TORONTO METROPOLITAN UNIVERSITY

PREDICTIVE MODELING OF CERVICAL CANCER RISK FACTORS

Initial Results and the Code

MONICA NGUYEN

Student #: 500839386

Supervisor: Tamar Abdou

Date: January 29, 2025

Introduction

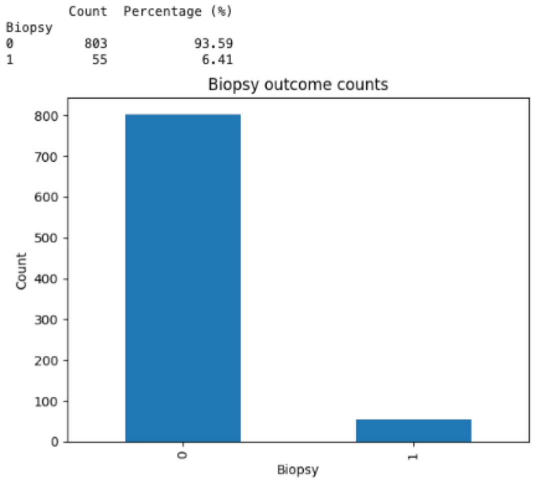
This project analyzes a clinical dataset comprising 858 patient records and 36 variables capturing demographic characteristics, sexual history, reproductive factors, contraceptive use, smoking behavior, and sexually transmitted disease related indicators. The primary objective of this analysis is to develop and evaluate predictive models for cervical cancer risk using the binary target variable 'Biopsy', which indicates whether biopsy findings suggest the presence of cervical cancer. Predicting biopsy outcomes is clinically relevant, as biopsies are invasive procedures and accurate risk stratification may support earlier identification of high-risk individuals and more efficient allocation of diagnostic resources.

Before model development, it is essential to understand the structure, quality, and distributional characteristics of the data. This includes assessing missingness patterns, identifying skewed or zero-inflated variables, and evaluating the plausibility and variability of key risk factors. The following sections therefore focus on exploratory data analysis and data preparation, with particular attention to missing data mechanisms and their implications for preprocessing and model stability.

Data Analysis

Outcome Variable Distribution (Biopsy)

The target variable for this analysis is Biopsy, a binary indicator representing whether biopsy results suggest cervical cancer risk. Examination of the outcome distribution reveals a substantial class imbalance within the dataset. Of the 858 observations, 803 cases (93.59%) correspond to negative biopsy outcomes (Biopsy = 0), while only 55 cases (6.41%) represent positive biopsy findings (Biopsy = 1).



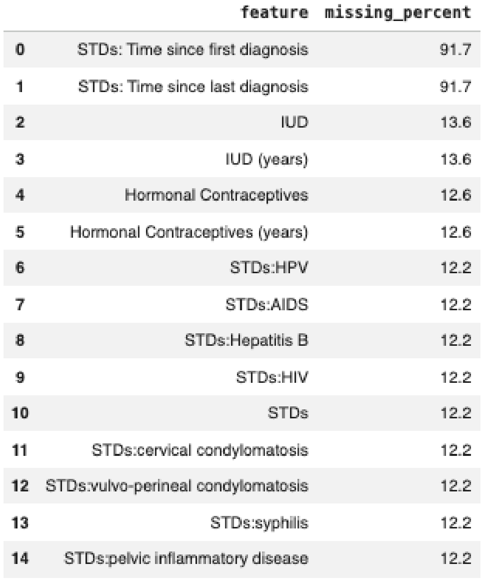
```
        Count  Percentage (%)
Biopsy
0         803           93.59
1          55            6.41
```

Figure 1: Biopsy Outcome

This pronounced imbalance has important implications for model development and evaluation. Models trained on this data may achieve high overall accuracy by predominantly predicting the majority class, while failing to correctly identify positive biopsy cases. Because false negatives in this context could delay identification of individuals at higher risk, evaluation metrics beyond accuracy are necessary. In later stages of this project, performance measures such as recall (sensitivity) and F1-score will be emphasized, and resampling or class-weighting strategies may be considered to improve the model's ability to detect minority-class cases.

Missing Data Assessment

Missingness analysis shows that several STD-related variables exhibit substantial data absence, with time-since-diagnosis measures missing in over 90% of observations. Other reproductive health variables, including IUD use and hormonal contraceptive use, display moderate missingness around 12–14%.

