

A Comprehensive Analysis of NFHS-5 data for TB in India

by

ABHISHEK MUKESHBHAI THAKKAR
202111023

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY




May 2023

Declaration


I hereby declare that

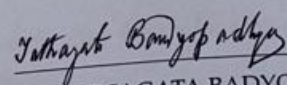
- i) The thesis comprises my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,
- ii) due acknowledgment has been made in the text to all the reference material used.


ABHISHEK M. THAKKAR

Certificate

This is to certify that the thesis work entitled **A Comprehensive Analysis of NFHS-5 data for TB in India** has been carried out by **Abhishek Mukeshbhai Thakkar** for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my/our supervision.


PROF. ARPIT RANA
Thesis Supervisor


PROF. TATHAGATA BADYOPADHYAY
Thesis Co-Supervisor

Acknowledgments

Undertaking this Master's has been a truly life-changing experience for me, and it would not have been possible without the support of my supervisor **Prof. Arpit Rana**. This work has taken a year to complete and has been a journey of ups and downs. I met many people who have given me important lessons, which helped me do new experiments in my thesis.

I would like to first say a very big thank you to my supervisor **Prof. Arpit Rana and Prof. Tathagata Bandyopadhyay**. From the beginning of our thesis timeline, he has completely supported me at every point, right from telling me where to begin till the end when I was writing my final report. **Prof. Arpit Rana** has also been very generous with my mistakes whenever I was not able to move forward on my problem. Not only is he an incredible thesis guide, but also a great motivator.

I would also like to thank **Prof. Tathagata Bandyopadhyay, Prof. Arpit Rana, Prof. Anuj Tawari, and Prof. Rachit Chaya** for their valuable feedback in various stage presentations that helped me elevate the quality of my work.

I would also like to thank other faculties of my institute who have taught me during my Masters here at DA-IICT.

I would also like to say a heartfelt thank you to my family for always believing in me and encouraging me to follow my dreams.

At last, I'd also like to thank all of my batch-mates and my seniors, who were always there in my moments of confusion. I would also like to extend my appreciation to all those people who have played a very important role in my life here at DA-IICT and also in some way or the other on my thesis.

Contents

Abstract	vi
List of Tables	vii
List of Figures	viii
1 Introduction	2
1.1 DHS and NFHS Program	2
1.2 NFHS-5 Data	3
1.3 Tuberculosis	4
2 Literature Review	5
2.1 Epidemiology and Control	5
2.2 Knowledge, Awareness, and Perception	8
2.3 Risk Factors and Comrbidities	10
2.4 TB and Gender	11
3 Methodology	12
3.1 Data Preprocessing	12
3.1.1 Data Cleaning	12
3.1.2 Data Classification	13
3.2 A Comprehensive Data Analysis	16
3.2.1 Sex Ratio	17
3.2.2 Bad Habits	19
3.2.3 Nutrition Level	25
3.2.4 Educational Level	32
3.2.5 Wealth Index and Agewise	35
3.2.6 State-wise	39
3.3 Implementation	41
3.3.1 Class Balancing Techniques	41
3.3.2 Prediction Model	44

4 Results	47
5 Conclusion	57
5.1 Policy Implications of This Research	58
References	60

Abstract

This study presents a comprehensive analysis of tuberculosis (TB) in India using data from the NFHS-5 (National Family Health Survey) program. The research begins by providing a thorough understanding of the DHS (Demographic and Health Survey) and NFHS programs, followed by an extensive literature review of TB-related studies and NFHS-related papers.

The findings from the literature review indicate that directing tuberculosis control initiatives toward the poorest 20% of the population may yield more successful outcomes compared to targeting the general population or the wealthiest 20%. Additionally, an examination of trends in TB incidence and mortality in India from 1990 to 2019, based on data from the Global Burden of Disease Study 2019, reveals significant insights into the country's TB burden.

One notable observation from the literature review is that a substantial proportion of TB patients over 60% have at least one comorbidity, with diabetes emerging as a prominent comorbidity. Furthermore, the study highlights a concerning lack of awareness regarding TB among Indian adults, with only 49.7% of participants reporting prior knowledge of the disease.

The research extensively utilizes complex and large-scale NFHS-5 data. A considerable amount of work is devoted to analyzing household-level data, which is categorized into three groups based on the Human Development Index, with each category representing five states. Python's CSV file processing capabilities are employed to handle and process a vast amount of data.

To identify the factors that most significantly affect TB, the study compares TB variables with 402 other variables. However, due to the limited number of TB cases within each category, the researchers calculate the number of TB and non-TB patients per 100,000 people for all variables. This approach provides a better understanding of the relationships between variables and TB incidence.

The study goes beyond analysis and prediction, incorporating the development of a model to predict an individual's likelihood of contracting TB. Notably, the data exhibit significant bias, as 99.7% of the cases are non-TB patients. To address this imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to generate synthetic data for the minority class. The researchers then focus on the most influential features associated with TB, resulting in a prediction model that achieves an accuracy of over 70% in accurately identifying individuals at risk of TB.

In summary, this comprehensive analysis of NFHS-5 data sheds light on the tuberculosis landscape in India. The findings emphasize the importance of targeting the most marginalized populations, highlighting the prevalence of comorbidities such as diabetes among TB patients, underscoring the need for increased public awareness and showcasing the potential of data-driven prediction models in improving TB control and prevention efforts.

List of Tables

3.1	Comparison of States in High HDI	40
3.2	Comparison of States in Medium HDI	40
3.3	Comparison of States in Low HDI	40
4.1	Comparison of Methods in High HDI	50
4.2	Comparison of Methods in Medium HDI	53
4.3	Comparison of Methods in Low HDI	56

List of Figures

2.1	Taxonomy	5
3.1	Percentage of TB patients per 100000 by Sex in Low HDI	17
3.2	Percentage of TB patients per 100000 by Sex in Medium HDI	18
3.3	Percentage of TB patients per 100000 by Sex in High HDI	18
3.4	Percentage of TB patients per 100000 by Smoking/Tobacco in Low HDI	20
3.5	Percentage of TB patients per 100000 by Smoking/Tobacco in Medium HDI	21
3.6	Percentage of TB patients per 100000 by Smoking/Tobacco in High HDI	21
3.7	Percentage of TB patients per 100000 by Alcohol Consumption in Low HDI	22
3.8	Percentage of TB patients per 100000 by Alcohol Consumption in Medium HDI	22
3.9	Percentage of TB patients per 100000 by Alcohol Consumption in High HDI	23
3.10	Percentage of TB patients per 100000 by Frequency of Smoking inside the house in Low HDI	23
3.11	Percentage of TB patients per 100000 by Frequency of Smoking inside the house in Medium HDI	24
3.12	Percentage of TB patients per 100000 by Frequency of Smoking inside the house in High HDI	24
3.13	Percentage of TB patients per 100000 by BMI Level of Female in Low HDI	26
3.14	Percentage of TB patients per 100000 by BMI Level of Female in Medium HDI	27
3.15	Percentage of TB patients per 100000 by BMI Level of Female in High HDI	27

3.16	Percentage of TB patients per 100000 by Anemia Level of Male in Low HDI	29
3.17	Percentage of TB patients per 100000 by Anemia Level of Male in Medium HDI	30
3.18	Percentage of TB patients per 100000 by Anemia Level of Male in High HDI	31
3.19	Percentage of TB patients per 100000 by Educational Level in Low HDI	33
3.20	Percentage of TB patients per 100000 by Educational Level in Medium HDI	33
3.21	Percentage of TB patients per 100000 by Educational Level in High HDI	34
3.22	Percentage of TB patients per 100000 by Wealth Index in Low HDI	36
3.23	Percentage of TB patients per 100000 by Wealth Index in Medium HDI	36
3.24	Percentage of TB patients per 100000 by Wealth Index in High HDI	37
3.25	Percentage of TB patients per 100000 by Age in Low HDI	37
3.26	Percentage of TB patients per 100000 by Age in Medium HDI . . .	38
3.27	Percentage of TB patients per 100000 by Age in High HDI	38

CHAPTER 1

Introduction

1.1 DHS and NFHS Program

Certainly! The Demographic and Health Surveys (DHS) [7] and the National Family Health Surveys (NFHS) are two well-known programs conducted in several countries, primarily in low- and middle-income regions. These surveys play a crucial role in collecting comprehensive and reliable data on various aspects of health, population, and development.

The DHS program, funded by the United States Agency for International Development (USAID), aims to provide data for monitoring and evaluating population and health programs. The surveys are conducted periodically, typically every five years, and cover a wide range of topics such as fertility, family planning, maternal and child health, nutrition, HIV/AIDS, and more. The DHS program employs standardized questionnaires and rigorous sampling techniques to ensure the data collected is representative and comparable across countries and over time.

Similarly, the NFHS program, implemented by the Ministry of Health and Family Welfare in India, focuses on gathering information related to reproductive and child health, family planning, nutrition, and other vital health indicators. These surveys are carried out at regular intervals, offering valuable insights into the health and well-being of the population, especially women and children.

Both the DHS and NFHS programs have contributed significantly to evidence-based decision-making, policy formulation, and program evaluation in the respective countries where they are conducted. The collected data helps identify health trends, assess the impact of interventions, and inform targeted strategies for improving healthcare services and outcomes.

1.2 NFHS-5 Data

The National Family Health Survey (NFHS)[12] is a large-scale household survey conducted in India to collect comprehensive and reliable data on various aspects of health, population, and nutrition. NFHS-5 refers to the fifth round of the survey, which was conducted between 2019 and 2020.

NFHS-5 aims to provide up-to-date information on key indicators related to reproductive and child health, maternal health, nutrition, family planning, HIV / AIDS, and other important health and social indicators. The survey covers a representative sample of households across all states and union territories of India, making it one of the most extensive data collection efforts in the country.

The NFHS-5 data set consists of a wide range of variables collected through interviews with household members, including women, men, and children. These variables capture information on demographic characteristics, household characteristics, reproductive health, maternal and child health, utilization of healthcare services, nutrition, HIV/AIDS awareness, and other relevant aspects of public health.

The data collected through NFHS-5 provides valuable insights into the health and well-being of individuals and communities in India. It serves as a critical resource for policymakers, researchers, and program implementers to assess health trends, monitor progress toward national and global health goals, and design evidence-based interventions and policies.

1.3 Tuberculosis

Tuberculosis (TB)[14] continues to be a major public health concern in India, with the country having the highest number of TB cases in the world. The National Family Health Survey (NFHS) is a nationally representative survey that provides valuable information on various health indicators in India, including TB. The most recent round of NFHS (NFHS-5) was conducted in 2019-20 and provides data on a range of socioeconomic and demographic factors that may be associated with TB incidence.

In this study, we conducted a comprehensive analysis of the NFHS-5 data for TB in India. Specifically, we analyzed the household member data, which contains information on more than 28 lac individuals across the country. We divided the data into three categories based on the human development index (HDI) - high, medium, and low - as per the United Nations Development Programme (UNDP) classification. We then performed data cleaning on each dataset and compared TB and non-TB patients with various factors, including nutrition (hemoglobin levels, anemia levels, glucose levels, BMI index), bad habits (smoking, drinking alcohol, tobacco consumption), wealth index, age group, educational level, and facilities at home (source of drinking water, type of toilet facilities, type of residence).

The aim of this study was to identify the factors that may be contributing to the incidence of TB in India and to provide insights that may inform future public health interventions to reduce the burden of TB in the country. By analyzing a wide range of factors and using advanced data visualization techniques, we provide a detailed and nuanced picture of the socio-economic and demographic factors associated with TB in India.

CHAPTER 2

Literature Review

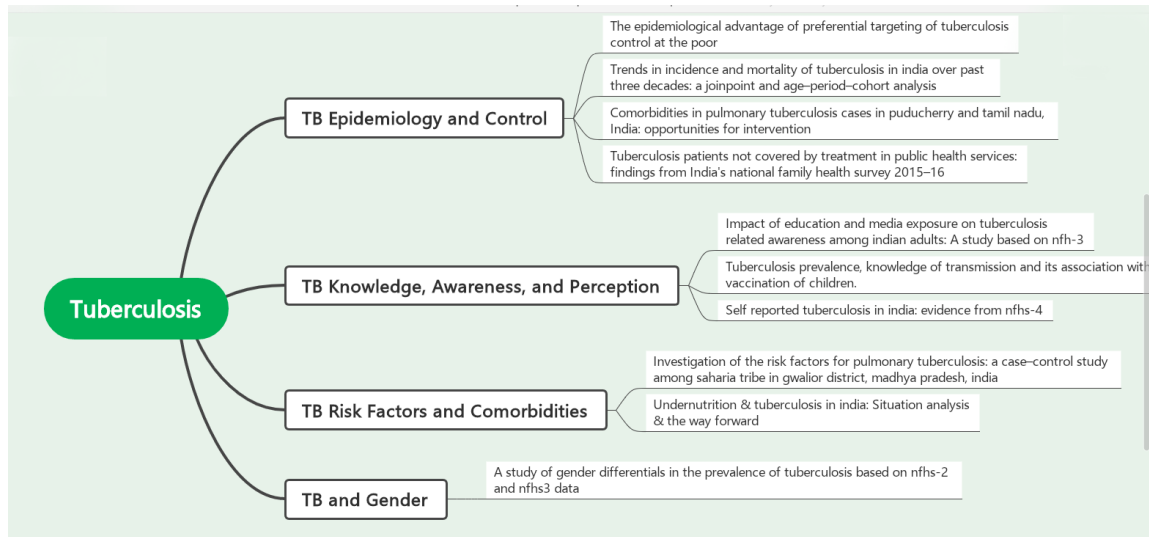


Figure 2.1: Taxonomy

2.1 Epidemiology and Control

In Epidemiology and Control,[9] look at the possible advantages of directing tuberculosis prevention efforts at the most underprivileged populations in low- and middle-income nations. The authors contend that focusing on the most vulnerable groups could significantly lessen the overall impact of tuberculosis in these nations. This is because the poorest populations have a higher probability of contracting tuberculosis, which is a major risk factor for the disease.

