CERVICAL CANCER RISK PREDICTION

Author: Nuona Chen

Introduction

Cervical cancer is a major health concern worldwide but can be prevented with early detection and treatment. Machine learning provides powerful tools to analyze health data and identify individuals at risk.

In this project, we use the UCI Cervical Cancer Risk Factors dataset, which contains demographic, behavioral, and medical attributes, to build classification models that predict biopsy outcomes.

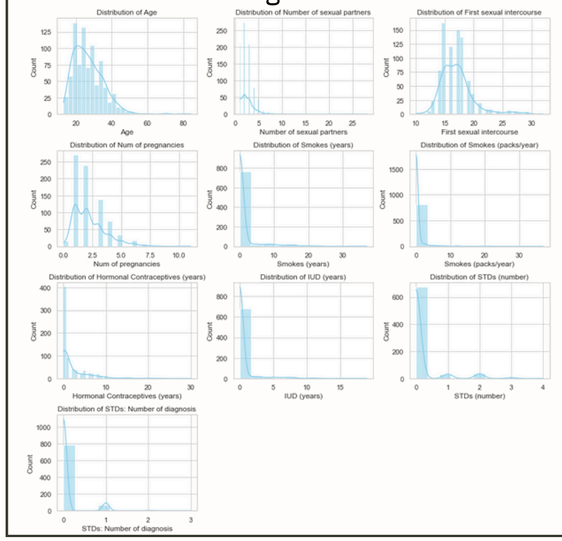
The goals are to **(1)** explore and understand the dataset, **(2)** develop predictive models for cervical cancer risk, and **(3)** evaluate performance using metrics such as accuracy, precision, recall, and ROC-AUC.

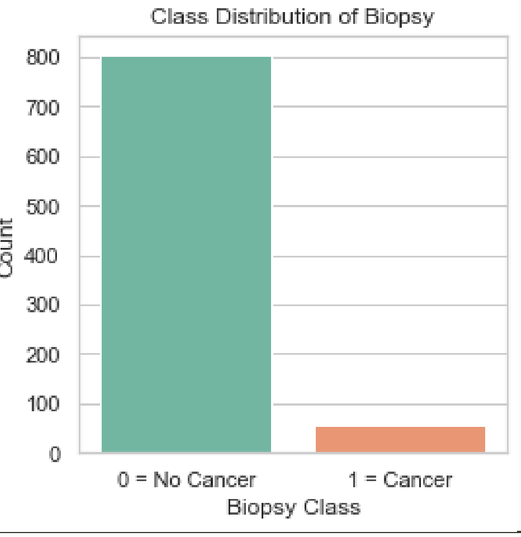
Data Description

The UCI Cervical Cancer Risk Factors dataset contains information collected from 858 patients. The dataset includes 36 attributes covering demographic details, lifestyle behaviors, medical history, and diagnostic test results. The target variables indicate whether a patient had abnormal findings from four screening methods: Hinselmann, Schiller, Cytology, and Biopsy. Among these, Biopsy is used as the main prediction target in this project.

EDA

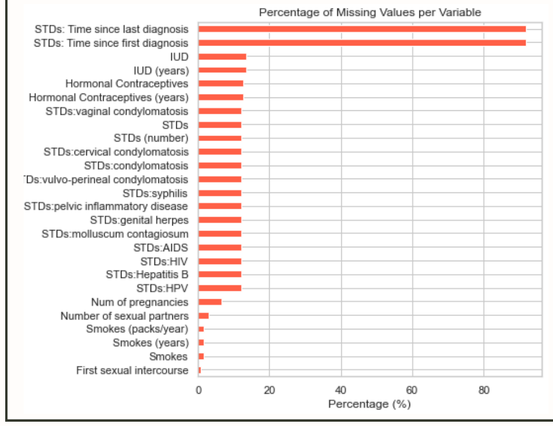
- Most continuous predictors, such as age, number of sexual partners, smoking years, contraceptive use, and number of STDs, are heavily right-skewed, with the majority of participants reporting low or zero values and only a few with high values. Categorical predictors reveal strong class imbalance: most participants do not smoke, use IUDs, or have STDs, and positive diagnoses for conditions such as HPV, cancer, and cytology tests are relatively rare. Overall, the dataset is dominated by younger participants with limited exposure to risk factors, and indicates strong skewness and imbalance.
- The outcome variable, Biopsy, exhibits strong class imbalance, with the majority of cases classified as no cancer and only a small proportion indicating cancer.

