

Cancer Incidence Prediction

A Project / Dissertation as a Course requirement for
Master of Science in Data Science and Computing

Sairaj Patro

23909



SRI SATHYA SAI INSTITUTE OF HIGHER LEARNING
(Deemed to be University)

Department of Mathematics and Computer Science
Muddenahalli Campus

March 2024



SRI SATHYA SAI INSTITUTE OF HIGHER LEARNING

(Deemed to be University)

Dept. of Mathematics & Computer Science
Muddenahalli Campus

CERTIFICATE

This is to certify that this Project / Dissertation titled **Cancer Incidence Prediction** submitted by **Sairaj Patro**, 23909, Department of Mathematics and Computer Science, Muddenahalli Campus is a bonafide record of the original work done under my/our supervision as a Course requirement for the Degree of Master of Science in Data Science and Computing.

.....
Dr. K Vengata Krishnan

.....
Sri Sathya Sai Mudigonda
Project / Dissertation Supervisor

Countersigned by

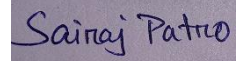
.....
Dr. (Ms.) Y Lakshmi Naidu

Head of the Department

Place: Muddenahalli
Date: 25th March 2024

DECLARATION

The Project / Dissertation titled **Cancer Incidence Prediction** was carried out by me under the supervision of Dr. K Vengata Krishnan Department of Mathematics and Computer Science, Muddenahalli Campus as a Course requirement for the Degree of Master of Science in Data Science and Computing and has not formed the basis for the award of any degree, diploma or any other such title by this or any other University.



Place : Muddenahalli

Date : 25th March 2024

.....
Sairaj Patro
23909
1st MSc
Muddenahalli Campus

Acknowledgements

First and foremost, I would like to express my most reverential gratitude to Bhagawan Sri Sathya Sai Baba for guiding me directly and indirectly via my guide. Bhagawan has been the driving force throughout my life. I dedicate this project at his Divine Lotus Feet.

I would like to thank my advisor, Sri Sathya Sai Mudigonda (SSSIHL, Prashanti Nilayam) for providing me with his guidance and motivation throughout the completion of this project. He has always been an inspiration in terms of work ethic, dedication and discipline. Always providing expert comments, analysis and support whenever necessary.

Also I would like to extend my gratitude to Dr. K Vengata Krishnan for Providing me external support in this due course of time.

This project wouldn't have been possible without Mother and Father's unwavering support. They constantly believed in me, even when I doubted myself. I'm so grateful for their encouragement.

Abstract

Cancer remains a significant public health concern worldwide. Accurately predicting future cancer incidence rates is crucial for effective resource allocation, informing public health strategies, and potentially guiding preventative measures. This study investigates the development of a model for predicting cancer incidence rates. We explore various factors potentially influencing cancer incidence, including demographics, environmental exposures, and lifestyle choices. The chosen modeling approach will be described, along with the data sources and pre-processing techniques employed.

The ultimate goal is to create a reliable and robust model capable of forecasting future cancer incidence rates. The predicted rates will be compared to observed data to assess the model's accuracy. This project contributes to the ongoing effort to combat cancer by providing valuable insights for healthcare professionals and policymakers. By anticipating future trends in cancer incidence, we can better prepare to address this critical global challenge.

The successful development of this model could offer several advantages. Firstly, it could identify populations at higher risk of developing specific cancers, allowing for targeted preventative measures and early detection efforts.

This could be achieved by analyzing the model's predictions alongside demographic and lifestyle data. Secondly, by predicting future healthcare needs related to cancer incidence, resource allocation within healthcare systems could be optimized. The model's forecasts could inform decisions on staffing, equipment, and facility requirements. Finally, the model's insights could inform public health campaigns and promote healthy lifestyle choices, potentially leading to a long-term reduction in cancer burden.

Content

Acknowledgement	4
Abstract	5
Content	6
1) Introduction	
a) Motivation	7
b) Objective	8
c) Domain Knowledge	9
2) Literature Review	
a) Traditional Statistical Methods	10
b) Methods	10
c) Planning/Monitoring	10
d) Machine Learning Approaches	10
e) Novel Data Sources	10
f) Results	11
g) Challenges and Limitations	11
h) Conclusion	11
3) Terminologies	12
4) Machine Learning	
a) Introduction	13
b) Models	14
c) Models Used and Accuracy	15
5) Visualization of Result	18
6) Conclusion	19
7) References	20

Introduction

Motivation

Cancer remains a leading cause of death globally, with a significant burden felt in India. The rising incidence of cancer cases poses a major challenge to public health systems and individual well-being. Predicting future cancer incidence rates in India holds immense value for several reasons.

