Adjusted R² Score: 0.1912
----------------------------------------
CatBoost Performance:
Mean Squared Error (MSE): 132629528.17306119
Root Mean Squared Error (RMSE): 11516.489403158464
Mean Absolute Error (MAE): 6623.659981326421
R² Score: 0.6686594433192503
Adjusted R² Score: 0.6417
----------------------------------------
XGBoost Performance:
Mean Squared Error (MSE): 189630855.73643163
Root Mean Squared Error (RMSE): 13770.651972090196
Mean Absolute Error (MAE): 8101.8254296875
R² Score: 0.5262563761700971
Adjusted R² Score: 0.5094
----------------------------------------
Gradient Boosting Performance:
Mean Squared Error (MSE): 124775139.79080595
Root Mean Squared Error (RMSE): 11170.27930674994
Mean Absolute Error (MAE): 6590.766318063166
R² Score: 0.6882816002763905
Adjusted R² Score: 0.6754
----------------------------------------
AdaBoost Performance:
Mean Squared Error (MSE): 149444568.7224416
Root Mean Squared Error (RMSE): 12224.752296976885
Mean Absolute Error (MAE): 7672.371799487616
R² Score: 0.6266514155973155
Adjusted R² Score: 0.6134
----------------------------------------
Bagging Performance:
Mean Squared Error (MSE): 135371397.65333334
Root Mean Squared Error (RMSE): 11634.92147173041
Mean Absolute Error (MAE): 6251.2
R² Score: 0.6618095919139593
Adjusted R² Score: 0.6497
----------------------------------------
```

Getting most important features

```
[10…] feature_importances=model.feature_importances_

[10…] importance_df = pd.DataFrame({
          'Feature': X.columns,
          'Importance': feature_importances
      })

      # Sort the DataFrame by importance scores in descending order
      importance_df = importance_df.sort_values(by='Importance', ascending=False)

      # Print the feature importance scores
      print(importance_df)
```

```
                                              Feature  Importance
124           79Total monthly income (approximately)    0.564276
118                78Expenditure on Household: Food    0.070328
130  80Main sources of Income in Household: Others:...    0.045837
90                                        63No of cows    0.042928
153  89What are different types of health risks tha...    0.019703
..                                                 ...         ...
183  106Participation rate (%) of the female member...    0.000000
17         32Problem resolution satisfactory Ordinal    0.000000
16   30Approached any elected representatives9mukhi...    0.000000
186    109Adult above 70 years undernourished (yes/no)    0.000000
300  127 What are the policy interventions required...    0.000000

[301 rows x 2 columns]
```

```
[10…] top_5_features = importance_df['Feature'].head(5)
      print(top_5_features)
```

```
124             79Total monthly income (approximately)
118                  78Expenditure on Household: Food
130    80Main sources of Income in Household: Others:...
90                                          63No of cows
153    89What are different types of health risks tha...
Name: Feature, dtype: object
```

Figure 24

The top 5 most important independent features were visualised using a bar plot:

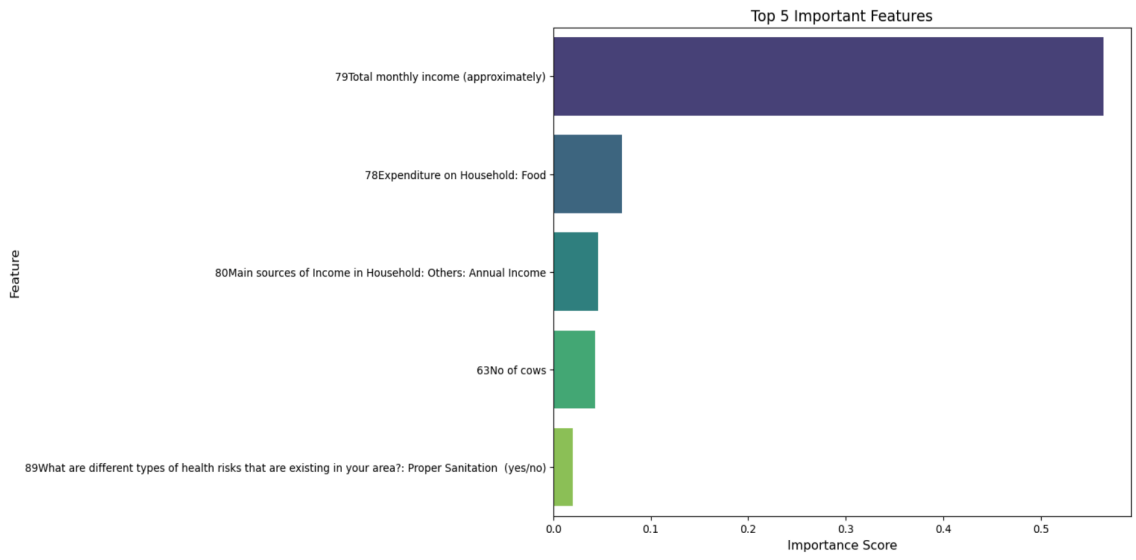


Figure 25

2. PDS Card Possession:

This variable is discrete in nature, so a classification model was developed, depicted in Figure 26

The results of the model are also depicted in Figure 28

The model with the best Accuracy score was selected and top 5 most important indepen dent features for this dependent variable were calculated:

Random Forest Performance:
Accuracy: 0.8533333333333334
F1-Score: 0.846973803071364
Precision: 0.8444444444444443
Recall: 0.8533333333333334

The top 5 features which effect PDS Card Possession are as follows:
1. Please specify how happy you are today
2. How do you feel about your life as a whole
3. Please specify how happy you were yesterday
4. What kind of interventions in policies of skill training can help to increase livelihood opportunities
5. Paddy crops area

Model building

```
[179]: from sklearn.metrics import accuracy_score, f1_score, confusion_matrix, roc_auc_score, precision_score, recall_score

# Importing ensemble classification models
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier, BaggingClassifier
from sklearn.tree import DecisionTreeClassifier
from catboost import CatBoostClassifier
from xgboost import XGBClassifier

# Create a dictionary to store classification models and their names
classification_models = {
    'Random Forest': RandomForestClassifier(n_estimators=100, random_state=42),
    'Decision Tree': DecisionTreeClassifier(random_state=42),
    'CatBoost': CatBoostClassifier(verbose=0, random_state=42),  # Verbose is set to 0 to suppress output
    'XGBoost': XGBClassifier(n_estimators=100, random_state=42),
    'Gradient Boosting': GradientBoostingClassifier(n_estimators=100, random_state=42),
    'AdaBoost': AdaBoostClassifier(n_estimators=100, random_state=42),
    'Bagging': BaggingClassifier(n_estimators=100, random_state=42)
}

# Iterate through models, train, predict, and evaluate each classification model
for model_name, model in classification_models.items():
    # Train the model
    model.fit(X_train, y_train)

    # Make predictions
    y_pred = model.predict(X_test)

    # Calculate classification metrics
    accuracy = accuracy_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred, average='weighted')  # 'weighted' accounts for label imbalance
    precision = precision_score(y_test, y_pred, average='weighted')
    recall = recall_score(y_test, y_pred, average='weighted')
    confusion = confusion_matrix(y_test, y_pred)


