**PROJECT REPORT**

**CANCER MORTALITY AND INCIDENCE RATE**

**USING MACHINE LEARNING**

**DOMAIN: APPLIED DATA SCIENCE**

**VELAMMAL INSTITUTE OF TECHNOLOGY,**

**PANCHETTI**

**TEAM ID: NM2023TMID20537**

**INDEX**

1. **INTRODUCTION**

1.1 Project Overview

1.2 Purpose

2. **IDEATION & PROPOSED SOLUTION**

2.1 Problem Statement Definition

2.2 Empathy Map Canvas

2.3 Ideation & Brainstorming

2.4 Proposed Solution

3. **REQUIREMENT ANALYSIS**

3.1 Functional requirement

3.2 Non-Functional requirements

**4. PROJECT DESIGN**

4.1 Data Flow Diagrams

4.2 Solution & Technical Architecture

4.3 User Stories

**5. CODING & SOLUTIONING**

5.1 Feature 1

5.2 Feature 2

5.3 Database Schema

**6. RESULTS**

6.1 Performance Metrics

**7. ADVANTAGES & DISADVANTAGES**

**8. CONCLUSION**

**9. FUTURE SCOPE**

**10. APPENDIX**

10.1 Source code

10.2 Github link

10.3 Demo link

**1.INTRODUCTION**

**1.1 PROJECT OVERVIEW**

**Introduction:**

The purpose of this project is to analyze cancer mortality and incidence rates to gain insights into the prevalence and impact of cancer within a specific population. By examining trends, patterns, and risk factors, the project aims to contribute to a better understanding of cancer and potentially inform public health interventions and strategies for cancer prevention and treatment.

**Objectives:**

The key objectives of the project are as follows:

* Analyse cancer mortality rates: Investigate the number of deaths caused by different types of cancer over a specific period. Identify trends, variations, and patterns in cancer mortality rates.
* Analyse cancer incidence rates: Study the number of new cancer cases diagnosed within the target population. Determine the types of cancer with the highest incidence rates and their distribution.
* Identify high-risk populations: Explore demographic factors, such as age, gender, ethnicity, socioeconomic status, and geographical location, to identify high-risk populations for specific types of cancer.
* Investigate risk factors: Examine potential risk factors associated with cancer, such as lifestyle choices (e.g., smoking, diet), environmental exposures, occupational hazards, and genetic predispositions.
* Compare international data: Compare cancer mortality and incidence rates across different countries or regions to identify variations and potential factors contributing to differences in cancer prevalence.

**Data Collection:**

To conduct the analysis, the project will require reliable and comprehensive data from relevant sources. Possible sources may include:

* National cancer registries: Obtain cancer incidence and mortality data from national databases or cancer registries, ensuring data accuracy, consistency, and representativeness.
* Demographic and socioeconomic data: Collect demographic information from census data, healthcare databases, or surveys to examine the relationship between cancer and population characteristics.
* Lifestyle and risk factor data: Gather information on lifestyle choices, environmental exposures, occupational hazards, and genetic factors from existing studies, surveys, or specialized databases.

**Data Analysis:**

* The collected data will undergo rigorous analysis using appropriate statistical and data mining techniques. The analysis may involve:
* Descriptive analysis: Calculate cancer mortality and incidence rates, summarize demographic characteristics, and identify key trends and patterns.
* Comparative analysis: Compare cancer rates across different populations, regions, or time periods to identify variations and potential influencing factors.
* Correlation analysis: Explore relationships between cancer rates and various risk factors to identify significant associations and potential causal links.
* Predictive modelling: Utilize machine learning algorithms or other modelling techniques to predict future cancer incidence rates based on available data and trends.