Firstly, accurate forecasts can inform resource allocation within healthcare systems. By anticipating the future demand for cancer treatment services, healthcare providers can optimize staffing levels, invest in necessary equipment, and ensure adequate infrastructure to handle the growing patient population. This proactive approach leads to improved patient outcomes and reduces strain on the healthcare system.

Secondly, predicting cancer incidence rates enables the development of targeted preventive measures. By identifying populations at higher risk for specific cancers based on demographics, lifestyle choices, and environmental exposures, public health initiatives can be tailored to address these risk factors. This could involve promoting healthy habits, encouraging early detection screenings, and implementing targeted vaccination programs where applicable.

Finally, reliable predictions of cancer incidence rates can guide the development of effective healthcare policies. Policymakers can utilize these forecasts to prioritize research funding, support preventative healthcare initiatives, and expand access to cancer treatment across diverse regions of India. This data-driven approach fosters a more efficient and equitable healthcare system prepared to address the growing cancer burden.

Objective

This study aims to develop a robust model for predicting future cancer incidence rates in India. The specific objectives are:

Identify Key Factors: To comprehensively understand the factors influencing cancer incidence in India, we will explore demographic data (age, sex, geographical location), environmental exposures (pollution levels, occupational hazards), and lifestyle choices (diet, tobacco use, physical activity).

Model Development and Selection: We will evaluate various modeling techniques, such as time series analysis, machine learning algorithms, or statistical regression models. The chosen model will be based on its ability to accurately capture historical trends and predict future cancer incidence rates effectively.

Model Validation and Evaluation: The developed model will be rigorously tested using existing cancer registry data. The predicted rates will be compared to observed data to assess the model's accuracy and reliability.

Identify High-Risk Populations: By analyzing the model's predictions alongside demographic and lifestyle data, we aim to identify populations with an increased risk of developing specific cancers. This information will be crucial for designing targeted prevention strategies.

Domain Knowledge

Understanding the existing body of knowledge on cancer epidemiology in India is crucial for developing effective strategies for prevention, early detection, and treatment. Leveraging data from the National Cancer Registry Programme (NCRP) is foundational to gaining insights into historical cancer incidence rates across various regions and population groups in India. The NCRP, established by the Indian Council of Medical Research (ICMR), collates data from Population Based Cancer Registries (PBCRs) spread across the country. These registries systematically collect information on cancer cases, including demographic details, tumor characteristics, and treatment outcomes. By analyzing NCRP data, researchers can identify trends, patterns, and disparities in cancer incidence, mortality, and survival rates over time.

In addition to NCRP data, recent research has focused on understanding cancer risk factors specific to the Indian context. Tobacco use, both smoking and smokeless forms, remains one of the leading causes of cancer in India, with a disproportionately high prevalence among men and women. Environmental factors, such as air and water pollution, industrial toxins, and exposure to carcinogens in the workplace, also contribute to the burden of cancer in certain regions. Dietary habits, including consumption of processed foods, low intake of fruits and vegetables, and high consumption of red and processed meats, have been linked to an increased risk of certain cancers.

Furthermore, socioeconomic factors play a significant role in shaping cancer incidence and outcomes in India. Socioeconomic disparities in access to healthcare, education, employment opportunities, and living conditions contribute to differential exposure to risk factors and access to

cancer screening, diagnosis, and treatment services. Studies have highlighted regional variations in cancer incidence rates, with higher rates observed in urban areas compared to rural regions. However, rural populations often face challenges in accessing quality cancer care due to limited healthcare infrastructure and financial constraints.

Genetic predisposition also influences cancer risk among Indian populations, with certain genetic mutations and familial cancer syndromes being more prevalent. Understanding the interplay between genetic susceptibility, environmental exposures, and lifestyle factors is essential for identifying high-risk populations and implementing targeted prevention and early detection strategies.

Additionally, the impact of public health policies, cancer control programs, and initiatives aimed at raising awareness, promoting healthy behaviors, and improving access to cancer care should be evaluated. Efforts to strengthen cancer surveillance, expand screening programs, enhance treatment facilities, and train healthcare professionals are critical for reducing the burden of cancer and improving outcomes for patients across India.

