ANALYSIS OF NFHS V DISTRICT-WISE DATA USING ML AND POWER BI

THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE

OF

MASTER OF SCIENCE

in

MATHEMATICS

by

Hrithik Singh

(2k22/MSCMAT/18)

Under the Supervision of

PROF. ANJANA GUPTA

Department Of Applied Mathematics

DELHI TECHNOLOGICAL UNIVERSITY

(FORMERLY DELHI COLLEGE OF ENGINEERING)

BAWANA ROAD, DELHI-110042

JUNE, 2024

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Hrithik Singh** student of M.Sc. Applied Mathematics, hereby declare that the Project Dissertation titled "**Analysis of NFHS V District-wise Data using ML and Power BI**" which is submitted by me to the Department of Applied Mathematics, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science, is original and not copied from any source without proper citation. The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Place: Delhi HRITHIK SINGH

Date: JUNE 2024 2K22/MSCMAT/18

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby clarify that the Project Dissertation titled "**Analysis of NFHS V District-wise Data using ML and Power BI**" which is submitted by (Hrithik Singh) 2K22/MSCMAT/18 Department of Applied Mathematics, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any degree or Diploma to this University or elsewhere.

Place: Delhi

Date : June 2024

Prof. Anjana Gupta

SUPERVISOR

DEPARTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

ABSTRACT

This study harnesses machine learning methodologies, with a focus on regression models, integrated into Power BI for the analysis of district-wise data extracted from the National Family Health Survey (NFHS) V. Emphasizing women's empowerment as the central theme, India is divided into zones to discern regional nuances. It determines the condition of women in society at different levels like districts, states and zone so that the required changes can be made at each level. The data has been taken from an official government website http://www.data.gov.in . which vouch for its authenticity. Also, finding the parameters which can lead to women empowerment in society. The insights from the data were analyzed using Python libraries like NumPy, Pandas, Scikit Learn and statsmodel.api.

The Power BI and MS Excel's Pivot table were used to visualize the findings. We have also analyzed how access to basic amenities affect literacy rate by using different regression models. The parameters considered in this study were the ones that would most likely contribute to the situation of women in society. This study will explore each point in detail and its impact on women's status.

DEPARTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

Achieving success in any endeavour requires the support and encouragement of many individuals, and this project is no exception. At the outset, I would like to express my sincere and heartfelt gratitude to everyone who contributed to the completion of this dissertation. Their active guidance, assistance, cooperation, and encouragement were invaluable in helping me achieve my goals. I am deeply indebted to Prof. Anjana Gupta for her dedicated guidance and motivation throughout this process. I am also immensely thankful to my project partner for the mutual coordination and for respecting each other's individuality, which was crucial in completing this project successfully. My heartfelt thanks go to Delhi Technological University for providing me with this opportunity. I am also profoundly grateful to my parents and family members for their unwavering moral and financial support. Lastly, I extend my gratitude to all my friends who directly or indirectly assisted me in completing this project. If I have omitted anyone in this brief acknowledgement, it is not due to a lack of gratitude.

Thanking You

Hrithik Singh

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Importance of women in society

Women play an integral role in society across various domains, including economics, education, healthcare, social activism, and politics. Economically, their participation in the workforce is essential for driving growth and innovation. Women entrepreneurs contribute to job creation and address societal needs, fostering economic development. Education is another critical area where women's influence is transformative. The right for women to education has been realized, and increasing girls' education is a core policy aim of the international development community and most governments of developing countries.[2] Gender equality in education is an important element of the Millennium Development Goals.[2] This international commitment is, in part, founded on a large literature that establishes the positive effects of women's education on a broad range of development outcomes, from reductions in fertility and child mortality to increased productivity and economic growth (World Bank 2001) [2] Educated women not only improve their own lives but also uplift entire communities by investing in their children's education and healthcare, breaking the cycle of intergenerational poverty. Education empowers women to engage more effectively at the decision-making level, defend their rights, and resist the existing social norms and stereotypes. Therefore, the literacy of women becomes a pivotal part of society. In the healthcare arena, women are the primary caregivers in families and are crucial in ensuring the health and well-being of their families. Most people concur that empowerment has multiple dimensions and involves not only agency but also the conditions under which resources can be obtained, the norms governing those conditions, and awareness of those standards. There is also agreement that it is a process and empowerment in some spheres can occur alongside disempowerment in other spheres. [1] Omen are also the proponents of social change and community development. They front grassroots movements, advocate for social justice, and push initiatives addressing issues such as gender inequality, health disparities, and sustainable environments. In political spaces, increasing women's representation is key to achieving gender equality and inclusive governance. Women bring a range of viewpoints and concerns to the table when making decisions, resulting in more equitable laws and regulations that represent the needs and interests of all societal members. In politics, increasing the representation of women is essential to achieving gender equality and inclusive governance. Women leadership relates more to the holistic and "bottom-ups" approach in community development.[3] Finally, women's political participation is also integral to democracy in that the voice of women and their rights are safeguarded. Despite women's important contributions in this regard, they do face big challenges: gender-based disparities in education, health, and political representation. These challenges require comprehensive research and evidence-based interventions. This study is, therefore, designed to make an analysis on the various aspects of the empowerment of women

and their well-being in the Indian diaspora with respect to the differences in gender regarding the aspects of education, health, and political representation.

The study shall, therefore, consider gender disparities in education through the understanding of access to basic facilities and women's literacy rates. Thus, the data on educational attainment and infrastructure development can be used in analyzing districts with disparities in gender literacy rates. Knowing these disparities shall be very important in the designing of focused interventions that will help improve the educational outcomes for women and also foster gender equality in education.

Second, As the importance of education has been reminded time and again. We are going to focus what are the factors which affects the literacy rate.

women's health status will be assessed on some main indicators of health, such as maternal health, diabetes, hypertension, obesity and anemia The analysis of data for these indicators will help the research identify the overall status of women's health and the key challenges or health disparities faced by women in India. This research will inform policy and intervention strategies aimed at improving women's health outcomes and reducing healthcare disparities.

In a nutshell, this study focuses on the empowerment and well-being of women in India, specifically in the areas of education, healthcare, and political representation. This thus contributes toward further policy formulation and intervention strategies for the promotion of gender equality and improvement of the general well-being of women.