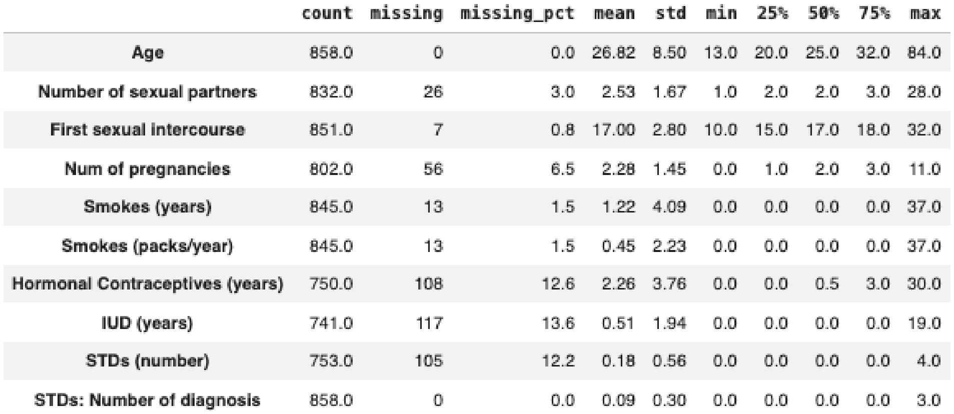


	feature	missing_percent
0	STDs: Time since first diagnosis	91.7
1	STDs: Time since last diagnosis	91.7
2	IUD	13.6
3	IUD (years)	13.6
4	Hormonal Contraceptives	12.6
5	Hormonal Contraceptives (years)	12.6
6	STDs:HPV	12.2
7	STDs:AIDS	12.2
8	STDs:Hepatitis B	12.2
9	STDs:HIV	12.2
10	STDs	12.2
11	STDs:cervical condylomatosis	12.2
12	STDs:vulvo-perineal condylomatosis	12.2
13	STDs:syphilis	12.2
14	STDs:pelvic inflammatory disease	12.2

Figure 2: Statistics of Missing

This pattern is consistent with structural missingness arising from questionnaire skip logic, as these follow-up timing questions are only applicable to participants who report a prior STD diagnosis. In addition, STD-related questions are inherently sensitive, and partial nonresponse may reflect respondent discomfort or reluctance to disclose personal health information. As a result, these patterns indicate that naïve imputation would be inappropriate for certain features and support the use of missingness indicators, variable exclusion, or modeling strategies that explicitly account for informative missingness.

Descriptive Statistics of Numeric Risk Factors



	count	missing	missing_pct	mean	std	min	25%	50%	75%	max
Age	858.0	0	0.0	26.82	8.50	13.0	20.0	25.0	32.0	84.0
Number of sexual partners	832.0	26	3.0	2.53	1.67	1.0	2.0	2.0	3.0	28.0
First sexual intercourse	851.0	7	0.8	17.00	2.80	10.0	15.0	17.0	18.0	32.0
Num of pregnancies	802.0	56	6.5	2.28	1.45	0.0	1.0	2.0	3.0	11.0
Smokes (years)	845.0	13	1.5	1.22	4.09	0.0	0.0	0.0	0.0	37.0
Smokes (packs/year)	845.0	13	1.5	0.45	2.23	0.0	0.0	0.0	0.0	37.0
Hormonal Contraceptives (years)	750.0	108	12.6	2.26	3.76	0.0	0.0	0.5	3.0	30.0
IUD (years)	741.0	117	13.6	0.51	1.94	0.0	0.0	0.0	0.0	19.0
STDs (number)	753.0	105	12.2	0.18	0.56	0.0	0.0	0.0	0.0	4.0
STDs: Number of diagnosis	858.0	0	0.0	0.09	0.30	0.0	0.0	0.0	0.0	3.0

Figure 3: Descriptive Statistics of Numeric Risk Factors

Descriptive statistics highlight important differences in completeness, variability, and distributional shape across predictors. Demographic and sexual history variables are largely complete and exhibit stable central tendencies, whereas behavioral and reproductive variables show greater dispersion and right-skewness, with some extreme values. Smoking, contraceptive use, and STD-related variables are characterized by substantial zero inflation and higher missingness, reflecting heterogeneous exposure and reporting patterns. These distributional features underscore the need for robust preprocessing strategies, including appropriate imputation methods and modeling approaches that can accommodate skewed and zero-inflated predictors.

