Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY, THANE
(All Programs Accredited by NBA)
Department of Information Technology

Academic Year: 2023-2024 **Name of Student:Avantika More**
Semester: VI **Student ID:21104033**
Class / Branch: TE-IT-A **Date of Performance: 03/04/2024**
Subject: BI Lab **Date of Submission:03/04/2024**
Name of Instructor: Prof.Geetanjali Kalme

Experiment No. 13

Aim: Business Intelligence Mini Project.

1. **Problem Definition:**The cervical cancer behavior risk project aims to investigate the various factors influencing behaviors related to cervical cancer prevention, screening, and treatment. This research endeavor seeks to understand the complex interplay between individual behaviors, socio-economic factors, cultural beliefs, healthcare access, and knowledge dissemination regarding cervical cancer. The project will define and analyze behavioral risks associated with cervical cancer, including low screening rates, inadequate vaccination coverage against HPV (Human Papillomavirus), delayed treatment-seeking behaviors, and misconceptions surrounding the disease. By elucidating these factors, the project aims to develop targeted interventions and strategies to promote early detection, increase vaccination uptake, improve access to screening services, and enhance overall cervical cancer prevention efforts. Through comprehensive data collection, analysis, and collaboration with healthcare providers, policymakers, and community stakeholders, this project endeavors to mitigate the burden of cervical cancer and reduce health disparities among at-risk populations.

2. **Data mining task to be performed:**For a cervical cancer behavior risk project, a data mining task would involve analyzing large datasets to identify patterns, trends, and associations related to various factors that influence the behavior and risk of cervical cancer. This could include demographic information such as age, ethnicity, socioeconomic status, as well as lifestyle factors like smoking, sexual behavior, contraceptive use, and medical history including previous screenings and vaccination status for HPV. The goal would be to uncover insights that could help in understanding the behavioral determinants of cervical cancer risk, identifying high-risk populations, and informing targeted interventions and prevention strategies. Data mining techniques such as classification, clustering, association rule mining, and predictive modeling could be employed to extract meaningful information from the data and uncover hidden patterns and relationships that may not be immediately apparent. Additionally, data visualization techniques could be used to present the findings in a clear and comprehensible manner, facilitating decision-making for healthcare professionals and policymakers.

3. **Dataset identified: Name of your dataset - UCI**

4. **Source of dataset: URL -**
   https://archive.ics.uci.edu/static/public/537/cervical+cancer+behavior+risk.zip

5. **Details of the dataset: Brief description of the dataset** - The UCI Cervical Cancer Behavior Risk dataset is a comprehensive repository containing anonymized data relevant to the behavioral risk factors associated with cervical cancer. This dataset comprises demographic information, lifestyle factors, medical history, and clinical findings of individuals, offering a holistic view of factors potentially influencing cervical cancer development. Variables such as age, sexual behavior, smoking habits, contraceptive usage, and various clinical parameters are included. This dataset serves as a valuable resource for researchers and healthcare professionals aiming to explore correlations between behavioral factors and cervical cancer incidence, potentially leading to enhanced preventive strategies and targeted interventions.

6. Algorithms to accomplish the task:

For the cervical cancer behavior risk project, various algorithms can be employed to predict the risk of cervical cancer based on different features. One popular dataset used for this task is the "Cervical Cancer Behavior Risk" dataset available on the UCI Machine Learning Repository. This dataset contains information on socio-demographic characteristics and behaviors of women that may influence the risk of cervical cancer.

One algorithm commonly used for classification tasks like this is the Random Forest algorithm. Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes for classification or mean prediction for regression tasks. It is robust to overfitting and tends to perform well on a variety of datasets.

To implement the Random Forest algorithm for the cervical cancer behavior risk project, the dataset can be preprocessed by handling missing values, encoding categorical variables, and splitting it into training and testing sets. Then, the Random Forest classifier can be trained on the training data and evaluated on the testing data using performance metrics such as accuracy, precision, recall, and F1-score.

After implementing the Random Forest algorithm, the results can be visualized using various techniques such as confusion matrices, ROC curves, and precision-recall curves. These visualizations provide insights into the performance of the algorithm and help in understanding its strengths and weaknesses in predicting cervical cancer behavior risk.