In order to compare the effects of various control measures, the research develops a computational formula that simulates the spread of tuberculosis in a community. The findings imply that directing tuberculosis control initiatives at the poorest 20% of the population compared to the population at large or the wealth-

iest 20% may be more successful.

In this [5], we analyze the trends in the incidence and mortality of tuberculosis (TB) in India from 1990 to 2019 using data from the Global Burden of Disease Study 2019. The study found that the age-standardized incidence and mortality rates of TB in India declined from 390.22 to 223.01 and from 121.72 to 36.11 per 100,000 population, respectively. The decline was observed for both males and females, but the decline was sharper in females.

The study also found that the incidence and mortality rates of TB decreased for both males and females across all ages during this period. The age effect showed that both incidence and mortality significantly increased with advancing age and decreased with advancing time period.

The study suggests that although there has been a significant decline in TB incidence and mortality rates, the annual rate of reduction is not enough to achieve the aim of India's National Strategic Plan 2017-2025. TB still remains a major public health problem in India, and the government needs to strengthen the four strategic pillars "Detect-Treat-PreventBuild" (DTPB) to achieve a TB-free India as envisioned in the National Tuberculosis Elimination Programme (2020). It is also important to address the unfavorable net age effect and focus on preventive measures for the aging population to control the continued increase in TB mortality.

In [8] Opportunities for Intervention," examines the prevalence and impact of comorbidities among patients with pulmonary tuberculosis (PTB) in Puducherry and Tamil Nadu, India. The study analyzed data from over 2,000 TB patients and found that more than 60% had at least one comorbidity, with diabetes and chronic obstructive pulmonary disease (COPD) being the most common.

Alcohol use in men and malnutrition are helping drive the TB epidemic in Southern India. They also emphasize the importance of addressing social determinants of health, such as poverty and malnutrition to improve the overall health of PTB patients.

In this paper, [11] The objective of this study was to describe the characteristics of tuberculosis (TB) patients in India who do not receive treatment from public health services. About 50% of TB patients in India seek care from private providers, which results in incomplete notification, variable quality of care, and

out-of-pocket expenditure.

The study used cross-sectional data from the National Family Health Survey-4 (2015-16) and logistic regression analysis to identify factors associated with seeking treatment from private providers and not seeking treatment at all. The results showed that the prevalence of self-reported TB was 308.17/100,000 population, and 38.8% of TB patients were outside the care of public health services. Of these, 3.3% did not seek treatment, while 35.3% accessed treatment from the private sector.

Factors associated with not seeking treatment were age less than 10 years, no/preschool education, the poorest wealth index, and the household's general rejection of the public sector when seeking health care. Factors associated with seeking treatment from private providers were female sex, younger age of the patient, higher education, and the household's general rejection of the public sector when seeking health care.

2.2 Knowledge, Awareness, and Perception

Now we will discuss the second category, which is Knowledge, Awareness, and perception. This paper [2] research study that aimed to investigate the impact of education and media exposure on tuberculosis (TB) awareness among Indian adults using data from the National Family Health Survey (NFHS) conducted in 2005-06.

The study found that overall awareness of TB among Indian adults was low, with only 49.7% of participants having heard of TB. However, participants with higher levels of education were more likely to have heard of TB, as were those who had been exposed to TB-related information through media sources such as television or radio.

Additionally, the study found that those who were aware of TB were more likely to have knowledge about the symptoms and transmission of the disease, as well as the importance of seeking medical care if they experienced symptoms. Overall, the study highlights the importance of education and media exposure in increasing awareness of TB in India and suggests that targeted health education campaigns could play an important role in improving TB-related knowledge and behaviors among the general population.

The paper [4] discusses a study conducted in India in 2015-2016 as part of the National Family Health Surveys (NFHS) to understand the perceptions of patients regarding tuberculosis (TB) and how it can help in designing a comprehensive, client-oriented program for the disease. The study found that the prevalence of TB remains significantly high, and a high percentage of people were unaware of the exact cause of disease proliferation. The majority of people believed that touching or sharing utensils can be a source of TB, which affected their responses about seeking diagnosis and treatment. However, most people knew that TB is a curable disease that can be prevented to some extent if immunization with the Bacillus Calmette-Guérin (BCG) vaccine is done at the correct stage.

Therefore, a large section of the population had their children vaccinated, and they would go for a diagnosis if they had symptoms suggestive of the disease. The study concludes that there is a need to investigate how this information could potentially be used to enhance the early seeking of appropriate services among TB

patients and health facilities can make a significant contribution to the treatment of tuberculosis.

We review the paper [9] This study aimed to estimate the prevalence of self-reported tuberculosis (TB) in India using data from the fourth round of the National Family Health Survey (NFHS-4). The authors analyzed data on TB self-reporting among adults aged 15-49 years, as well as demographic and socioeconomic factors associated with self-reporting.

The results showed that the overall prevalence of self-reported TB in India was 0.47%, with significant variation across different states and demographic subgroups. Factors associated with higher self-reporting rates included being male, living in urban areas, having lower levels of education, and belonging to disadvantaged socioeconomic groups.

The authors suggest that the low self-reporting rate of TB in India may be due to poor awareness and stigma associated with the disease, as well as limited access to diagnostic and treatment services. They call for greater investment in TB awareness and control programs, particularly in disadvantaged communities, to improve TB detection and control in India.

2.3 Risk Factors and Comorbidities

Now we will discuss the third category, which is risk factors and comorbidities. This case-control study [3] aimed to investigate the risk factors for TB among the Saharia tribe. The study included 140 cases (patients with confirmed TB) and 140 controls (healthy individuals from the same community).

The results showed that several factors were associated with an increased risk of TB, including low income, poor housing conditions, lack of education, and malnutrition. In addition, smoking, alcohol consumption, and indoor air pollution from cooking with biomass fuel were identified as significant risk factors for TB. The study also found that individuals who had a family member with TB were more likely to develop the disease themselves, indicating a potential genetic susceptibility to TB within the Saharia tribe.

Overall, the study highlights the need for targeted interventions to address the social determinants of health and reduce the burden of TB among the Saharia tribe and other vulnerable populations in India.

[10] This paper discusses the co-occurrence of undernutrition and tuberculosis (TB) in India, highlighting the impact of undernutrition on TB incidence, disease progression, and treatment outcomes. The authors provide a comprehensive overview of the current situation in India, including prevalence rates and risk factors for undernutrition and TB, as well as the challenges faced in addressing these issues in the country's healthcare system.

They also discuss the potential solutions and interventions that can be implemented to reduce the burden of undernutrition and TB in India, such as nutritional supplementation, improved case detection and management, and integrated care models. The paper concludes with a call to action for policymakers, healthcare providers, and the public to work together to address this important public health challenge.

2.4 TB and Gender

Now we will discuss the last category, which is TB and Gender. In [13], The study aimed to investigate gender differences in the prevalence of tuberculosis (TB) in India, using data from the National Family Health Survey NFHS-2 and NFHS-3. The study found that the prevalence of TB was higher among men than women, and this difference was significant in both surveys.

The study also found that the proportion of women reporting TB increased significantly from NFHS-2 to NFHS-3, whereas the proportion of men remained relatively stable. The study suggested that there is a need for gender-specific interventions to address TB in India, as well as further research to understand the underlying factors contributing to the gender differential in TB prevalence.

It is noted that the increase in the gender gap is more in rural areas than in urban areas. 88% of the entire growth is linked to rural areas, and 12% is linked to urban areas.

Hindus, other caste groups, and high SLI categories contribute the most in urban areas.

CHAPTER 3

Methodology

3.1 Data Preprocessing

3.1.1 Data Cleaning

In our analysis of the NFHS-5 dataset, we focused on the household-level data, which consisted of 6,482 variables encompassing questions asked to household members. This data set was particularly extensive, comprising approximately 637,000 household records. Due to the size and complexity of the dataset, we undertook several steps to ensure its usability and relevance to our research on tuberculosis (TB).

To facilitate our analysis, we initially converted the data from its original format into CSV files, which allowed us to work with the data efficiently using Python programming language. By employing Python's data manipulation and analysis capabilities, we were able to conduct various preprocessing steps aimed at refining the dataset.

One of the initial preprocessing steps involved the removal of rows containing missing values, as our focus was specifically on households and variables related to TB. Removing such rows ensured that our analysis would be based on complete and relevant data, thereby enhancing the accuracy and reliability of our findings.

By eliminating data points unrelated to TB, we further streamlined the dataset, enabling us to concentrate solely on the variables and information pertinent to our research objectives. This meticulous approach helped us to avoid any confounding factors and maintain a clear focus on TB-related indicators within the household-level data.

Through these initial data preprocessing steps, we have established a robust foundation for our subsequent analyses, enabling us to delve deeper into the NFHS-5 dataset and extract meaningful insights regarding the prevalence and associated factors of TB within households.

3.1.2 Data Classification

Following the data cleaning process, we encountered the challenge of working with an extensive dataset exceeding 7 GB in size. To overcome this obstacle, we adopted a data classification approach to facilitate our analysis and streamline our focus. By dividing the dataset into distinct categories, we were able to efficiently work with manageable subsets of data.

We categorized the dataset based on the Human Development Index (HDI) provided by the United Nations Development Programme (UNDP). This index serves as a measure of overall development and encompasses various socio-economic indicators. In our classification, we created three categories: High HDI, Medium HDI, and Low HDI, each containing five states, resulting in a total of 15 states out of the 37 states covered in the NFHS-5 dataset.

The High HDI category comprised the following states: Delhi, Goa, Puducherry, Kerala, and Chandigarh. These states were selected based on their high HDI scores, reflecting advanced levels of human development.

The Medium HDI category included the states of Rajasthan, Arunachal Pradesh, Karnataka, Uttarakhand, and Telangana. These states demonstrated HDI scores falling within an average range.

Lastly, the Low HDI category encompassed Bihar, Jharkhand, Madhya Pradesh, Odisha, and Uttar Pradesh. These states exhibited comparatively lower HDI scores, signifying a need for targeted interventions to enhance human development indicators.

By stratifying the data into these categories, we were able to focus our analysis on specific sets of states, enabling a more manageable and targeted examination of the data. This approach allowed us to gain deeper insights into the variations and patterns within each category, providing a more nuanced understanding of the relationship between HDI and the variables related to our research objectives.

Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are two popular techniques for dimensionality reduction in machine learning and data analysis. Both methods aim to reduce the dimensionality of a dataset while preserving as much relevant information as possible. Here's a brief description of each method:

1. Principal Component Analysis (PCA):

- PCA is a statistical technique used to transform a high-dimensional dataset into a lower-dimensional space.
- It identifies the directions (principal components) in the data that capture the most significant variations.
- The first principal component explains the largest variance in the data, and subsequent components explain the remaining variances in decreasing order.
- PCA achieves dimensionality reduction by projecting the original data onto a new coordinate system defined by the principal components.
- The transformed dataset contains fewer dimensions, with each dimension being a linear combination of the original features.
- The reduced dimensions retain the most important patterns and structure of the original data.
- PCA is an unsupervised method and can be applied to both numerical and continuous variables.

2. Singular Value Decomposition (SVD):

- SVD is a matrix factorization technique commonly used in linear algebra and data analysis.
- It decomposes a matrix into three matrices: U , Σ , and V , where U and V are orthogonal matrices and Σ is a diagonal matrix containing the singular values.

- SVD can be applied to any matrix, including rectangular and square matrices.
- In the context of dimensionality reduction, SVD is used to reduce the dimensionality of a dataset by retaining the most significant singular values and their corresponding columns in the U and V matrices.
- The resulting lower-dimensional representation captures the essential structure and relationships in the data.
- SVD can be employed for both numerical and categorical data, but it often requires numerical transformations or encoding for non-numeric variables.

Both PCA and SVD have various applications in data preprocessing, feature extraction, and data compression. They are useful for visualizing high-dimensional data, removing noise, finding latent factors, and improving computational efficiency in subsequent analysis tasks. The choice between PCA and SVD depends on the specific context and requirements of the problem at hand.

3.2 A Comprehensive Data Analysis

The Human Development Index (HDI) [1] is a composite measure that provides an overview of a country's overall level of human development. It takes into account multiple factors across three broad dimensions: health, education, and standard of living. The specific factors that contribute to the calculation of HDI include:

1. **Life expectancy at birth:** This factor reflects the average number of years a newborn is expected to live, indicating the health status and access to health-care within a country.
2. **Education:** HDI considers two indicators related to education:
 - a. **Expected years of schooling:** It measures the number of years of education an average child is expected to receive.
 - b. **Mean years of schooling:** This indicator represents the average number of years of education completed by the adult population, providing insights into the overall educational attainment within a country.
3. **Gross national income (GNI) per capita:** This factor assesses the income level and economic well-being of individuals in a country, taking into account the total income generated by residents and non-residents.