1.2. WHAT IS NFHS?

The National Family Health Survey (NFHS) 2019-21, the fifth in its series, offers comprehensive insights into population, health, and nutrition across India, each state/union territory (UT), and 707 districts as of March 31, 2017. [23] Administered by the Ministry of Health and Family Welfare (MoHFW), Government of India, the survey is facilitated by the International Institute for Population Sciences (IIPS), Mumbai, with funding from MoHFW and technical support from ICF, USA. .[23] Additional assistance for the Dried Blood Sample (DBS) component was provided by the Indian Council of Medical Research (ICMR) and the National AIDS Research Institute (NARI), Pune.[23]

Fieldwork for NFHS-5 was conducted in two phases: Phase I, which ran from June 17, 2019, to January 30, 2020, covered 17 states and 5 UTs, and Phase II, which ran from January 2, 2020, to April 30, 2021, covered 11 states and 3 UTs.[4] 636,699 homes, 724,115 women, and 101,839 males provided data for the survey. Building on previous iterations, NFHS-5 maintains continuity in content and methods while introducing new topics like preschool education, disability, access to sanitation facilities, menstruation practices, and abortion methods and reasons. National Health Policy2017 and National Health Programs related to child health were also analysed.[4]

NFHS serves as a crucial tool for setting benchmarks, monitoring health sector progress, evaluating program effectiveness, and identifying areas and demographics

requiring targeted interventions. The NFHS-5 data provides evidence to support current and future initiatives, helping to develop policies that are customized to meet unique needs and provide fair access to key services. In summary, NFHS-5 maintains the practice of offering reliable data that is crucial for making well-informed decisions in public health and family welfare efforts throughout India.

Aims of the project

- Analyzing the sex ratio at birth for children born in the last five years to identify districts with significant gender disparities.
- To identify how the access to basic facilities affects the women literacy rate using machine learning models.
- Insights about the health of women in India

CHAPTER-2

Literature Review

The National Family Health Survey (NFHS) is a vital data repository that provides valuable information on several aspects of health, education, and socioeconomic position throughout India. The introduction of NFHS5 has made district-wise data more helpful for comprehending local dynamics and guiding policy initiatives. This literature review examines previous studies on women's empowerment, health inequalities, education, and economic progress to provide context for the analysis of district-wise data from NFHS5. Machine learning techniques are employed inside the Power BI framework for this analysis. The published article examines the enduring problem of inadequate female presence in politics, attributing it to societal structure, political dynamics, and ideological factors. By introducing a more precise measure of national gender ideology, the study highlights its significant impact on the number of women in national legislatures.[6] Women's Empowerment and Gender Disparities: Kadam (2012) emphasizes women's empowerment as pivotal for reducing gender disparities and fostering societal development. Women's pivotal role in family planning and economic development underscores the need for enhancing their socioeconomic status. [9] Becker, Hubbard, and Murphy (2010) corroborate this by highlighting the global trend of women's higher education attainment, a critical factor in empowerment and economic participation.[10]

Health and Socioeconomic Development: Antony and Laxmaiah (2008) provide insights into India's progress in human development and poverty reduction, crucial for understanding health outcomes. India has made a study progress in the HDI value. Extreme poverty is concentrated in rural areas of northern States while income growth has been dynamic in southern States and urban areas. [11] However, Garg et al. (2010) draw attention to the rising prevalence of obesity among Indian women, indicating complex health challenges that require targeted interventions. While hunger is still an issue in many parts of India, the rise in obesity rates brought on by sedentary lifestyles and junk food habits in many metropolitan and economically stable areas is particularly concerning. Prevention and control of this serious problem through awareness programs to adopt diversified nutritional food and healthy lifestyle are strongly recommended [12] Institutional births ensure deliveries happen under the supervision of skilled healthcare personnel in an enabling environment. For countries like India, with high neonatal and maternal mortalities, achieving 100% coverage of institutional births is a top policy priority.[20] In this respect, public health institutions have a key role, given that they remain the preferred choice by most of the population, owing to the existing barriers to healthcare access.[20] While research in this domain has focused on private health institutions, there are limited studies, especially in the Indian context, that look at the enablers of institutional births in public health facilities.[20] In this study, we look to identify the significant predictors of institutional birth in public health facilities in India.[20] Although nearly 3 in 5 women in India utilized a minimum mandated≥4 ANC visits during

their last pregnancy, only one in five of those received adequate quality of ANC services indicating suboptimal content.[21]

Safety and social media: Kumar and Aggarwal (2019) Examine how social media can be used to make Indian cities safer places for women, suggesting avenues for leveraging technology in data analysis and policy formulation. Their study underscores the importance of incorporating social media data in analyzing NFHS5 findings related to women's safety and well-being.[13]

Social Determinants of Health: Braveman, Egerter, and Williams (2011) discuss the social determinants of health, emphasizing the influence of socioeconomic factors on health outcomes.[14] Short and Zacher (2022) further elaborate on this by examining patterns of women's health disparities in the US, providing valuable insights applicable to interpreting NFHS5 data within the Indian context.[15]

Data Analysis Techniques: Palocsay, Markham, and Markham (2010) introduce data analysis techniques using Excel, offering practical tools for processing and interpreting NFHS5 data. Their approach thus provides a ground for research on complex relationships embedded within district-wise data sets. [16] Yuchi et al. and Tso and Yau provide some predictive modeling approaches relevant to the analysis of NFHS5 data in indoor air pollution and electricity consumption, respectively. The two studies thus demonstrate the potential of using machine learning techniques in extracting meaningful insights from large-scale data sets, a point in line with Power BI framework objectives. [17][18] In order to combat the issue of growth failure in children in high priority districts, district-level direct public intervention programs that enhance parents' education, the standard of living in their homes, and the health care facilities for children should be given more importance.[19] This study will be valuable for policymakers and health workers in understanding the dynamic nature of prevalence and changes of child growth failure indices across the districts in West Bengal. [19]

These researches enhance our comprehensive comprehension of the various aspects presented in the NFHS5 district-wise data. To do data analysis on Power BI, researchers must possess a comprehensive understanding of various aspects, including women's empowerment, health education, and socio-economic development. Thus, employing machine learning approaches to analyze NFHS5 district-wise data offers a unique opportunity to uncover complex patterns, thereby facilitating evidence-based decision making. By utilizing insights from several fields of study, policymakers will design focused strategies to tackle issues such as women's empowerment, healthcare inequalities, and socio-economic development challenges. India's progress towards equitable growth and sustainable development can be greatly enhanced by implementing data-driven strategies at the grassroots level, utilizing Power BI. This has the potential to bring about significant positive transformation.

CHAPTER 3

Data and Methodology

The data has been sourced from the portal data.gov.in, which is managed and hosted by the National Informatics Centre (NIC) under the Ministry of Electronics & Information Technology, Government of India. The data available on this portal is owned by the respective Ministries, States, Departments, or Organizations and is licensed under the Government Open Data License – India. From this comprehensive dataset, we have extracted key attributes to analyze and understand the condition of women in society. The dataset includes various indicators related to health, education, and infrastructure across different districts.

Here's a brief overview of the data structure:

District Names: Name of the district.

State/UT: State or Union Territory the district belongs to.

Zones: Geographical zone.

Number of Households surveyed: Total households surveyed.

Number of Women age 15-49 years interviewed: Number of women interviewed.

Number of Men age 15-54 years interviewed: Number of men interviewed.

Sex ratio: Sex ratio in the district.

Literacy: Literacy rate.

Schooling: Women (age 15-49) with 10 or more years of schooling

Child marriage: Percentage of child marriages.

