TORONTO METROPOLITAN UNIVERSITY

PREDICTIVE MODELING OF CERVICAL CANCER RISK FACTORS

Final Results and Project Report

MONICA NGUYEN

Student #: 500839386

Supervisor: Tamar Abdou

Date: January 29, 2025

ABSTRACT

Cervical cancer affects women worldwide, yet an estimated 94% of deaths occur in low- and middle-income countries, where access to HPV vaccination, screening, and treatment remains limited (World Health Organization, 2025). Using the UCI Cervical Cancer Risk Factors dataset (n = 858), this study trained classification models on routinely collected lifestyle and medical history variables, including sexual and reproductive history, contraceptive use, smoking exposure, and STD indicators (Fernandes et al., 2017). The outcome of interest was biopsy result, coded as 1 for biopsy-positive and 0 for biopsy-negative. Because the dataset exhibits severe class imbalance (about 6.4% positive cases) and substantial missingness across multiple predictors, preprocessing focused on standardizing missing value encodings, applying imputation, and removing diagnostic or post-screening variables to reduce information leakage. Variables with extreme missingness were also dropped to improve reliability. Feature importance results suggested that duration of hormonal contraceptive use, along with key reproductive and sexual history measures, ranked among the strongest predictors of biopsy positivity.

Table of Contents

1. Introduction

Most cervical cancer cases are linked to persistent infection with high-risk human papillomavirus (HPV). HPV is commonly transmitted through sexual contact, and in many people the immune system clears the infection. However, in some cases the infection persists and can lead to cancer over time. Screening and early detection can prevent many cases and reduce deaths by detecting precancerous changes before they progress (World Health Organization, 2025). Because risk is influenced by a combination of behaviors and long-term exposures, risk prediction based on patient history can support more focused screening strategies.

In recent years, data mining and machine learning have been increasingly used to support medical decision-making by identifying patterns in health data. Prior work applying classification methods to cervical cancer risk factor data consistently highlights two major challenges. First, strong class imbalance can bias model learning toward the majority class and reduce classification quality (Sun et al., 2022). Second, high levels of missingness can distort both training and evaluation if not handled carefully (Muraru et al., 2024). These challenges are also present in the dataset used for this capstone.

This study tests whether machine learning models can provide useful risk predictions for screening, with the goal of supporting women in low- and middle-income countries where screening and follow-up resources are limited. These models could help identify higher-risk women earlier and prioritize them for testing and referral, so more people receive the right care sooner. The analysis is guided by three research questions:

I. Which lifestyle and medical history variables are most strongly associated with cervical cancer risk based on statistical significance and feature importance score from predictive models?

II. How do selected machine learning models compare in classifying individuals into high-risk and low-risk groups under severe class imbalance?

III. To what extent can machine learning models identify nonlinear interactions or previously unrecognized variable combinations that are not captured by traditional clinical risk scoring systems?

This paper evaluates multiple classification models, including logistic regression, decision tree, random forest, and XGBoost, for predicting cervical cancer risk. It first summarizes related work on cervical cancer risk prediction and common data mining approaches, then describes the dataset and methodology, including preprocessing, model training, and evaluation metrics. It then presents and interprets the results and closes with the main conclusions and directions for future research.

2. Literature Review and Integration

Across models and feature importance results, the strongest predictors in this study were duration of hormonal contraceptive use, number of pregnancies, and sexual history measures. This pattern aligns with prior machine learning studies using the UCI Cervical Cancer Risk Factors dataset, where reproductive and sexual history variables often emerge as top contributors because they reflect long term HPV exposure. Tree-based and ensemble methods are also favoured for modeling nonlinear relationships that involve thresholds or plateaus, where risk increases sharply at certain points and then stabilizes (Ijaz et al., 2020; Breiman, 2001; Hastie et al., 2009; Al Mudawi & Alazeb, 2022).

Comparisons across prior studies similarly emphasize reproductive and sexual history variables as key drivers of prediction, and they frame performance gains as partly dependent on how preprocessing addresses rare outcomes and noisy observations (Ijaz et al., 2020). Sun et al. (2022) report a comparable ordering of influential predictors in the same Venezuelan screening dataset, with hormonal contraceptive duration and pregnancy history among the most important variables, alongside smoking exposure and sexual partner measures, as shown in Figure 1. They

also explain that stronger ensemble strategies can improve overall discrimination relative to simpler baselines, reinforcing that model choice and imbalance handling can meaningfully change screening performance (Sun et al., 2022).