Distributional Patterns and Outliers

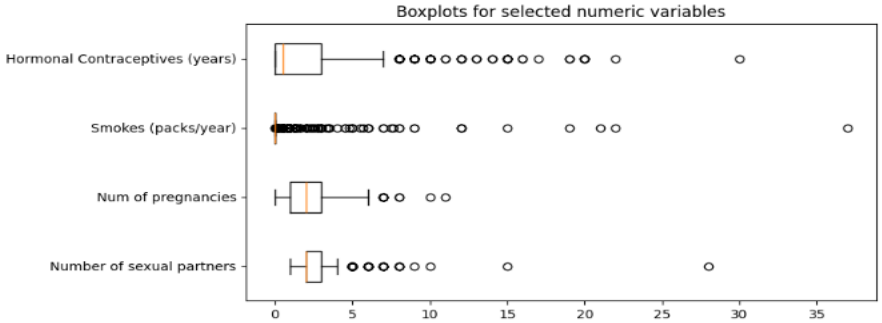


Figure 4: Boxplots for Selected Numeric Variables

Boxplots were generated for selected numeric variables that are continuous or count-based, clinically meaningful, and expected to exhibit skewness and outliers. This visualization enables efficient assessment of distributional characteristics prior to modeling. The variables 'Hormonal Contraceptives (years)', 'smokes (packs/year)', 'number of pregnancies', and 'number of sexual partners' have been identified in the epidemiologic and predictive modeling literature as important risk factors or predictors for cervical cancer (Sun et al, 2022). In the machine learning study by Sun *et al.* (2022), duration of hormonal contraceptive use was the most important predictive feature, followed by number of pregnancies, smoking exposure, and number of sexual partners, when identifying women at high risk of cervical cancer. The boxplots reveal substantial right skewness and high-end outliers in these measures, with most observations clustered at low values and a subset of individuals exhibiting markedly higher exposures. The narrow interquartile ranges relative to the full value ranges indicate limited typical exposure alongside extreme but infrequent cases. These patterns suggest that the observed outliers are likely meaningful rather than data errors and underscore the need for modeling approaches that are robust to skewed distributions and influential observations.

Implications for Modeling

Tree-based models are well suited to this dataset given both its empirical characteristics and prior evidence from cervical cancer risk modeling literature. As shown in the descriptive and visual analyses, many predictors exhibit pronounced right-skewness, zero inflation, and extreme values, particularly for smoking exposure, contraceptive duration, and STD-related variables. Tree-based methods do not assume linearity or normality and instead rely on threshold-based splits, allowing them to naturally accommodate skewed distributions and heterogeneous exposure levels without extensive transformation (Hastie et al., 2009).

This modeling choice is further supported by prior studies using the UCI Cervical Cancer Risk Factors dataset (Fernandes et al., 2017). Several comparative analyses report superior performance of decision trees and ensemble methods relative to linear and distance-based classifiers. For example, Asadi et al. (2017) and Alsmariy et al. (2020) found that decision tree–based models outperformed logistic regression and k-nearest neighbors when predicting cervical cancer outcomes, largely due to their ability to capture nonlinear relationships among behavioral and reproductive factors. Random Forest models, introduced by Breiman (2001), further improve predictive performance by aggregating multiple decision trees, reducing variance, and increasing robustness in noisy or small clinical datasets.

Ensemble methods such as Random Forest and gradient boosting have also been shown to handle imbalanced medical data effectively, particularly when class weighting or cost-sensitive learning is applied, as noted by Chen and Guestrin (2016) in the development of XGBoost. These properties are especially relevant in cervical cancer screening contexts, where positive outcomes are rare and minimizing false negatives is a clinical priority.

From a clinical perspective, tree-based models align well with the multifactorial nature of cervical cancer risk. Risk is unlikely to arise from isolated predictors and instead reflects interactions among age, sexual history, smoking behavior, and STI exposure. Decision trees and their ensembles can implicitly identify such interaction effects without explicit specification, supporting Research Question 3, which focuses on uncovering nonlinear or previously unrecognized combinations of risk factors. Similar conclusions regarding interaction-driven risk modeling in cervical cancer have been highlighted in epidemiological reviews by Bosch et al. (2002) and subsequent predictive modeling studies.