By incorporating insights from research on cancer epidemiology, risk factors, socioeconomic determinants, and healthcare infrastructure, policymakers, healthcare providers, and public health experts can develop evidence-based strategies to address the growing challenge of cancer in India effectively.

Literature Review

Cancer remains a significant global health burden, and accurately predicting future incidence rates is crucial for resource allocation, healthcare planning, and potentially early intervention strategies. This review explores the current landscape of research on cancer incidence rate prediction.

1)Traditional Statistical Methods:

Age-Period-Cohort (APC) models are widely used for cancer prediction. These models account for variations in incidence due to age, historical period (secular trends), and birth cohort effects. Additionally, time series analysis and regression techniques incorporating established risk factors (e.g., smoking) have shown promise.

2)Methods:

The National Cancer Registry Programme Report 2020, reported the cancer incidence from 28 Population-Based Cancer Registries (PBCRs) for the years 2012-2016. This was used as the basis to calculate cancer estimates in India. Information pertaining to the population at risk was extracted from the Census of India (2001 and 2011) for the estimation of age–sex stratified population. PBCRs were categorised into the respective State and regions of the country to understand the epidemiology of cancer. The age-specific incidence rate for each specific anatomical site of cancer was applied to the estimated population to derive the number of cancer cases in India for 2022.

3)Planning/Monitoring:

Accurate and up-to-date cancer data is crucial for planning, monitoring, and improving cancer control efforts in any region. Population-Based Cancer Registries (PBCRs) are the primary tool for collecting this data in India. Trained personnel must manually collect information from various sources (hospitals, government offices, labs) using standardized forms. This retrospective approach creates a time lag between when cancer occurs and when the data becomes available. Similar delays are observed globally, with a typical gap of 2-4 years between data collection and publication (e.g., US cancer registries, GLOBOCAN).

To address this delay and ensure timely information for cancer control programs, this proposal suggests using recently collected data to generate more frequent cancer incidence estimates. This approach would provide a clearer picture of current trends and allow for more effective cancer control measures.

4)Machine Learning Approaches:

The rise of big data and advancements in machine learning offer exciting possibilities for cancer incidence prediction. Studies have explored the use of artificial neural networks and support vector machines to analyze large datasets encompassing demographics, environmental factors, and genetic information. These models can potentially capture complex relationships not readily identifiable with traditional methods.

5)Novel Data Sources:

Beyond traditional data sources like cancer registries, researchers are exploring novel approaches. A recent study utilized Google Trends data as a proxy for public interest in cancer, demonstrating potential for early detection of rising incidence trends. Integrating such data sources with traditional methods could enhance prediction accuracy.

According to a 2020 report by the National Cancer Registry Programme (NCRP), data on cancer incidence in India was collected from 28 Population-Based Cancer Registries (PBCRs) between 2012 and 2016. The report details cancer cases categorized by sex and age groups (0-4, 5-9, etc. up to 75+ years). Population data for risk assessment was obtained from the Census of India (2001 & 2011) categorized by state/union territory and sex.

6) Results:

A study in India estimates nearly 1.5 million new cancer cases (14,61,427) for 2022, with a rate of 100.4 per 100,000 people. The research suggests that approximately one in nine individuals in India will develop cancer during their lifetime. Lung cancer was the most prevalent type diagnosed in men, while breast cancer topped the list for women. Among children (aged 0-14), lymphoid leukemia was the most common cancer, affecting boys slightly more (29.2%) than girls (24.2%). The study also predicts a concerning rise of 12.8% in cancer cases by 2025 compared to 2020.

7) Challenges and Limitations:

Despite the progress, challenges remain. The accuracy of predictions can be limited by data quality, the inherent complexity of cancer biology, and the emergence of new risk factors. Additionally, ethical considerations regarding data privacy and potential misuse of predictions need careful consideration.

Future research should focus on:

Incorporating novel data sources (e.g., environmental, genetic) for more comprehensive models. Developing interpretable machine learning models to understand the factors driving predicted trends.

Refining existing methods to improve prediction accuracy and account for regional and demographic variations.

8) Conclusion:

Predicting cancer incidence rates is a crucial step in the fight against the disease. By integrating traditional statistical methods with advanced machine learning and exploring novel data sources, researchers can develop increasingly accurate prediction models. Addressing the challenges and limitations is vital to translate these advancements into improved healthcare strategies.

The cancer incidence is continuing to increase in India. The new estimates will be helpful in planning cancer prevention and control activities through the intervention of early detection, risk reduction and management.