Performing TB data analysis across different categories, such as nutrition-based, bad habits, education level, and sex, can provide valuable insights into the factors associated with TB prevalence and its variations. By examining these categories, we can gain a deeper understanding of the relationships between TB and various demographic and behavioral factors. Here's an overview of the potential analyses we have conducted.

In order to conduct a comprehensive analysis, it is essential to consider the prevalence of tuberculosis (TB) cases in relation to the population size. Given that the overall prevalence of TB cases was a mere 0.3% across all three categories, the absolute number of TB patients would be relatively low. Therefore, to ensure meaningful comparisons, we have opted to express the number of TB patients per 100,000 individuals for both the non-TB and TB patient groups within each feature. This approach allows for a more standardized and representative assessment of the impact of TB within the studied population. By employing this

methodology, we aim to provide a robust basis for our analysis and facilitate accurate comparisons between different features.

3.2.1 Sex Ratio

In our analysis across all three categories of High, Medium, and Low HDI, a distinct pattern emerges regarding the prevalence of tuberculosis (TB) among males and females. It is evident that the proportion of males affected by TB exceeds 60%, whereas females comprise less than 40% of TB cases. Conversely, when considering non-TB patients, both genders exhibit a relatively equal distribution, with a near 50-50% ratio. This stark contrast in gender distribution leads us to a compelling conclusion: males have a higher susceptibility to contracting TB compared to females.

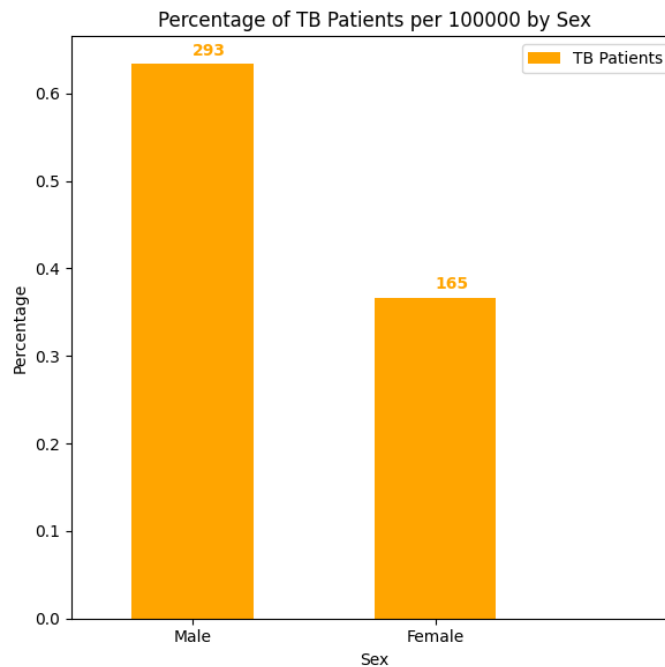


Figure 3.1: Percentage of TB patients per 100000 by Sex in Low HDI

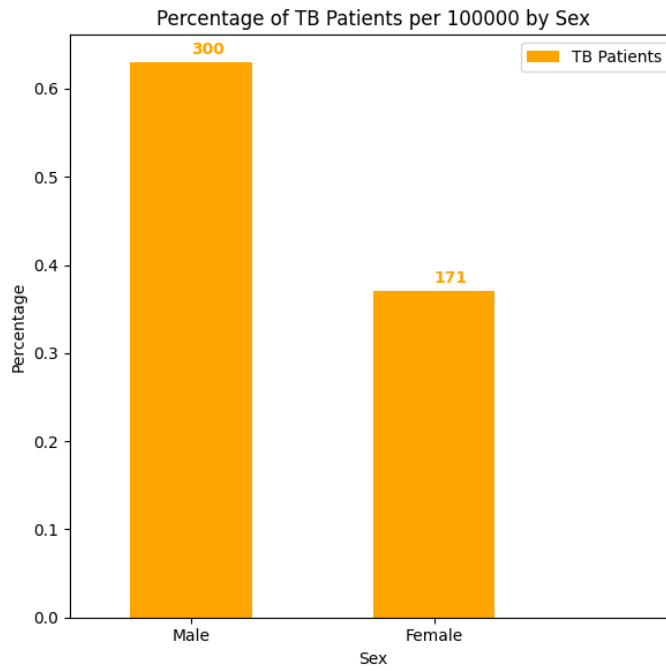


Figure 3.2: Percentage of TB patients per 100000 by Sex in Medium HDI

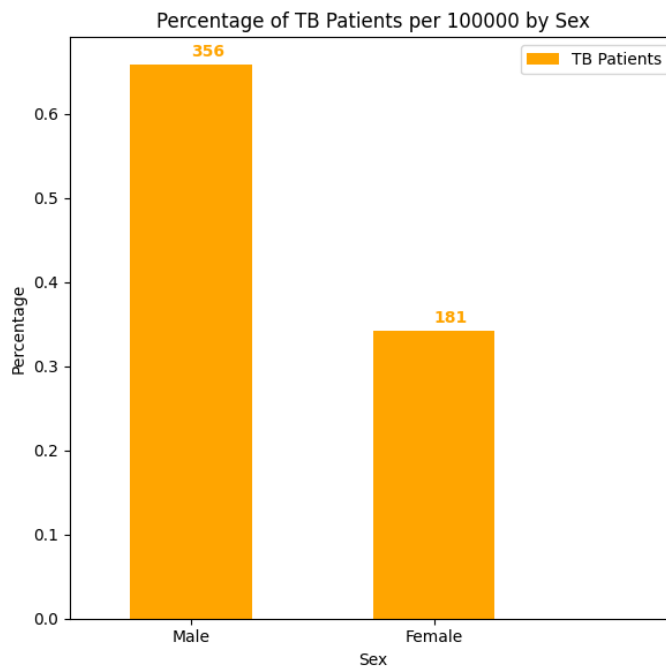


Figure 3.3: Percentage of TB patients per 100000 by Sex in High HDI

3.2.2 Bad Habits

In our analysis of the impact of bad habits, specifically smoking and alcohol consumption, on tuberculosis (TB) prevalence, a significant finding emerges. We observe that the number of TB patients who engage in these habits is nearly three times higher compared to non-TB patients. This compelling correlation allows us to draw a clear conclusion: smoking and alcohol consumption have a profound impact on the occurrence and severity of TB.

The association between smoking, alcohol consumption, and TB is well-documented in numerous studies and supported by existing scientific evidence. These habits can weaken the immune system, making individuals more susceptible to TB infection and increasing the likelihood of developing active TB disease. Additionally, smoking and alcohol consumption can impede the efficacy of TB treatment, leading to poor treatment outcomes and prolonged infectiousness.

In the analysis of the frequency of smoking inside the house and its correlation with tuberculosis (TB) cases, a notable pattern emerges. When comparing the frequency of smoking with the occurrence of TB, it becomes evident that individuals who smoke daily have the second-highest percentage of TB cases, accounting for approximately 33% of the overall cases. The highest percentage of TB cases, around 40%, is observed among individuals who never smoke. As we examine the other categories of smoking frequency, including weekly, monthly, and rarely, the percentage of TB cases decreases to less than 10% in all three categories. This trend holds true for the low human development index (HDI) category.

Similarly, in the medium HDI category, the highest number of TB patients is found among individuals who smoke daily, accounting for 37% of the cases. Surprisingly, this surpasses the percentage of TB cases among individuals who never smoke, which stands at 35%. As we move through the categories of weekly, monthly, and rarely, the percentage of TB cases gradually decreases.

These findings underscore the significant impact of daily smoking on the prevalence of TB cases, particularly in the low and medium HDI categories.

It is evident that smoking habits play a crucial role in the occurrence of TB, with higher frequencies of smoking correlating with a higher risk of contracting the disease. These observations highlight the importance of addressing smoking cessation and promoting awareness of the adverse health effects associated with tobacco use in order to combat the incidence of TB.

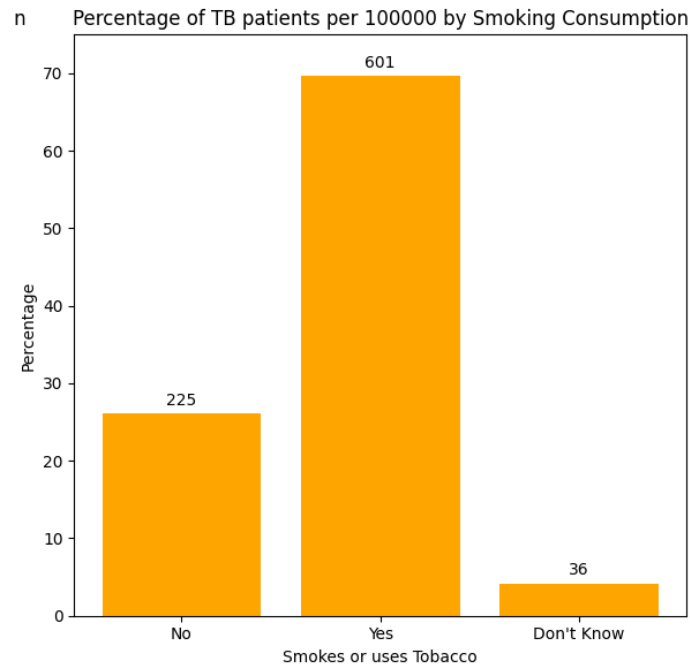


Figure 3.4: Percentage of TB patients per 100000 by Smoking/Tobacco in Low HDI

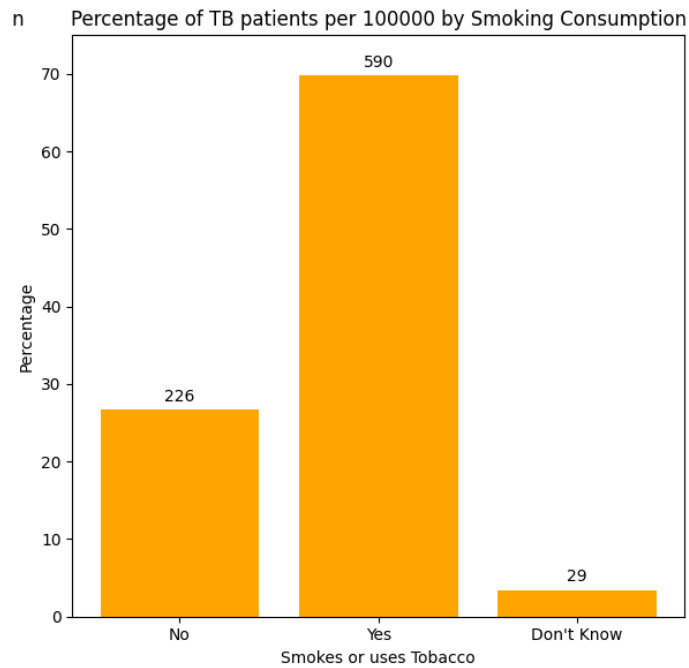


Figure 3.5: Percentage of TB patients per 100000 by Smoking/Tobacco in Medium HDI

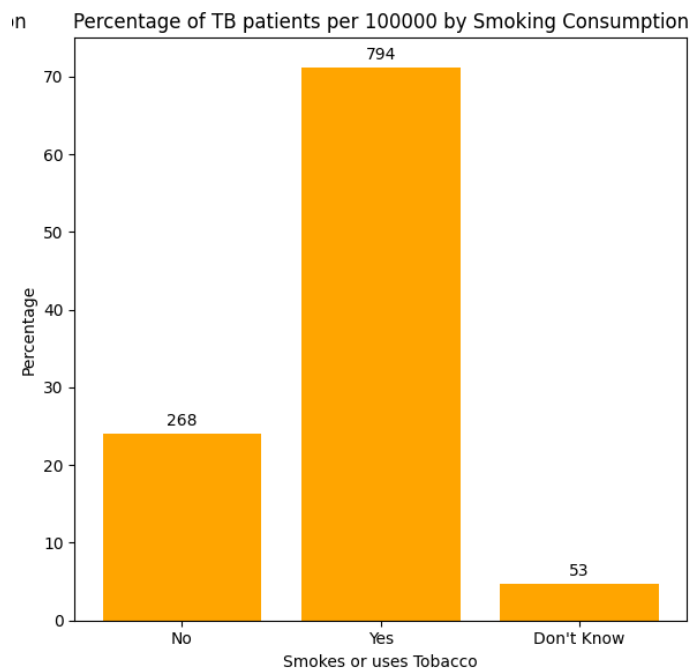


Figure 3.6: Percentage of TB patients per 100000 by Smoking/Tobacco in High HDI

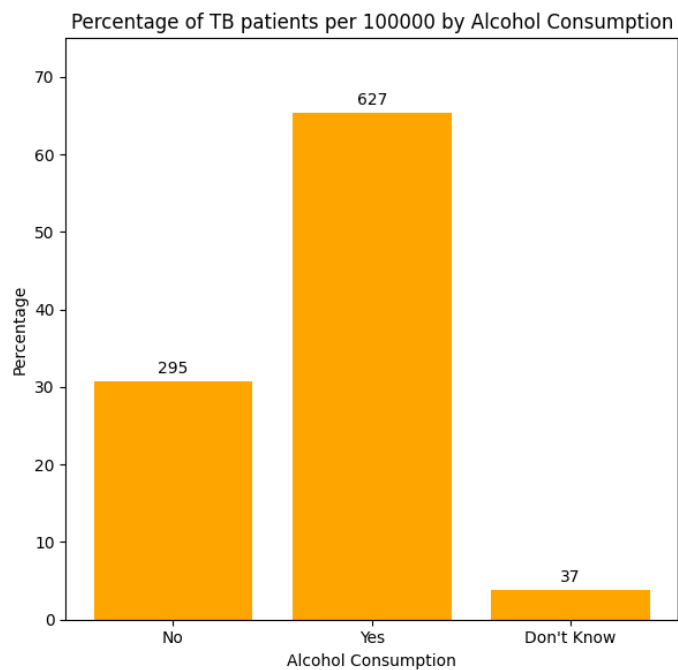


Figure 3.7: Percentage of TB patients per 100000 by Alcohol Consumption in Low HDI