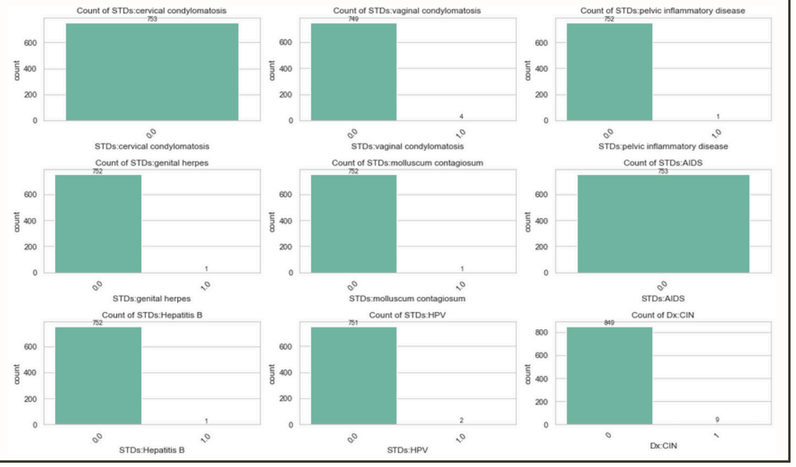




Data Preprocessing

- The variables ***STDs: Time since last diagnosis and STDs: Time since first diagnosis*** had over 80% missing values and were therefore excluded from the predictors. For the remaining features, missing values were imputed using the median for continuous variables and the mode for categorical variables.
- The categorical variables ***Count of STDs: cervical condylomatosis, vaginal condylomatosis, pelvic inflammatory disease, genital herpes, molluscum contagiosum, AIDS, Hepatitis B, and HPV, and Dx:CIN*** indicate little to no variation, with most values being the same. Because they contribute noise and have zero or near-zero variance, these variables were excluded from the predictors





Data Modeling Method

The dataset was partitioned into 80% for training and 20% for testing. To address the class imbalance in the outcome variable, **the Synthetic Minority Oversampling Technique (SMOTE)** was applied to the training set. **Decision Tree, Random Forest, and AdaBoost** models were then trained on the oversampled data. Hyperparameter tuning was conducted using 5-fold cross-validation, with recall chosen as the optimization criterion.

To optimize classification for the positive class, which represents cervical cancer cases and is of primary clinical importance, predicted probabilities from the trained models were used to generate a precision–recall curve on the test set. This approach is particularly relevant for imbalanced datasets, where the positive class occurs infrequently and standard thresholds may under-detect critical cases. The F1-score, which balances precision and recall, was calculated at each probability threshold, and the threshold that maximized the F1-score was selected as the optimal decision cutoff. This ensures that the model identifies as many true positive cases as possible while controlling the rate of false positives.

Model performance was then evaluated on the test set using accuracy, sensitivity (recall), precision, F1-score, specificity, and the area under the ROC curve (ROC-AUC).

Data Modeling Results

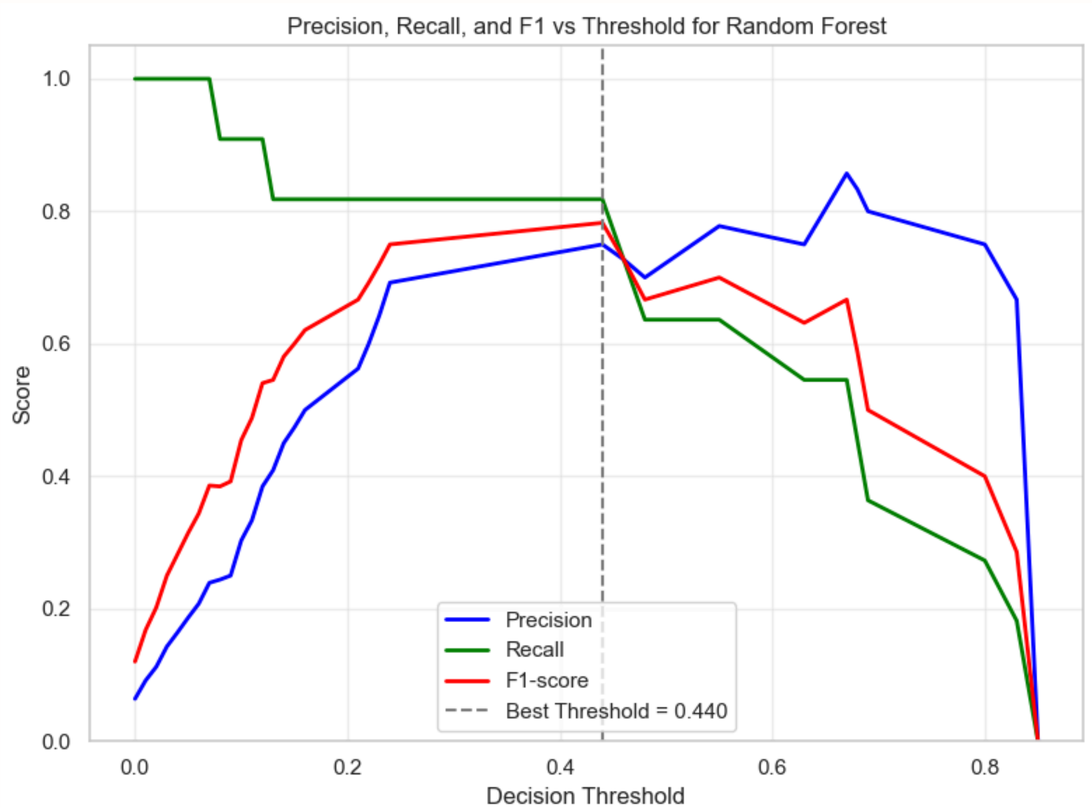
All models achieved high overall accuracy (95.9 - 97.1%) and specificity (97.5 - 98.8%). However, their performance on identifying positive cases varied. Decision Tree achieved high precision (80%) but moderate recall (72.7%), while AdaBoost showed lower recall (72.7%) and precision (66.7%). Random Forest achieved the highest recall (81.8%) with precision (75%), resulting in the highest F1-score (78.2%) and ROC AUC (0.965).