Terminologies

- **Carcinogenesis:** The process by which normal cells transform into cancer cells.
- **Benign:** A benign tumor is a non-cancerous mass of abnormal cells. These cells grow at a slower rate than cancerous cells and typically stay localized in their original location.
- **Malignant:** A malignant tumor is cancerous. These cells grow and divide rapidly, have the ability to invade nearby tissues, and can spread to other parts of the body through the bloodstream or lymphatic system.
- **Tumor:** An abnormal mass of tissue resulting from uncontrolled cell division; can be benign or malignant.
- **Oncogene:** A gene that has the potential to cause cancer.
- **Metastasis:** The spread of cancer cells from their original location to other parts of the body.
- **Biopsy:** The removal of a sample of tissue for examination under a microscope to determine if cancer or other abnormal cells are present.
- **Chemotherapy:** The use of drugs to kill cancer cells or stop them from growing.
- **Radiation therapy:** The use of high-energy radiation to kill cancer cells.

- **Incidence:** The number of new cases of a disease (cancer) in a population over a specific period (usually per year).
- **Incidence Rate:** Incidence expressed as a rate per unit of population (e.g., per 100,000 people). It provides a more informative measure than just the number of cases.
- **Prediction:** An estimate of the future occurrence of an event (cancer incidence) based on past data and trends.
- **Risk Factors:** Factors that increase the likelihood of developing cancer (e.g., smoking, genetics).
- **Age-Standardized Rate (ASR):** Incidence rate that adjusts for the age distribution of the population, allowing for fairer comparisons across different populations.
- **Crude Rate:** The crude incidence rate (CR) provides an overview of cancer burden within a population. It's calculated by dividing the total number of new cancer cases diagnosed in a specific period by the total population size, typically expressed per 100,000 people.
- **Survival Rate:** The percentage of people who live for a specific period (e.g., 5 years) after being diagnosed with cancer.
- **Mortality rate:** It refers to the number of deaths (in general or due to a specific cause) in a particular population, scaled to the size of that population, per unit of time. It's a key metric used to understand the burden of disease and track population health trends.
- **Cancer Prevention:** Strategies to reduce the risk of developing cancer.

Machine Learning

1)Introduction:

Traditionally, statistical models and historical data formed the backbone of cancer incidence prediction. However, the explosion of big data and the concurrent advancements in machine learning are ushering in a new era for this field. Let's explore how machine learning is revolutionizing cancer incidence prediction:

Untangling Complexity: Cancer is a multifaceted disease influenced by a multitude of factors, making it challenging for traditional models to capture all the nuances. Machine learning algorithms, with their prowess in identifying hidden patterns within vast datasets, have the potential to overcome this hurdle and shed light on these intricate relationships.

Leveraging Diverse Data Sources: Machine learning models can effectively integrate a broader spectrum of data sources beyond the confines of traditional cancer registry information. This opens doors to exploring the influence of:

Genetics: Identifying genetic mutations that correlate with an increased risk of developing cancer.
Environmental Factors: Unveiling potential links between pollution levels and specific types of cancer.

Socioeconomic Factors: Understanding how socioeconomic status can contribute to cancer incidence.

Enhanced Accuracy and Generalizability: By analyzing large and diverse datasets, machine learning models hold the promise of generating predictions with greater accuracy compared to traditional methods. Furthermore, training these models on data from various populations can lead to more generalizable insights applicable across broader contexts.

Unearthing Novel Risk Factors: Machine learning's proficiency in identifying complex patterns within data may lead to the discovery of new risk factors for cancer development. This knowledge can be instrumental in refining early detection and prevention strategies.

Personalized Predictions: The future might witness the use of machine learning to pave the way for personalized cancer risk assessments. By incorporating individual genetic and lifestyle data, models could predict a person's specific susceptibility to different cancers, allowing for tailored preventive measures.

Challenges and Considerations:

Despite its immense potential, machine learning for cancer incidence prediction faces certain challenges:

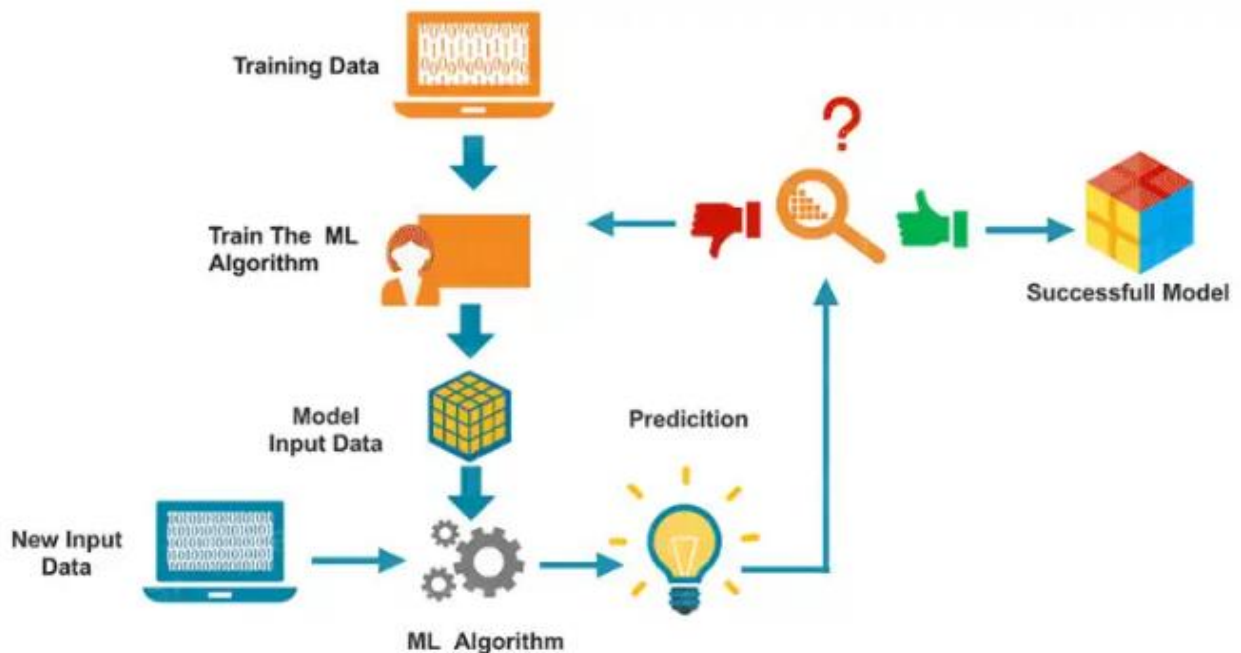
Data Quality: The accuracy of the predictions heavily relies on the quality and completeness of the data used to train the models.

Interpretability: The inner workings of some machine learning models can be intricate, hindering transparency and trust in the predictions.

Ethical Concerns: Data privacy and potential biases within algorithms necessitate careful consideration to ensure responsible and ethical application.

2)Models:

Machine learning algorithms excel at learning from data, enabling them to identify patterns and make predictions without explicit programming. This makes them valuable tools in various fields, including cancer incidence prediction. Here's an overview of common models used in this context:



Supervised Learning Techniques:

These models train on data where each point has a corresponding outcome or label. In cancer prediction, the label might indicate whether someone developed cancer.

Logistic Regression: A foundational algorithm for situations with two possible outcomes (cancer vs. no cancer). It estimates the likelihood of an event (cancer) occurring based on its characteristics (age, genetics, etc.).

Decision Tree Learning: These branching structures categorize data points based on a series of sequential questions about their features. Each question leads to a specific branch, ultimately reaching a final node representing the predicted outcome (cancer risk).

Random Forest Approach: This method combines multiple decision trees, leading to improved accuracy and robustness compared to a single tree. Each tree "votes" on the outcome, with the final prediction based on the majority vote.

Support Vector Machines (SVMs): These models create a dividing line (hyperplane) in a high-dimensional space to separate data points belonging to different categories (cancer vs. no cancer). They are well-suited for high-dimensional data with clear distinctions between classes.

Unsupervised Learning Techniques:

These models identify hidden patterns within unlabeled data, where data points lack predefined outcomes. They can be useful for data exploration and feature extraction in cancer incidence prediction.

K-Means Clustering: This algorithm groups data points into a predetermined number of clusters (k) based on their similarities. It can help identify subgroups within the data that might have different cancer risk factors.

Principal Component Analysis (PCA): This technique reduces the complexity of high-dimensional datasets by identifying the most significant features that capture most of the data's variation. This can be helpful for preparing data for supervised learning models.

Selecting the Optimal Model:

The choice of the most suitable machine learning model for cancer incidence prediction hinges on various factors, including the type of data available, the desired result (classification or prediction), and the intricacy of the relationships between features. Additionally, combining supervised and unsupervised learning approaches or employing ensemble methods like Random Forests can be powerful strategies.