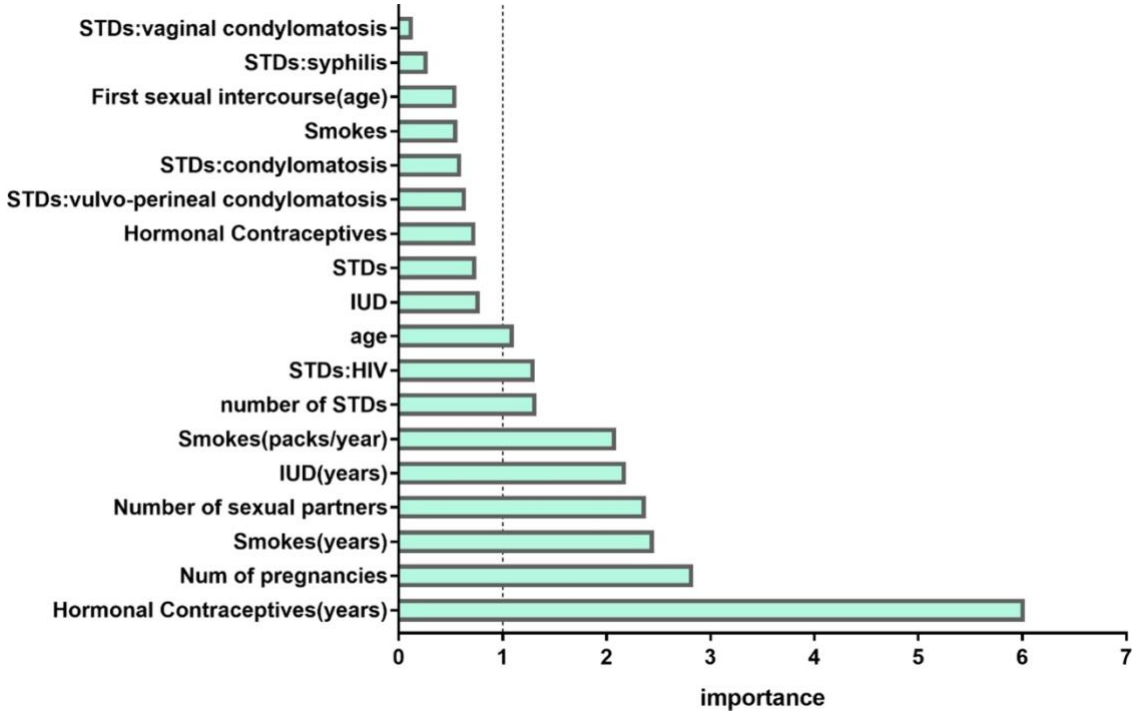

Figure 1. Feature Importance Measures by Sun et al., 2022

In contrast, individual STD indicator variables contributed less to prediction in this analysis, which is also consistent with prior discussions of this dataset (Ijaz et al., 2020; Sun et al., 2022). A likely explanation is that STD related questions are sensitive and can be underreported, unknown, or missing, which reduces their reliability for models.

Overall, the literature helps explain why results can differ across papers using the same dataset. Performance and feature rankings can shift slightly depending on how missingness is handled, whether class imbalance is addressed through weighting or resampling, and whether models are evaluated with screening-relevant metrics that emphasize sensitivity and minority-class detection (Ijaz et al., 2020; Sun et al., 2022). Because this dataset contains substantial missingness and a

small positive class, models that do not explicitly account for imbalance often appear strong on overall accuracy while failing to detect biopsy-positive cases, which reinforces the need to interpret results using recall and ROC–AUC rather than accuracy alone.

3. Methodology

Data Source

This study uses the UCI Cervical Cancer (Risk Factors) dataset, which contains records from 858 patients screened at Hospital Universitario de Caracas in Caracas, Venezuela. The dataset includes demographic, lifestyle, and clinical history variables such as sexual and reproductive history, contraceptive use, smoking exposure, and indicators related to sexually transmitted diseases. The primary outcome is the biopsy result, treated as a binary target where 1 indicates biopsy-positive and 0 indicates biopsy-negative (Fernandes, 2017). Figure 2 summarizes the end-to-end workflow used in this study.