Antenatal, Postnatal, Neonatal: Health services indicators.

Average expenditure: Average out-of-pocket expenditure per delivery in a public health facility (for last birth in the 5 years before the survey) (Rs.).

Institutional births: Percentage of births in medical institutions.

Cancer screening (cervical, breast, oral): Cancer screening rates.

Tobacco and alcohol usage: Usage rates.

Nutritional status (Underweight, Overweight, Anaemic): Nutritional health indicators.

Chronic conditions (Diabetic, High BP): Prevalence of chronic conditions.

We are going to visualize the data and find the key factors which affects the literacy rate and analyse the health of women in society using MS EXCEL, Python and Power BI.

Machine learning is a rapidly expanding field that focuses on using algorithms to analyze data, learn from it, and uncover relationships among variables. These relationships can be used for predicting outcomes, classifying information, or identifying patterns within the data. In this study, however, it's important to clarify that our data is cross-sectional, which means we can only test associations rather than establish causality. Therefore, when we refer to "prediction," we are actually examining the correlations between variables. [7]

The use of correlation matrix to find the key attributes for literacy. Using the key features to make machine learning models to predict the literacy rate. We are about to apply linear regression, random forest regressor and decision tree models to use the best of them with minimum error.

A correlation matrix is a table that displays the correlation coefficients for a certain number of variables. The correlation coefficient, in general, works by pairing each variable in the matrix with every other variable. The correlation coefficient is subsequently utilized to determine the magnitude and orientation of the linear relationship between two variables. This scale varies from -1 to 1, where:

1 = perfect positive linear relationship

0 = no linear relationship and

−1 = perfect negative linear relationship.

The correlation matrix is an essential instrument in data analysis and statistics used for understanding the correlations between variables in a particular dataset. They could help in showing the pattern of relationships, dependency, and potential multicollinearity issues with the data.

3.1 PYTHON LIBRARIES

3.1.1 Pandas module

It is an open-source data analysis and manipulation library that is both robust and widely used. Data structures as well as techniques for structured data processing are supplied by it making it an indispensable library for data science and analysis. Large datasets can be efficiently stored and manipulated. Processing tabular data from csv files has been taken care of in a convenient Python library Pandas, which is tailored to the high-level processing of tabular data [21]

3.1.2 NumPy stands for Numerical Python, and that's about what it is—a library for numerical computing in Python. It provides support for complex, multi-dimensional arrays and matrices, along with a wide range of advanced mathematical algorithms for manipulating these arrays. It is arguably the most crucial library for scientific and data-related tasks and acts as the base for most other Python libraries, especially those within the realm of machine learning, data science, and scientific computing. The NumPy package continues the work of the successful Numeric array object.

Its purpose is to create the cornerstone for a useful environment for scientific computing.[20]

3.1.3 Scikit-learn is an open-source library for general use in machine learning using the Python language. It offers tools for data analysis and modeling, making it applicable to both people starting and experienced practitioners. Scikit-learn is built upon NumPy, SciPy, and Matplotlib and extends their capabilities to provide comprehensive tools for a variety of tasks in machine learning. [22]

3.1.4 Matplotlib is an overall package and a very widely used data visualization library in Python. It provides a versatile and scalable framework for creating a wide range of plots and charts, including static, animated, and interactive ones. Matplotlib is commonly employed in disciplines such as machine learning, data science, finance, and scientific study.

3.2 MACHINE LEARNING MODELS

ML approaches are often used to predict values of an outcome Y given a set of input variables X (X = X1, X2, …Xp).[32] Supervised ML techniques solve this problem by first using a dataset for which Y and X are both known to estimate a function f that captures the relationship between Y and X. [32] This function f is then used to predict Y for new or unseen data where X is known, but Y is unknown.[32] While there are both parametric and non-parametric approaches to estimate f, we focus on parametric approaches in this paper.[32] The parametric approach first involves making an assumption about the functional form of f (for example, f is a linear model). [32] The next step involves fitting or training the model using observed data (or training data).[32] Finally, the performance of the model is evaluated on a test dataset.[32]

3.2.1 Linear Regression

Linear regression is a statistical tool employed to understand how one or more factors influence an outcome. It works by fitting a straight line to observed data, aiming to capture the overall trend while minimizing the differences between the predicted values and the actual data points. Its objective is to uncover the most accurate relationship between variables. In simple linear regression, Simple linear regression lives up to its name: it is a very straightforward simple linear regression approach for predicting a quantitative response y on the basis of a single predictor variable x. [24] It assumes that there is approximately a linear relationship between x and y. [24] We can show this linear relationship mathematically as

$$y=mx+b$$

Where:

y = Dependent variable

x = Independent variable.

m = slope of the line.

b = y-intercept.

Simple linear regression is extended to incorporate more than one explanatory variable by multiple linear regression. In both cases, we still use the term 'linear' because we assume that the response variable is directly related to a linear combination of the explanatory variables.[25]

The equation for multiple linear regression is similar to that of simple linear regression but includes additional terms.

$y=b_0+b_1x_1+b_2x_2+...+b_nx_n$

The variable y is dependent. The $x_1,x_2,...,x_n$ variables are considered to be independent, whereas b_0 represents the intercept. The coefficients b_1, $b_2,...,b_n$ describe the effect on the dependent variable y when each independent variable changes by one unit, while keeping all other variables constant.

We use ordinary least squares for estimation of coefficients—the slope and intercept—of the linear equation. Essentially, OLS operates by reducing the sum of the squared discrepancies between the observed and predicted values in the training data. It's like finding the best fit line through a scatterplot of points, trying to get as close as possible to each of the data points, without veering too far off. Linear regression is a commonly employed statistical technique that is utilized across several disciplines to make predictions, anticipate future outcomes, and gain insights into the connections between different variables. It's also a fundamental building block for more complex machine learning algorithms.

3.2.2 Decision Trees:

A decision tree, derived from machine learning theory, is a very effective tool for solving classification and regression issues. Unlike other classification approaches that use a set of features (or bands) jointly to perform classification in a single decision step, the decision tree is based on a multistage or hierarchical decision scheme or a tree like structure.[26] A Decision Tree Regressor is a machine learning method specifically designed for regression tasks. While decision trees are often associated with classification problems, where the goal is to classify data into categories, decision tree regressors are used when the target variable is continuous rather than categorical. [5] Decision trees are employed to facilitate decision making and are extensively utilized for regression problems.[5] Apart from their applicability, decision trees are also transparent, meaning non-experts are easily able to use and understand them.[5]

Decision tree regression is an adaptation of decision tree classification designed to estimate real-valued functions, such as predicting continuous outcomes. The construction of a regression tree involves binary recursive partitioning, an iterative process that splits the data into segments. At first, the entire set of training samples is utilized to establish the framework of the tree.[26] The method proceeds by performing a binary split on the data at every conceivable point, and then chooses the split that divides the data into two parts in such a way that the total of the squared differences from the mean in each half is minimized.[26] The procedure of splitting is subsequently done to each of the newly formed branches. The process continues until each node reaches a user-specified minimum node size. [26]