**Interpretation and Reporting:**

The findings will be interpreted, and meaningful insights will be derived from the analysis. The project will generate a comprehensive report summarizing the key findings, including:

* Overview of cancer mortality and incidence rates: Provide an overview of the current state of cancer prevalence within the target population.
* Identified risk factors: Highlight the significant risk factors associated with cancer development or mortality.
* High-risk populations: Identify specific demographic groups or regions with elevated cancer rates or unique risk profiles.
* Recommendations: Provide evidence-based recommendations for cancer prevention, early detection, and intervention strategies based on the findings.

**Limitations and Ethical Considerations:**

The project may have certain limitations, such as data availability, accuracy, and potential biases. Ethical considerations, including data privacy and confidentiality, will be adhered to throughout the project, ensuring the protection of individuals' personal information.

* 1. **PURPOSE**

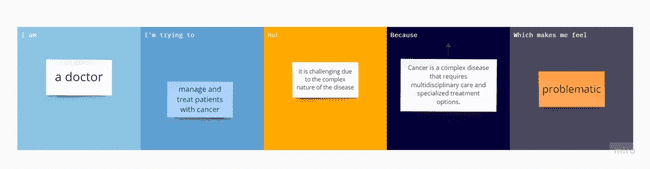
The purpose of a cancer mortality and incidence rate project is to systematically analyse and interpret data on cancer deaths and new cancer cases within a specific population, with the aim of understanding the burden of cancer, identifying high-risk groups, exploring potential risk factors, informing prevention and treatment strategies, and ultimately reducing the impact of cancer on public health. By examining trends, patterns, and associations, the project contributes to the body of knowledge surrounding cancer, supports evidence-based decision-making, and facilitates targeted interventions to improve outcomes and save lives.

**2.** **IDEATION & PROPOSED SOLUTION**

**2.1 PROBLEM STATEMENT DEFINITION**

The goal of this project is to predict the status of cancer incidence or mortality rate based on a set of features to provide important insights into the prevalence and impact of cancer within a given population, as well as how this burden is changing over time.