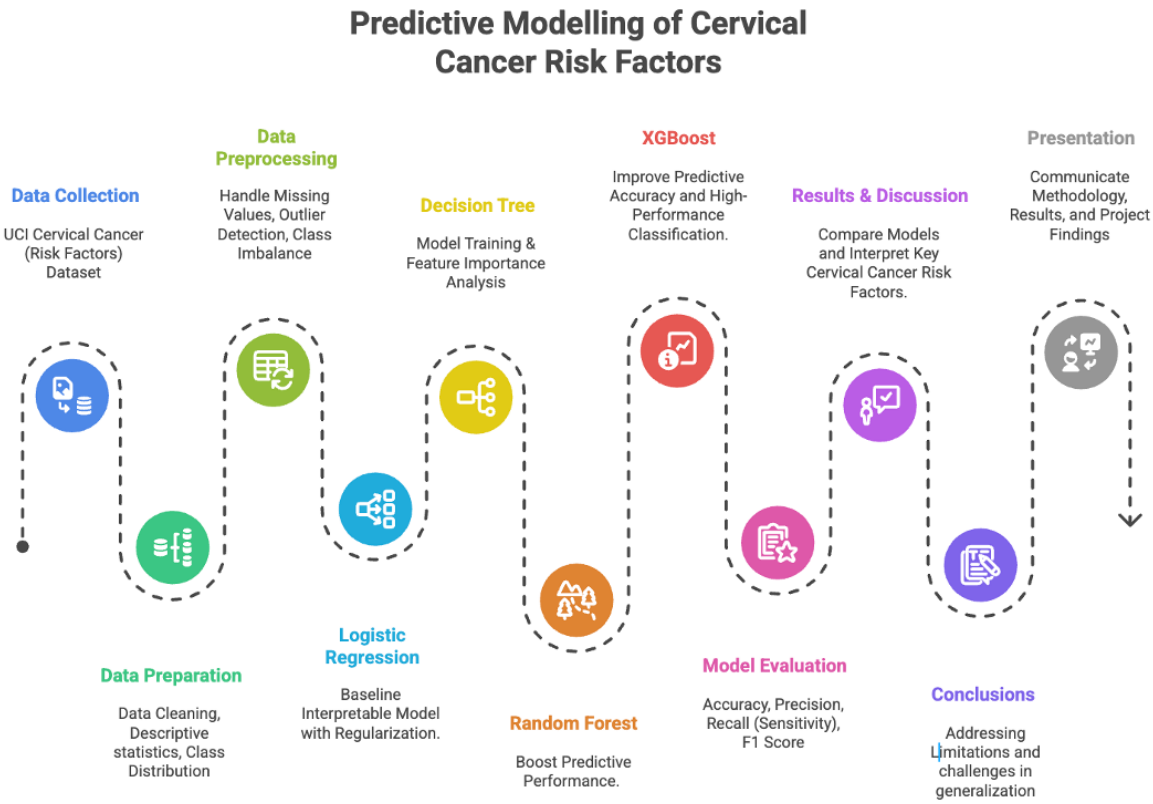

Figure 2. Methodology

The process begins with data preparation and preprocessing to address missing values, outliers, and severe class imbalance. Several supervised learning models were then trained, including logistic regression as an interpretable baseline and tree-based methods (decision tree, random forest, and XGBoost) to capture nonlinear relationships. Model performance was evaluated using classification metrics, and findings were interpreted through feature importance and comparative discussion. This workflow ensures that all models are assessed consistently and supports the study's goals of identifying influential risk factors and comparing predictive performance.

Data Preprocessing

Preprocessing addressed two key challenges in the dataset: substantial missingness across predictors and severe class imbalance in the target variable. Figure 3 highlights this imbalance, with biopsy-positive cases representing only 55 of 858 observations (6.41%), compared with 803 biopsy-negative cases (93.59%). This matters because a model can achieve high overall accuracy by predicting the majority class most of the time while still failing to detect many true positive cases.

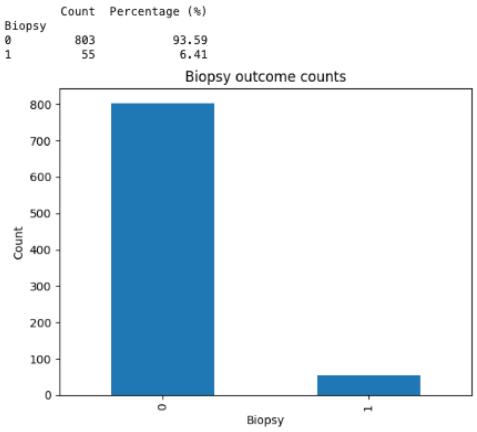

Figure 3: Class Distribution of the Target Variable (Biopsy)

Missing values were first standardized and quantified for each variable to ensure consistent handling across the dataset. Variables with extreme missingness, defined as more than 90% missing observations, were removed due to low reliability and increased risk of introducing noise. Table 1 summarizes missingness by variable and identifies the features removed using this threshold.