3.2.3 Random Forest Regressor:

Random Forest is an ensemble of unpruned classification or regression trees created by using bootstrap samples of the training data and random feature selection in tree induction. [28] Prediction is achieved by combining the predictions of the ensemble, either through majority voting or averaging. Initially introduced by Breiman, this approach entails the creation of multiple decision trees during the training phase, amalgamating their individual predictions to generate a collective outcome. One of the most striking advantages of Random Forest lies in its capacity to address the prevalent issue of overfitting, a common pitfall encountered in conventional decision trees. By leveraging the collective wisdom of numerous trees, it effectively diminishes the risk of fixating on idiosyncrasies within the data, thus ensuring mor e robust and reliable predictions. Furthermore, its resilience to noisy data and outliers underscores its adaptability across diverse datasets, rendering it a formidable tool for real-world applications. Beyond its predictive prowess, Random Forest offers valuable insights into feature importance, facilitating informed decision-making in feature selection and data interpretation. Moreover, its scalability and parallelizability make it particularly well-suited for handling voluminous datasets and distributed computing environments, thereby enhancing its appeal in contemporary data-centric endeavors. In essence, Random Forest epitomizes a sophisticated yet accessible approach to machine learning, embodying the culmination of theoretical innovation and practical efficacy in the pursuit of data-driven insights.

3.3 RMSE AND CROSS VALIDATION SCORE

RMSE, or Root Mean Square Error, is a commonly used metric in regression analysis to evaluate the accuracy of a predictive model.[27] The root mean square error (RMSE) has been used as a standard statistical metric to measure model performance in meteorology, air quality, and climate research studies.[27]The metric quantifies the mean discrepancy between the model's predicted values and the actual observed values in the dataset.[27]The root mean square error (RMSE) is calculated by taking the square root of the average of the squared differences between the predicted and actual values. A lower RMSE indicates that the model's predictions are more accurate and closely match the actual data, demonstrating better performance.

The Mean Cross Validation Score is a quantitative metric used to assess the effectiveness of a machine learning model through cross-validation. Cross-validation involves dividing a dataset into multiple subsets or folds to evaluate the model's generalization ability. In k-fold cross-validation, the data is split into k equal-sized folds, where each fold is used once as the validation set while the remaining k-1 folds are used for training. This process is repeated k times, with each fold serving as the validation set once. The Mean Cross Validation Score is computed as the average evaluation score obtained across all iterations of this process. This metric provides insight into how well the model performs across different data subsets on average. The Mean Cross Validation Score is the average of the evaluation scores obtained from each iteration of the cross-validation process. It provides an overall measure of how well the model performs across different subsets of the data. A higher Mean Cross Validation Score indicates better generalization performance of the model, suggesting that it is less likely to overfit to the training data. Cross-validation (CV) is

a widely used technique for evaluating the performance of machine learning models. Among the various types of CV, k-fold cross-validation is particularly popular. [29] However, choosing the value of k randomly can lead to suboptimal performance and increased computational complexity. Studies have shown that the optimal value of k and the resulting model validation performance can vary across different machine learning algorithms.[29] Therefore, it is essential to select the appropriate value of k in k-fold CV to ensure the best possible model performance. [29] However, to the best of our knowledge, a few if any studies have explored in detail with extensive empirical results how values of k (number of subsets) affect validation results off different machine learning algorithms.[29]

3.4 PIVOT TABLES

Pivot tables are a powerful tool for data analysis and visualization in Excel. Pivot Tables enable you to extract meaning from large amount of data. This description is deceptively simple because in fact Pivot tables are powerful and sophisticated tool that enables you to do things that would be impossible or difficult to do any other way. It enables you to take what seems to be indecipherable mass of facts and extract any trend or patterns buried in the data.[30] They allow you to summarize, group, and rearrange large datasets quickly and easily. For example, one can utilize pivot tables to filter data according to various criteria so as to see sales activity for some given period or by product category. Creation of custom calculations: It is enabled in this way. In other words, profit margins for each product class can be calculated using a pivot table or return on investment for every department can be worked out. You are able to make charts and graphs representing your data through it.

3.5 POWER BI

Data Visualization is a process of making understand the significance of data through visual context, and it is a part of analytics, there are several techniques to visualize the data such as Interactive and Dynamic in nature and coming to visual context, there are a number of things such as plots, graphs, slicers, stacked column charts, Histogram, Bar Charts, tables, matrix and other forms of visual contexts; In this paper we focussed on interactive data visualization through Microsoft Power BI tool, Microsoft Power BI is a suite of business intelligence and analytics tool for analyze data and share insights and gets answers quickly with the help of interactive data visualization using dashboard available on every device such as Applications, Desktops, Mobiles…etc.[31]With the help of visuals and filters, the user or person gets convenient and easier to understand the data and it has an architecture of five main components as discussed below and follows Power BI Services, Power BI Gateways, Power BI Desktop, Power BI Apps and Power BI Connectors.[31]

Microsoft has a service called Business Analytics known as Power BI. It is user-friendly with simple interface that gives end-users powerful interactive visualizations and business intelligence capabilities. Here is some more information:

- Data Connectivity: Power BI connects to multiple data sources, for example: Excel Spreadsheets, Databases like SQL Server, MySQL, and so on; cloud services like Azure, Google Analytics, among others.

- Data Preparation: This tool has a built-in feature known as Power Query which helps clean, reshape and combine different data from various sources before analysis. This enables users to prepare their data for analysis and visualization without the need of any other tools.

- Data Modelling: Power BI enables users to link different datasets, set up calculations, and organize data into hierarchies. It uses DAX (Data Analysis Expressions), which is similar to Excel formulas.

- Visualization: Users of Power BI are able to generate multiple interactive visualizations including; charts, graphs, maps & tables. These visualizations can be tailored according to specific needs and merged into interactive reports and dashboards thus becoming part of dynamic system dashboards.