Conclusion

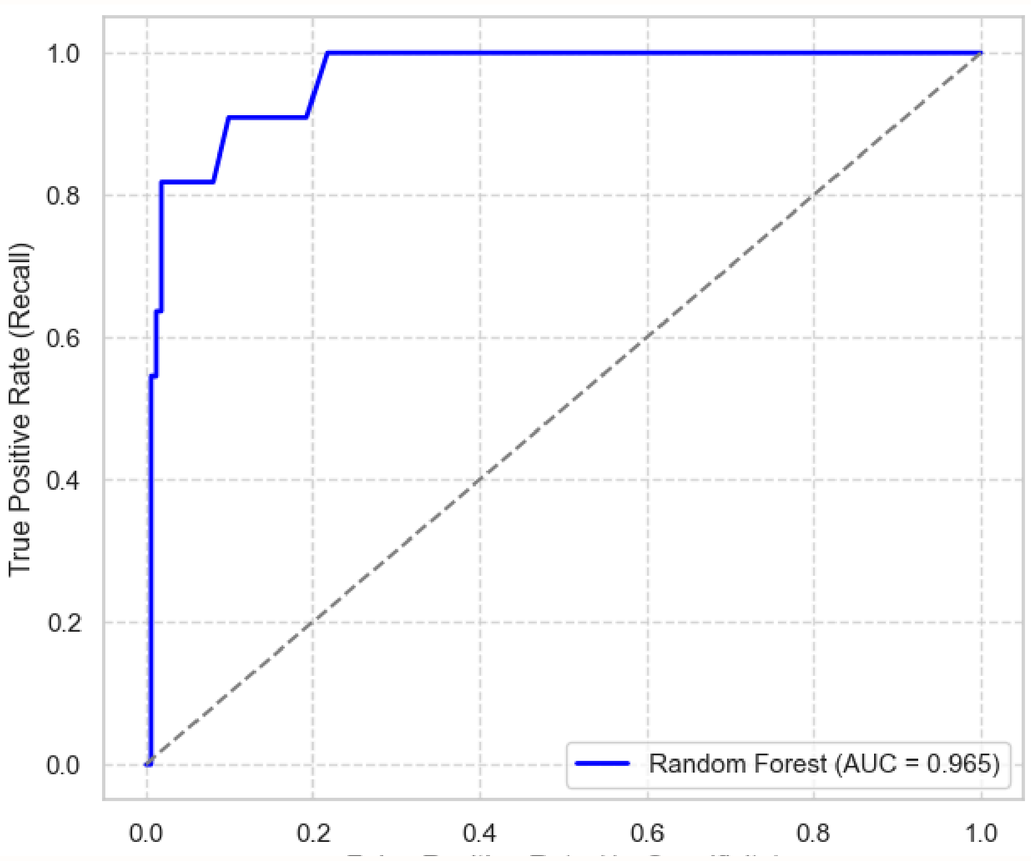
Given the primary goal of the project - to reliably identify positive cervical cancer cases, the Random Forest model was selected as the preferred classifier. Its higher sensitivity ensures that the maximum number of true positive cases are detected, which is crucial in a clinical context where missing cases can have serious consequences. At the same time, it maintains a good balance of precision and overall performance.

	Optimal Threshold	Accuracy	Recall	Specficity	Precision	F1	ROC AUC
Decision Tree	0.974	0.971	0.727	0.988	0.800	0.762	0.819
Random Forest	0.440	0.971	0.818	0.988	0.750	0.782	0.965
AdaBoost	0.488	0.959	0.727	0.975	0.667	0.696	0.940

Appendix



Appendix: ROC Curve for Random Forest



Appendix