By understanding these different machine learning models, you gain valuable insights into how researchers leverage them to improve cancer incidence prediction, ultimately leading to the development of more effective preventive and control measures.

3)Models Used: (Here for prediction)

- Linear Regression:

Here I used this Linear Regression model for the prediction of the number of cancer cases for the later years. As here I had collected the data of 2018-2022 (state wise) the number of cases of cancer. And then I had to predict the number of cancer cases for the immediate next year, after 5 years and after 20 years.

At first I choose the data of 2018 to 2021 as my independent Features and the data of 2022 as the dependent feature, then I predicted for 2023 then through iterative prediction I predicted the number of cases after 5 year (i.e, 2028)and similarly for after 20 years(i.e for 2043).After the division of the model into training and test set, we train the data and tested on the test data as follows:

```
# Train the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)
```

Here in the above code snippet, it is just taking the data and train itself and try to find the trend available in the data and then we will test the training of the model in the test data that we divided at the beginning of the process as follows:

```
X_2023 = dataset[['2019', '2020', '2021', '2022']].values
predictions_2023 = model.predict(X_2023)
```

In the above code snippet, we are predicting the values for the year of 2023 from the last four year data that is available with us.

The prediction for all the states for the year of 2023,2028 and 2043 looks like from given input of data of year 2018-2022 is:

	State	2018	2019	2020	2021	2022	2023	2028	2043
0	Jammu & Kashmir	12073	12396	12726	13060	13395	NaN	NaN	NaN
1	Ladakh	271	279	286	294	302	NaN	NaN	NaN
2	Himachal Pradesh	8412	8589	8799	8978	9164	NaN	NaN	NaN
3	Punjab	36888	37744	38636	39521	40435	NaN	NaN	NaN
4	Chandigarh	966	994	1024	1053	1088	NaN	NaN	NaN

The above is the input we provided and the below one is the output we received.

```

Predicted number of cases for 2023, 2028, and 2043:
Jammu & Kashmir: 2023 - 13741, 2028 - 15566, 2043 - 22091
Ladakh: 2023 - 310, 2028 - 350, 2043 - 470
Himachal Pradesh: 2023 - 9357, 2028 - 10340, 2043 - 13545
Punjab: 2023 - 41358, 2028 - 46188, 2043 - 62632
Chandigarh: 2023 - 1121, 2028 - 1306, 2043 - 2011
Uttaranchal: 2023 - 12361, 2028 - 13921, 2043 - 19419
Haryana: 2023 - 31686, 2028 - 36173, 2043 - 52611
Delhi: 2023 - 27544, 2028 - 31860, 2043 - 48210
Rajasthan: 2023 - 76653, 2028 - 86878, 2043 - 123594
Uttar Pradesh: 2023 - 215888, 2028 - 241774, 2043 - 330892
Bihar: 2023 - 112137, 2028 - 127383, 2043 - 182487
Sikkim: 2023 - 516, 2028 - 656, 2043 - 1243
Arunachal Pradesh: 2023 - 1113, 2028 - 1247, 2043 - 1723
Nagaland: 2023 - 1899, 2028 - 2140, 2043 - 3010
Manipur: 2023 - 2195, 2028 - 2720, 2043 - 4918
Mizoram: 2023 - 2059, 2028 - 2462, 2043 - 4087
Tripura: 2023 - 2785, 2028 - 3192, 2043 - 4692
Meghalaya: 2023 - 3099, 2028 - 3505, 2043 - 4949
Assam: 2023 - 40769, 2028 - 45926, 2043 - 64080
West Bengal: 2023 - 116254, 2028 - 130226, 2043 - 178412
Jharkhand: 2023 - 36841, 2028 - 42054, 2043 - 61161
Orissa: 2023 - 54114, 2028 - 60110, 2043 - 80045
Chhattisgarh: 2023 - 29991, 2028 - 33882, 2043 - 47714
Madhya Pradesh: 2023 - 83972, 2028 - 94911, 2043 - 133857
Gujarat: 2023 - 75305, 2028 - 85489, 2043 - 122237
Daman: 2023 - 162, 2028 - 239, 2043 - 599
Dadra & Nagar Haveli: 2023 - 256, 2028 - 361, 2043 - 833
Maharashtra: 2023 - 124593, 2028 - 139652, 2043 - 191705
Telangana: 2023 - 51196, 2028 - 57605, 2043 - 80079
Andhra Pradesh: 2023 - 75128, 2028 - 83383, 2043 - 110720
Karnataka: 2023 - 92604, 2028 - 104522, 2043 - 146751
Goa: 2023 - 1738, 2028 - 1960, 2043 - 2770
Lakshadweep: 2023 - 29, 2028 - 34, 2043 - 49
Kerala: 2023 - 60145, 2028 - 65168, 2043 - 79294
Tamil Nadu: 2023 - 95943, 2028 - 108677, 2043 - 154301
Pondicherry: 2023 - 1733, 2028 - 2030, 2043 - 3200
Andaman & Nicobar Islands: 2023 - 406, 2028 - 480, 2043 - 777
Total: 2023 - 1496998, 2028 - 1684339, 2043 - 2340731