    # Print performance metrics
    print(f"{model_name} Performance:")
    print(f"Accuracy: {accuracy}")
    print(f"F1-Score: {f1}")
    print(f"Precision: {precision}")
    print(f"Recall: {recall}")
    print("Confusion Matrix:")
    print(confusion)
    print("-" * 40)
```

Figure 26

The top 5 most important independent features were visualised using a bar plot:

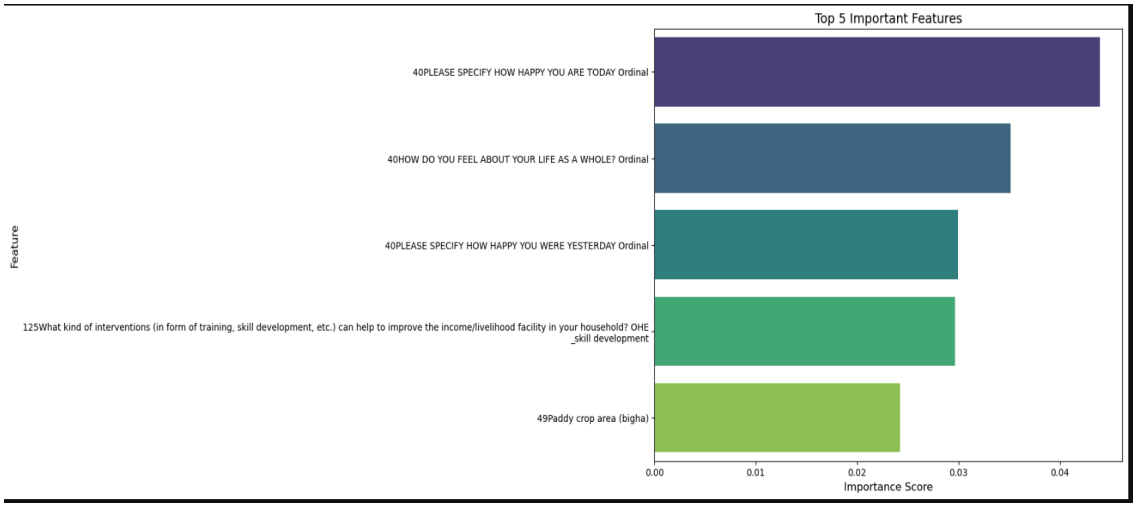


Figure 27

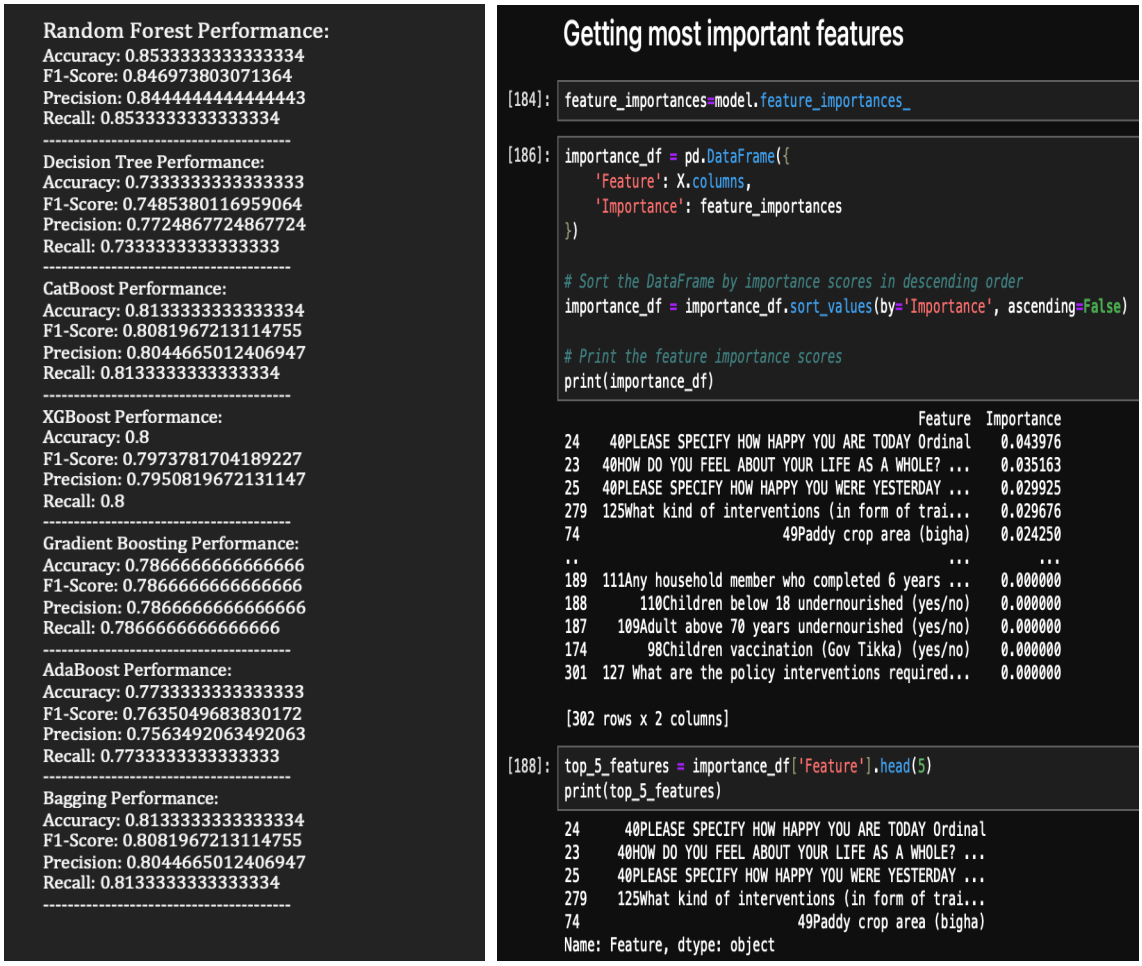


Figure 28

3. Expenditure on Agricultural Activities in Family: Labour

This variable is continuous in nature, so a regression model was developed, depicted in Figure 29

The results of the model are also depicted in Figure 30

AdaBoost Performance:

Mean Squared Error (MSE): 1763594.9713803893

Root Mean Squared Error (RMSE): 1328.0041307843849

Mean Absolute Error (MAE): 904.1240939951242

R^2 Score: 0.8988629708092158

Adjusted R^2 Score: 0.8950

The model with the best Adjusted R^2 Score and R^2 Score was selected and top 5 most im portant independent features for this dependent variable were calculated:

The top 5 features which effect Agricultural Activities in Family: Labour are as follows:

1. Expenditure in Agricultural Activities: Seeds
2. Main Source of Income: Animal Livestock
3. Have you ever been hospitalised for maternity
4. Expenditure in Agricultural Activities: Insecticides
5. Children Nutrition Intake in a day

Model building

[175]:

```
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error

# Importing ensemble models
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor, AdaBoostRegressor, BaggingRegressor
from sklearn.tree import DecisionTreeRegressor
from catboost import CatBoostRegressor
from xgboost import XGBRegressor


# Create a dictionary to store models and their names
models = {
    'Random Forest': RandomForestRegressor(n_estimators=100, random_state=42),
    'Decision Tree': DecisionTreeRegressor(random_state=42),
    'CatBoost': CatBoostRegressor(verbose=0, random_state=42),  # Verbose is set to 0 to suppress output
    'XGBoost': XGBRegressor(n_estimators=100, random_state=42),
    'Gradient Boosting': GradientBoostingRegressor(n_estimators=100, random_state=42),
    'AdaBoost': AdaBoostRegressor(n_estimators=100, random_state=42),
    'Bagging': BaggingRegressor(n_estimators=100, random_state=42)
}

# Function to calculate Adjusted R²
def adjusted_r2(r2, n, p):
    #return 1 - (1 - r2) * (n - 1) / (n - p - 1)
    return 1 - ((1-r2)*(n-1)/(n-p-1))

# Iterate through models, train, predict, and evaluate each model
for model_name, model in models.items():
    # Train the model
    model.fit(X_train, y_train)

    # Make predictions
    y_pred = model.predict(X_test)

    # Calculate evaluation metrics
    mse = mean_squared_error(y_test, y_pred)
    rmse = np.sqrt(mse)
    mae = mean_absolute_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    # Calculate Adjusted R²
    n = X_test.shape[0]  # number of observations
    p = X_test.shape[1]  # number of features
    adj_r2 = adjusted_r2(r2, n, p)

    # Print performance metrics
    print(f"{model_name} Performance:")
    print(f"Mean Squared Error (MSE): {mse}")
    print(f"Root Mean Squared Error (RMSE): {rmse}")
    print(f"Mean Absolute Error (MAE): {mae}")
    print(f"R² Score: {r2}")
    print(f"Adjusted R² Score: {adj_r2}")
    print("-" * 40)
```

Figure 29