CHAPTER 4
RESULTS AND DISCUSSION

4.1 GENDER DISPARITY:

There are two parameters to understand the gender disparity in the districts.

i) Present Sex ratios of the districts

Average sex ratio of India is 1020.71 per 1000 males

Lowest: 755/1000 Daman, Dadra and Nagar Haveli & Daman and Diu

Highest :1332/1000 Diu, Dadra and Nagar Haveli & Daman and Diu

District Names	State/UT	Average of Sex ratio
Daman	Dadra and Nagar Haveli & Daman and Diu	755.00
Dadra & Nagar Haveli	Dadra and Nagar Haveli & Daman and Diu	817.00
Sonipat	Haryana	844.00
New Delhi	NCT of Delhi	859.00
Bathinda	Punjab	861.00

Table 1: Districts with lowest sex ratio

ii) Sex ratio at birth for children born in the last five years (females per 1,000 males)

Average sex ratio at birth for children born in last five years for India is 944.70

Lowest District: 658/1000 Datia, Madhya Pradesh

Highest District:1485/1000, Alappuzha, Kerala

District Names	State/UT	Average of Sex ratio 5 years
Datia	Madhya Pradesh	658.00
Satna	Madhya Pradesh	658.00
Muzaffarpur	Bihar	685.00
West District	Sikkim	685.00
Kinnaur	Himachal Pradesh	691.00
North District	Sikkim	693.00
Warangal Rural	Telangana	698.00

Table 2: Districts with lowest sex ratio at birth for children born in last 5 years

As per the data given, we have divided the districts into three parts:

Sex ratio greater than 1000 is high

Sex ratio less than 1000 and greater than 900 is considered medium

Sex ratio less than 900 will be taken as low

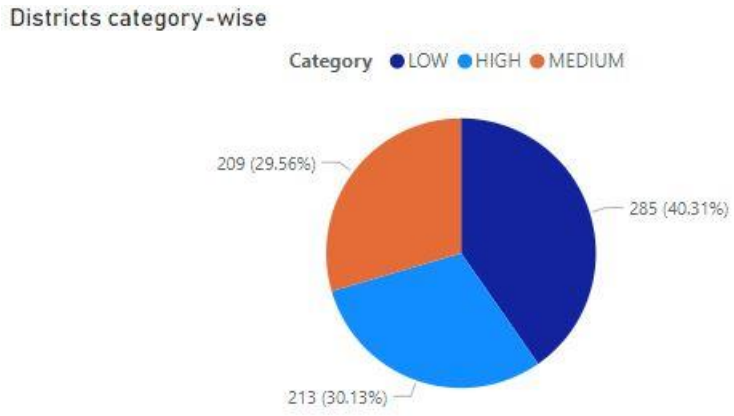

Fig 4.1: Distribution of districts on the basis of sex ratio of children born in last 5 years

40% i.e. 285 districts out of 707 surveyed districts of India have sex ratio less than 900 in the last five years.

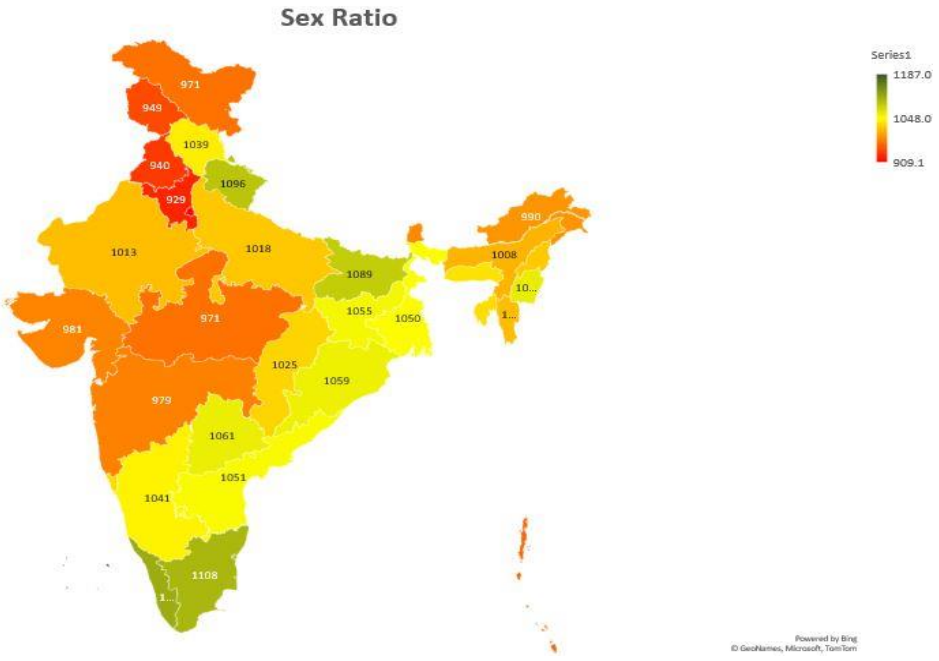

Fig 4.2: Geographical heatmap of sex ratio

Zones	Average of Sex ratio 5 years	Average of Sex ratio
Southern	937.28	1073.86
Eastern	922.43	1066.71
North-Eastern	982.93	1014.12
Northern	929.14	992.78
Central	977.23	989.51
Western	947.58	980.91
Total	**944.70**	**1020.71**

Table 3: Zone-wise Distribution of Sex ratio

As per the table and the map, southern states of India have the highest sex ratio and the western states have lowest sex ratio.

4.2 LITERACY RATE PREDICTION MODELS

4.2.1 Literacy Rate: Average literacy rate is 74.33% with a standard deviation of 12.24, indicating variability across districts.

Highest: 99.30% Ernakulam, Kerala

Lowest: 38.60% Jhabua, Madhya Pradesh

District Names	State/UT	Average of Literacy
Jhabua	Madhya Pradesh	38.60
Sukma	Chhattisgarh	40.50
Shrawasti	Uttar Pradesh	40.80
Nabarangapur	Odisha	41.00
Malkangiri	Odisha	41.30

Table 4: Districts with lowest literacy rate

4.2.2 Schooling: Average citizens who have had attended school for at least 10 years is 40.31 with variability indicated by a standard deviation of 14.22.

Highest: 88.20% Mahe, Puducherry

Lowest: 38.60% Pakur, Jharkhand

State/UT	District Names	Average of Schooling
Jharkhand	Pakur	13.60
Haryana	Mewat	13.90
Tripura	Dhalai	13.90
Odisha	Malkangiri	14.00
Uttar Pradesh	Bahraich	14.40
Bihar	Kishanganj	15.00

Table 5: Districts with lowest schooling rate

This indicates the high drop-out rate from school. There are only 40.31% females who have attended school for more than 10 years. The escalation of dropout rates among girls worsens the gap in education between genders, upholding societal norms that prioritize male education. This sustains systemic gender inequality across multiple domains. Furthermore, offspring of mothers with limited education often encounter educational hindrances, perpetuating a cycle of poverty and constrained prospects across successive generations.

Checking the dependence of literacy rate on the basic amenities like sanitation facility, clean fuel, drinking water, electricity and consumption of iodine salt using corelation matrix. Following are the results for the same.