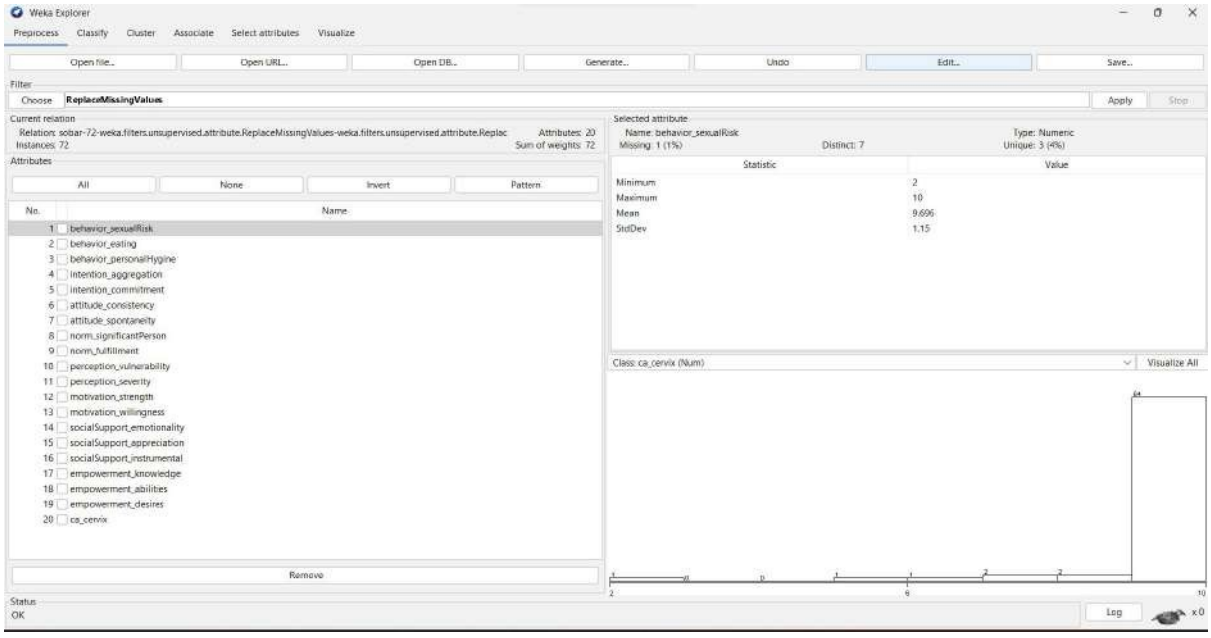
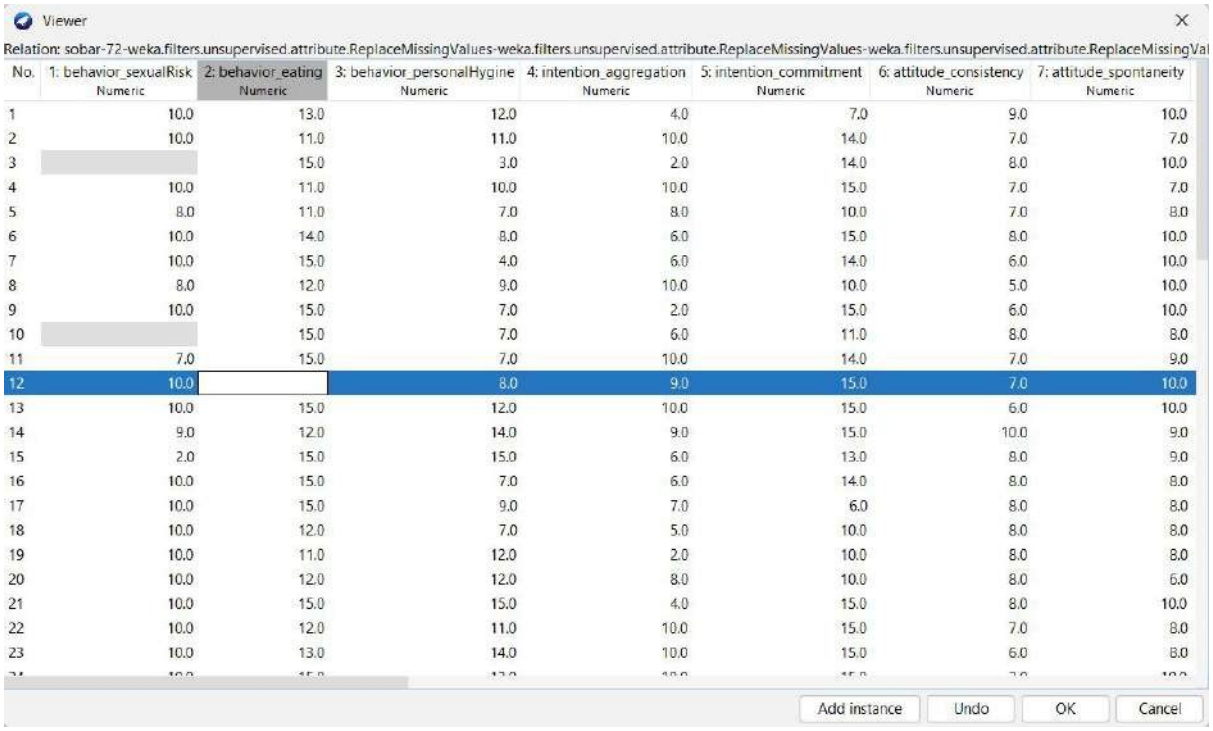