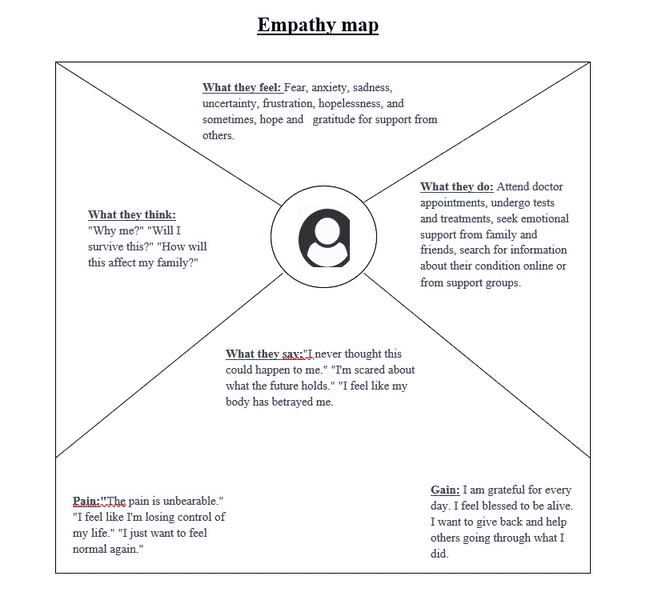
**Problem Statement 1:**



**Problem Statement 2:**

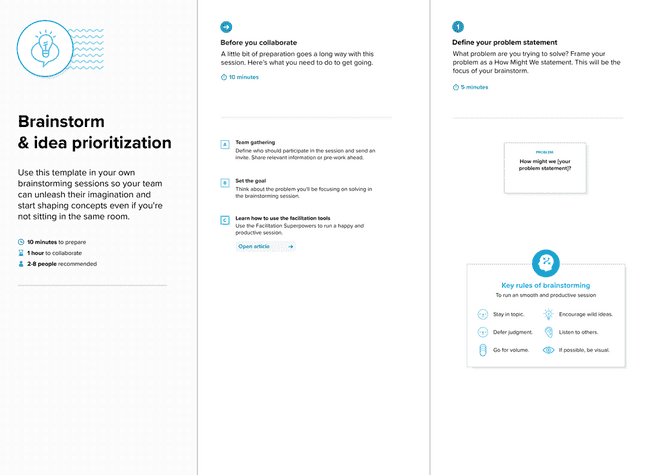


**2.2 EMPATHY MAP CANVAS**



**2.3 IDEATION & BRAINSTORMING**

* Providing more access to education and resources for patients to learn about their specific type of cancer, treatment options, and how to manage side effects.
* Creating a support network for patients to connect with others going through similar experiences and share information and resources.
* Implementing technology solutions, such as telemedicine, to increase access to care and reduce barriers to treatment.
* Increasing funding for cancer research to develop more effective treatments and improve patient outcomes.
* Providing more mental health support and counselling services for patients and their families.



**2.4 PROPOSED SOLUTION**

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Parameter** | **Description** |
| 1. | Problem Statement (Problem to be solved) | The goal of this project is to predict the status of cancer incidence or mortality rate based on a set of features to provide important insights into the prevalence and impact of cancer within a given population, as well as how this burden is changing over time. |
| 2. | Idea / Solution description | Our project aims to develop a machine learning model that can accurately predict cancer mortality and incidence rates based on a set of input variables. |
| 3. | Novelty / Uniqueness | This project has a good number of reasons to be called unique. Such as use of Machine Learning, combination of multiple data sources, public health impact, open source, etc. |
| 4. | Social Impact / Customer Satisfaction | It has the potential to improve public health outcomes and customer satisfaction by providing more accurate and reliable predictions of cancer rates. By identifying populations at high risk of cancer and areas with limited healthcare access, policymakers can develop targeted strategies to improve prevention, early detection, and treatment, ultimately improving health outcomes for the population. |
| 5. | Business Model (Revenue Model) | We aim to do our project as service oriented and hence we have not yet thought about a revenue path. But as the project flows and if we get an idea that could benefit both common people and us, we would like to consider it. |
| 6. | Scalability of the Solution | By leveraging distributed computing frameworks, cloud computing services, algorithm optimization and AutoML tools, the project can be scaled to handle larger and more complex datasets, making it more useful to healthcare professionals and policymakers. But for now, we have been considering implementing effective algorithms regarding scalability. But would like to further implement all the above-mentioned tools if more scalability is required. |

**3. REQUIREMENT ANALYSIS**

**3.1 FUNCTIONAL REQUIREMENT**

Following are the functional requirements of the proposed solution:

|  |  |  |
| --- | --- | --- |
| **FR No.** | **Functional Requirement (Epic)** | **Description** |
| FR-1 | **User Registration:** | Registration through Form  Registration through Gmail  Registration through LinkedIN |
| FR-2 | **User Confirmation:** | Confirmation via Email Confirmation via OTP |
| FR-3 | **Accurate data collection:** | The data used to calculate cancer mortality and incidence rates must be accurate and reliable. This may involve ensuring that data is collected consistently across time and geographic regions, and that it is based on robust diagnostic and reporting methods. |
| FR-4 | **Standardized calculation methods:** | There should be standardized methods for calculating cancer mortality and incidence rates, in order to ensure that the rates are comparable across different time periods and regions. |
| FR-5 | **Timely reporting:** | Cancer mortality and incidence rates should be reported in a timely manner, so that they can be used to inform public health decisions and interventions. |
| FR-6 | **Regular updates:** | Cancer mortality and incidence rates should be updated regularly, in order to track changes over time and inform ongoing public health efforts. |