```

In the above snippet the prediction is validated after compared with the prediction compared WHO's prediction.

State	2022-2023	2022-2028	2022-2043
Jammu & Kashmir	2.58%	16.21%	64.92%
Ladakh	2.65%	15.89%	55.63%
Himachal Pradesh	2.11%	12.83%	47.81%
Punjab	2.28%	14.23%	54.90%
Chandigarh	3.03%	20.04%	84.83%
Uttaranchal	2.45%	15.38%	60.95%
Haryana	2.71%	17.25%	70.53%
Delhi	3.03%	19.17%	80.33%
Rajasthan	2.58%	16.26%	65.40%
Uttar Pradesh	2.34%	14.61%	56.85%
Bihar	2.62%	16.57%	67.00%
Sikkim	4.03%	32.26%	150.60%
Arunachal Pradesh	2.39%	14.72%	58.51%
Nagaland	2.43%	15.43%	62.35%
Manipur	4.67%	29.71%	134.53%

The above code snippet shows the percentage increase of the cancer rate from 2022 to each year i.e, 2023, 2028 and 2043.

For the prediction of the cancer for the 20 years I used something called as Iterative prediction, the code snippet looks like:

```
# Predict the number of cases for 2028 using iterative prediction
for i in range(20): # Predict for 20 years (2024-2043)
    # Use previous predictions as features
    X_pred = dataset.iloc[:, -4:].values
    prediction = model.predict(X_pred)
    prediction = prediction.round().astype(int)
    col = int(dataset.columns[-1]) + 1
    dataset[str(col)] = prediction
```

The above code created new columns and saves the new prediction one after the another by itself.

- **Random Forest:**

While Random Forest is powerful, it needs a lot of data to learn complex patterns. With limited cancer rate data, the trees in the forest become too similar, leading to overly simple predictions (under fitting) for future years.

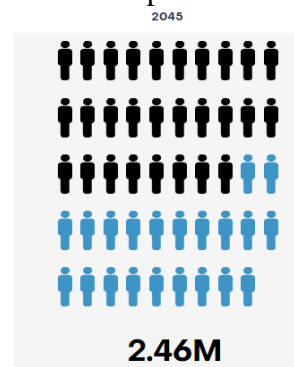
From the above two models **Linear Regression is found to be accurate** and the best one for this type of data, so I choose that model for the prediction.

Prediction Validation(for 2045):

My Prediction(2045):

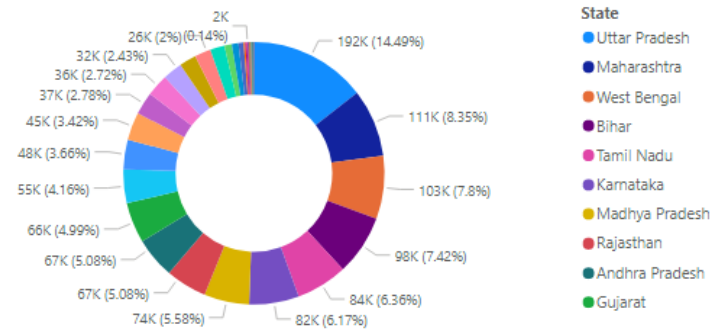
2438449

WHO prediction:



Visualization of Results

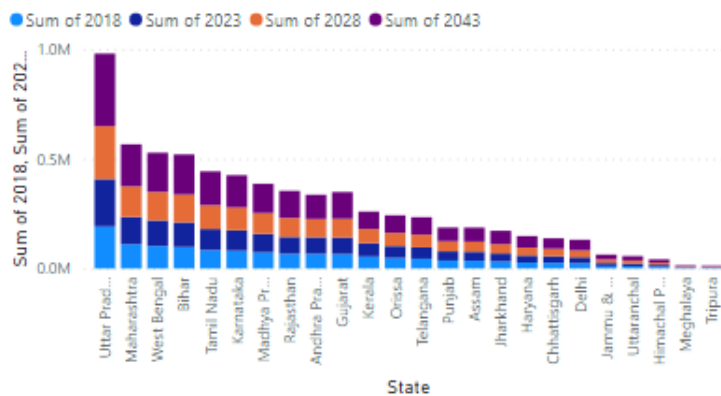
1.