	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

Table 1: Top Variables by Missingness (%)

To reduce information leakage and ensure the model reflects a realistic screening setting, variables representing diagnostic test results or post-screening outcomes were removed prior to model training. Specifically, the columns Hinselmann, Schiller, Citology, Dx, Dx:Cancer, Dx:CIN, and Dx:HPV were excluded because they contain information collected after or alongside clinical assessment that would not be available at the time of initial risk screening and could artificially inflate performance. This resulted in a final feature set of 27 predictors used for subsequent train-test splitting, imputation, and model evaluation.

Model Selection

Multiple models were used to balance interpretability with predictive performance and to test whether nonlinear patterns improve prediction under severe class imbalance. **Unweighted logistic regression** was included as a baseline because it is easy to interpret and shows how a standard model behaves without any imbalance adjustments. **Class-weighted logistic regression** was included to improve detection of biopsy-positive cases by giving more importance to the minority class during training, which is crucial for a screening-style problem.

Tree-based models were included because they can capture nonlinear patterns and cutoff effects without strong distribution assumptions. A **decision tree (max depth = 5)** was used as a simple, interpretable nonlinear baseline. The depth limit keeps the tree readable and reduces overfitting, since deeper trees can memorize noise, especially when positive cases are rare. A **balanced random forest** was included to improve performance and stability by averaging many trees and reducing bias toward the majority class through balancing during training. **Class-weighted XGBoost** was included because boosting can capture complex nonlinear relationships and interactions, and class weighting helps the model pay more attention to the rare positive class.

Tree-based models are especially suitable for this dataset because several predictors show right skewness, many zeros, and wide ranges, particularly for contraceptive duration, smoking exposure, number of pregnancies, and number of sexual partners, as shown in Figure 4. These characteristics can weaken linear model assumptions, while split-based methods adapt naturally to uneven distributions.

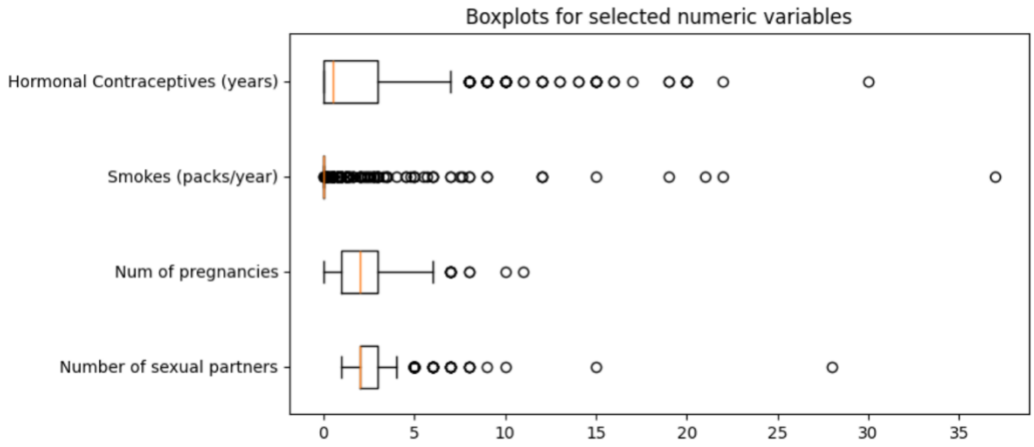

Figure 4: Distribution and Outliers for Selected Numeric Variables

Training & Testing

The dataset was split into training and testing sets using a 75/25 ratio to provide enough data for model learning while reserving a sufficiently large holdout set for an unbiased estimate of performance. With 858 observations and a low positive rate (about 6.4%), allocating 25% to testing preserves a meaningful number of positive cases in the test set for evaluation while still leaving most observations for training. More extreme splits can be less stable in this setting. For example, an 80/20 split reduces the number of positive cases available for testing, making recall and precision estimates noisier, while a 70/30 split further reduces the training data available to learn minority-class patterns. The 75/25 split is a practical compromise that supports both model fitting and reliable assessment of generalization.