```
Random Forest Performance:
Mean Squared Error (MSE): 6414716.453333333
Root Mean Squared Error (RMSE): 2532.7290524912714
Mean Absolute Error (MAE): 1219.76
R² Score: 0.6321347158959184
Adjusted R² Score: 0.6190
----------------------------------------
Decision Tree Performance:
Mean Squared Error (MSE): 8874400.0
Root Mean Squared Error (RMSE): 2978.9931184881916
Mean Absolute Error (MAE): 1178.6666666666667
R² Score: 0.4910790366179227 4
Adjusted R² Score: 0.4794
----------------------------------------
CatBoost Performance:
Mean Squared Error (MSE): 3445159.359669771
Root Mean Squared Error (RMSE): 1856.1140481311409
Mean Absolute Error (MAE): 985.6068695615351
R² Score: 0.8024301563679888
Adjusted R² Score: 0.7943
----------------------------------------
XGBoost Performance:
Mean Squared Error (MSE): 5200459.178543567
Root Mean Squared Error (RMSE): 2280.451529531721
Mean Absolute Error (MAE): 870.8706750488282
R² Score: 0.7017688299858517
Adjusted R² Score: 0.6935
----------------------------------------
Gradient Boosting Performance:
Mean Squared Error (MSE): 5470922.829223348
Root Mean Squared Error (RMSE): 2339.0003910267624
Mean Absolute Error (MAE): 904.019334986659
R² Score: 0.6862585282568573
Adjusted R² Score: 0.6783
----------------------------------------
AdaBoost Performance:
Mean Squared Error (MSE): 1763594.9713803893
Root Mean Squared Error (RMSE): 1328.0041307843849
Mean Absolute Error (MAE): 904.1240939951242
R² Score: 0.8988629708092158
Adjusted R² Score: 0.8950
----------------------------------------
Bagging Performance:
Mean Squared Error (MSE): 6013109.053333334
Root Mean Squared Error (RMSE): 2452.164157093349
Mean Absolute Error (MAE): 1205.8
R² Score: 0.6551657292499898
Adjusted R² Score: 0.6466
----------------------------------------
```

Getting most important features

```
[180]: feature_importances=model.feature_importances_

[182]: importance_df = pd.DataFrame({
           'Feature': X.columns,
           'Importance': feature_importances
       })

       # Sort the DataFrame by importance scores in descending order
       importance_df = importance_df.sort_values(by='Importance', ascending=False)

       # Print the feature importance scores
       print(importance_df)
```

	Feature	Importance
114	78Expenditure on Agricultural Activities in Fa...	0.120261
127	80Main sources of Income in Household: Livesto...	0.051343
167	96Have you ever been hospitalized for maternit...	0.042783
113	78Expenditure on Agricultural Activities in Fa...	0.040851
162	94Rate your children nutrition intake in a day...	0.040053
..	...	...
198	121Are you doing any extra job/business to imp...	0.000000
199	122Do female members participate in family inc...	0.000000
78	55Count of medicinal plants	0.000000
201	3Religion OHE_hindu	0.000000
219	26Decsion maker in family OHE_female mambers	0.000000

[302 rows x 2 columns]

```
[184]: top_5_features = importance_df['Feature'].head(5)
       print(top_5_features)
```

```
114    78Expenditure on Agricultural Activities in Fa...
127    80Main sources of Income in Household: Livesto...
167    96Have you ever been hospitalized for maternit...
113    78Expenditure on Agricultural Activities in Fa...
162    94Rate your children nutrition intake in a day...
Name: Feature, dtype: object
```

Figure 30

The top 5 most important independent features were visualised using a bar plot:

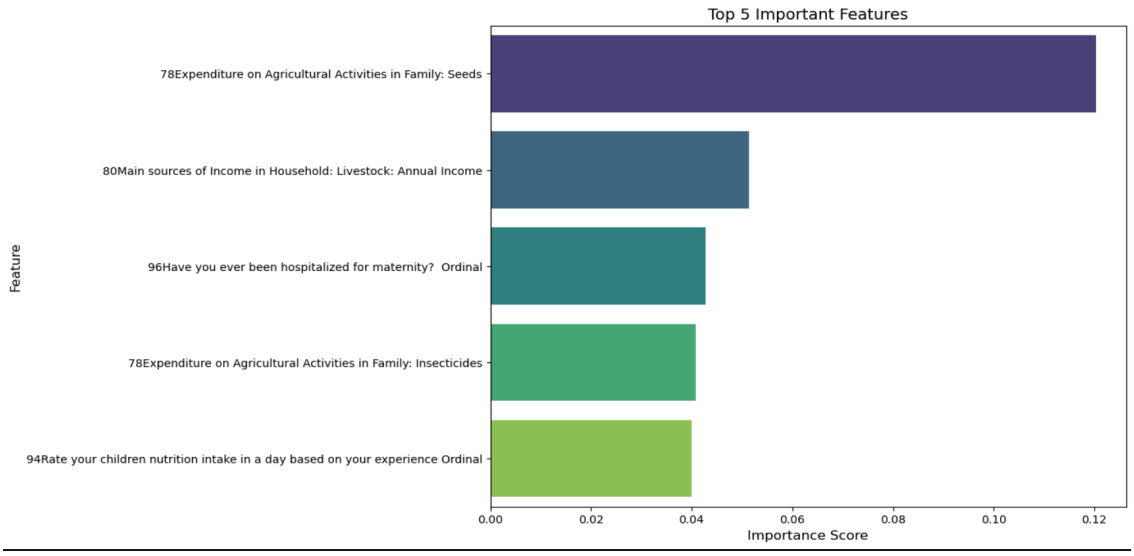


Figure 31

4. Expenditure on Household: Food

This variable is continuous in nature, so a regression model was developed, depicted in Figure 32

The results of the model are also depicted in Figure 34

The model with the best Adjusted R^2 Score and R^2 Score was selected and top 5 most im portant independent features for this dependent variable were calculated:

CatBoost Performance:
Mean Squared Error (MSE): 40577028.99341722
Root Mean Squared Error (RMSE): 6370.010125063948
Mean Absolute Error (MAE): 4475.445225635981
R^2 Score: 0.669017848586507
Adjusted R^2 Score: 0.6590

The top 5 features which effect Expenditure on Household: Food are as follows:
1. Expenditure in Household: Clothes
2. Expenditure in Agricultural Activities: Insecticides
3. Household Income
4. Expenditure in Household: Medicine
5. Expenditure in Agricultural Activities: Labour

Model building

[174]:

```
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error

# Importing ensemble models
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor, AdaBoostRegressor, BaggingRegressor
from sklearn.tree import DecisionTreeRegressor
from catboost import CatBoostRegressor
from xgboost import XGBRegressor


# Create a dictionary to store models and their names
models = {
    'Random Forest': RandomForestRegressor(n_estimators=100, random_state=42),
    'Decision Tree': DecisionTreeRegressor(random_state=42),
    'CatBoost': CatBoostRegressor(verbose=0, random_state=42),  # Verbose is set to 0 to suppress output
    'XGBoost': XGBRegressor(n_estimators=100, random_state=42),
    'Gradient Boosting': GradientBoostingRegressor(n_estimators=100, random_state=42),
    'AdaBoost': AdaBoostRegressor(n_estimators=100, random_state=42),
    'Bagging': BaggingRegressor(n_estimators=100, random_state=42)
}

# Function to calculate Adjusted R²
def adjusted_r2(r2, n, p):
    #return 1 - (1 - r2) * (n - 1) / (n - p - 1)
    return 1 - ((1-r2)*(n-1)/(n-p-1))

# Iterate through models, train, predict, and evaluate each model
for model_name, model in models.items():
    # Train the model
    model.fit(X_train, y_train)

    # Make predictions
    y_pred = model.predict(X_test)

    # Calculate evaluation metrics
    mse = mean_squared_error(y_test, y_pred)
    rmse = np.sqrt(mse)
    mae = mean_absolute_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    # Calculate Adjusted R²
    n = X_test.shape[0]  # number of observations
    p = X_test.shape[1]  # number of features
    adj_r2 = adjusted_r2(r2, n, p)