Sanitation	0.665475
Clean fuel	0.447324
Electricity	0.335629
I salt	0.247692
Drinking Water	-0.029476

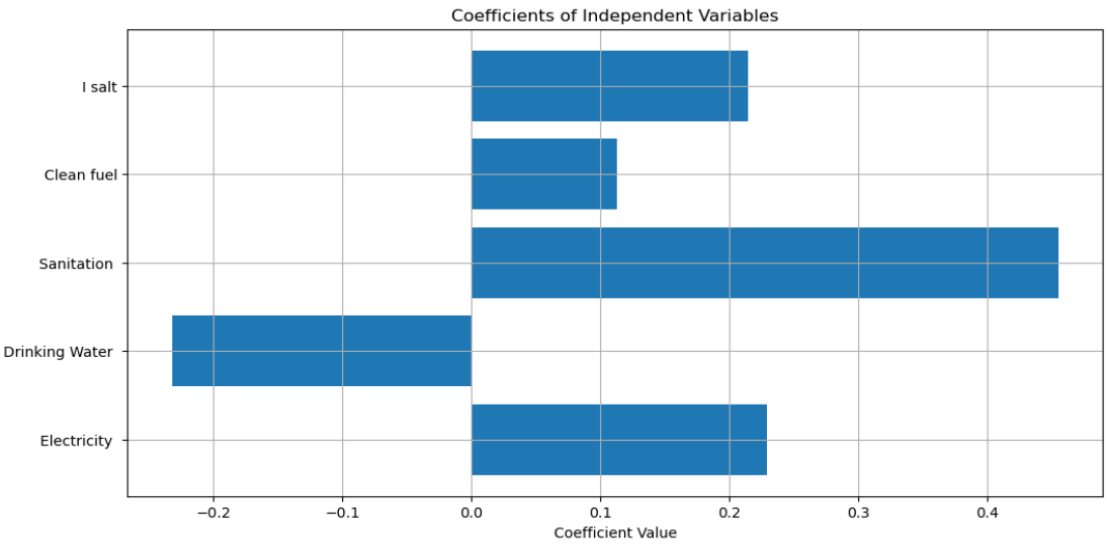

Fig 4.3: Coefficient value of independent variables

Sanitation is most correlated among the chosen attributes.

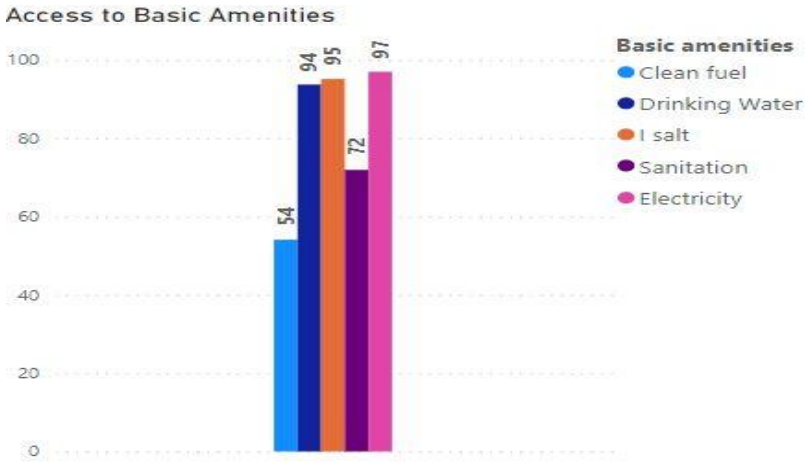

Fig 4.4: Access to basic amenities for women

As per the bar graph,72% people have the access to sanitation facility which is most correlated as per the study. The state must take required actions to make conditions better for the citizens.

As drinking water attribute is in negative. We used feature scaling so that we can have all the attributes in the same range or scale, making them directly comparable and facilitating accurate analysis and model training.

We will split the district-wise data in two parts using sklearn's TrainTestSplit. We will call the 80% data as training data on which we will train the models and we will use 20% data for testing the accuracy of the model.

First, we applied linear regression on the same where the targeted value was literacy rate and also checked root mean squared error of the predicted value and mean square of cross validation number.

4.2.3 Linear Regression

RMSE: 0.14888456456599877

Mean Cross Validation score:0.14505961766588577

The linear regression model achieved a relatively low RMSE score, indicating good predictive performance. The mean cross-validation score is close to the RMSE, suggesting consistent performance across different folds of the dataset.

Similarly, we have applied same for Decision Tree regressor

4.2.4 Decision Tree Regressor

RMSE: 0.18472681389375822

RMSE cross-validation score: 0.2123720124126113

The decision tree regressor yielded a higher RMSE compared to the linear regression model, indicating slightly inferior predictive accuracy. The RMSE cross-validation

score is higher than the RMSE, suggesting potential variability in model performance across different folds.

4.2.5 Random Forest Regressor

RMSE: 0.14654124177982325

RMSE cross-validation score: 0.15730631027245592

The random forest regressor produced a comparable RMSE to the linear regression model, indicating similar predictive accuracy. The RMSE cross-validation score is close to the RMSE, indicating stable performance across different folds with minimal variability.

Overall, both linear regression and random forest regression appear to be suitable for predicting the target variable in this context, with the random forest regressor potentially offering a slight edge in terms of stability and robustness.

4.3 WOMEN HEALTHCARE

4.3.1 Institutional Births: 88.68% of births occur in institutions, with 64.95% in public facilities and 24% in private facilities.

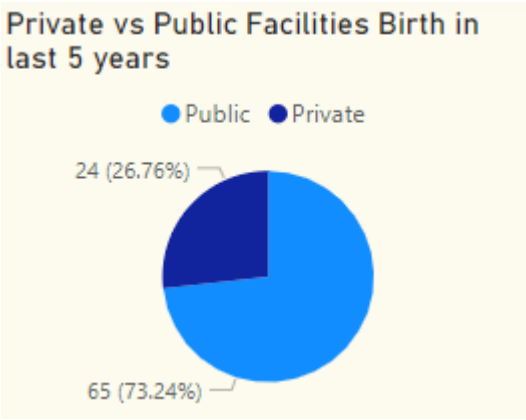

Fig 4.5: Distribution of birth in private and public facilities

4.3.2 Maternal Health:

Since the beginning of the Safe Motherhood Initiative, India has accounted for at least a quarter of maternal deaths reported globally.[33] India's goal is to lower maternal mortality to less than 100 per 100,000 livebirths but that is still far away despite its programmatic efforts and rapid economic progress over the past two decades.[33] Geographical vastness and sociocultural diversity mean that maternal mortality varies across the states, and uniform implementation of health-sector reforms is not possible.[33] Maternal health is one of the essential parts. There are

certain initiatives taken by the state with respect to the maternal care like Mother Child Protection Card (MCP). On an average,95.9% women have been benefitted from the MCP card.

Average out-of-pocket expenditure per delivery in a public health facility (for last birth in the 5 years before the survey) (Rs.) is 3270.

4.3.3 Nutritional Status

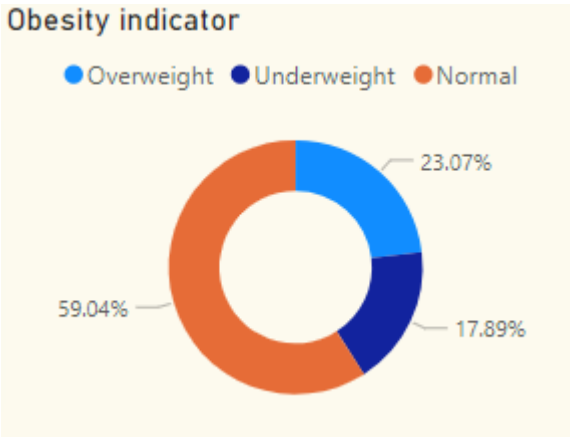

Fig 4.6: Distribution of women based on BMI

Nutritional challenges for women can be understood by diving them into following categories 17.89% underweight, 23.07% overweight, and 55.95% anemic, highlighting nutritional challenges.