Evaluation Metrics

Models were evaluated using accuracy, precision, recall (sensitivity), F1 score, and ROC–AUC. Accuracy was reported for completeness, but it was not treated as the primary indicator of success because high accuracy can occur even when a model fails to identify any biopsy-positive cases. Recall was prioritized because the task resembles a screening context where missing true positives is especially costly. Precision was included to measure how many predicted positives were correct, and F1 score was used to summarize the balance between precision and recall. ROC–AUC was reported to show how well the model separates biopsy-positive and biopsy-negative cases across many different decision thresholds, which is important because it indicates whether the model can truly distinguish the two groups even when the positive class is rare.

4. Model Evaluation

Model evaluation was conducted across three dimensions: effectiveness, efficiency, and stability. Effectiveness assesses how well each model identifies biopsy-positive cases using screening-relevant metrics. Efficiency evaluates computational cost in terms of runtime and memory usage. Stability examines how consistent model performance remains when the data are resampled through cross-validation and repeated train test splits.

Effectiveness

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 (Balanced)	0.9349	0.50	0.0714	0.125	0.6953
XGBoost (Class-Weighted)	0.8558	0.0952	0.1429	0.1143	0.5437

Table 2: Model Performance for Cervical Cancer Risk Prediction

As shown in Table 2, the unweighted logistic regression achieved high accuracy (0.9349) but produced zero recall and zero F1 score, indicating that it effectively predicted the majority class but failed to identify any positive cases. Applying class weights changed model behavior substantially. Class-weighted logistic regression reduced accuracy (0.7442) but achieved the highest recall (0.50), demonstrating stronger sensitivity to biopsy-positive cases at the expense of increased false positives. Among the tree-based models, the balanced random forest maintained high accuracy (0.9349) and achieved the highest ROC–AUC (0.6953), suggesting stronger overall discrimination than other models, although its recall remained low (0.0714). XGBoost (class-weighted) produced intermediate results, with moderate accuracy (0.8558) but limited recall (0.1429) and a lower ROC–AUC (0.5437), indicating weaker separation between classes in this configuration. Decision tree (max depth = 5) also produced zero recall and a low ROC–AUC (0.4534), which is consistent with underfitting.

Overall, these results show that accuracy alone is not sufficient for this imbalanced screening task. Class weighting improved sensitivity to positive cases, and the balanced random forest showed the best overall separation, but additional threshold tuning would likely be needed to improve recall.

Efficiency

	model	train_time_s	predict_time_s	train_peak_mem_mb	predict_peak_mem_mb
0	LogReg (unweighted)	0.091671	0.007873	0.265723	0.097777
1	LogReg (class-weighted)	0.135978	0.006325	0.259101	0.088286
2	Decision Tree	0.016990	0.006177	0.198825	0.066485
3	Random Forest	3.747880	0.152885	0.602404	0.087801
4	XGBoost	0.224381	0.010030	0.063367	0.054156

Table 3: Model Runtime and Memory Usage

Table 3 summarizes computational efficiency across the four classification models by reporting training time, prediction time, and peak memory usage during both training and inference. These measures help quantify the practical cost of each approach, which is important when selecting a model for repeated evaluation, cross-validation, or deployment.

Decision Tree and both versions of Logistic Regression were the fastest and used relatively little memory. Random Forest was the most expensive, with the longest training time and the slowest prediction time, reflecting the cost of building many trees. XGBoost fell in between, training much faster than Random Forest while keeping prediction time low and using the least memory overall, highlighting the trade-off between performance and efficiency when choosing a final model.

Stability

Stability was evaluated to see how consistent each model's performance is when the data are resampled. This matters in a highly imbalanced dataset because there are so few biopsy-positive cases that small changes in which positives fall into each split can noticeably change recall, precision, and F1 score. We assessed robustness using stratified cross-validation and repeated train–test splits, and summarized results using the mean and standard deviation across runs. The

mean shows average performance, while the standard deviation shows how much results vary, with lower values indicating greater stability.

	model	accuracy_mean	accuracy_std	precision_mean	precision_std	recall_mean	recall_std	f1_mean	f1_std	roc_auc_mean	roc_auc_std
1	LogReg (class-weighted)	0.728390	0.021952	0.082694	0.026268	0.327273	0.123315	0.131686	0.042845	0.542609	0.076904
4	XGBoost	0.875350	0.023467	0.126905	0.068245	0.163636	0.120605	0.136862	0.079436	0.619462	0.035673
3	Random Forest	0.933578	0.008619	0.373333	0.396877	0.072727	0.068030	0.115476	0.108039	0.671104	0.038674
2	Decision Tree	0.931253	0.009898	0.200000	0.400000	0.018182	0.036364	0.033333	0.066667	0.447991	0.113336
0	LogReg (unweighted)	0.934734	0.002257	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.542087	0.108757