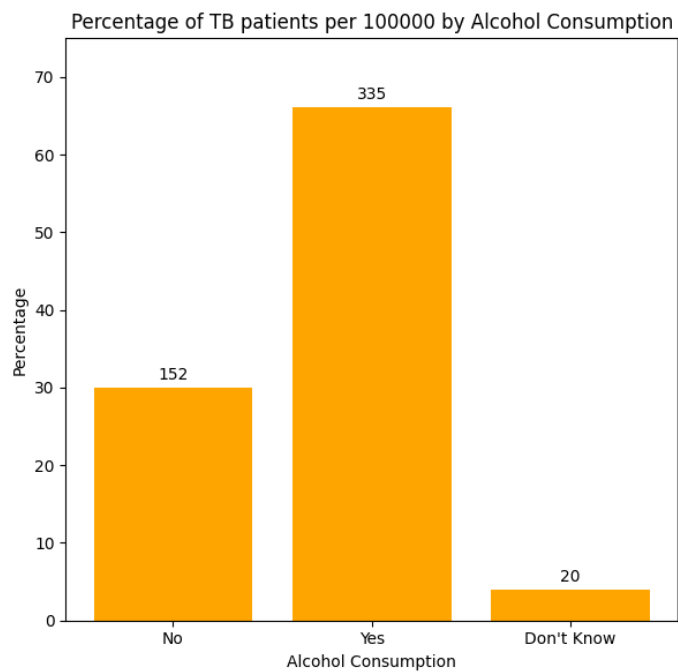


Figure 3.8: Percentage of TB patients per 100000 by Alcohol Consumption in Medium HDI

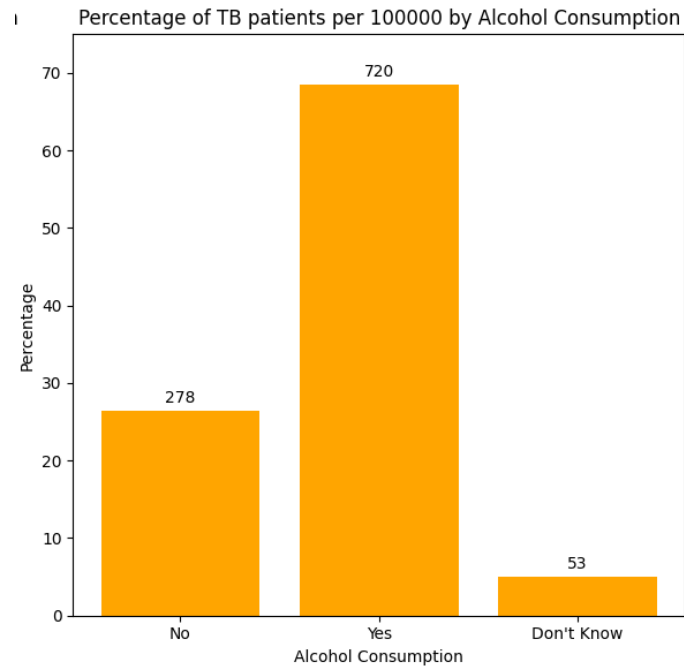


Figure 3.9: Percentage of TB patients per 100000 by Alcohol Consumption in High HDI

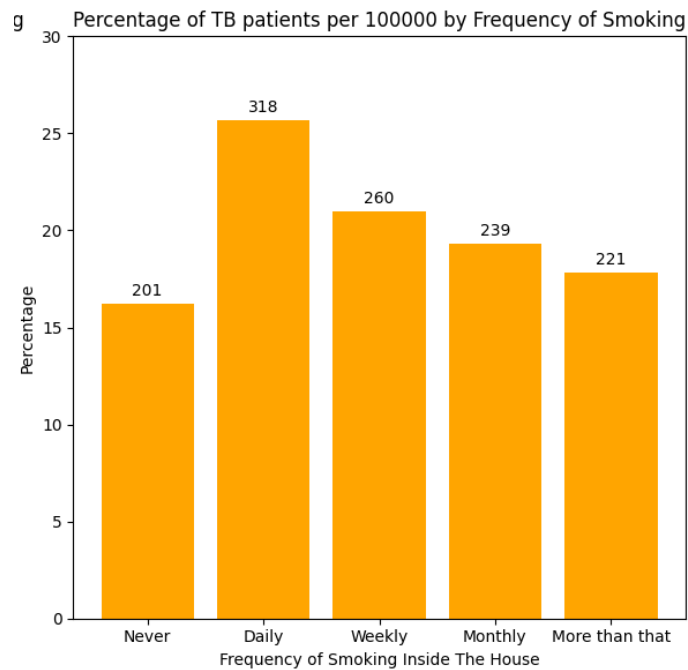


Figure 3.10: Percentage of TB patients per 100000 by Frequency of Smoking inside the house in Low HDI

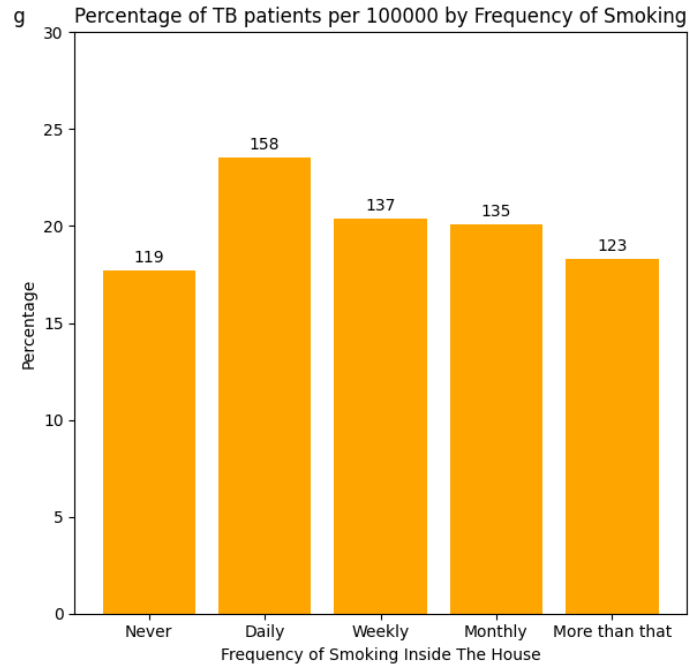


Figure 3.11: Percentage of TB patients per 100000 by Frequency of Smoking inside the house in Medium HDI

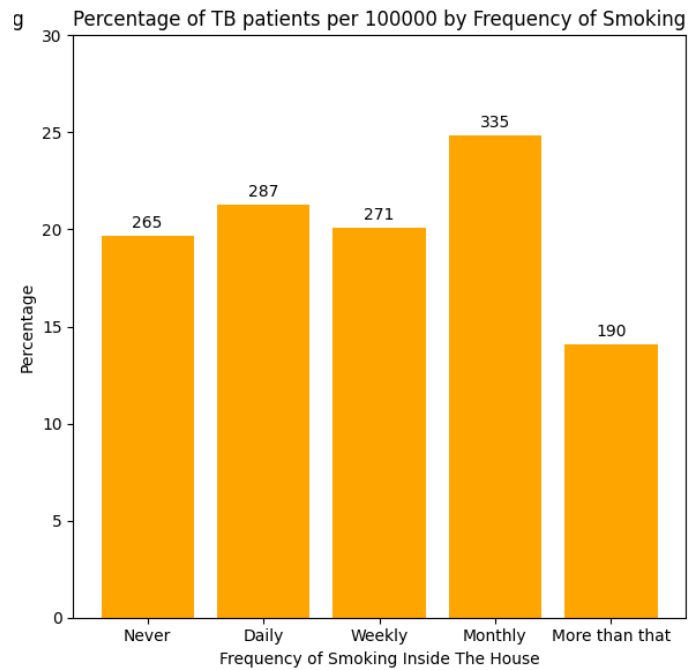


Figure 3.12: Percentage of TB patients per 100000 by Frequency of Smoking inside the house in High HDI

3.2.3 Nutrition Level

In the analysis of the impact of nutrition on tuberculosis (TB) prevalence, two key indicators were considered: Body Mass Index (BMI) and anemia level among males. The findings revealed interesting patterns across different HDI categories.

BMI level: In the High HDI category, there was a higher incidence of TB cases between the BMI levels of 1900-2000. Conversely, in the Medium HDI category, a majority of TB cases were observed at BMI levels below 1800. It is important to note that these variations occur because the data is calculated per 100,000 people, meaning that even a small number of TB cases can significantly affect the data. Furthermore, as BMI levels increased, the number of TB cases generally decreased.

Upon comparing the body mass index (BMI) levels of female TB patients in the high and medium human development index (HDI) categories, a notable trend emerges. In both categories, the highest number of TB patients is observed in two specific BMI ranges, namely, <1800 and >2500. These two ranges account for the majority of TB cases among female patients, while the remaining BMI levels have relatively low occurrences, each representing less than 10% of the total cases.

This observation highlights the significance of BMI levels in relation to TB prevalence among female individuals in both the high and medium HDI categories. The higher number of cases in the <1800 range suggests a possible association between undernutrition or low BMI and an increased vulnerability to TB infection. On the other hand, the elevated number of cases in the >2500 range may indicate a potential link between overweight or obesity and TB incidence among females.

These findings emphasize the importance of addressing nutritional factors and maintaining a healthy BMI in efforts to prevent and manage TB cases among females. Public health interventions should focus on promoting balanced nutrition and raising awareness about the potential risks associated with both undernutrition and overweight/obesity in relation to TB infection.

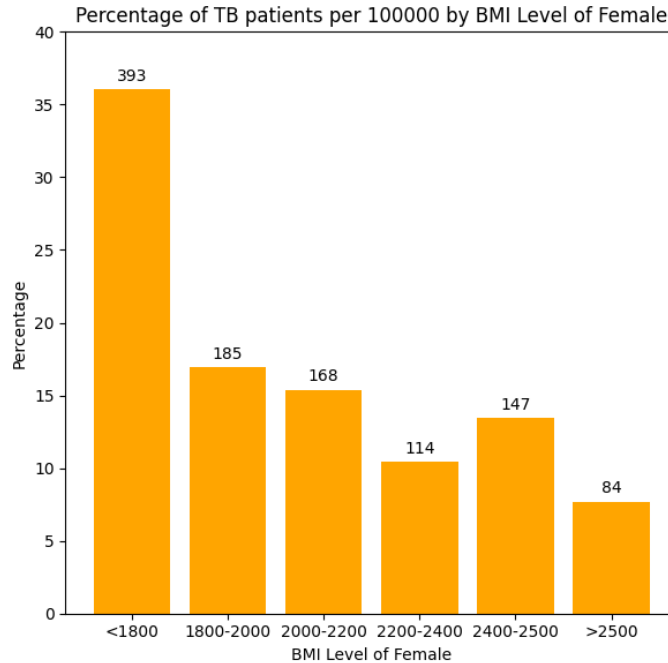


Figure 3.13: Percentage of TB patients per 100000 by BMI Level of Female in Low HDI

Anemia level: When examining the anemia levels of male individuals in the medium and low human development index (HDI) categories, striking similarities can be observed. In both categories, the highest number of TB cases is found among individuals with anemia levels falling within the range of 0-1. This specific range consistently demonstrates the highest prevalence of TB cases among male patients.

Conversely, an interesting pattern emerges when considering the anemia levels beyond this range. Male individuals with anemia levels greater than 4 show no recorded cases of TB, indicating a potential protective effect against TB infection in this group. Additionally, anemia levels ranging from 1 to 4 exhibit a relatively lower percentage, with less than 20% of TB cases observed within this range.