Obesity leads to many diseases due to its impact on various physiological processes and systems within the body. The development of chronic ailments like type 2 diabetes, cardiovascular diseases, hypertension, stroke, some forms of cancer, and musculoskeletal disorders like osteoarthritis is directly linked to it. Furthermore, metabolic syndrome, sleep apnea, infertility, and psychological problems like melancholy and low self-esteem can all be brought on by fat. Managing and preventing obesity is crucial for reducing the burden of these associated diseases and improving overall health outcomes.

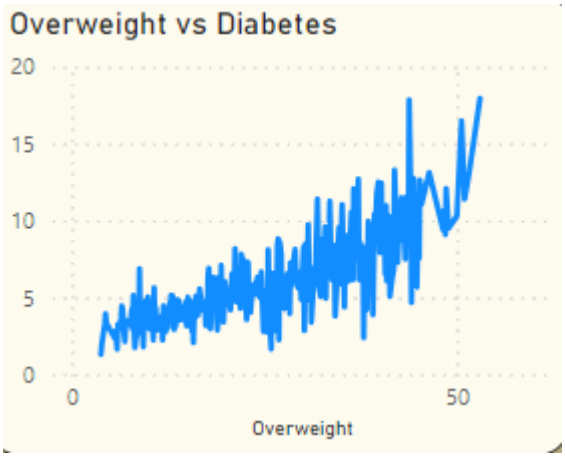

Fig 4.7: Overweight vs diabetes

The graph shows as the obesity will increase so does the diabetes which might further lead to Chronic Conditions (Diabetic, High BP):

5.59% diabetic and 12.46% have high blood pressure.

Diabetes and its complications are considered as a major cause of morbidity and mortality in India. It is one of the unpredictable diseases that may occur due to the lack of insulin in Human Body. Globally, diabetes directly caused 1.5 million deaths. Between 2000 and 2019, there was a 3% increase in age-standardized mortality rates from diabetes. The estimates in 2019 showed that 77 million individuals had diabetes in India, which is expected to rise to over 134 million by 2045.[8]

Highest :53% women of Kanyakumari, Tamil Nadu who are surveyed are found out to be overweight

Lowest: 3% women of Sukma, Chhattisgarh.

Zones	Average of Overweight
Southern	33.50
Northern	25.46
Western	21.50
North-Eastern	19.30
Eastern	16.99
Central	14.75
Total	**23.07**

Table 6: Zone-wise distribution of obese women

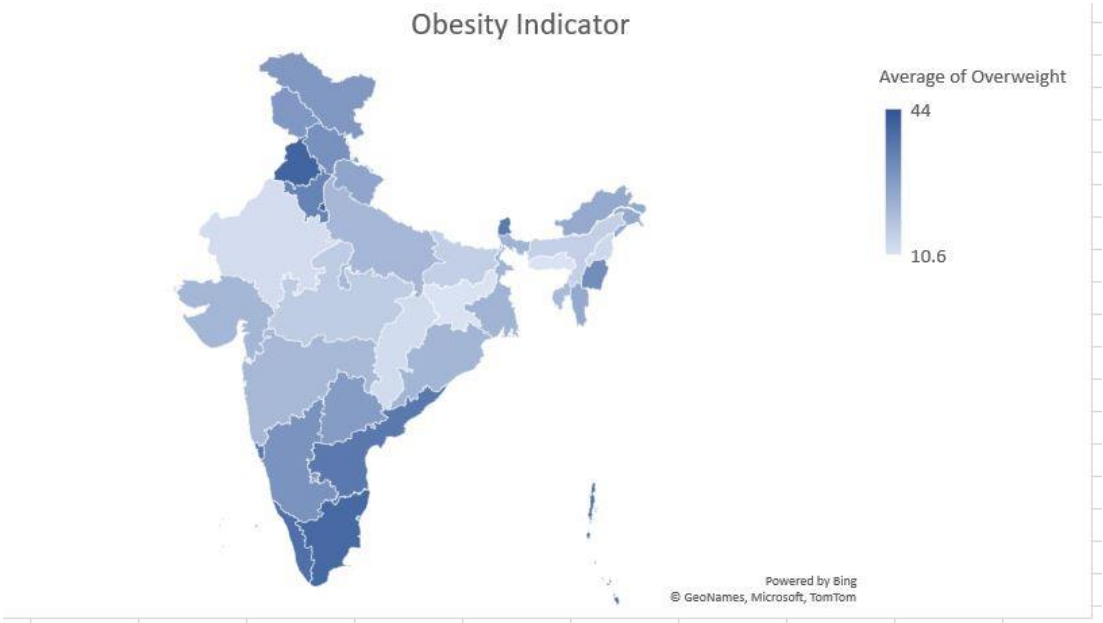

Fig 4.8: Geographical heatmap of India showing the distribution of women who are overweight.

As we can see from the heatmap the southern part of India has more obese women population and in the north Punjab has high obesity rate.as compared to neighboring states. Addressing obesity in India on a large scale requires a comprehensive and multi-faceted strategy involving the government, healthcare providers, communities, and individuals.

CHAPTER 5

CONCLUSIONS

The research conducted delves into the multifaceted aspects of women's significance in society, elucidating their pivotal roles across economic, educational, healthcare, social, and political domains. Empowering women emerges as a critical pathway towards inclusive development and sustainable progress. The research used data from the National Family Health Survey (NFHS) to analyze various indicators related to women's health, education, and socio-economic status across different districts in India.

Important discoveries highlight the urgent need for focused measures to address gender inequality, particularly concerning literacy rates, access to basic facilities, and health outcomes. The research also indicates that in particular as it has been shown by study results about 40.31% of the female population have spent more than ten years in school, indicating a need for increased educational opportunities for women. On the other hand, this can be understood from the increasing dropout rate among girls that 285 districts had sex ratio less than 900 while there are institutional challenges which perpetuate gender inequity that have to be overcome through concerted efforts towards education and gender balance promotion. Southern India has better women literacy rate which leads to more participation of women in the development of India.

Additionally, relations between literacy levels and factors such as sanitation facilities expose the interplay of socio-economic variables with respect to learning outcomes. It is evident from data analysis that higher literacy rates tend to correspond to better access to critical infrastructure like household toilets, clean cooking fuel and electricity this implies that infrastructural development has a role in promoting education attainment.