**3.2 NON-FUNCTIONAL REQUIREMENTS**

Following are the non-functional requirements of the proposed solution:

|  |  |  |
| --- | --- | --- |
| **FR No.** | **Non-Functional Requirement** | **Description** |
| NFR-1 | **Usability** | The information related to cancer mortality and incidence rates should be easily accessible and understandable by the intended audience, which may include healthcare professionals, public health officials, policymakers, and the general public. |
| NFR-2 | **Security** | Cancer mortality and incidence rates should be calculated and reported in compliance with applicable laws and regulations related to data privacy and security |
| NFR-3 | **Reliability** | The methods used to calculate cancer mortality and incidence rates should be reliable and valid, meaning that they produce consistent and accurate results. This may involve using accepted standards and protocols, verifying data quality, and conducting regular audits and evaluations. |
| NFR-4 | **Performance** | The systems and processes used to collect, store, and analyse data related to cancer mortality and incidence rates should perform reliably and quickly, in order to support real-time decision making and action. |
| NFR-5 | **Availability** | The information related to cancer mortality and incidence rates should be available 24/7, in order to enable access by healthcare professionals, public health officials, policymakers, and the general public at any time. |
| NFR-6 | **Scalability** | The systems and processes used to collect and analyse data related to cancer mortality and incidence rates should be scalable and flexible, in order to accommodate changing needs and requirements over time. |

**4. PROJECT DESIGN**

**4.1 DATA FLOW DIAGRAMS**

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.

P

atients

Healthcare

providers

Labs

Patient

Information

Diagnosis and

Treatment Planning

Lab Test

Results

Treatment Plans

Treatment and

Monitoring

Treatment

Updates

Medications

Medication

Management

Test Results

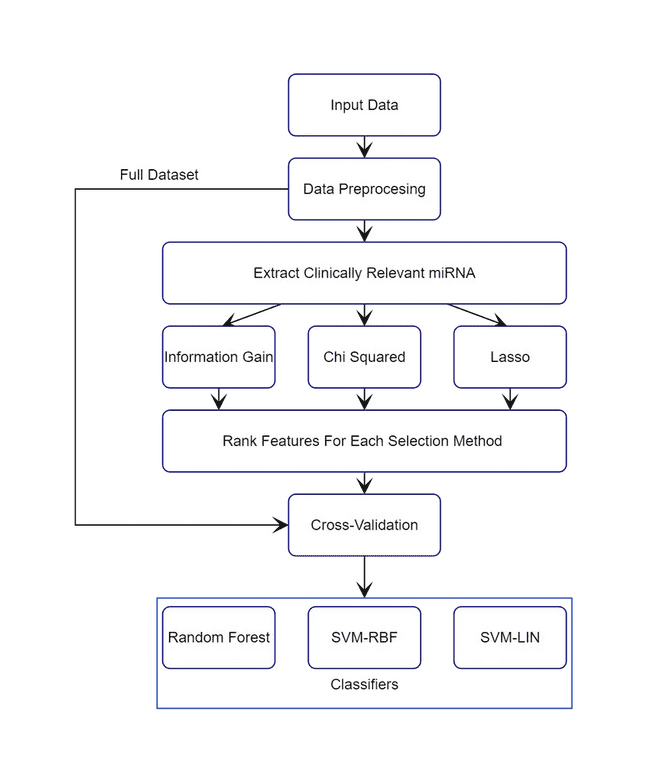
**4.2 SOLUTION & TECHNICAL ARCHITECTURE**

**Solution Architecture:**

Solution architecture is a complex process – with many sub-processes – that bridges the gap between business problems and technology solutions. Its goals are to:

* Find the best tech solution to solve existing business problems.
* Describe the structure, characteristics, behaviour, and other aspects of the software to project stakeholders.
* Define features, development phases, and solution requirements.
* Provide specifications according to which the solution is defined, managed, and delivered.