These findings highlight the correlation between anemia levels and TB incidence among males in both the medium and low HDI categories. The higher number of TB cases in the 0-1 range suggests a possible association between lower hemoglobin levels and increased susceptibility to TB infection. However, it is crucial to note that further investigation is required to

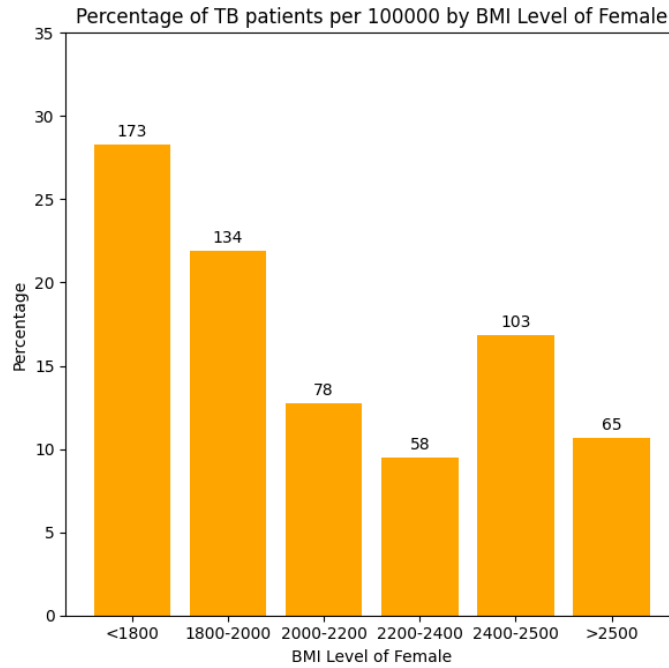


Figure 3.14: Percentage of TB patients per 100000 by BMI Level of Female in Medium HDI

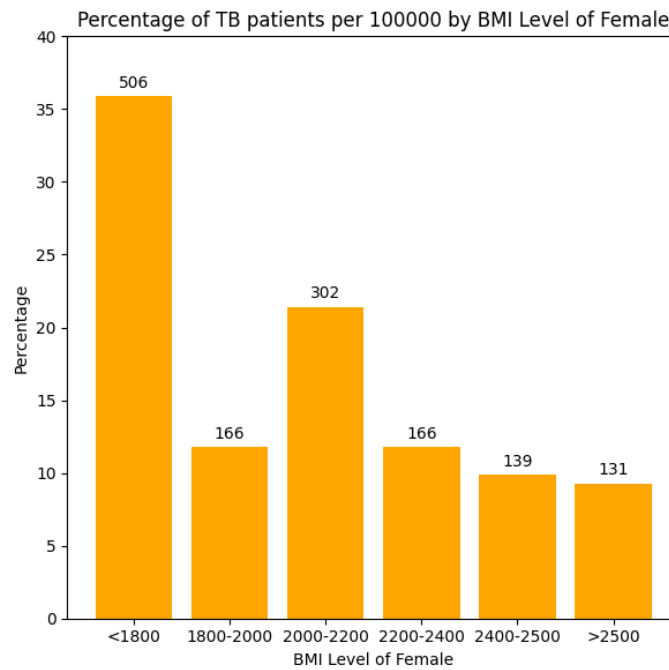


Figure 3.15: Percentage of TB patients per 100000 by BMI Level of Female in High HDI

fully understand the underlying mechanisms and causative factors driving this relationship.

The observed pattern underscores the importance of addressing anemia as a potential risk factor for TB and advocating for interventions aimed at improving overall iron and hemoglobin status among males. Effective strategies for preventing and managing anemia may contribute to the reduction of TB cases in these populations.

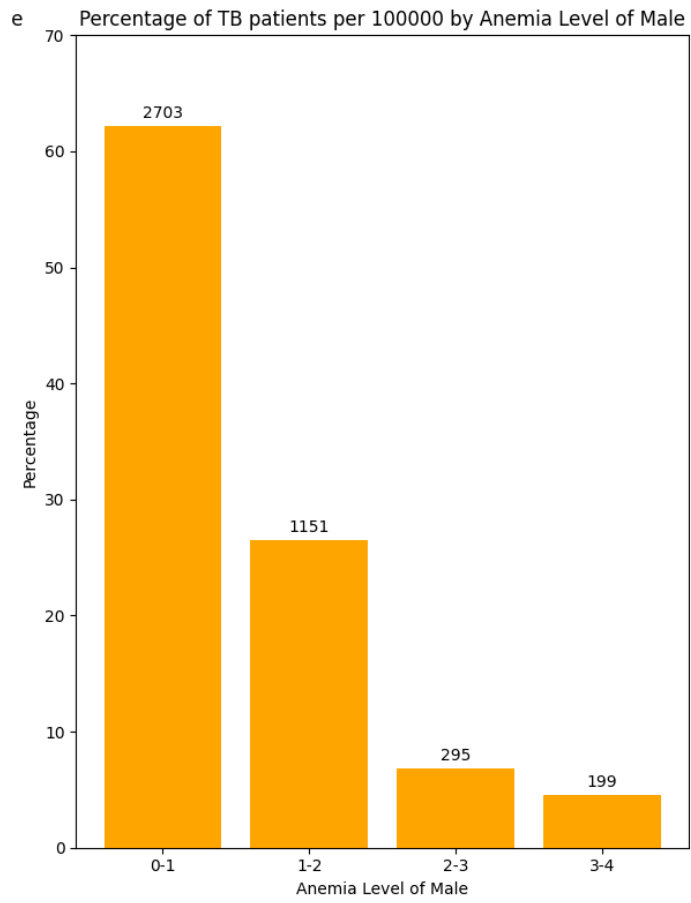


Figure 3.16: Percentage of TB patients per 100000 by Anemia Level of Male in Low HDI

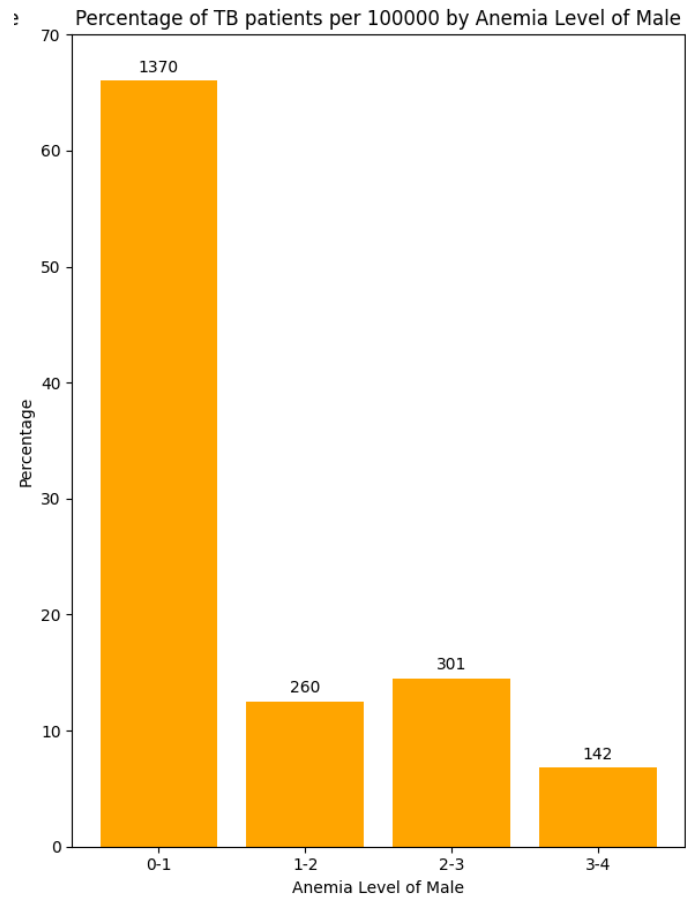


Figure 3.17: Percentage of TB patients per 100000 by Anemia Level of Male in Medium HDI

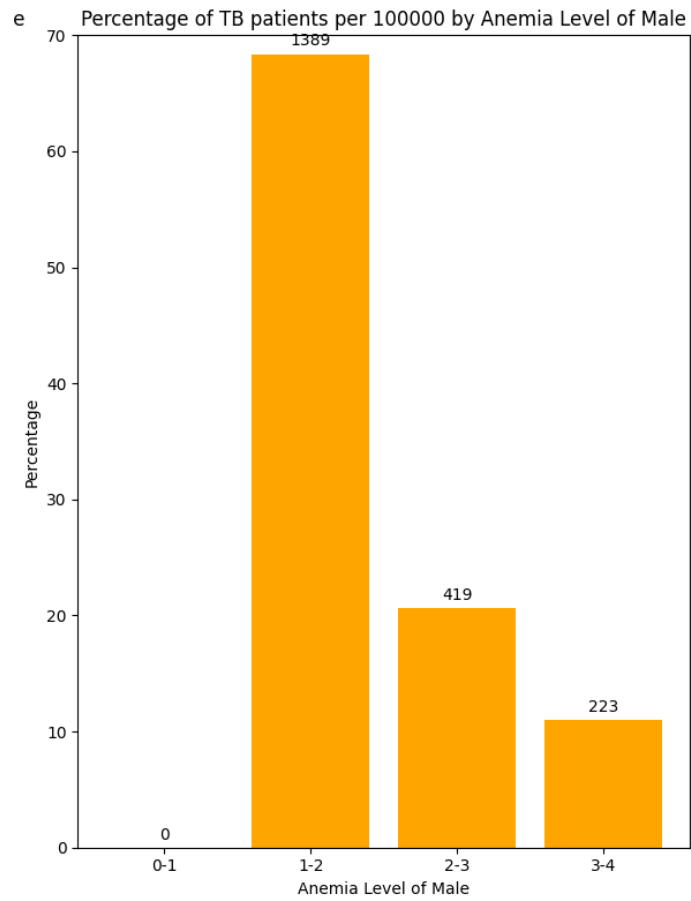


Figure 3.18: Percentage of TB patients per 100000 by Anemia Level of Male in High HDI

3.2.4 Educational Level

In the analysis of the highest educational level achieved by individuals and its impact on tuberculosis (TB) prevalence, compelling patterns emerge across different Human Development Index (HDI) categories. These findings underscore the importance of education as a factor influencing TB cases in India.

In the Medium HDI category, it is noteworthy that individuals with no education represent the highest percentage of TB cases, accounting for over 50% of the affected population. Furthermore, those with secondary education contribute to approximately 25% of TB cases, while individuals with higher education levels exhibit a significantly lower incidence of TB.

Similarly, in the High HDI category, individuals with only secondary education account for approximately 50% of TB cases. Those with primary education contribute to around 30% of TB cases, whereas individuals with higher education levels experience a substantially lower prevalence of TB, constituting merely 10% of the affected population.

These findings lead us to a crucial conclusion: education plays a pivotal role in influencing TB cases. Access to education equips individuals with knowledge about TB prevention, awareness of healthcare practices, and the ability to make informed decisions regarding their health. Higher education levels are associated with increased health literacy, which includes understanding the importance of preventive measures, early detection, and timely treatment-seeking behaviors.

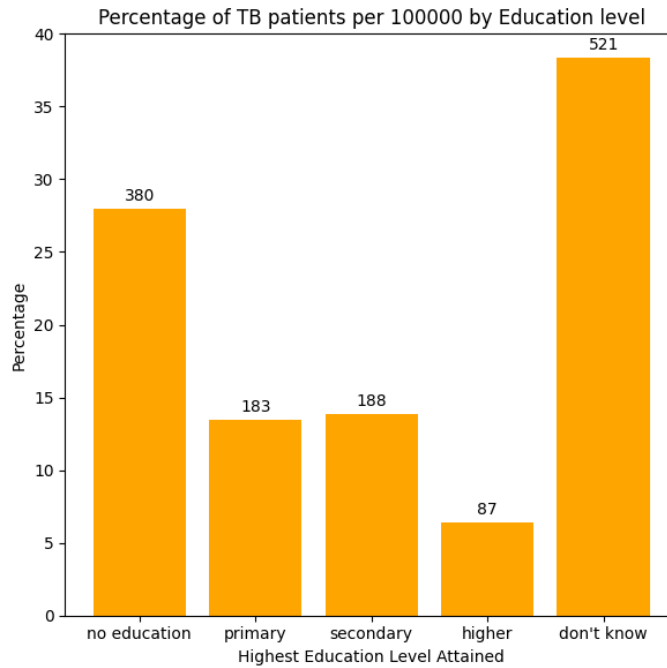


Figure 3.19: Percentage of TB patients per 100000 by Educational Level in Low HDI

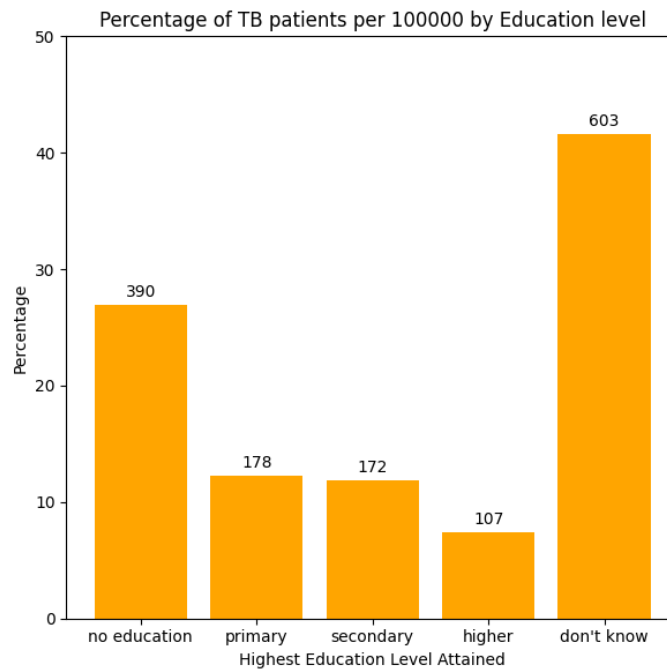


Figure 3.20: Percentage of TB patients per 100000 by Educational Level in Medium HDI

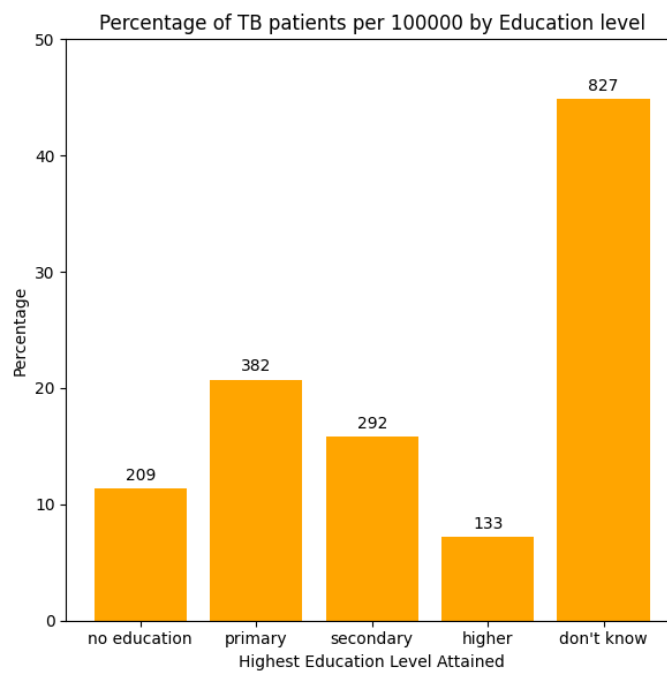


Figure 3.21: Percentage of TB patients per 100000 by Educational Level in High HDI

3.2.5 Wealth Index and Agewise

In terms of the wealth index, a distinct pattern emerges. In both the High and Medium HDI categories, individuals belonging to the poorest segment exhibit the highest percentage of TB cases, accounting for approximately 35% of the affected population. Conversely, individuals in the richest segment have the lowest TB incidence, with less than 10% of cases reported. Moreover, as we move from the poorest to the poorer, middle, and richer segments, the number of TB cases progressively decreases. These findings suggest a clear relationship between wealth index and TB prevalence, with a higher socioeconomic status being associated with a lower risk of contracting TB.