    # Print performance metrics
    print(f"{model_name} Performance:")
    print(f"Mean Squared Error (MSE): {mse}")
    print(f"Root Mean Squared Error (RMSE): {rmse}")
    print(f"Mean Absolute Error (MAE): {mae}")
    print(f"R² Score: {r2}")
    print(f"Adjusted R² Score: {adj_r2}")
    print("-" * 40)
```

Figure 32

The top 5 most important independent features were visualised using a bar plot:

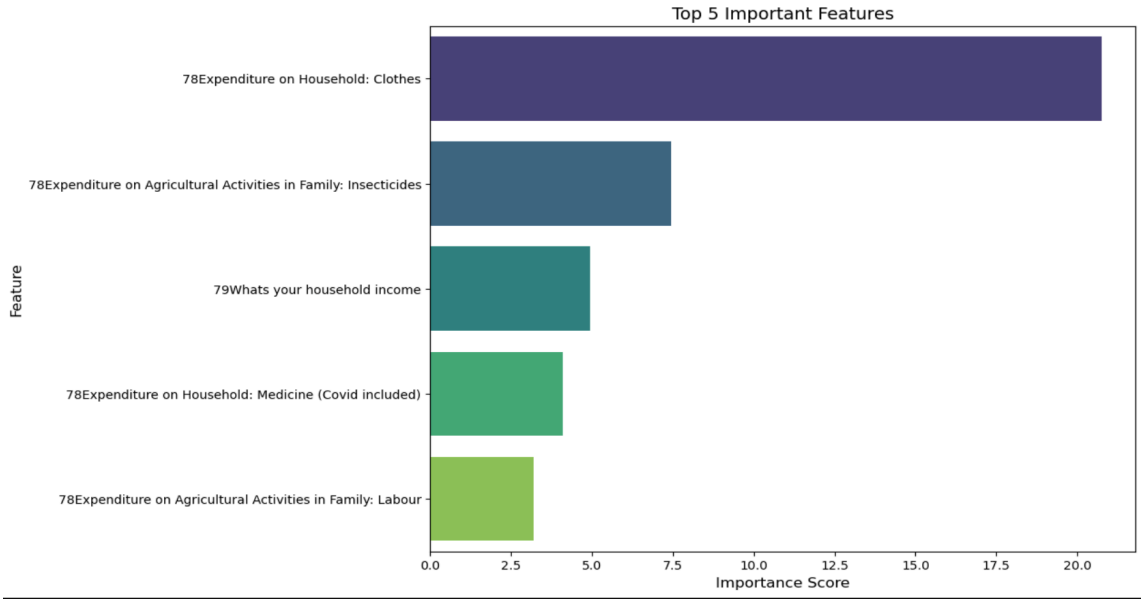


Figure 33

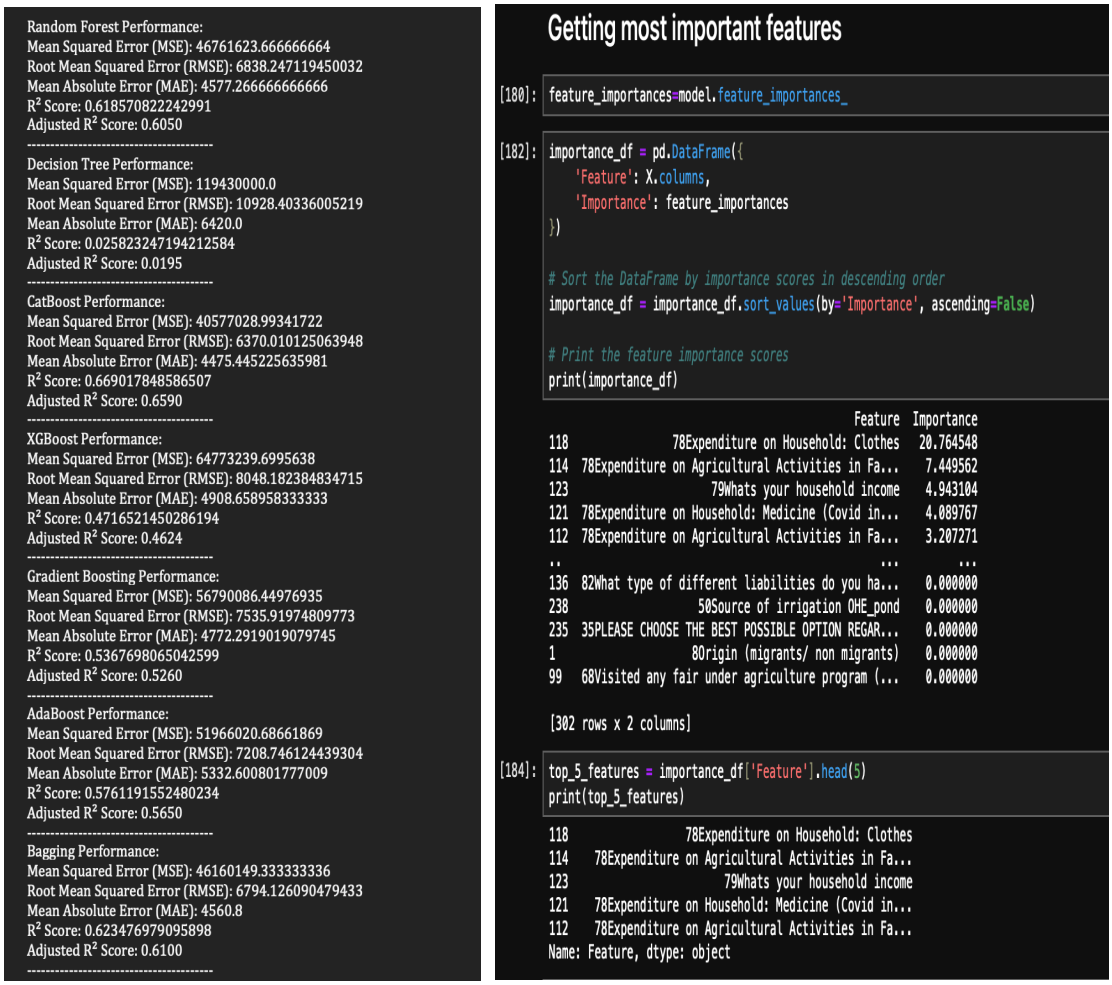


Figure 34

5. Bank Account Possession

This variable is discrete in nature, so a classification model was developed, depicted in Figure 35

The results of the model are also depicted in Figure 36

The model with the best Accuracy score was selected and top 5 most important indepen dent features for this dependent variable were calculated:

CatBoost Performance:

Accuracy: 0.88

F1-Score: 0.873591235878124

Precision: 0.8716923076923077

Recall: 0.88

The top 5 features which effect Bank Account Possession are as follows:

1. LPG Connection
2. PDS Card
3. Electricity Connection
4. Social responsibilities are supportive and rewarding
5. Expenditure in Agricultural Activities: Insecticides

Model building

```
[100]: from sklearn.metrics import accuracy_score, f1_score, confusion_matrix, roc_auc_score, precision_score, recall_score

       # Importing ensemble classification models
       from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier, BaggingClassifier
       from sklearn.tree import DecisionTreeClassifier
       from catboost import CatBoostClassifier
       from xgboost import XGBClassifier

       # Create a dictionary to store classification models and their names
       classification_models = {
           'Random Forest': RandomForestClassifier(n_estimators=100, random_state=42),
           'Decision Tree': DecisionTreeClassifier(random_state=42),
           'CatBoost': CatBoostClassifier(verbose=0, random_state=42),  # Verbose is set to 0 to suppress output
           'XGBoost': XGBClassifier(n_estimators=100, random_state=42),
           'Gradient Boosting': GradientBoostingClassifier(n_estimators=100, random_state=42),
           'AdaBoost': AdaBoostClassifier(n_estimators=100, random_state=42),
           'Bagging': BaggingClassifier(n_estimators=100, random_state=42)
       }

       # Iterate through models, train, predict, and evaluate each classification model
       for model_name, model in classification_models.items():
           # Train the model
           model.fit(X_train, y_train)

           # Make predictions
           y_pred = model.predict(X_test)

           # Calculate classification metrics
           accuracy = accuracy_score(y_test, y_pred)
           f1 = f1_score(y_test, y_pred, average='weighted')  # 'weighted' accounts for label imbalance
           precision = precision_score(y_test, y_pred, average='weighted')
           recall = recall_score(y_test, y_pred, average='weighted')
           confusion = confusion_matrix(y_test, y_pred)


           # Print performance metrics
           print(f"{model_name} Performance:")
           print(f"Accuracy: {accuracy}")
           print(f"F1-Score: {f1}")
           print(f"Precision: {precision}")
           print(f"Recall: {recall}")
           print("Confusion Matrix:")
           print(confusion)
           print("-" * 40)
```

Figure 35