Fig 6.1:Replace Missing Values



No.	1: behavior_sexualRisk Numeric	2: behavior_eating Numeric	3: behavior_personalHygine Numeric	4: intention_aggregation Numeric	5: intention_commitment Numeric	6: attitude_consistency Numeric	7: attitude_spontaneity Numeric
1	10.0	13.0	12.0	4.0	7.0	9.0	10.0
2	10.0	11.0	11.0	10.0	14.0	7.0	7.0
3		15.0	3.0	2.0	14.0	8.0	10.0
4	10.0	11.0	10.0	10.0	15.0	7.0	7.0
5	8.0	11.0	7.0	8.0	10.0	7.0	8.0
6	10.0	14.0	8.0	6.0	15.0	8.0	10.0
7	10.0	15.0	4.0	6.0	14.0	6.0	10.0
8	8.0	12.0	9.0	10.0	10.0	5.0	10.0
9	10.0	15.0	7.0	2.0	15.0	6.0	10.0
10		15.0	7.0	6.0	11.0	8.0	8.0
11	7.0	15.0	7.0	10.0	14.0	7.0	9.0
12	10.0		8.0	9.0	15.0	7.0	10.0
13	10.0	15.0	12.0	10.0	15.0	6.0	10.0
14	9.0	12.0	14.0	9.0	15.0	10.0	9.0
15	2.0	15.0	15.0	6.0	13.0	8.0	9.0
16	10.0	15.0	7.0	6.0	14.0	8.0	8.0
17	10.0	15.0	9.0	7.0	6.0	8.0	8.0
18	10.0	12.0	7.0	5.0	10.0	8.0	8.0
19	10.0	11.0	12.0	2.0	10.0	8.0	8.0
20	10.0	12.0	12.0	8.0	10.0	8.0	6.0
21	10.0	15.0	15.0	4.0	15.0	8.0	10.0
22	10.0	12.0	11.0	10.0	15.0	7.0	8.0
23	10.0	13.0	14.0	10.0	15.0	6.0	8.0