Moreover, what this means is that a comprehensive health care system must be implemented focusing on women’s needs using key health indicators including child marriage rates, institutional births and nutritional status. The findings from our data analysis shed light on the strong link between obesity and chronic illnesses, Highlighting the immediate necessity for proactive healthcare efforts to mitigate the effects of non-communicable disorders. Public health initiatives should aim to raise awareness about the benefits of balanced diets and regular physical activity. Educational institutions and workplaces can encourage healthier lifestyles by incorporating nutritional education and physical exercise into their programs. Policy measures such as regulating junk food advertisements and taxing sugary drinks can help curb unhealthy eating habits. Additionally, enhancing access to affordable, nutritious foods through subsidies and local farming initiatives is essential. Effective reduction of obesity rates across the country hinges on the collaboration of all stakeholders and a sustained commitment to health promotion. In essence, this research highlights how crucial it is for women to be at the forefront of societal advancement. It calls for comprehensive strategies to tackle the various obstacles hindering women's empowerment and overall welfare. By harnessing insights derived from data and enacting focused interventions, policymakers and stakeholders

have the opportunity to promote inclusive progress and propel us closer to gender parity. This approach lays the groundwork for a fairer and more prosperous society for all.

BIBLIOGRAPHY

[1] Saha, S., & Narayanan, S. (2022). A simplified measure of nutritional empowerment: Using machine learning to abbreviate the Women's Empowerment in Nutrition Index (WENI). World Development, 154, 105860.

[2] Pande, R., Malhotra, A., & Grown, C. (2005, July). Impact of investments in female education on gender equality. In XXV IUSSP International Population Conference, Tours, France.

[3] Hassan, Z., & Silong, A. D. (2008). Women leadership and community development. European Journal of Scientific Research, 23(3), 361-372.

[4] Dhirar, N., Dudeja, S., Khandekar, J., & Bachani, D. (2018). Childhood morbidity and mortality in India—analysis of national family health survey 4 (NFHS-4) findings. Indian pediatrics, 55, 335-338.

[5] Nijeboer, M. Enhancing Gender Equality in Machine Learning: Optimal Discrimination-Aware Decision Trees using Integer Optimization.

[6] Paxton, P., & Kunovich, S. (2003). Women's political representation: The importance of ideology. Social forces, 82(1), 87-113.

[7] Raj, A., Dehingia, N., Singh, A., McDougal, L., & McAuley, J. (2020). Application of machine learning to understand child marriage in India. SSM-population health, 12, 100687.

[8]Jha, U. M. (2024). A Comparison of the Classification Models of Machine Learning for Predicting the Risks of Diabetes: An Analysis of National Family Health Survey data of Delhi. International Journal of Research and Analytical Reviews (IJRAR), 11(2).

".[9] Kadam, R. N. (2012). Empowerment of Women in India-An Attempt to Fill the Gender Gap (June, 2012). International Journal of Scientific and research publications, 2(6)

[10] Becker, G. S., Hubbard, W. H., & Murphy, K. M. (2010). Explaining the worldwide boom in higher education of women. Journal of Human Capital, 4(3), 203-241.

[11] Antony, G. M., & Laxmaiah, A. (2008). Human development, poverty, health & nutrition situation in India. Indian Journal of Medical Research, 128(2), 198-205.

[12] Garg, C., Khan, S. A., Ansari, S. H., & Garg, M. (2010). Prevalence of obesity in Indian women. Obesity reviews, 11(2), 105-108.

[13] Kumar, D., & Aggarwal, S. (2019, February). Analysis of women safety in indian cities using machine learning on tweets. In 2019 Amity International Conference on Artificial Intelligence (AICAI) (pp. 159-162). IEEE.

[14] Braveman, P., Egerter, S., & Williams, D. R. (2011). The social determinants of health: coming of age. Annual review of public health, 32, 381-398.

[15] Short, S. E., & Zacher, M. (2022). Women's Health: Population Patterns and Social Determinants. Annual Review of Sociology, 48, 277-298

[16] Palocsay, S. W., Markham, I. S., & Markham, S. E. (2010). Utilizing and teaching data tools in Excel for exploratory analysis. Journal of Business Research, 63(2), 191-206.

[17] Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., ... & Allen, R. W. (2019). Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. Environmental pollution, 245, 746-753.

[18] Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 32(9), 1761-1768.

[19] Das, P., Roy, R., Das, T., & Roy, T. B. (2021). Prevalence and change detection of child growth failure phenomena among under-5 children: a comparative scrutiny from NFHS-4 and NFHS-5 in West Bengal, India. Clinical Epidemiology and Global Health, 12, 100857.

[20] Oliphant, T. E. (2006). Guide to numpy (Vol. 1, p. 85). USA: Trelgol Publishing.

[21] Lemenkova, P. (2019). Processing oceanographic data by Python libraries NumPy, SciPy and Pandas. Aquatic Research, 2(2), 73-91.

[22] Kramer, O., & Kramer, O. (2016). Scikit-learn. Machine learning for evolution strategies, 45-53.

[23] International Institute for Population Sciences (IIPS) and ICF. 2021. National Family Health Survey (NFHS-5), 2019-21: India. Mumbai: IIPS.

[24] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear regression. In An introduction to statistical learning: With applications in python (pp. 69-134). Cham: Springer International Publishing.

[25] Tranmer, M., & Elliot, M. (2008). Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR), 5(5), 1-5.

[26] Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. (2005). Decision tree regression for soft classification of remote sensing data. Remote Sensing of Environment, 97(3), 322-336.

[27] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). Geoscientific model development discussions, 7(1), 1525-1534.

[28] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences, 43(6), 1947-1958.

[29] Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of machine learning algorithms with different K values in K-fold cross-validation. International Journal of Information Technology and Computer Science, 13(6), 61-71.

[30] Aitken, P. G. (2006). Excel pivot tables and charts. John Wiley & Sons.

[31] Bhargava, M. G., Phani, K. T., Kiran, S., & Rao, D. R. (2018). Analysis and design of visualization of educational institution database using power bi tool. Global Journal of Computer Science and Technology, 18(C4), 1-8.

[32] Dey, A. K., Dehingia, N., Bhan, N., Thomas, E. E., McDougal, L., Averbach, S., ... & Raj, A. (2022). Using machine learning to understand determinants of IUD use in India: Analyses of the National Family Health Surveys (NFHS-4). SSM-Population Health, 19, 101234.

[33] Vora, K. S., Mavalankar, D. V., Ramani, K. V., Upadhyaya, M., Sharma, B., Iyengar, S., ... & Iyengar, K. (2009). Maternal health situation in India: a case study. Journal of health, population, and nutrition, 27(2), 184.

PAPER NAME
thesis-2.pdf

AUTHOR
. .

WORD COUNT
8440 Words

CHARACTER COUNT
47560 Characters

PAGE COUNT
35 Pages

FILE SIZE
1.1MB

SUBMISSION DATE
May 29, 2024 12:56 PM GMT+5:30

REPORT DATE
May 29, 2024 12:56 PM GMT+5:30

12% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 10% Internet database
- 5% Publications database
- Crossref database
- Crossref Posted Content database
- 11% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 10 words)