```
Random Forest Performance:
Accuracy: 0.8133333333333334
F1-Score: 0.7415686274509804
Precision: 0.6814414414414415
Recall: 0.8133333333333334
----------------------------------------
Decision Tree Performance:
Accuracy: 0.8
F1-Score: 0.8078654887432317
Precision: 0.8184322033898306
Recall: 0.8
----------------------------------------
CatBoost Performance:
Accuracy: 0.88
F1-Score: 0.873591235878124
Precision: 0.8716923076923077
Recall: 0.88
----------------------------------------
XGBoost Performance:
Accuracy: 0.8133333333333334
F1-Score: 0.8133333333333334
Precision: 0.8133333333333334
Recall: 0.8133333333333334
----------------------------------------
Gradient Boosting Performance:
Accuracy: 0.8133333333333334
F1-Score: 0.8133333333333334
Precision: 0.8133333333333334
Recall: 0.8133333333333334
----------------------------------------
AdaBoost Performance:
Accuracy: 0.8
F1-Score: 0.7968000000000001
Precision: 0.7939153439153439
Recall: 0.8
----------------------------------------
Bagging Performance:
Accuracy: 0.84
F1-Score: 0.8444028103044496
Precision: 0.8502222222222222
Recall: 0.84
----------------------------------------
```

Getting most important features

```
[104]: feature_importances=model.feature_importances_
```

```
[105]: importance_df = pd.DataFrame({
           'Feature': X.columns,
           'Importance': feature_importances
       })

       # Sort the DataFrame by importance scores in descending order
       importance_df = importance_df.sort_values(by='Importance', ascending=False)

       # Print the feature importance scores
       print(importance_df)
```

```
                                               Feature  Importance
11                          17LPG connection (yes/no)   24.085413
8                             12PDS Card (yes/no)    5.148746
10                   16Electricty connection (yes/no)    2.804681
49   42MY SOCIAL RELATIONSHIPS ARE SUPPORTIVE AND R...    2.787385
113  78Expenditure on Agricultural Activities in Fa...    1.844470
..                                                 ...         ...
209                  19Dialect spoken OHE_santhali    0.000000
210              20Traditional folk dance OHE_santhal    0.000000
241                  50Source of irrigation OHE_well    0.000000
102        70Source of drinking water: Others (yes/no)    0.000000
1                    80rigin (migrants/ non migrants)    0.000000

[302 rows x 2 columns]
```

```
[106]: top_5_features = importance_df['Feature'].head(5)
       print(top_5_features)
```

```
11                          17LPG connection (yes/no)
8                             12PDS Card (yes/no)
10                   16Electricty connection (yes/no)
49     42MY SOCIAL RELATIONSHIPS ARE SUPPORTIVE AND R...
113    78Expenditure on Agricultural Activities in Fa...
Name: Feature, dtype: object
```

Figure 36

The top 5 most important independent features were visualised using a bar plot:

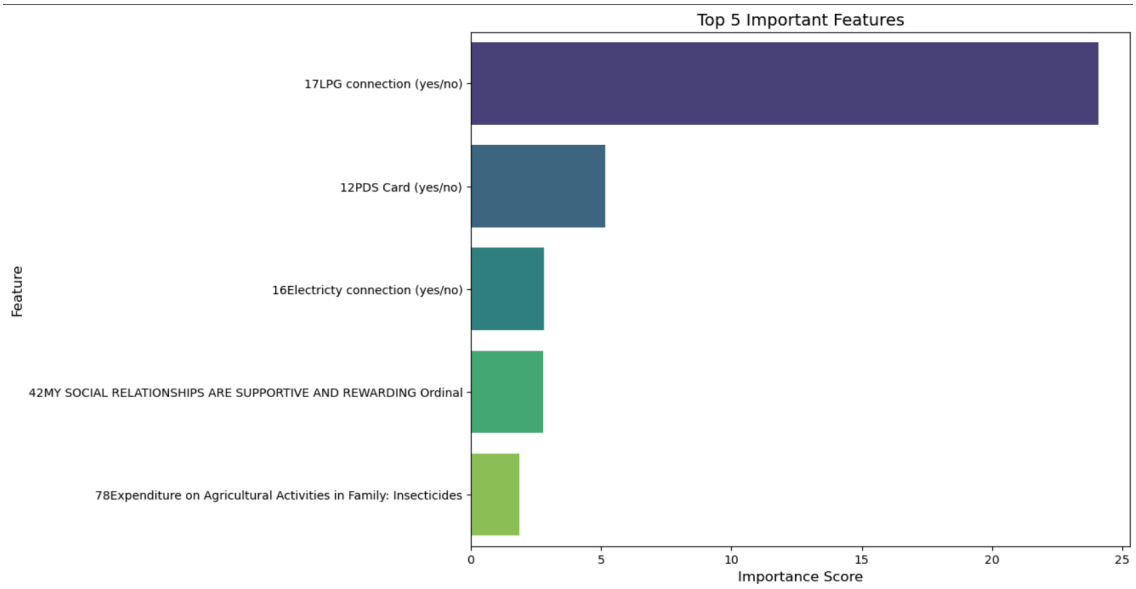


Figure 37

6. LPG Connection

This variable is discrete in nature, so a classification model was developed, depicted in Figure 38

The results of the model are also depicted in Figure 40

The model with the best Accuracy score was selected and top 5 most important indepen dent features for this dependent variable were calculated:

XGBoost Performance:
Accuracy: 0.84
F1-Score: 0.8308771929824562
Precision: 0.8360000000000001
Recall: 0.84

The top 5 features which effect LPG Connection are as follows:
1. Bank Account
2. Non Irrigated Land Area
3. Paddy Crop Area
4. Activity done in free time: Painting
5. Expenditure in Agricultural Activities: Seeds

Model building

```
[100]: from sklearn.metrics import accuracy_score, f1_score, confusion_matrix, roc_auc_score, precision_score, recall_score

# Importing ensemble classification models
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier, BaggingClassifier
from sklearn.tree import DecisionTreeClassifier
from catboost import CatBoostClassifier
from xgboost import XGBClassifier

# Create a dictionary to store classification models and their names
classification_models = {
    'Random Forest': RandomForestClassifier(n_estimators=100, random_state=42),
    'Decision Tree': DecisionTreeClassifier(random_state=42),
    'CatBoost': CatBoostClassifier(verbose=0, random_state=42),  # Verbose is set to 0 to suppress output
    'XGBoost': XGBClassifier(n_estimators=100, random_state=42),
    'Gradient Boosting': GradientBoostingClassifier(n_estimators=100, random_state=42),
    'AdaBoost': AdaBoostClassifier(n_estimators=100, random_state=42),
    'Bagging': BaggingClassifier(n_estimators=100, random_state=42)
}

# Iterate through models, train, predict, and evaluate each classification model
for model_name, model in classification_models.items():
    # Train the model
    model.fit(X_train, y_train)

    # Make predictions
    y_pred = model.predict(X_test)

    # Calculate classification metrics
    accuracy = accuracy_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred, average='weighted')  # 'weighted' accounts for label imbalance
    precision = precision_score(y_test, y_pred, average='weighted')
    recall = recall_score(y_test, y_pred, average='weighted')
    confusion = confusion_matrix(y_test, y_pred)