Fig 6.2:Remove

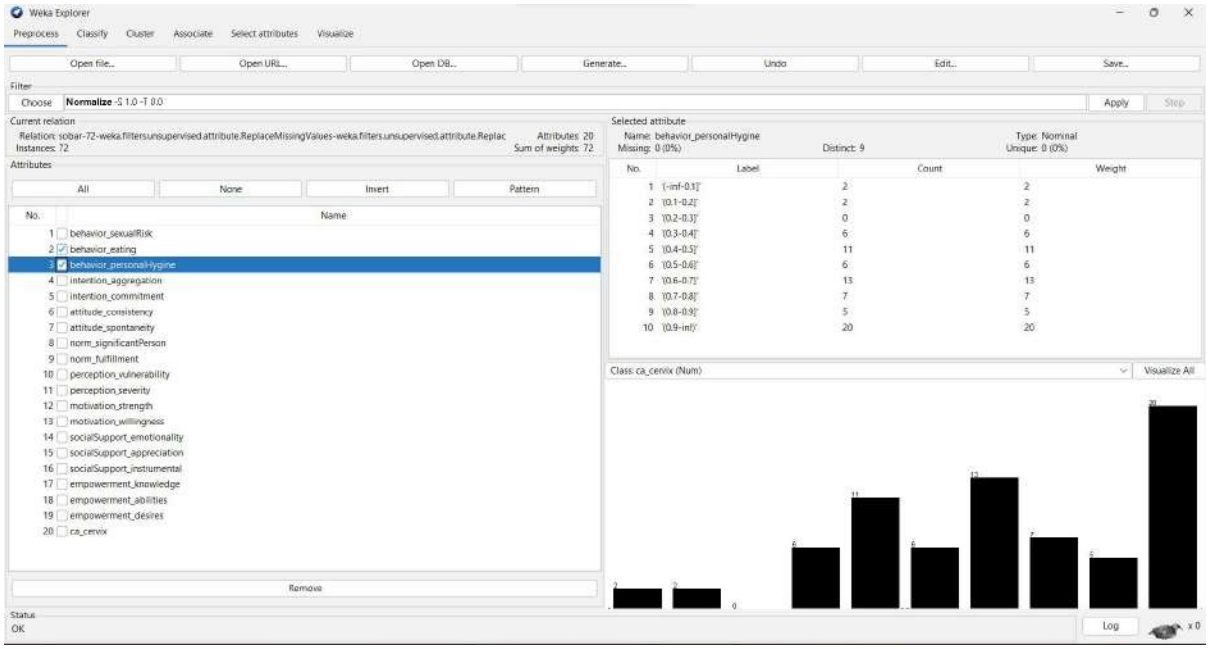


Fig 6.3:Normalize

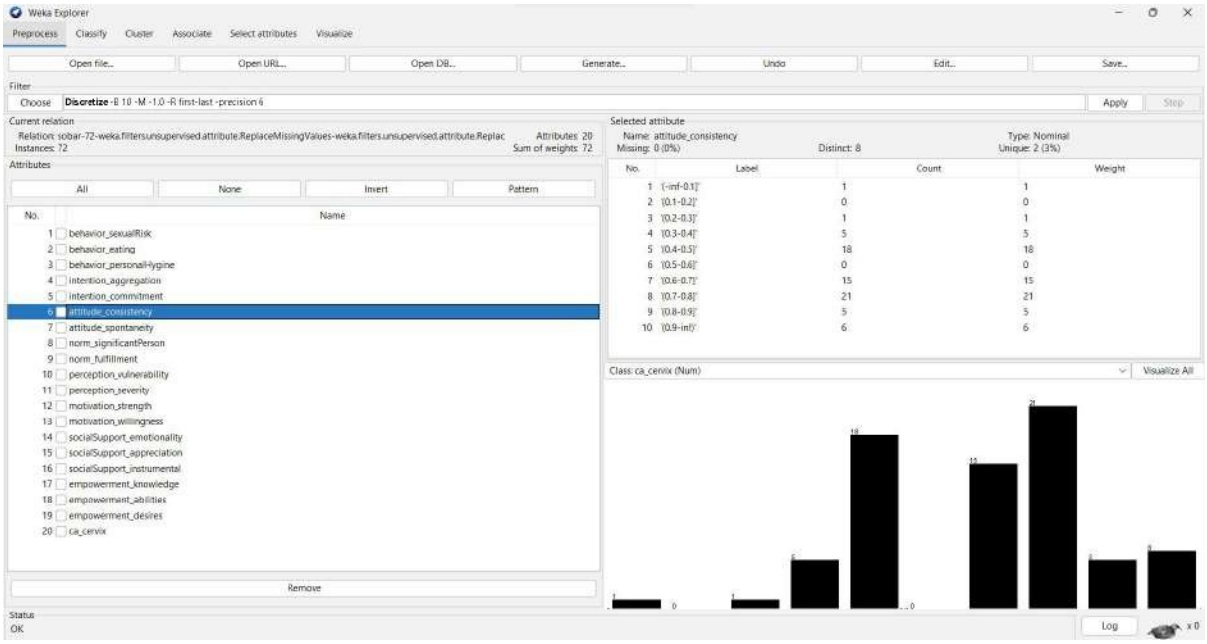


Fig 6.4:Descretize

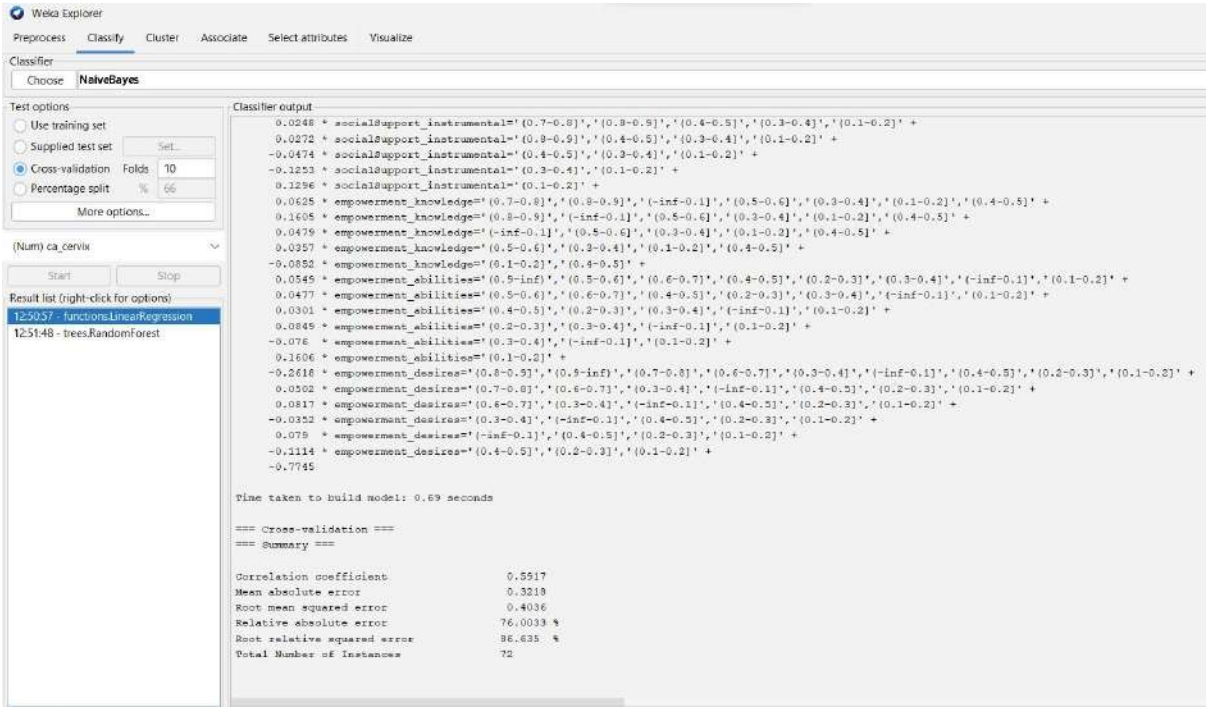


Fig 6.5:Linear Regression

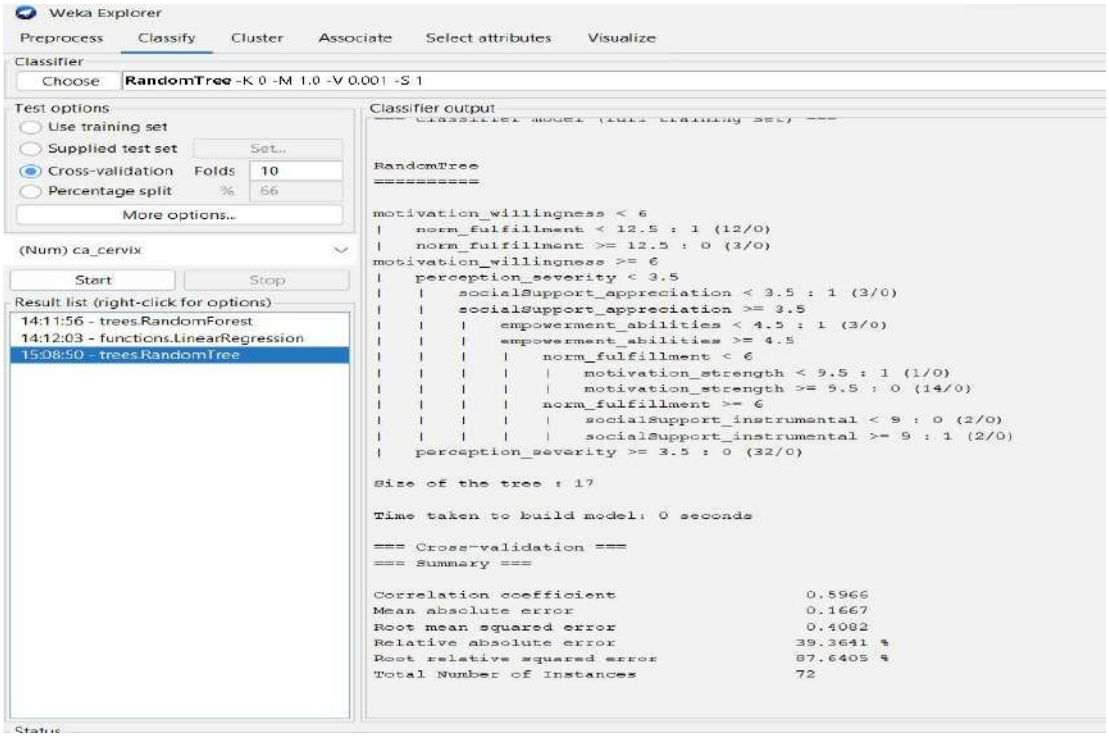


Fig 6.6:Random Forest

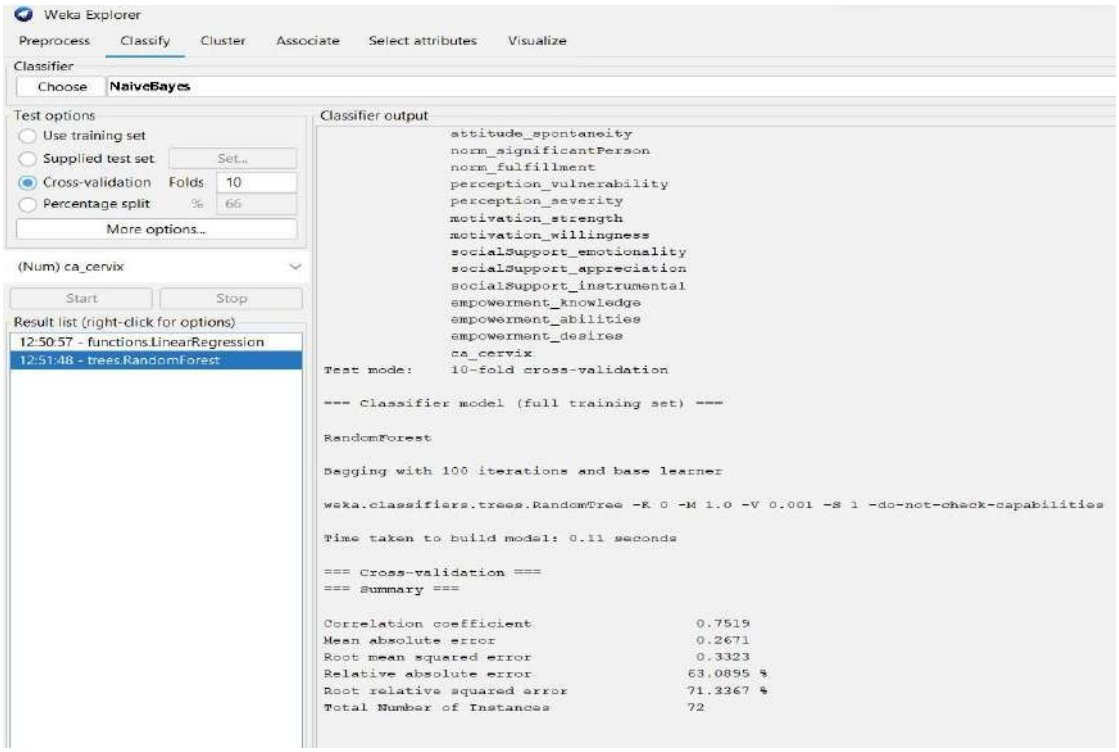


Fig 6.7:Random Tree

7. Conclusion:

In conclusion, for the cervical cancer behavior risk project utilizing the dataset from the UCI repository, it is imperative to select the most suitable algorithm to ensure accurate and reliable results. After careful consideration and analysis of the dataset characteristics such as size, complexity, and the nature of the target variable, it is recommended to employ a machine learning algorithm that can handle classification tasks effectively. Among the various algorithms available, decision trees, random forests, or support vector machines (SVMs) are promising choices due to their ability to handle both numerical and categorical data, as often found in medical datasets like the one provided by UCI. Additionally, ensemble methods such as random forests can provide robustness against overfitting and handle noisy data effectively, which is crucial for ensuring the model's generalizability. Furthermore, considering the importance of interpretability in medical decision-making, decision trees can offer insights into the factors contributing to cervical cancer behavior risk, aiding medical professionals in better understanding the underlying patterns. Therefore, based on the characteristics of the dataset and the objectives of the project, employing decision trees or ensemble methods like random forests would be the best algorithmic approach for predicting cervical cancer behavior risk accurately.