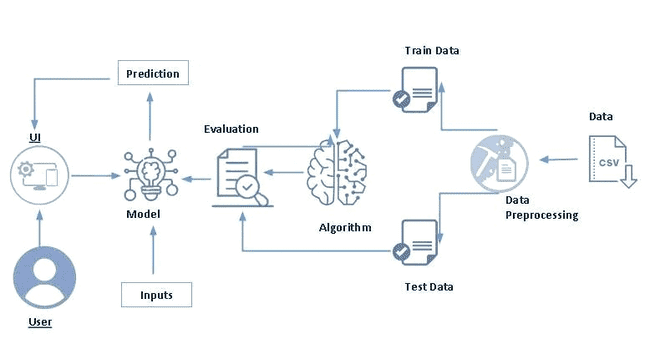
**Example - Solution Architecture Diagram:**



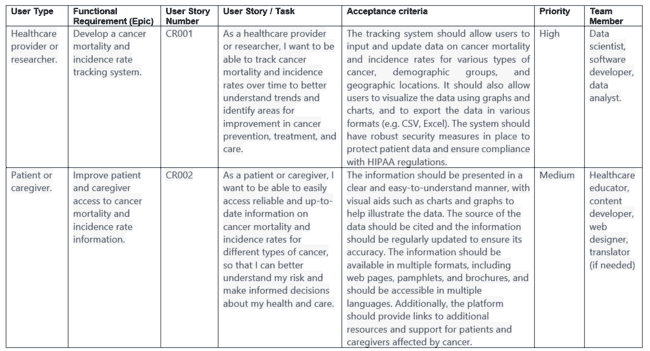
**Technical Architecture:**

Technical Architecture (TA) is a form of IT architecture that is used to design computer systems. It involves the development of a technical blueprint regarding the arrangement, interaction, and interdependence of all elements so that system-relevant requirements are met.