The Above pie chart shows the cancer cases of all the states in India and highlighting the top 10 states in India.

2.

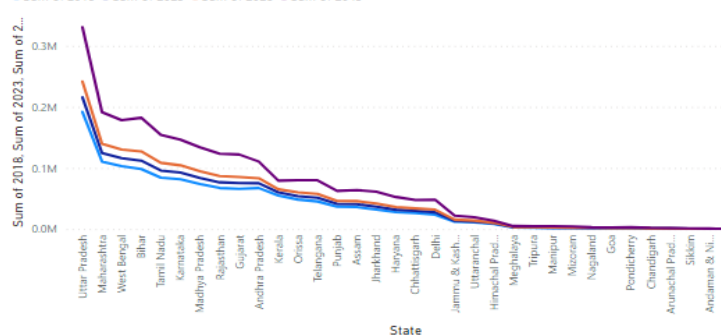
Sum of 2018, Sum of 2023, Sum of 2028 and Sum of 2043 by State



The Above stacked bar graph shows the growth of cancer across all the states in India, which helps us to find out in which state there is a boon in the cancer cases in the years.

3.

Sum of 2018, Sum of 2023, Sum of 2028 and Sum of 2043 by State



The above Line graph shows the trend in different States in India for different years.

Conclusion

This project explored the prediction of cancer incidence rates for various states in India. Here's a possible conclusion summarizing the key points:

Key Findings:

- The project evaluated different machine learning models (likely linear regression, Random Forest, and Gradient Boosting) trained on historical data (2018-2021) to predict cancer cases for future years (2023, 2028, 2043).
- The analysis resulted in visualizations (stacked bar chart, Pie Chart, Line Graph) depicting the predicted cases for each state across the specified years.
- You might have also calculated and visualized the percentage increase in cases from 2022 to each predicted year, providing insights into potential trends.

Model Selection:

- To choose the best prediction model, I compared the performance of linear regression, Random Forest, and Gradient Boosting on the historical data.
- Based on evaluation metrics (mean absolute error), linear regression emerged as the model with the most accurate predictions for our specific dataset.

Limitations:

- While linear regression provided the best fit for our data, it might not capture complex non-linear relationships that could influence cancer rates.
- The accuracy of predictions still depends on the quality and completeness of historical data used for training.
- External factors influencing cancer rates (e.g., lifestyle changes, access to healthcare) are not directly accounted for in any of the models explored.

Future Work:

- Incorporate additional data sources (e.g., demographics, socio-economic factors) to enhance the model's ability to capture real-world influences on cancer rates.
- Develop interactive dashboards or reports to facilitate further exploration and analysis of the predicted data.

Overall Significance:

This project provides valuable insights into predicting cancer incidence rates across Indian states. While limitations exist, the chosen model (linear regression) offered accurate predictions based on available data for the next 20 years. The generated visualizations can inform public health officials and researchers working towards understanding and potentially mitigating future cancer burdens.

Also this will help in improving the techniques adopted by the state governments and the central government to tackle this issue of Cancer across the country.

References

- <https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services>
- <https://www.icmr.nic.in/>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10231735/>
- <https://www.google.co.in/>
- <https://ascopubs.org/doi/10.1200/GO.20.00122>
- [https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045\(18\)30447-9/fulltext](https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(18)30447-9/fulltext)
- <https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2023.1065737/full>
- https://scholar.google.com/scholar_lookup?title=Global+cancer+observatory:Cancer+today&author=J+Ferlay&author=M+Ervik&author=F+Lam&author=M+Colombet&author=L+Mery&publication_year=2020&
- https://gco.iarc.fr/tomorrow/en/dataviz/isotype?types=0&sexes=0&mode=population&group_populations=0&multiple_populations=0&multiple_cancers=0&cancers=39&populations=356&single_unit=50000&years=2045
- <https://actuarial-dashboard.shinyapps.io/CancerPoolPricing/#section-tmh-463>
- <https://www.healthdata.org/research-analysis/health-by-location/disease-burden-initiative-india>