    # Print performance metrics
    print(f"{model_name} Performance:")
    print(f"Accuracy: {accuracy}")
    print(f"F1-Score: {f1}")
    print(f"Precision: {precision}")
    print(f"Recall: {recall}")
    print("Confusion Matrix:")
    print(confusion)
    print("-" * 40)
```

Figure 38

The top 5 most important independent features were visualised using a bar plot:

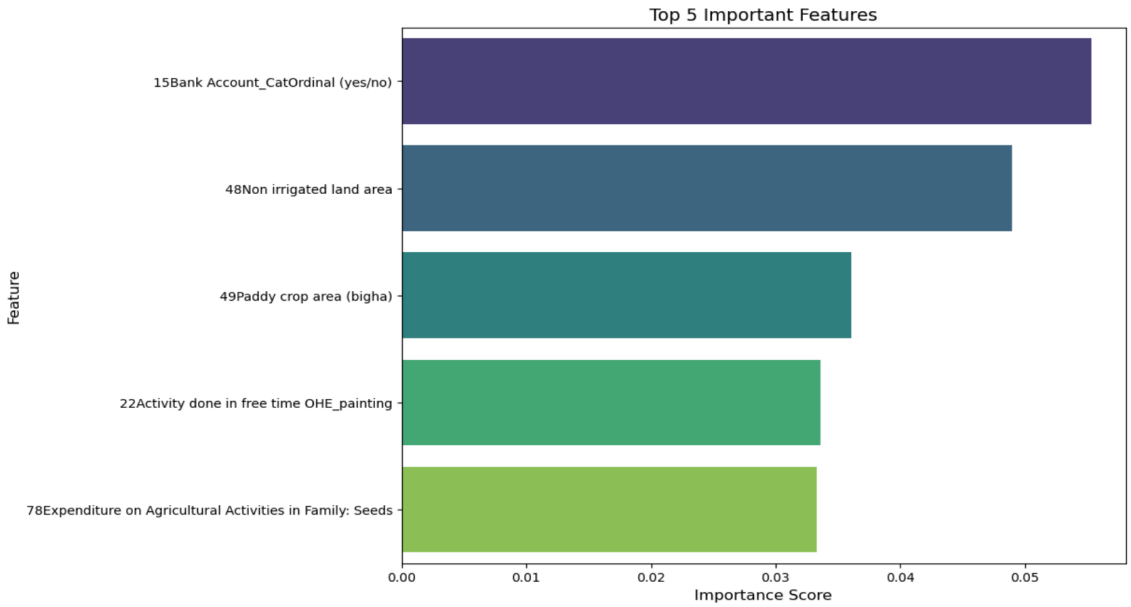


Figure 39

Figure 40

7. Please specify the extent of your satisfaction level regarding the existing educational facilities in your area?

This variable is discrete in nature, so a classification model was developed, depicted in Figure 41

The results of the model are also depicted in Figure 42

The model with the best Accuracy score was selected and top 5 most important indepen dent features for this dependent variable were calculated:

Random Forest Performance:

Accuracy: 0.84

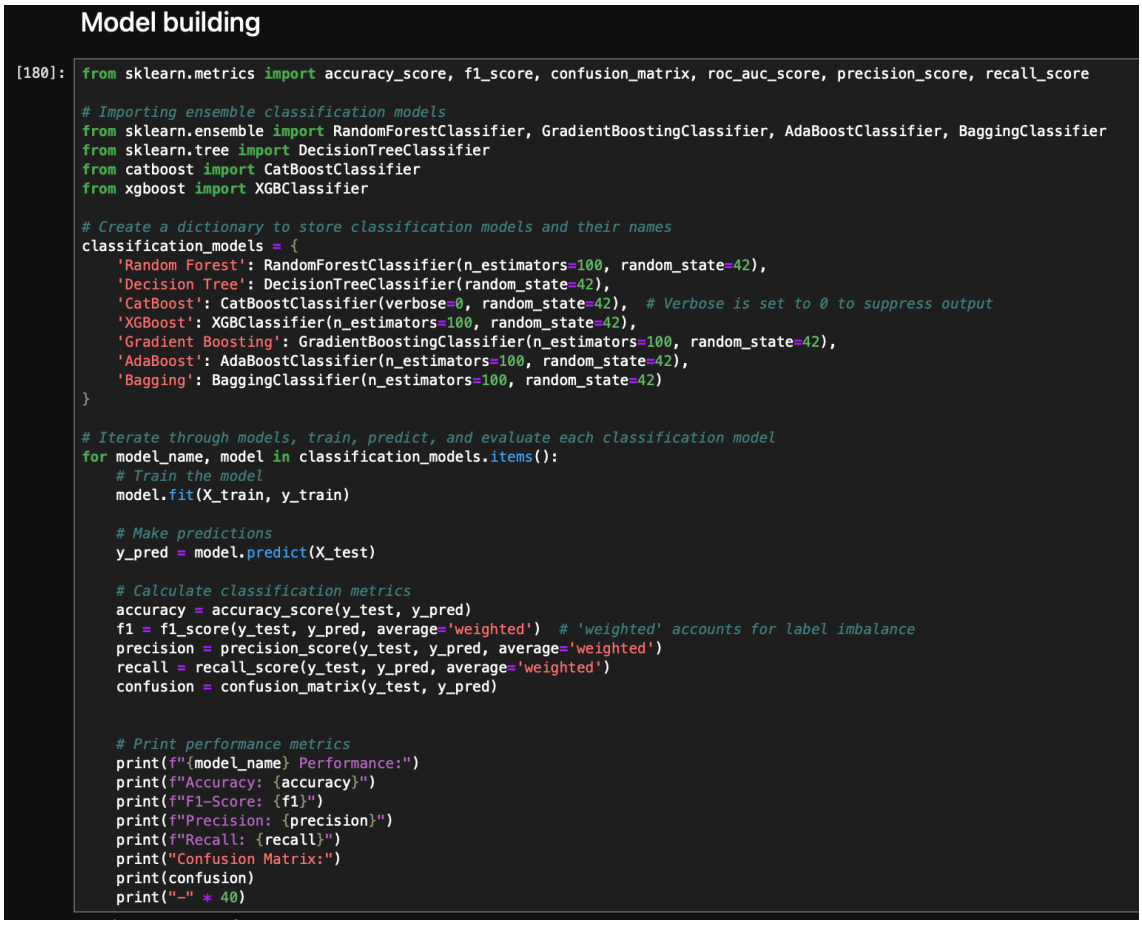
F1-Score: 0.8480740217987925

Precision: 0.9207267645003494

Recall: 0.84

The top 5 features(with their feature importance scores), which effect Satisfaction with E ducation Facilities in nearby area are as follows:

1. Always Optimistic about my future
2. When things go wrong it takes long time to get normal
3. On most days sense of accomplishment is felt in daily chores
4. Things I do in life are valuable and worthwhile
5. How much satisfied were you with your health 6 months ago?



Model building

[180]:

```
from sklearn.metrics import accuracy_score, f1_score, confusion_matrix, roc_auc_score, precision_score, recall_score

# Importing ensemble classification models
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier, BaggingClassifier
from sklearn.tree import DecisionTreeClassifier
from catboost import CatBoostClassifier
from xgboost import XGBClassifier

# Create a dictionary to store classification models and their names
classification_models = {
    'Random Forest': RandomForestClassifier(n_estimators=100, random_state=42),
    'Decision Tree': DecisionTreeClassifier(random_state=42),
    'CatBoost': CatBoostClassifier(verbose=0, random_state=42),  # Verbose is set to 0 to suppress output
    'XGBoost': XGBClassifier(n_estimators=100, random_state=42),
    'Gradient Boosting': GradientBoostingClassifier(n_estimators=100, random_state=42),
    'AdaBoost': AdaBoostClassifier(n_estimators=100, random_state=42),
    'Bagging': BaggingClassifier(n_estimators=100, random_state=42)
}

# Iterate through models, train, predict, and evaluate each classification model
for model_name, model in classification_models.items():
    # Train the model
    model.fit(X_train, y_train)

    # Make predictions
    y_pred = model.predict(X_test)

    # Calculate classification metrics
    accuracy = accuracy_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred, average='weighted')  # 'weighted' accounts for label imbalance
    precision = precision_score(y_test, y_pred, average='weighted')
    recall = recall_score(y_test, y_pred, average='weighted')
    confusion = confusion_matrix(y_test, y_pred)


    # Print performance metrics
    print(f"{model_name} Performance:")
    print(f"Accuracy: {accuracy}")
    print(f"F1-Score: {f1}")
    print(f"Precision: {precision}")
    print(f"Recall: {recall}")
    print("Confusion Matrix:")
    print(confusion)
    print("-" * 40)
```

Figure 41

```
Random Forest Performance:
Accuracy: 0.84
F1-Score: 0.8480740217987925
Precision: 0.9207267645003494
Recall: 0.84
----------------------------------------
Decision Tree Performance:
Accuracy: 0.7733333333333333
F1-Score: 0.7956743871055127
Precision: 0.8231372549019608
Recall: 0.7733333333333333
----------------------------------------
CatBoost Performance:
Accuracy: 0.8266666666666667
F1-Score: 0.8387160493827159
Precision: 0.918272604588394
Recall: 0.8266666666666667
----------------------------------------
XGBoost Performance:
Accuracy: 0.8133333333333334
F1-Score: 0.8301350540655225
Precision: 0.8589667565139265
Recall: 0.8133333333333334
----------------------------------------
Gradient Boosting Performance:
Accuracy: 0.8
F1-Score: 0.8229207639171069
Precision: 0.8582483660130718
Recall: 0.8
----------------------------------------
AdaBoost Performance:
Accuracy: 0.6666666666666666
F1-Score: 0.7008547008547008
Precision: 0.8440860215053764
Recall: 0.6666666666666666
----------------------------------------
Bagging Performance:
Accuracy: 0.8
F1-Score: 0.8121707103724506
Precision: 0.9142483660130718
Recall: 0.8
----------------------------------------
```

Getting most important features

```
[187]: feature_importances=model.feature_importances_

[189]: importance_df = pd.DataFrame({
           'Feature': X.columns,
           'Importance': feature_importances
       })