Examining TB cases by age, distinct age-related patterns are observed across different HDI categories. In the High HDI category, individuals aged 50-60 and 70 and above demonstrate a higher prevalence of TB cases. Conversely, individuals under the age of 18 exhibit a lower incidence of TB. This trend suggests that middle-aged and elderly populations are at a relatively higher risk of TB infection within the High HDI category.

In the Medium HDI category, individuals aged 70 and above have the highest number of TB cases. This finding emphasizes the vulnerability of the elderly population to TB infection within the Medium HDI category.

In conclusion, the analysis of wealth index and age in relation to TB cases reveals important insights. Addressing socioeconomic disparities and implementing age-specific interventions can contribute to reducing TB prevalence. By prioritizing efforts to reach vulnerable populations, such as those in lower wealth index categories and specific age groups, we can make significant progress in mitigating the burden of TB in society.

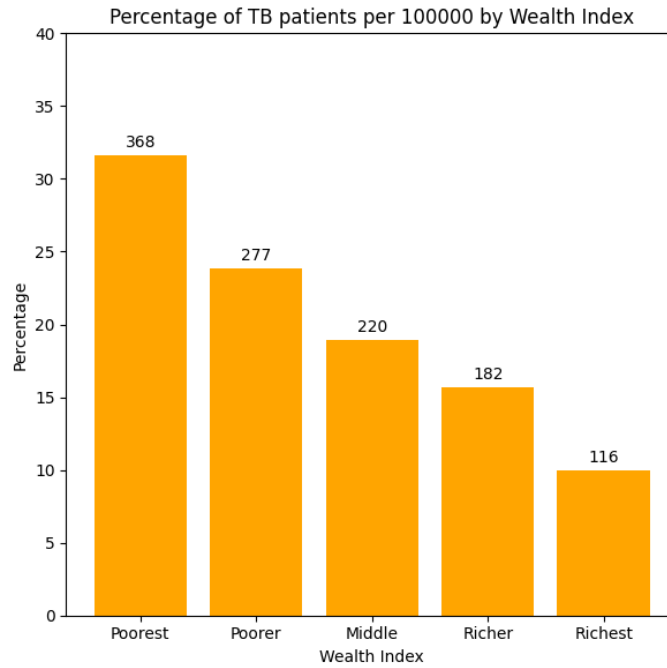


Figure 3.22: Percentage of TB patients per 100000 by Wealth Index in Low HDI

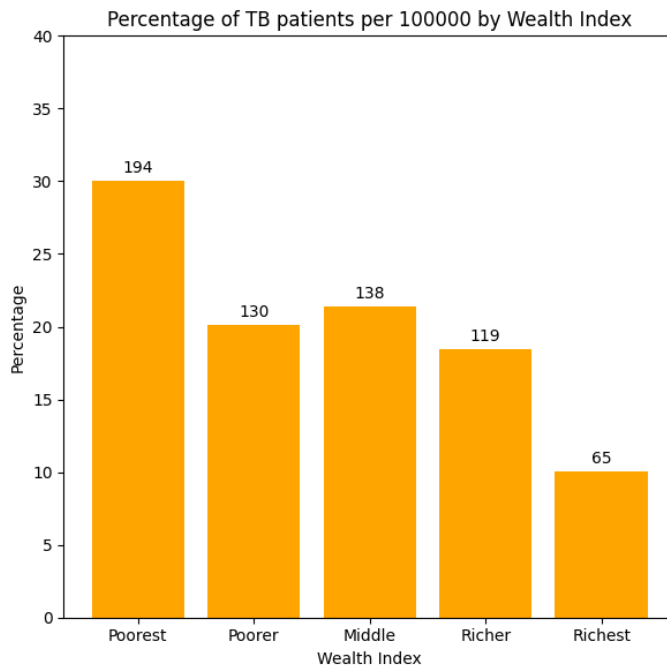


Figure 3.23: Percentage of TB patients per 100000 by Wealth Index in Medium HDI

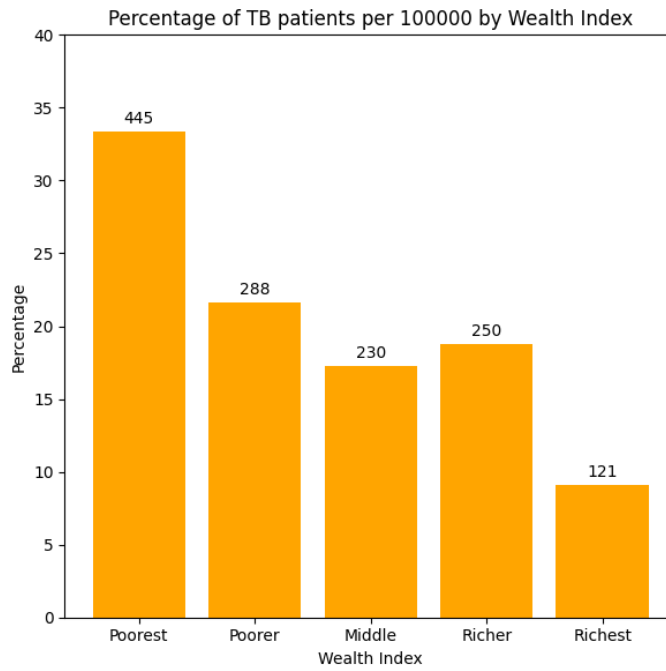


Figure 3.24: Percentage of TB patients per 100000 by Wealth Index in High HDI

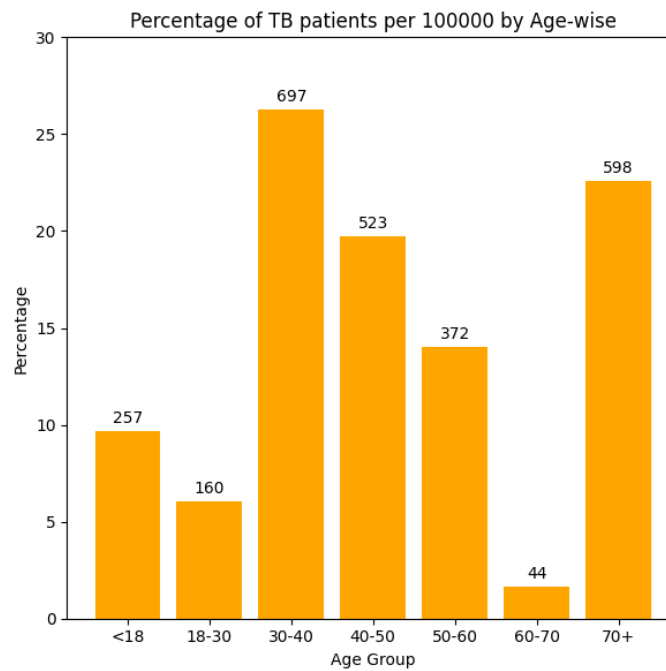


Figure 3.25: Percentage of TB patients per 100000 by Age in Low HDI

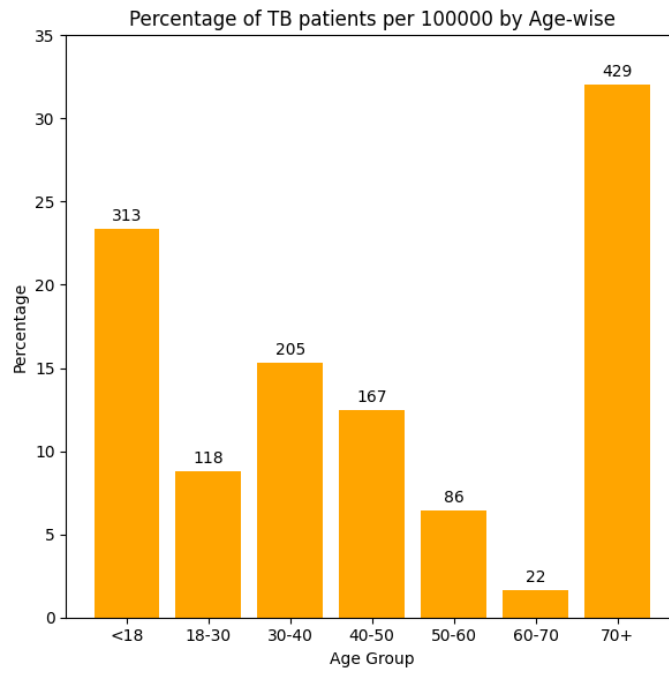


Figure 3.26: Percentage of TB patients per 100000 by Age in Medium HDI

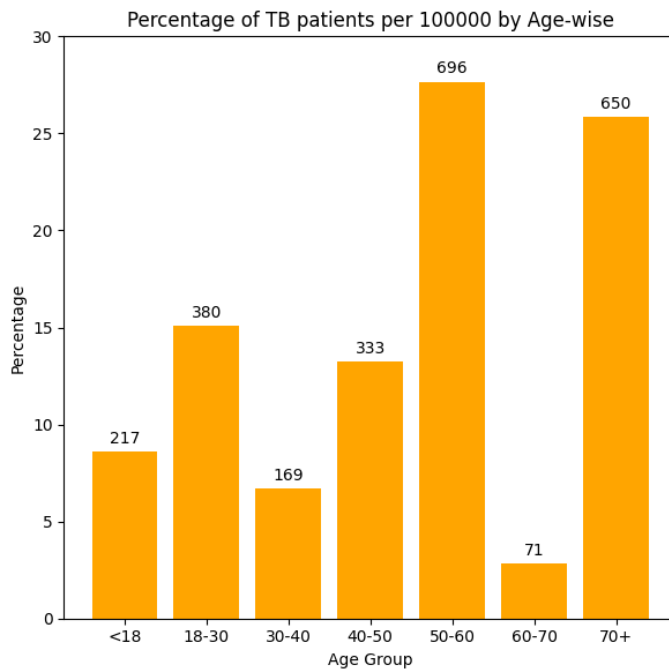


Figure 3.27: Percentage of TB patients per 100000 by Age in High HDI

3.2.6 State-wise

The table provides a comprehensive summary of tuberculosis (TB) cases in different states and gender categories. It breaks down the data by state (including Chandigarh, Delhi, Goa, Kerala, and Puducherry) and gender (Male, Female, Transgender).

For instance, in Delhi, the table shows that there are 31 TB cases among Females and 28 among Males, while there are no recorded Transgender TB cases. In Kerala, the number of TB cases is higher, with 60 among Females and 145 among Males, again with no recorded Transgender TB cases. The table allows for a quick overview of the distribution of TB cases across various demographic groups within each state.

In Bihar, there are a substantial number of TB cases among both males and females, with 289 TB cases among females and 454 among males. There are also 11 transgender individuals in the dataset, but none of them have TB. In Jharkhand, there are fewer TB cases compared to Bihar. Among females, there are 63 TB cases, and among males, there are 167 TB cases. There are 6 transgender individuals, and none of them have TB.

Madhya Pradesh has a similar pattern with more TB cases among males 151 compared to females 88. There are 53 transgender individuals, and none of them have TB. In Odisha, there are 93 TB cases among females and 140 among males. There are 3 transgender individuals, and none of them have TB.

Uttar Pradesh has a relatively high number of TB cases, with 278 TB cases among females and 489 among males. There are 15 transgender individuals, and none of them have TB.