Data Preparation

Data preparation was conducted to ensure data integrity, prevent information leakage, and create a modeling-ready dataset aligned with the predictive objective. Diagnostic and post-screening variables that would not be available at the time of risk prediction were first removed to prevent data leakage. Features with excessive missingness (greater than 90%) were also excluded, as they offered limited utility and were likely subject to structural or sensitivity-driven missingness. Following feature exclusion, the dataset was separated into predictors and the target outcome, and remaining features were categorized as binary or numeric based on their measurement properties.

Missing values were then imputed using feature-type–specific strategies. Numeric variables were imputed using the median to reduce sensitivity to skewed distributions and extreme values, while binary variables were imputed using the most frequent value to preserve categorical meaning. Finally, numeric features were standardized using z-score normalization to ensure comparability across predictors and prevent scale-driven bias during model training. Together, these preprocessing steps produced a cleaned, stable, and interpretable dataset suitable for downstream classification and evaluation.

Model Evaluation

Model	Accuracy	Precision (Positive)	Recall (Positive)	F1 Score (Positive)	ROC-AUC
Logistic Regression (Unweighted)	0.9349	0.00	0.00	0.00	0.6542
Logistic Regression (Class-Weighted)	0.7442	0.1273	0.50	0.2029	0.6624
Decision Tree (max depth = 5)	0.9116	0.00	0.00	0.00	0.4534
Random Forest (Balance)	0.9349	0.50	0.0714	0.125	0.6953
XGBoost (Class-Weighted)	0.8558	0.0952	0.1429	0.1143	0.5437

Table 1: Model Performance for Cervical Cancer Risk Prediction

Summary of Results

Table 1 summarizes the performance of all evaluated classification models using accuracy, precision, recall, F1 score, and ROC–AUC. Recall was emphasized due to severe class imbalance, with biopsy-positive cases representing only 6.4% of the dataset.

The unweighted logistic regression model achieved high overall accuracy (93.5%) but failed to identify any positive biopsy cases, resulting in zero recall for the minority class. This outcome demonstrates the inadequacy of accuracy as a primary evaluation metric in highly imbalanced clinical datasets.

Introducing class weighting substantially improved sensitivity. Class-weighted logistic regression achieved a recall of 0.50, correctly identifying half of positive cases, although this improvement came at the cost of reduced accuracy and increased false positives. Cross-validated recall for this model averaged approximately 0.31, indicating moderate and variable sensitivity across folds.

Tree-based models showed mixed performance. A shallow decision tree and a balanced random forest achieved high accuracy but exhibited very low recall for the positive class, suggesting continued bias toward majority-class predictions under default thresholds. XGBoost with class weighting demonstrated modest improvements in recall compared to other tree-based models but did not outperform class-weighted logistic regression in its current configuration.

Overall, class-weighted logistic regression provided the best balance between sensitivity and interpretability among the models evaluated.

References

Alsmariy, R., Healy, G., & Abdelhafez, H. (2020). Predicting Cervical Cancer using Machine Learning Methods. *International Journal of Advanced Computer Science and Applications, 11(7).* http://dx.doi.org/10.14569/IJACSA.2020.0110723

Asadi, H., Dowling, R., Yan, B., & Mitchell, P. (2014). Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PloS one*, *9*(2), e88225. https://doi.org/10.1371/journal.pone.0088225

Bosch, F. X., Lorincz, A., Muñoz, N., Meijer, C. J., & Shah, K. V. (2002). The causal relation between human papillomavirus and cervical cancer. *Journal of clinical pathology*, *55*(4), 244–265. https://doi.org/10.1136/jcp.55.4.244

Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Fernandes, K., Cardoso, J., & Fernandes, J. (2017). Cervical Cancer (Risk Factors) [Dataset]. *UCI Machine Learning Repository.* https://doi.org/10.24432/C5Z310

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). *Springer*.

Sun, L., Yang, L., Liu, X., Tang, L., Zeng, Q., Gao, Y., Chen, Q., Liu, Z., & Peng, B. (2022). Optimization of Cervical Cancer Screening: A Stacking-Integrated Machine Learning Algorithm Based on Demographic, Behavioral, and Clinical Factors. *Frontiers in Oncology*, *12*, 821453. https://doi.org/10.3389/fonc.2022.821453