       # Sort the DataFrame by importance scores in descending order
       importance_df = importance_df.sort_values(by='Importance', ascending=False)

       # Print the feature importance scores
       print(importance_df)
```

	Feature	Importance
59	43I am always optimisic about my future. Ordinal	0.073392
61	43when things go wrong in my life it generally...	0.063079
57	43most days I feel a sense of accomplishment f...	0.049805
58	43I generally feel that what I do in my life i...	0.032460
65	43how much satisfied were you with your health...	0.030460
..	...	...
212	21 Traditional Festival OHE_others	0.000000
97	67Participation in skilled development program...	0.000000
98	68Visited any fair under agriculture program (...	0.000000
209	19Dialect spoken OHE_santhali	0.000000
301	127 What are the policy interventions required...	0.000000

[302 rows x 2 columns]

```
[191]: top_5_features = importance_df['Feature'].head(5)
       print(top_5_features)
```

```
59    43I am always optimisic about my future. Ordinal
61    43when things go wrong in my life it generally...
57    43most days I feel a sense of accomplishment f...
58    43I generally feel that what I do in my life i...
65    43how much satisfied were you with your health...
Name: Feature, dtype: object
```

Figure 42

The top 5 most important independent features were visualised using a bar plot:

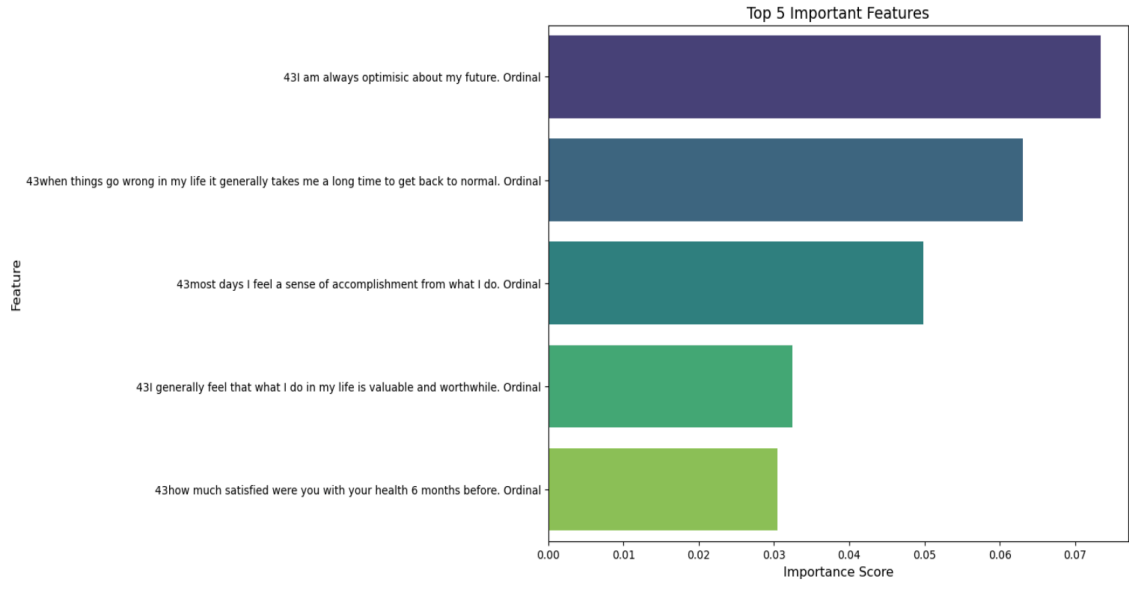


Figure 43

Conclusion and Future Scope

1. Conclusion

This thesis focused on identifying the socio-economic factors that significantly influence economic outcomes for tribal communities in Jharkhand, using advanced analytical techniques and machine learning models. By building ensemble-based regression and classification models, we were able to assess the relationships between various independent variables and crucial dependent variables like annual income, PDS card possession, agricultural labour, household food expenditure, bank account possession, LPG connection, and satisfaction with educational facilities. Ensemble methods provided interpretability, allowing us to understand the impact of each feature on these economic indicators and to identify the top factors affecting each outcome.

My analysis highlights several key factors that strongly influence economic conditions. For instance, variables like total monthly income, household expenditure on food, sources of income, and livestock ownership were found to have a substantial impact on annual income. Similarly, variables related to happiness, skill training, and agricultural land influenced PDS card possession. Such insights can guide policymakers in focusing resources and designing initiatives that address the most influential factors, thus enhancing the overall socio-economic status of these communities.

2. Future Scope

In future research, several directions could be explored to further develop the insights generated in this study and support policy formulation:

2.1. Policy Recommendations Based on Top Features

- Annual Income: Policies could focus on improving household monthly income, increasing food security, supporting diverse income sources, promoting livestock

ownership, and enhancing sanitation facilities. For instance, providing financial assistance or subsidies to small farmers and livestock owners could directly raise their income levels. Health awareness campaigns could address sanitation-related health risks, thus potentially reducing medical expenses and increasing disposable income.

- PDS Card Possession: Since happiness and mental well-being were major influences on PDS card possession, the government could consider mental health programs and happiness-focused interventions in these communities. Skill development programs aimed at increasing livelihood opportunities could also be prioritized to improve satisfaction with life, which indirectly supports PDS card possession.

- Agricultural Activities in Family (Labour): Providing subsidies or assistance for agricultural inputs like seeds and insecticides can reduce household expenditure, allowing more investment in labour. Additionally, maternity-related healthcare and child nutrition support programs would help families manage labour in agricultural activities more effectively.

- Household Food Expenditure: Policies aimed at reducing the cost of essential items like clothes and medicines can help allocate more resources toward food expenditure. Financial assistance for medical expenses and promoting agricultural self-sufficiency could also reduce dependence on external resources for food.

- Bank Account Possession: Expanding access to essential services like LPG and electricity could positively influence bank account ownership. Moreover, fostering community support programs could encourage individuals to open bank accounts by enhancing their trust in formal financial services.

- LPG Connection: Providing financial support for agricultural inputs like seeds and non-irrigated land management could help increase LPG connection uptake. Additionally, promoting creative activities like painting and offering incentives could lead to greater resource allocation for energy needs.

- Satisfaction with Education Facilities: Programs that promote optimism, a sense of accomplishment, and health satisfaction could improve satisfaction with educational facilities. Community-based initiatives that empower individuals to view education as valuable and worthwhile could indirectly enhance perceptions of educational quality.

2.2. Use of Deep Learning Models

While ensemble methods provided interpretability and allowed us to identify the most impactful features, future studies could consider applying deep learning (DL) models to potentially improve predictive accuracy. Although DL models, such as Artificial Neural Networks (ANNs), might lack interpretability, they could be valuable for understanding complex, non-linear relationships within the data. By analyzing feature interactions at a deeper level, DL models could reveal intricate patterns that ensemble methods may overlook. However, the trade-off between accuracy and interpretability must be carefully considered, particularly for applications that require transparency, such as government policy recommendations.

2.3. Advanced Feature Importance Analysis and Future Scope

This study lays the groundwork for understanding key socio-economic determinants by identifying influential features for each dependent variable. To deepen this analysis in future work, advanced interpretability techniques such as SHAP (SHapley Additive exPlanations) could be applied to further refine feature importance insights. SHAP is particularly beneficial as it attributes a specific contribution value to each feature for individual predictions, allowing a more precise understanding of how each variable impacts the model's outcomes across varying scenarios.

By employing SHAP, future analyses could capture complex interactions and non-linear relationships between variables, providing a clearer picture of underlying socio-economic influences. For instance, while the current study identifies general patterns, SHAP could uncover subtle, context-dependent variations in feature importance, such as how household income and expenditure contribute differently to economic outcomes across

demographic groups. Additionally, SHAP values allow for visualizations that highlight each feature's impact on individual predictions, which could enhance communication with policymakers and stakeholders, making it easier to target interventions.