State	Gender	Non TB patients	TB patients
Chandigarh	Female	1635	1
	Male	1751	0
Delhi	Female	20689	31
	Male	22630	28
Goa	Female	3686	8
	Male	3592	10
Puducherry	Female	6971	4
	Male	6200	17
Kerala	Female	24349	60
	Male	21778	145

Table 3.1: Comparison of States in High HDI

State	Gender	Non TB patients	TB patients
Arunachal pradesh	Female	35089	150
	Male	35619	198
Telangana	Female	51625	81
	Male	48482	152
Uttarakhand	Female	27293	15
	Male	25082	33
Karnataka	Female	59224	86
	Male	57184	136
Rajasthan	Female	83124	108
	Male	81829	229

Table 3.2: Comparison of States in Medium HDI

State	Gender	Non TB patients	TB patients
Bihar	Female	93012	289
	Male	85068	454
Jharkhand	Female	53561	63
	Male	50619	167
Madhya Pradesh	Female	100491	88
	Male	102895	151
Odisha	Female	55802	93
	Male	52915	140
Uttar Pradesh	Female	188596	278
	Male	185145	489

Table 3.3: Comparison of States in Low HDI

3.3 Implementation

3.3.1 Class Balancing Techniques

Data imbalance is a common issue [6] in machine learning and data analysis, where the distribution of classes or categories in the dataset is heavily skewed. This imbalance can negatively impact the performance and accuracy of predictive models, as the model may become biased towards the majority class and struggle to effectively learn from the minority class. To address this challenge, several techniques can be employed to mitigate data imbalance:

1. Undersampling:

Undersampling involves reducing the number of instances from the majority class (or classes) to balance the class distribution. The goal is to match the number of instances in the minority class with that of the majority class. This technique helps prevent the model from being biased towards the majority class and ensures that both classes receive equal representation during training. Undersampling can be done randomly, where instances from the majority class are randomly selected and removed, or it can be performed using more advanced methods, such as Tomek links or Cluster Centroids, which identify and remove specific instances based on their proximity to other instances.

2. Oversampling:

Oversampling, on the other hand, involves increasing the number of instances in the minority class to balance the class distribution. The aim is to generate synthetic instances for the minority class to match the number of instances in the majority class.

This helps to provide sufficient representation and prevent the model from being biased toward the majority class. The most commonly used oversampling technique is the Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic examples by interpolating between the feature vectors of existing minority class instances. SMOTE generates new samples by considering the k nearest neighbors of each minority class instance and creating synthetic instances along the line segments connecting them

3. **Cost-Sensitive Learning:**

Cost-sensitive learning involves assigning different costs or penalties to different types of misclassifications based on their relative importance. This approach is particularly useful when the misclassification costs of different classes are uneven or when the consequences of false positives and false negatives differ significantly. By incorporating these costs into the learning algorithm, the model can prioritize minimizing the overall cost instead of simply optimizing accuracy. Cost-sensitive learning algorithms adjust the classification threshold or modify the loss function to account for the associated costs, thus allowing for more effective decision-making in scenarios where certain errors are more costly than others.

4. **Ensemble Methods:**

Ensemble methods combine multiple individual models to create a stronger and more accurate predictive model. By leveraging the diversity of the constituent models, ensemble methods can reduce overfitting, increase generalization, and improve overall performance. Common ensemble methods include:

- (a) **Bagging (Bootstrap Aggregating):** It involves training multiple models on different bootstrap samples of the training data and averaging their predictions to make final decisions.
- (b) **Boosting:** It builds an ensemble of models sequentially, with each subsequent model focusing on instances that were misclassified by the previous models, thereby improving overall performance.
- (c) **Random Forest:** It is an ensemble of decision trees, where each tree is trained on a random subset of features, and the final prediction is determined by majority voting.

Ensemble methods can enhance model stability, handle complex relationships, and capture diverse patterns in the data, leading to improved predictive accuracy.

5. Data Augmentation:

Data augmentation techniques are used to artificially increase the size of the training dataset by creating additional synthetic samples. This is particularly useful when the available dataset is limited or imbalanced. Data augmentation techniques introduce variations or perturbations to the existing data, such as rotation, scaling, flipping, adding noise, or applying transformations while preserving the original class labels. By augmenting the training data, the model learns from a more diverse set of examples, which can help improve generalization and reduce overfitting.

Data augmentation is commonly applied in computer vision tasks, such as image classification and object detection, but can also be used in other domains. It allows models to learn from a more comprehensive range of data variations and can lead to improved performance and robustness.

6. Algorithmic Approaches:

Some algorithms inherently handle imbalanced data better than others. For example, support vector machines (SVM) with appropriate class weights or decision trees with balanced weightage to class samples can be effective in handling imbalanced datasets.

These machine learning models can be applied to imbalanced datasets: Logistic Regression, Decision Trees, Support Vector Machines (SVM), Random Forests, Gradient Boosting Models (e.g., XGBoost, LightGBM, CatBoost), Neural Networks (including CNN and RNN), Naive Bayes, K-nearest Neighbors (KNN), Anomaly Detection Algorithms (e.g., One-Class SVM, Isolation Forest), Ensemble Methods (e.g., AdaBoost, Bagging, Stacking)

These models can be adapted and used in various ways to address the class imbalance, such as adjusting class weights, applying resampling techniques, or using specialized algorithms for imbalanced classification. The selection of the most suitable model will depend on the specific problem and dataset characteristics.

3.3.2 Prediction Model

In our study, we encountered the challenge of data imbalance with our real-time dataset. Due to the nature of the data, traditional approaches such as undersampling and data augmentation techniques were not viable options. Undersampling would have resulted in a loss of valuable data, while data augmentation techniques did not offer meaningful transformations applicable to our dataset.

To address this issue, we opted to utilize the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is an effective method [15] for generating synthetic samples of the minority class by utilizing the K-nearest neighbors (KNN) algorithm. By applying SMOTE, we successfully balanced the distribution of our target and non-target classes, ensuring a more representative dataset for model training and evaluation.

Having achieved a balanced dataset, we proceeded to apply various classification algorithms, leveraging the analyzed features that proved to be most effective in predicting tuberculosis (TB) cases. By utilizing these features and employing machine learning models, we were able to accurately predict whether an individual had TB or not.

It is important to note that the choice of SMOTE as the data imbalance solution was driven by the real-time nature of our dataset and the limitations of other techniques. By achieving class balance through SMOTE, we could mitigate the bias towards the majority class and enhance the performance of our predictive models.

This approach not only allowed us to leverage the full potential of our data but also improved the accuracy and reliability of our TB predictions. By incorporating the analyzed features into our prediction models, we could identify significant factors associated with TB and utilize them to make informed predictions about an individual's TB status.

The successful implementation of SMOTE in addressing data imbalance in our real-time dataset demonstrates the importance of employing tailored techniques to overcome specific challenges. This methodology ensures that our analysis remains robust and provides valuable insights into the prediction and prevention of TB, contributing to the existing body of knowledge in the field.

In order to address the data imbalance issue in our study across all three categories (high, medium, and low HDI), we employed the Synthetic Minority Over-sampling Technique (SMOTE). By applying SMOTE, we successfully balanced the distribution of our target and non-target classes, ensuring a more representative dataset for analysis.

After performing SMOTE, we separated the dataset into features and target columns. Subsequently, we utilized the logistic regression model as our classifier. The dataset was randomly split into training and testing sets, with 80% allocated for training and 20% for testing purposes. Upon fitting the logistic regression model, we generated a classification report to evaluate the model's performance.

In the second approach, we implemented Principal Component Analysis (PCA) before applying logistic regression. The purpose of PCA was to reduce the dimensionality of the feature space while retaining the most relevant information. Similar to the previous technique, we split the dataset into training and testing sets, and a classification report was generated based on the logistic regression model applied to the transformed data.

In the third approach, instead of PCA, we utilized our own analyzed features. These features were carefully selected based on their significance and impact on the target variable. We compared the performance of logistic regression using these analyzed features with the previous techniques. Notably, this technique yielded the highest accuracy among all three categories.

In the fourth approach, we employed the combination of SMOTE, analyzed features, and Adaboost, which is an ensemble technique known for its ability to combine weak classifiers to create a stronger, more accurate model. Despite its promising potential, this approach yielded lower accuracy in comparison to the previous three methods.

Continuing our analysis, we explored the fifth approach, which incorporated SMOTE, analyzed features, and Gradient Boosting. Gradient Boosting is another ensemble technique that sequentially trains multiple models, with each subsequent model focusing on correcting the mistakes made by the previous models. The implementation of this approach resulted in improved accuracy across all three categories, surpassing the performance achieved by Adaboost.

Lastly, we investigated the effectiveness of the sixth approach, which combined SMOTE, analyzed features, and Random Forest. Random Forest is a powerful ensemble learning method that constructs a multitude of decision trees and combines their outputs to make predictions. This approach yielded exceptional results, with accuracy rates ranging from 98% to 99% in all three categories. These findings indicate that the SMOTE, analyzed features, and Random Forest approach produced the most accurate predictions among all the methods evaluated in our study.

These outcomes highlight the importance of carefully selecting and combining different techniques to optimize prediction accuracy in the analysis of tuberculosis cases. The superior performance of the SMOTE analyzed features, and Random Forest approach demonstrates its potential to enhance the accuracy of tuberculosis prediction models, making it a valuable method for future research and clinical applications.

CHAPTER 4

Results

Table 4.1 presents a comparison of different methods used in the analysis of tuberculosis (TB) cases within regions characterized by high Human Development Index (HDI). The table displays performance metrics, including precision, recall, F1 score, support, and accuracy, for each method and category.

First of all, we discuss about the definition and formulas of the terms which we have used in our performance metrics.

Precision is a metric that calculates the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive.

$$\text{Precision} = \text{True Positives} / (\text{True Positives} + \text{False Positives})$$

Recall measures the proportion of correctly predicted positive instances (true positives) out of all actual positive instances.

$$\text{Recall} = \text{True Positives} / (\text{True Positives} + \text{False Negatives})$$

The F1 score is a harmonic mean of precision and recall. It's useful when you want to balance both precision and recall.

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

Support refers to the number of actual occurrences of a class in the dataset. It can be seen as the "true" count of instances belonging to a particular class.

Accuracy is a measure of how many instances are correctly classified overall. It's the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances.

$$\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / \text{Total Instances}$$

True Positives (TP) are the instances that are correctly predicted as positive. False Positives (FP) are the instances that are predicted as positive but are actually negative.

False Negatives (FN) are the instances that are predicted as negative but are actually positive.

True Negatives (TN) are the instances that are correctly predicted as negative.

The first two methods, "SMOTE and Logistic Regression" and "SMOTE, PCA, and Logistic Regression," utilized logistic regression in combination with SMOTE (Synthetic Minority Over-sampling Technique) and principal component analysis (PCA). Both methods achieved similar results, with precision, recall, and F1 scores of around 0.63 to 0.65 for both the non-TB and TB categories. The accuracy for these methods was 65%.

The next approach, "SMOTE, Analyzed Features, and Logistic Regression," incorporated analyzed features in addition to SMOTE and logistic regression. This method demonstrated improved performance, with precision, recall, and F1 scores of approximately 0.72 to 0.79 for the non-TB category and 0.74 to 0.81 for the TB category. The overall accuracy increased to 76%.

Moving forward, the "SMOTE, Analyzed Features, and AdaBoost" method utilized AdaBoost, an ensemble technique, along with SMOTE and analyzed features. This approach showed further enhancement, with precision, recall, and F1 scores ranging from 0.80 to 0.85 for the non-TB category and 0.82 to 0.84 for the TB category. The overall accuracy reached 83%.

Continuing the analysis, the "SMOTE, Analyzed Features, and Gradient Boosting" method employed Gradient Boosting, another ensemble technique, along with SMOTE and analyzed features. This approach yielded even higher precision, recall, and F1 scores, ranging from 0.83 to 0.89 for the non-TB category and 0.83 to 0.88 for the TB category. The overall accuracy increased to 86%.

Finally, the "SMOTE, Analyzed Features, and Random Forest" method utilized Random Forest, a powerful ensemble learning technique, in combination with SMOTE and analyzed features. This approach showcased outstanding performance, with precision, recall, and F1 scores of 0.97 to 0.99 for both the non-TB and TB categories. The accuracy of this method reached an impressive 98%.

These results demonstrate the effectiveness of utilizing different methods, including ensemble techniques, in the analysis of TB cases within high HDI regions. The SMOTE, Analyzed Features, and Random Forest approach exhibited the highest accuracy and can be considered a valuable method for tuberculosis prediction models in similar contexts.

Method	Cases	Precision	Recall	F1 score	Support	Accuracy
SMOTE and Logistic Regression	Non-TB	0.65	0.64	0.63	22541	65%
	TB	0.64	0.66	0.65	22773	
SMOTE, PCA and Logistic Regression	Non-TB	0.65	0.64	0.63	22541	65%
	TB	0.64	0.67	0.63	22773	
SMOTE, Analyzed Features and Logistic Regression	Non-TB	0.79	0.72	0.75	17749	76%
	TB	0.74	0.81	0.77	17719	
SMOTE, Analyzed Features and AdaBoost	Non-TB	0.81	0.85	0.83	17749	83%
	TB	0.84	0.80	0.82	17719	
SMOTE, Analyzed Features and Gradient Boosting	Non-TB	0.84	0.89	0.86	17749	86%
	TB	0.88	0.83	0.85	17719	
SMOTE, Analyzed Features and Random Forest	Non-TB	0.99	0.97	0.98	17749	98%
	TB	0.97	0.99	0.98	17719	

Table 4.1: Comparison of Methods in High HDI

Table 4.2 presents a comparison of different methods used in the analysis of tuberculosis (TB) cases within regions characterized by the medium Human Development Index (HDI). The table provides performance metrics, including precision, recall, F1 score, support, and accuracy, for each method and category.

The first two methods, "SMOTE and Logistic Regression" and "SMOTE, PCA, and Logistic Regression," utilized logistic regression in combination with SMOTE (Synthetic Minority Over-sampling Technique) and principal component analysis (PCA). Both methods yielded similar results, with precision, recall, and F1 scores of around 0.60 for both the non-TB and TB categories. The accuracy for these methods was 60%.

Moving on, the "SMOTE, Analyzed Features, and Logistic Regression" approach incorporated analyzed features along with SMOTE and logistic regression. This method demonstrated improved performance, with precision, recall, and F1 scores of approximately 0.77 to 0.80 for the non-TB category and 0.76 to 0.80 for the TB category. The overall accuracy increased to 78%.