Table 4: Stability Results

Table 4 shows that unweighted logistic regression achieved high mean accuracy (0.935) but had zero precision, recall, and F1 score, indicating it mostly predicted the negative class. In contrast, class-weighted logistic regression substantially improved sensitivity, achieving the highest mean recall (0.327), although accuracy decreased (0.728). Decision Tree and Random Forest maintained high mean accuracy (about 0.931–0.934) but had low recall (0.018 and 0.073), meaning they still missed most positive cases. Random Forest produced the highest mean ROC–AUC (0.671), suggesting the strongest overall separation between classes even though default-threshold recall remained low. XGBoost showed a middle trade-off, with moderate recall (0.164), the highest F1 (0.137), and ROC–AUC of 0.619.

Overall, these stability results suggest that class-weighted logistic regression is the best option when the priority is detecting biopsy-positive cases, while random forest provides the strongest overall discrimination and can be useful for understanding which predictors contribute most to risk ranking.

5. Findings and Interpretation

Key findings and Results Analysis

Across all evaluated models, unweighted approaches achieved high overall accuracy but failed to reliably identify biopsy-positive cases. This finding reinforces that accuracy alone is insufficient

for evaluating screening-oriented models. When class imbalance was addressed, model behavior shifted substantially. Class-weighted models consistently improved recall and F1 score, indicating a greater ability to detect high-risk individuals. Although this improvement was accompanied by an increase in false positives, the trade-off aligns with the priorities of early detection and screening applications.

To interpret which predictors drove model separation, permutation importance was computed for the balanced random forest because it had the highest ROC-AUC score. This approach measures how much performance drops when one feature is randomly shuffled, breaking its relationship with the outcome while keeping its distribution unchanged. Table 5 indicates that the highest feature importance score was observed for hormonal contraceptive duration, number of pregnancies, number of sexual partners, and age at first sexual intercourse, suggesting the model relied primarily on reproductive and sexual history variables to distinguish biopsy-positive from biopsy-negative cases. STD-related measures contributed modest additional information, while several specific STD subtype indicators showed near-zero importance, implying minimal value in this fitted model, likely due to high missingness in the dataset.

	Feature	Permutation_Importance
8	Hormonal Contraceptives (years)	0.081734
3	Num of pregnancies	0.044670
1	Number of sexual partners	0.041693
2	First sexual intercourse	0.039064
0	Age	0.031743
7	Hormonal Contraceptives	0.021464
25	STDs: Number of diagnosis	0.020833
22	STDs:HIV	0.013291
11	STDs	0.009337
10	IUD (years)	0.008866
12	STDs (number)	0.003429
17	STDs:syphilis	0.002674
15	STDs:vaginal condylomatosis	0.000355
19	STDs:genital herpes	0.000000
21	STDs:AIDS	0.000000

Table 5: Permutation Feature Importance Results

Interpretation in Relation to Research Questions

I. Which lifestyle and medical history variables are most strongly associated with cervical cancer risk based on statistical significance and feature importance score from predictive models?

Across exploratory analysis and model-based importance, the variables that are most strongly associated with cervical cancer risk are duration of hormonal contraceptive use, reproductive history, and sexual history measures. The permutation importance results reinforce this interpretation by showing that shuffling these variables produces the largest decreases in ROC–AUC, indicating that the model relies on them most for separating biopsy-positive from biopsy-negative cases.

II. How do selected machine learning models compare in classifying individuals into high-risk and low-risk groups under severe class imbalance?

The results show that standard classifiers can achieve high accuracy while failing to identify any positives, as seen in unweighted logistic regression and the decision tree. Class-weighted logistic regression produced the strongest sensitivity on the test set (recall = 0.50) and maintained the highest average recall across resampling (0.327), making it the most aligned with a screening goal. However, a recall of 0.50 still indicates that roughly half of biopsy-positive cases would be missed in this test split, highlighting substantial room for improvement before deployment in real screening workflows. In practical terms, this level of sensitivity suggests the model could be considered only as a supportive risk-flagging tool that requires conservative thresholds and continued clinical oversight rather than a standalone decision system. The balanced random forest offered the best overall discrimination, with the highest ROC–AUC on the test set (0.6953) and on

average (0.671), which supports its use for ranking patients by risk. XGBoost fell between these approaches, showing modest recall and weaker discrimination in the tested configuration.