2.4. Longitudinal and Broader Geographic Studies

Extending this research to a longitudinal study could help observe changes in feature importance over time, providing insights into how socio-economic improvements evolve. Additionally, applying these models to other regions or states with tribal populations could assess the generalizability of these findings and help design region-specific policies.

References

EFFECTIVENESS OF PUBLIC DISTRIBUTION SYSTEM (PDS) IN QUALITY OF LIFE IMPROVEMENT – A STUDY OF THE TRIBAL POPULATION OF PURULIA DISTRICT IN WEST BENGAL – The Social Science Review a Multidisciplinary Journal. (n.d.). https://tssreview.in/?article=effectiveness-of-public-distribution-system-pds-in-quality-of-life-improvement-a-study-of-the-tribal-population-of-purulia-district-in-west-bengal

Kumari, M. & Central University of Jharkhand, India. (2018). Land and property rights among tribal communities in Jharkhand. In *International Journal of Science and Research (IJSR)*. https://www.ijsr.net/archive/v7i8/ES24320153140.pdf

Kumari, P. (2024). Enhancing the Socio-Economic Empowerment of Tribal Communities with Entrepreneurship and Skill Training Program in Jharkhand. *International Journal of Novel Research and Development*, *9*(7), b124-c124. https://www.ijnrd.org/papers/IJNRD2407113.pdf

Majhi, A. (2018). A COMPARATIVE STUDY OF TRADITIONAL FISHING PRACTICE AMONG TRIBAL PEOPLE AT SOME SELECTED REGIONS OF PURULIA DISTRICT. In Sidho-Kanho-Birsha University, *International Journal for Research Under Literal Access* (Vol. 1, Issue 7, pp. 206–208). https://d1wqtxts1xzle7.cloudfront.net/58140637/A_COMPARATIVE_STUDY_OF_TRADITIONAL_FISHING_PRACTICE_AMONG_TRIBAL_PEOPLE_AT_SOME_SELECTED_REGIONS_OF_PURULIA_DISTRICT-libre.pdf?1546966802=&response-content-disposition=inline%3B+filename%3DA_COMPARATIVE_STUDY_OF_TRADITIONAL_FISHI.pdf&Expires=1730745274&Signature=DwEi33zZ2w6mQrKAooZSZE3UGKIxTmfavATyExv7V6eDZ6OmF89DOtHX54vPvZ0rySn~VgU20hhAGIjREf-yLmdCM2-JVqNUyf8JJPqZuD3ikrRG2gfuxbTrb9H1qkj8MW6vD0MU~I0mn10oJO6t3-Ne67KpTuxq8odgaloW5gJdFSlBxvAMo-

SpVAgJ2zMqJ~DQHCZZRxRRE9CivCGuX5wzsogUYhrSBbnubDW1jDL09jKix2r93x3Qdg-ut7AQ4NJebpCdVVe-GWY3oQiyoKpHmqEHAp0o2YCgG352XAkhXkTXCBcr0QICzL0uCPGS9SzUMiYQF7VOWsjT4RRBQg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Nandru, P., & Rentala, S. (2019). Demand-side analysis of measuring financial inclusion. *International Journal of Development Issues*, *19*(1), 1–24. https://doi.org/10.1108/ijdi-06-2018-0088

Physical Growth and Nutritional Status of the Shabar Tribal Adolescents of Orissa, India: a Cross-sectional Study. (2008, March 1). PubMed. https://pubmed.ncbi.nlm.nih.gov/22691768/

Potdar, K., S, T., & D, C. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, *175*(4), 7–9. https://doi.org/10.5120/ijca2017915495

Rashmi, R., & Paul, R. (2022). Early childhood circumstances and educational wellbeing inequality among tribal and non-tribal children in India: evidence from a panel study. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-13889-5

Raul, M., Majumdar, K., & Chatterjee, D. (2021). Indigenous Knowledge and Endogenous Development: Exploring survival strategy of a tribal community of Jharkhand, India. *International Journal of Research and Review*, *8*(8), 156–170. https://doi.org/10.52403/ijrr.20210822

Ray, D., Rout, H. S., & Ray, P. (2020). How do households utilize banking services and what are the determinants of it? An empirical analysis from the rural and tribal areas of an eastern Indian state. *Journal of Public Affairs*. https://doi.org/10.1002/pa.2520

Saxena, V., & Bhattacharya, P. C. (2017). Inequalities in LPG and electricity consumption in India: The role of caste, tribe, and religion. *Energy Sustainable Development/Energy for Sustainable Development*, *42*, 44–53. https://doi.org/10.1016/j.esd.2017.09.009

Shekhar, S. (2016). socio economic issues of the tribals in Jharkhand. *Nls*. https://www.academia.edu/24924148/socio_economic_issues_of_the_tribals_in_Jharkhand

Tsai, C., & Hu, Y. (2022). Empirical comparison of supervised learning techniques for missing value imputation. *Knowledge and Information Systems*, *64*(4), 1047–1075. https://doi.org/10.1007/s10115-022-01661-0

Xing. (2024). Global Happiness Ranking: An Ensemble Regression Vs. Traditional Approach Investigation. http://arno.uvt.nl/show.cgi?fid=172634

ThesisFinalDraft02Nov24.docx

ORIGINALITY REPORT

7%	5%	5%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Indian School of Mines Student Paper	1%
2	ebin.pub Internet Source	<1%
3	fastercapital.com Internet Source	<1%
4	www.projectpro.io Internet Source	<1%
5	Submitted to University of Carthage Student Paper	<1%
6	Submitted to Monash University Student Paper	<1%
7	Azizur Rahman, Faruq Abdulla, Md. Moyazzem Hossain. "Scientific Data Analysis with R - Biostatistical Applications", CRC Press, 2024 Publication	<1%
8	Submitted to University of Hertfordshire Student Paper	<1%

9	ideas.repec.org Internet Source	<1 %
10	Sohrab Mokhtari, Kang K. Yen. "False Data Injection Attack Detection, Isolation, and Identification in Industrial Control Systems Based on Machine Learning: Application in Load Frequency Control", Electronics, 2024 Publication	<1 %
11	medium.com Internet Source	<1 %
12	www-emerald-com-443.webvpn.sxu.edu.cn Internet Source	<1 %
13	Di Wu. "Data Mining with Python - Theory, Application, and Case Studies", CRC Press, 2024 Publication	<1 %
14	Dothang Truong. "Data Science and Machine Learning for Non-Programmers - Using SAS Enterprise Miner", CRC Press, 2024 Publication	<1 %
15	Submitted to University of Surrey Student Paper	<1 %
16	opus.bsz-bw.de Internet Source	<1 %
17	www.coursehero.com Internet Source	<1 %

18	Submitted to Whitecliffe College of Art & Design Student Paper	<1%
19	huggingface.co Internet Source	<1%
20	www.fastercapital.com Internet Source	<1%
21	"Advanced Network Technologies and Intelligent Computing", Springer Science and Business Media LLC, 2024 Publication	<1%
22	"Front Matter", 2023 8th International Conference on Computer Science and Engineering (UBMK), 2023 Publication	<1%
23	Submitted to University of Bolton Student Paper	<1%
24	Submitted to University of Sydney Student Paper	<1%
25	www.researchgate.net Internet Source	<1%
26	link.springer.com Internet Source	<1%
27	Submitted to RMIT University Student Paper	<1%

28	Submitted to Sydney Polytechnic Institute Student Paper	<1 %
29	www.yumpu.com Internet Source	<1 %
30	Submitted to British University in Egypt Student Paper	<1 %
31	Submitted to University of Exeter Student Paper	<1 %
32	publications.excas.org Internet Source	<1 %
33	scholarworks.bwise.kr Internet Source	<1 %
34	Muhammad Tayyeb Bukhari. "Efficacy of lightweight Vision Transformers in diagnosis of pneumonia", Cold Spring Harbor Laboratory, 2024 Publication	<1 %
35	brightideas.houstontx.gov Internet Source	<1 %
36	core-cms.prod.aop.cambridge.org Internet Source	<1 %
37	www.goldenratio.id Internet Source	<1 %
38	www.mdpi.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		

ThesisFinalDraft02Nov24.docx

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46

PAGE 47

PAGE 48

PAGE 49

PAGE 50

PAGE 51

PAGE 52

PAGE 53

PAGE 54

PAGE 55

PAGE 56

PAGE 57

PAGE 58

PAGE 59

PAGE 60

PAGE 61

PAGE 62

PAGE 63

PAGE 64

PAGE 65

PAGE 66

PAGE 67

PAGE 68

PAGE 69

PAGE 70

PAGE 71

PAGE 72

PAGE 73

PAGE 74

PAGE 75

PAGE 76

PAGE 77

PAGE 78

PAGE 79

PAGE 80

PAGE 81

PAGE 82

PAGE 83

PAGE 84

PAGE 85

PAGE 86

PAGE 87

PAGE 88

PAGE 89

PAGE 90

PAGE 91

PAGE 92