**Example: Technical Architecture Diagram**



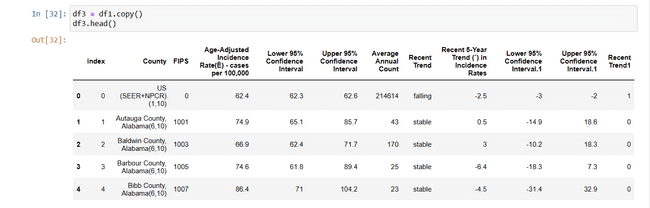
**4.3 USER STORIES**

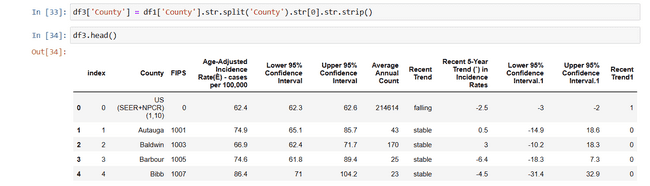


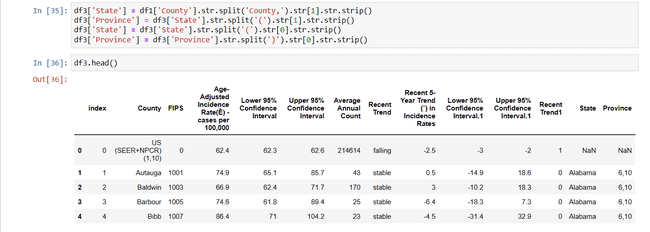
**5. CODING & SOLUTIONING**

**5.1 FEATURE 1**

**Divided the categorical columns into sections to improve accuracy**

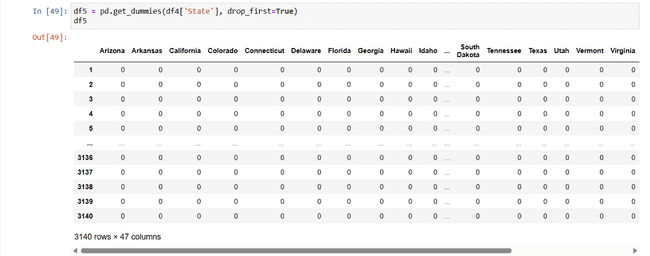


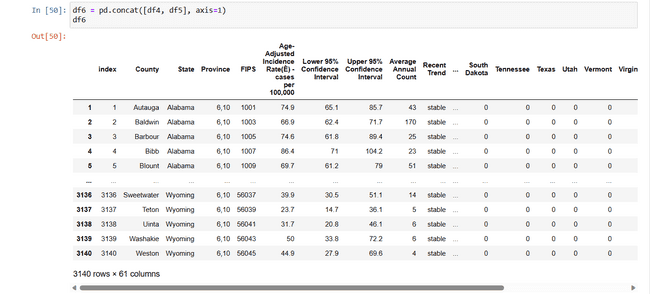




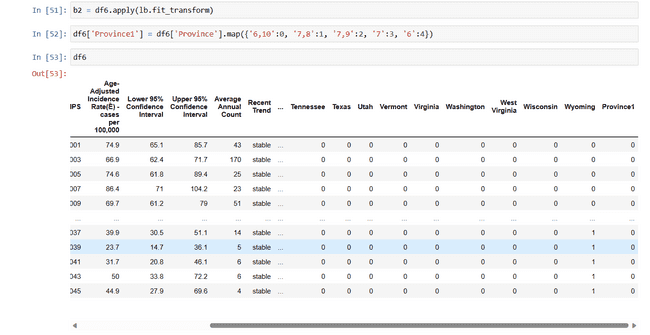
**5.2 FEATURE 2**

**Encoded State Categorical column using OneHotEncoding**





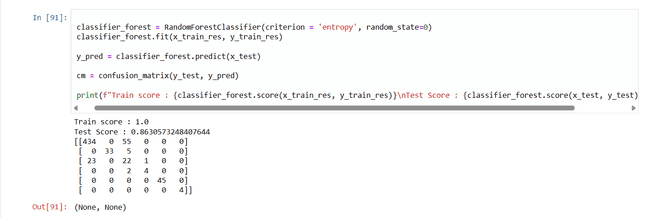
**Encoded Province Categorical column using Label Encoding**



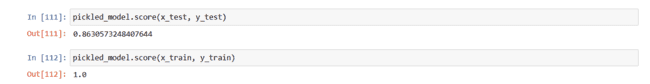
**6. RESULTS**

**6.1 PERFORMANCE METRICS**

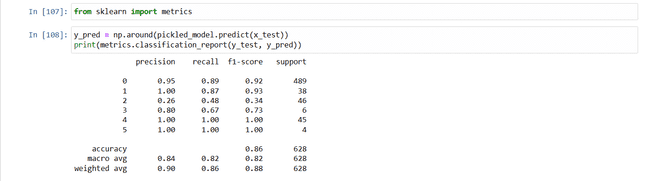
**Confusion Matrix**



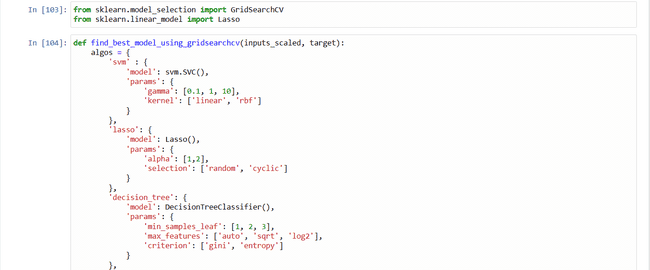
**Accuracy Score**

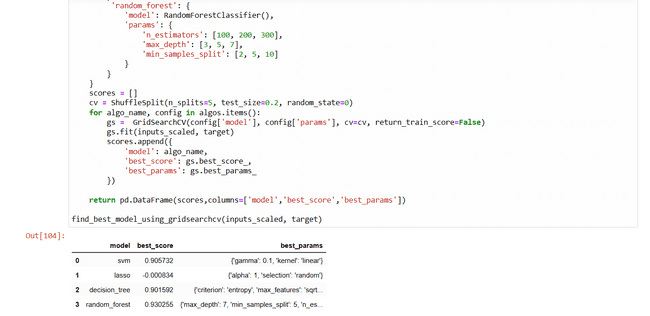


**Classification Report**

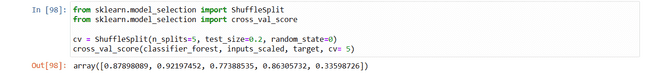


**1.Hyperparameter tuning: GridSearchCV**





**2.Validation Method: Cross Validation**



**7. ADVANTAGES & DISADVANTAGES**

**7.1 ADVANTAGES**

There are several advantages of incorporating machine learning techniques into a cancer mortality and incidence rate project:

**Improved Accuracy:** Machine learning algorithms can effectively analyze large volumes of data, identify complex patterns, and make accurate predictions. By leveraging machine learning, researchers can improve the accuracy of cancer mortality and incidence rate predictions, leading to more reliable insights for decision-making.

**Early Detection:** Machine learning algorithms can be trained to detect early signs and patterns associated with cancer development. By analyzing diverse data sources, such as medical records, genetic information, and lifestyle data, machine learning models can assist in early cancer detection, potentially leading to timely interventions and improved patient outcomes.

**Risk Assessment and Stratification:** Machine learning models can assess individual risk profiles for developing specific types of cancer based on a range of factors, such as demographics, lifestyle choices, and genetic markers. This enables personalized risk assessment and stratification, allowing healthcare professionals to target preventive measures and allocate resources more effectively.

**Identification of Risk Factors:** Machine learning algorithms can identify hidden relationships and correlations between various risk factors and cancer mortality or incidence rates. By analyzing a wide range of variables simultaneously, machine learning models can uncover complex interactions, contributing to a deeper understanding of the factors contributing to cancer development and progression.

**Data-driven Insights**: Machine learning models provide data-driven insights that can guide evidence-based decision-making. By analysing historical data on cancer mortality and incidence rates, machine learning algorithms can identify patterns and trends that may not be readily apparent through traditional statistical analysis. This enables researchers to derive actionable insights and recommendations to inform cancer prevention and treatment strategies.

**Enhanced Research Capabilities:** Machine learning enables researchers to process and analyze large datasets efficiently. By automating tasks such as data preprocessing, feature selection, and model training, machine learning algorithms can accelerate the research process, allowing researchers to explore multiple hypotheses and conduct more in-depth analyses.

**7.2 DISADVANTAGES**

**Data Quality and Bias:** Machine learning models heavily rely on the quality and representativeness of the input data. If the data used for training the models is incomplete, inaccurate, or biased, it can lead to biased predictions and unreliable insights. Biases in the data, such as underrepresentation of certain demographics or geographic regions, can result in skewed results and limited generalizability.

**Interpretability and Explain ability:** Machine learning models, particularly complex ones like deep learning models, can be challenging to interpret and explain. Understanding the underlying factors and features driving the predictions may be difficult, making it harder to gain insights and trust in the model's results. Lack of interpretability may hinder the ability to explain findings to healthcare professionals, policymakers, and the general public.

**Overfitting and Generalizability:** Machine learning models can be prone to overfitting, which occurs when a model performs well on the training data but fails to generalize to new, unseen data. Overfitting can lead to overly optimistic predictions and limited practical utility. Ensuring the generalizability of machine learning models requires careful validation and testing on independent datasets.

**Data Privacy and Security:** Cancer mortality and incidence rate projects often involve sensitive and personal health data. Maintaining data privacy and security is crucial to protect individuals' privacy and comply with ethical and legal standards. Machine learning projects must implement robust security measures to safeguard the data from unauthorized access or breaches.

**Lack of Contextual Understanding:** Machine learning models excel at identifying patterns and correlations within the data, but they may lack contextual understanding. Understanding the underlying biological, environmental, or socioeconomic factors driving cancer mortality and incidence rates often requires domain knowledge and expert interpretation. Machine learning should be used as a tool in conjunction with domain expertise to derive meaningful insights.