Next, the "SMOTE, Analyzed Features, and AdaBoost" method utilized AdaBoost, an ensemble technique, in combination with SMOTE and analyzed features. This approach showed further enhancement, with precision, recall, and F1 scores ranging from 0.87 to 0.90 for the non-TB category and 0.87 to 0.90 for the TB category. The overall accuracy improved to 88%.

Continuing the analysis, the "SMOTE, Analyzed Features, and Gradient Boosting" approach employed Gradient Boosting, another ensemble technique, along with SMOTE and analyzed features. This method yielded higher precision, recall, and F1 scores, ranging from 0.88 to 0.93 for the non-TB category and 0.87 to 0.93 for the TB category. The overall accuracy reached 90%.

Finally, the "SMOTE, Analyzed Features, and Random Forest" method utilized Random Forest, a powerful ensemble learning technique, along with SMOTE and analyzed features. This approach showcased exceptional performance, with precision, recall, and F1 scores of 0.98 to 1.00 for both the non-TB and TB categories. The accuracy for this method reached an impressive 99%.

These results demonstrate the effectiveness of utilizing different methods, including ensemble techniques, in the analysis of TB cases within medium HDI regions. The SMOTE, Analyzed Features, and Random Forest approach exhibited the highest accuracy and can be considered a valuable method for tuberculosis prediction models in similar contexts.

Method	Cases	Precision	Recall	F1 score	Support	Accuracy
SMOTE and Logistic Regression	Non-TB	0.60	0.60	0.60	100919	60%
	TB	0.60	0.61	0.60	100910	
SMOTE, PCA and Logistic Regression	Non-TB	0.60	0.60	0.60	100919	60%
	TB	0.60	0.61	0.60	100919	
SMOTE, Analyzed Features and Logistic Regression	Non-TB	0.77	0.80	0.79	75105	78%
	TB	0.80	0.76	0.78	75177	
SMOTE, Analyzed Features and AdaBoost	Non-TB	0.87	0.90	0.88	75105	88%
	TB	0.90	0.87	0.88	75177	
SMOTE, Analyzed Features and Gradient Boosting	Non-TB	0.88	0.93	0.91	75105	90%
	TB	0.93	0.87	0.90	75177	
SMOTE, Analyzed Features and Random Forest	Non-TB	1.00	0.98	0.99	75105	99%
	TB	0.98	1.00	0.99	75177	

Table 4.2: Comparison of Methods in Medium HDI

Table 4.3 presents a comparison of different methods used in the analysis of tuberculosis (TB) cases within regions characterized by low Human Development Index (HDI). The table provides performance metrics, including precision, recall, F1 score, support, and accuracy, for each method and category.

The first two methods, "SMOTE and Logistic Regression" and "SMOTE, PCA, and Logistic Regression," utilized logistic regression in combination with SMOTE (Synthetic Minority Over-sampling Technique) and principal component analysis (PCA). Both methods yielded similar results, with precision, recall, and F1 scores ranging from 0.57 to 0.63 for the non-TB category and 0.61 to 0.65 for the TB category. The accuracy for these methods was 62%.

Moving on, the "SMOTE, Analyzed Features, and Logistic Regression" approach incorporated analyzed features along with SMOTE and logistic regression. This method demonstrated improved performance, with precision, recall, and F1 scores of approximately 0.77 to 0.81 for the non-TB category and 0.76 to 0.80 for the TB category. The overall accuracy increased to 79%.

Next, the "SMOTE, Analyzed Features, and AdaBoost" method utilized AdaBoost, an ensemble technique, in combination with SMOTE and analyzed features. This approach showed further enhancement, with precision, recall, and F1 scores ranging from 0.83 to 0.85 for the non-TB category and 0.83 to 0.85 for the TB category. The overall accuracy improved to 84%.

Continuing the analysis, the "SMOTE, Analyzed Features, and Gradient Boosting" approach employed Gradient Boosting, another ensemble technique, along with SMOTE and analyzed features. This method yielded higher precision, recall, and F1 scores, ranging from 0.82 to 0.90 for the non-TB category and 0.82 to 0.89 for the TB category. The overall accuracy reached 86%.

Finally, the "SMOTE, Analyzed Features, and Random Forest" method utilized Random Forest, a powerful ensemble learning technique, along with SMOTE and analyzed features. This approach showcased exceptional performance, with precision, recall, and F1 scores of 0.95 to 0.99 for both the non-TB and TB categories. The accuracy for this method reached an impressive 97%.

These results demonstrate the effectiveness of utilizing different methods, including ensemble techniques, in the analysis of TB cases within low HDI regions. The SMOTE, Analyzed Features, and Random Forest approach exhibited the highest accuracy and can be considered a valuable method for tuberculosis prediction models in similar contexts.

Method	Cases	Precision	Recall	F1 score	Support	Accuracy
SMOTE and Logistic Regression	Non-TB	0.62	0.58	0.60	193240	62%
	TB	0.61	0.65	0.63	194037	
SMOTE, PCA and Logistic Regression	Non-TB	0.63	0.57	0.60	193240	62%
	TB	0.61	0.66	0.63	194037	
SMOTE, Analyzed Features and Logistic Regression	Non-TB	0.77	0.81	0.79	133551	79%
	TB	0.80	0.76	0.78	134445	
SMOTE, Analyzed Features and AdaBoost	Non-TB	0.83	0.85	0.84	133551	84%
	TB	0.85	0.83	0.84	134445	
SMOTE, Analyzed Features and Gradient Boosting	Non-TB	0.83	0.90	0.86	133551	86%
	TB	0.89	0.82	0.85	134445	
SMOTE, Analyzed Features and Random Forest	Non-TB	0.99	0.95	0.97	133551	97%
	TB	0.95	0.99	0.97	134445	

Table 4.3: Comparison of Methods in Low HDI

CHAPTER 5

Conclusion

In this study, we conducted a comprehensive analysis of household member data, focusing on identifying the factors that significantly impact tuberculosis (TB) patients compared to non-TB individuals. Through this analysis, we were able to identify key features associated with TB cases, shedding light on the underlying factors contributing to TB occurrence. In order to obtain accurate data, we normalized the number of tuberculosis (TB) and non-TB patients per 100,000 people in each category. This approach was adopted due to the relatively low number of TB cases available for analysis. By considering the number of cases per 100,000 individuals, we were able to account for population variations and derive meaningful insights from the data.

To develop a prediction model for TB, we addressed the issue of data imbalance using the Synthetic Minority Over-sampling Technique (SMOTE). Subsequently, we applied three different approaches to evaluate the effectiveness of our prediction model. The first approach involved applying SMOTE and logistic regression. The second approach incorporated SMOTE, Principal Component Analysis (PCA), and logistic regression. Finally, in the third approach, we employed SMOTE, our own analyzed features, and logistic regression. Upon comparing the results of these approaches, we found that the third approach yielded the highest accuracy, with approximately 80% accuracy in predicting whether a person has TB across all three categories of high, medium, and low Human Development Index (HDI).

This finding emphasizes the significance of incorporating our own analyzed features in conjunction with SMOTE and logistic regression for TB prediction. The utilization of these carefully selected features allowed for a more accurate prediction of TB cases, surpassing the accuracy achieved by the other two approaches. Our study contributes to the existing literature by providing insights

into the prediction of TB cases using a combination of data-balancing techniques and logistic regression. The high accuracy achieved in the third approach highlights the potential of our methodology for effective TB prediction across different HDI categories.

However, it is important to acknowledge certain limitations of our study. These include the reliance on a specific dataset and the need for further validation on larger and more diverse datasets. Additionally, the generalizability of our findings to other populations and settings should be considered.

In conclusion, our study demonstrates the importance of addressing data imbalances and employing a comprehensive approach to predict TB cases. The findings support the use of SMOTE, our analyzed features, and logistic regression as an effective methodology for accurate TB prediction. These results contribute to the understanding of TB epidemiology and can inform public health interventions and strategies to combat TB effectively. Further research is warranted to validate and refine our approach and explore its applicability in real-world healthcare settings.

5.1 Policy Implications of This Research

Early Detection and Intervention

The predictive models developed in this study have the potential to revolutionize tuberculosis (TB) control by enabling early detection of individuals at a higher risk of developing the disease. This advancement could lead to timely medical intervention and treatment, consequently reducing both the severity and transmission of TB.

Public Health Campaigns

The correlation analysis conducted on features such as smoking, alcohol consumption, and education level can serve as the foundation for targeted public health campaigns. These initiatives can focus on raising awareness about the risk factors associated with TB, promoting healthier behaviors, and ultimately contributing to disease prevention.

Policy Recommendations

Our research contributes to evidence-based policy recommendations for combating TB. By uncovering significant associations between socio-economic factors and TB prevalence, policymakers can consider interventions that improve education and income distribution, thereby reducing TB risk among vulnerable populations.

Healthcare Equity

Through the examination of wealth index, education, and sex ratio, our study highlights healthcare disparities linked to TB. This insight can guide the development of strategies aimed at promoting equity in healthcare access and quality of treatment across diverse socio-economic backgrounds.

International Collaboration

As TB remains a global health concern, our findings can contribute to international efforts aimed at tackling the disease. Sharing insights and models with organizations such as the World Health Organization (WHO) facilitates the creation of coordinated strategies that transcend borders.

Research and Further Studies

This study lays the groundwork for future research in the field of TB. Researchers can build upon our findings to explore nuanced relationships between different variables and their intricate impact on TB prevalence, fostering a deeper understanding of the disease dynamics.

Health Education Programs

The analysis conducted on factors like education, smoking, and alcohol can provide the basis for targeted health education initiatives. These programs can educate individuals on the risks associated with specific behaviors and their correlation with TB, promoting informed decision-making.

Data-Driven Decision-Making

In essence, this research advocates for data-driven decision-making in the healthcare sector. It emphasizes the importance of employing quantitative insights to guide policies, interventions, and the allocation of resources effectively, ultimately contributing to more efficient TB control strategies.

References

- [1] S. Anand and A. Sen. Human development index: Methodology and measurement. 1994.
- [2] P. Barman. Impact of education and media exposure on tuberculosis related awareness among indian adults: A study based on nfh-3. *SAARC Journal of Tuberculosis, Lung Diseases and HIV/AIDS*, 17(2):8–14, 2019.
- [3] J. Bhat, V. Rao, R. Sharma, M. Muniyandi, R. Yadav, and M. K. Bhondley. Investigation of the risk factors for pulmonary tuberculosis: a case–control study among saharia tribe in gwalior district, madhya pradesh, india. *The Indian journal of medical research*, 146(1):97, 2017.
- [4] A. Das, T. Lakhan, and S. Unisa. Tuberculosis prevalence, knowledge of transmission and its association with vaccination of children. *Journal of Infection Prevention*, 22(6):259–268, 2021.
- [5] D. Dhamnetiya, P. Patel, R. P. Jha, N. Shri, M. Singh, and K. Bhattacharyya. Trends in incidence and mortality of tuberculosis in india over past three decades: a joinpoint and age–period–cohort analysis. *BMC pulmonary medicine*, 21(1):1–14, 2021.
- [6] I. Domingues, J. P. Amorim, P. H. Abreu, H. Duarte, and J. Santos. Evaluation of oversampling data balancing techniques in the context of ordinal classification. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.
- [7] A. A. Fisher and A. A. Way. The demographic and health surveys program: An overview. *International Family Planning Perspectives*, pages 15–19, 1988.
- [8] N. S. Hochberg, S. Sarkar, C. R. Horsburgh Jr, S. Knudsen, J. Pleskunas, S. Sahu, R. W. Kubiak, S. Govindarajan, P. Salgame, S. Lakshminarayanan, et al. Comorbidities in pulmonary tuberculosis cases in puducherry and tamil nadu, india: opportunities for intervention. *PLoS One*, 12(8):e0183195, 2017.

- [9] S. Mazumdar, S. Satyanarayana, and M. Pai. Self-reported tuberculosis in india: evidence from nfhs-4. *BMJ global health*, 4(3):e001371, 2019.
- [10] C. Padmapriyadarsini, M. Shobana, M. Lakshmi, T. Beena, and S. Swaminathan. Undernutrition & tuberculosis in india: Situation analysis & the way forward. *The Indian journal of medical research*, 144(1):11, 2016.
- [11] G. Pardeshi, A. Deluca, S. Agarwal, and J. Kishore. Tuberculosis patients not covered by treatment in public health services: findings from india’s national family health survey 2015–16. *Tropical Medicine & International Health*, 23(8):886–895, 2018.
- [12] S. Pattnaik, J. Murmu, R. Agrawal, T. Rehman, S. Kanungo, and D. S. Pati. Prevalence, pattern, and determinants of disabilities in india: Insights from nfhs-5 (2019-21). *Frontiers in Public Health*, 11:608, 2023.
- [13] P. Sharma, A. Kumar, and P. Singh. A study of gender differentials in the prevalence of tuberculosis based on nfhs-2 and nfhs-3 data. *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine*, 35(2):230, 2010.
- [14] P. Sudre, G. Ten Dam, and A. Kochi. Tuberculosis: a global overview of the situation today. *Bulletin of the World Health Organization*, 70(2):149, 1992.
- [15] J. Wang, M. Xu, H. Wang, and J. Zhang. Classification of imbalanced data by using the smote algorithm and locally linear embedding. In *2006 8th international Conference on Signal Processing*, volume 3. IEEE, 2006.