III. To what extent can machine learning models identify nonlinear interactions or previously unrecognized variable combinations that are not captured by traditional clinical risk scoring systems?

The balanced random forest achieved a higher ROC–AUC than logistic regression and the shallow decision tree, suggesting that some risk factors relate to biopsy outcome in nonlinear ways. This implies that risk may increase after certain cutoffs and may also depend on combinations of factors rather than any single variable alone. Although this study did not directly quantify specific interactions, the results indicate that ensemble models can capture added structure in the data that simpler models miss, and this structure could be explored further to clarify which patterns of risk factors most consistently elevate predicted risk.

Research Impact and Practical Implications

When the priority is to avoid missed biopsy-positive cases, evaluation should focus on sensitivity and false negatives, and models should be tuned with a lower decision threshold to achieve a target recall. In this context, class-weighted approaches are useful because they shift learning toward the minority class, making them more appropriate for high-stakes screening support than models that optimize overall accuracy.

From a broader research perspective, this study supports the value of using routinely collected risk-factor histories to inform risk stratification, while also showing the limits of current performance under strong class imbalance and missingness. Any translation to real settings would require validation on external populations, checks of model performance across patient groups to

avoid errors, and clear governance for privacy and consent, especially because key predictors involve sensitive sexual and reproductive history.

6. Limitations and Ethical Considerations

Study Limitations

a. Extreme Class Imbalance

A major limitation of this study is the severe imbalance in the biopsy outcome, with only 6.41% of cases labeled positive. This imbalance can bias many classifiers toward predicting the majority negative class, which inflates overall accuracy while failing to detect true positives. As a result, accuracy alone is not a reliable indicator of screening usefulness in this setting. The imbalance also makes results less consistent across different data splits, because the small number of positive cases means that recall and F1 can change a lot depending on which positives end up in the training or test set. Overall, this limits how confidently the models can be viewed as reliable tools for identifying high-risk individuals.

b. Missing and Self-Reported Data

Several predictors contain substantial missingness and many features are based on self-reported behavioral and medical history. When data are missing, there is less usable information for some predictors, and results can change depending on how missing values are filled in, especially if the missingness is not random. Self-reported information can also be inaccurate because people may forget details or avoid reporting sensitive topics. Together, these issues add noise to the dataset, which can weaken model performance and make feature-importance results less reliable.

c. Limited Generalizability

The dataset reflects a specific population and data collection context, which may not represent other populations due to geographic, demographic, and healthcare system differences. Because these factors are not fully captured, model performance and identified risk drivers may not transfer directly to other regions or clinical settings without additional validation.

Ethical Considerations

a. False Negatives

In a screening context, false negatives are the most ethically concerning error because they represent missed high-risk individuals who may benefit from earlier follow-up, diagnostic testing, or intervention. When a model incorrectly classifies a biopsy-positive case as low risk, the likely consequence is delayed diagnosis and treatment, which can worsen outcomes and undermine trust in screening programs. For this reason, model evaluation should prioritize sensitivity-oriented measures such as recall and should treat low recall as a serious limitation, even when overall accuracy appears high.

b. False Positives

False positives also carry ethical and practical costs. Labeling low-risk individuals as high risk can lead to unnecessary anxiety, avoidable follow-up testing, and increased burden on clinical services. In real-world settings, excessive false positives may reduce program efficiency, increase costs, and create downstream harms such as stress, stigma, or reduced willingness to participate in future screening. This trade-off reinforces the need to interpret model outputs in context and to select decision thresholds that balance the harms of missed cases against the harms of over-referral.

c. Misuse of Predictive Models

Risk prediction models should be used as decision-support tools, not as standalone diagnostic systems. In a screening context, the biggest ethical risk is missing true positive cases, so model trust depends heavily on sensitivity. Models are easier to rely on when recall is closer to 1.0 because they identify most high-risk patients and reduce harmful false negatives, while still allowing clinicians to confirm results with appropriate follow-up testing. To prevent misuse, predictions should only be applied within the population and setting where the model has been validated, and performance should be checked across patient subgroups to avoid unequal error rates. Responsible use also requires clear documentation of what the model is for, what it cannot do, how thresholds were chosen, and what actions should follow a high-risk flag, along with ongoing monitoring to ensure performance does not degrade over time.

Overall, acknowledging these limitations and ethical considerations reinforces the importance of responsible model development and cautious interpretation when applying data-driven approaches in healthcare contexts.