**Human Bias Amplification:** Machine learning models can inadvertently amplify human biases present in the training data. If the data used for training contains biases, such as racial or gender disparities in healthcare access or diagnoses, the resulting predictions may perpetuate or exacerbate those biases. Careful consideration and mitigation strategies are necessary to address and minimize biased outcomes.

**Resource Requirements:** Developing and training machine learning models can be computationally intensive and may require significant computational resources, time, and expertise. Implementing and maintaining the necessary infrastructure, computational power, and technical skills may pose challenges for smaller research teams or organizations with limited resources.

**8. CONCLUSION**

In conclusion, studying cancer mortality and incidence rates is essential for understanding the prevalence, impact, and risk factors associated with cancer within a specific population. By analyzing these rates, researchers and healthcare professionals can gain valuable insights that contribute to evidence-based decision-making, targeted interventions, and improved outcomes in cancer prevention and treatment. While incorporating machine learning techniques offers advantages such as improved accuracy, early detection, and personalized risk assessment, there are also potential disadvantages to consider, including data quality and bias, interpretability challenges, generalizability issues, and privacy concerns. Mitigating these limitations through careful data management, validation processes, interpretability techniques, and ethical considerations is crucial. By combining the strengths of machine learning with domain expertise, comprehensive cancer mortality and incidence rate projects can provide valuable insights to inform public health strategies, reduce the burden of cancer, and improve the well-being of affected individuals and communities.

**9. FUTURE SCOPE**

The future scope for cancer mortality and incidence rate projects is promising, with several potential areas for further exploration and development. Some of the future directions and opportunities include:

**Integration of Multi-omics Data:** Incorporating multi-omics data, such as genomics, transcriptomics, proteomics, and metabolomics, can provide a more comprehensive understanding of cancer development and progression. Integrating these data types with mortality and incidence rates can facilitate the identification of molecular signatures, biomarkers, and personalized treatment strategies.

**Application of Deep Learning:** Deep learning techniques, including convolutional neural networks and recurrent neural networks, hold promise for analyzing complex medical imaging data, such as radiological images and pathology slides. Leveraging deep learning algorithms can enhance early detection, diagnosis, and prognosis prediction, leading to improved patient outcomes.

**Real-Time Monitoring and Predictive Analytics**: Incorporating real-time monitoring of cancer mortality and incidence rates, combined with predictive analytics, can enable timely interventions and resource allocation. By continuously analyzing data streams from healthcare systems and population health databases, healthcare providers can proactively identify emerging trends, assess the effectiveness of interventions, and adapt strategies accordingly.

**Social Determinants of Health:** Exploring the influence of social determinants of health, such as socioeconomic status, education, and access to healthcare, on cancer mortality and incidence rates is crucial. Integrating these factors into the analysis can provide a more comprehensive understanding of health disparities and guide targeted interventions to address inequities in cancer outcomes.

**Integration of Digital Health Technologies**: Embracing digital health technologies, such as wearable devices, mobile health applications, and telemedicine, can generate vast amounts of real-time health data. Integrating these data sources with cancer mortality and incidence rates can support remote monitoring, personalized interventions, and population-level health surveillance.

**Ethical and Privacy Considerations:** With the increasing use of sensitive health data in cancer research, addressing ethical concerns and ensuring privacy protection is paramount. Future projects should prioritize robust data governance, secure data sharing frameworks, and transparent consent processes to maintain public trust and safeguard individuals' privacy.

**10. APPENDIX**

**10.1. SOURCE CODE**







**10.2. GITHUB LINK**

[**https://github.com/naanmudhalvan-SI/PBL-NT-GP--7213-1681036940**](https://github.com/naanmudhalvan-SI/PBL-NT-GP--7213-1681036940)

**10.3. DEMO LINK**

[https://drive.google.com/drive/folders/1qXO7g1BNTTTwEEeBmMpEsa8xf3aawTIa?usp=share\_link](%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20https://drive.google.com/drive/folders/1qXO7g1BNTTTwEEeBmMpEsa8xf3aawTIa?usp=share_link)