7. Future Work and Recommendations

a. Expand and Diversify the Dataset

Future work should focus on testing the approach on larger, more representative screening data. Adding more biopsy-positive cases would make training more reliable and reduce the swing in recall and F1 score seen when positives are rare. Using data from different clinics, regions, and patient groups would also show whether the model generalizes beyond the original sample. A strong next step is to run the same preprocessing and evaluation pipeline on multi-site data, compare results across sites, and confirm that both performance and the most important predictors remain stable in new settings.

b. Explore Feature Engineering and Interaction Effects

Improved performance may also come from building features that better reflect how risk factors combine in real life. Cervical cancer risk is rarely driven by one variable alone, so future work should examine whether pairs or groups of predictors together provide clearer separation than single predictors. One approach is to create and test interaction features in interpretable models, then compare them with flexible methods that can learn interactions automatically. This should be done with strict validation and regularization so that added complexity improves generalization rather than fitting noise.

c. Model Calibration and Risk Stratification

Future versions of this approach should move beyond simple yes or no outputs and instead provide well-calibrated risk scores that clinicians can act on. In real screening settings, risk is typically managed through categories such as low, moderate, and high risk, because each level can be tied to different follow-up steps, timelines, or referrals. To support this, predicted probabilities must be calibrated so that a higher score truly means a higher chance of biopsy positivity, and decision thresholds should be chosen to match clinical priorities and available resources. Model evaluation should therefore include calibration checks and threshold-based decision analysis, not only classification metrics, so the output is interpretable and supports practical decision-making.

Overall, the most useful next steps are to improve generalizability, make predictions easier to explain, and design outputs that fit clinical workflow. Larger and more diverse validation data, clearer modeling of combined risk patterns, and calibrated risk categories would make the system more reliable and more aligned with how screening decisions are made in practice.

8. Conclusion

This report investigated three questions: (1) which lifestyle and medical history variables are most strongly linked to biopsy outcomes, (2) how different machine learning models compare when classifying high-risk versus low-risk individuals under severe class imbalance, and (3) whether machine learning models provide evidence of nonlinear patterns that may be missed by simpler clinical scoring approaches.

Overall, the results show that accuracy alone is not meaningful in this dataset because biopsy-positive cases are rare and models can achieve high accuracy by predicting mostly negatives. This limitation was clearest for unweighted logistic regression, which maintained high accuracy but failed to identify biopsy-positive cases, making it unsuitable for a screening-oriented goal. Addressing imbalance shifted model behavior. Class-weighted logistic regression produced the highest recall, making it the most aligned with prioritizing sensitivity, although recall remained modest and indicates substantial room for improvement before any real screening use. In contrast, the balanced random forest achieved the strongest overall discrimination (highest ROC–AUC), supporting its use as risk prioritization when follow-up resources are limited.

The most influential predictors were hormonal contraceptive duration, pregnancy history, and sexual history indicators, while individual STD indicators contributed less consistently, likely due to missingness. Since the tree-based ensemble performed better than the simpler models, it suggests that risk patterns may not be purely linear and may increase significantly after certain cutoffs, even though this study did not directly test specific interaction effects.

In summary, machine learning can highlight meaningful risk patterns, but detecting biopsy-positive cases remained difficult because positives were rare and many variables had missing values. To make this approach more practical for screening, future work should use larger and

more diverse datasets, apply stronger methods for class imbalance, and produce calibrated risk scores that can be grouped into clear low-, moderate-, and high-risk categories for follow-up decisions.

9. GitHub Repository Link

https://github.com/monicascode/CIND820-Initial-Results-and-the-Code

References

Al Mudawi, N., & Alazeb, A. (2022). A Model for Predicting Cervical Cancer Using Machine Learning Algorithms. *Sensors (Basel, Switzerland)*, *22*(11), 4132. https://doi.org/10.3390/s22114132

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*.

Ijaz, M. F., Attique, M., & Son, Y. (2020). *Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods*. *Sensors, 20*(10), 2809. https://doi.org/10.3390/s20102809

Muraru, M. M., Simó, Z., & Iantovics, L. B. (2024). Cervical Cancer Prediction Based on Imbalanced Data Using Machine Learning Algorithms with a Variety of Sampling Methods. Applied Sciences, 14(22), 10085. https://doi.org/10.3390/app142210085

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

World Health Organization. (2025). *Cervical cancer.* https://www.who.int/news-room/fact-sheets/detail/cervical-cancer