Cervical Cancer: Classification Model

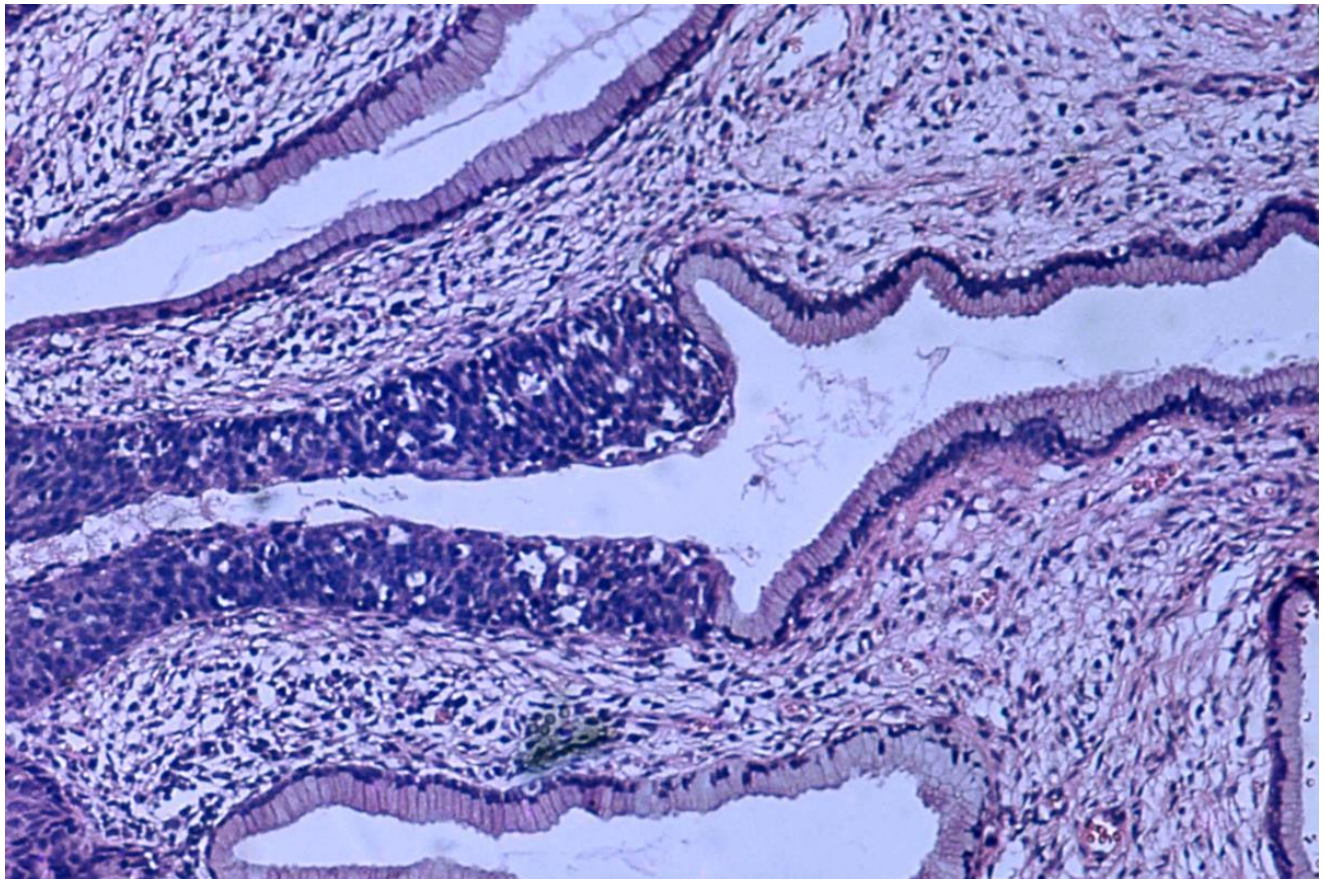
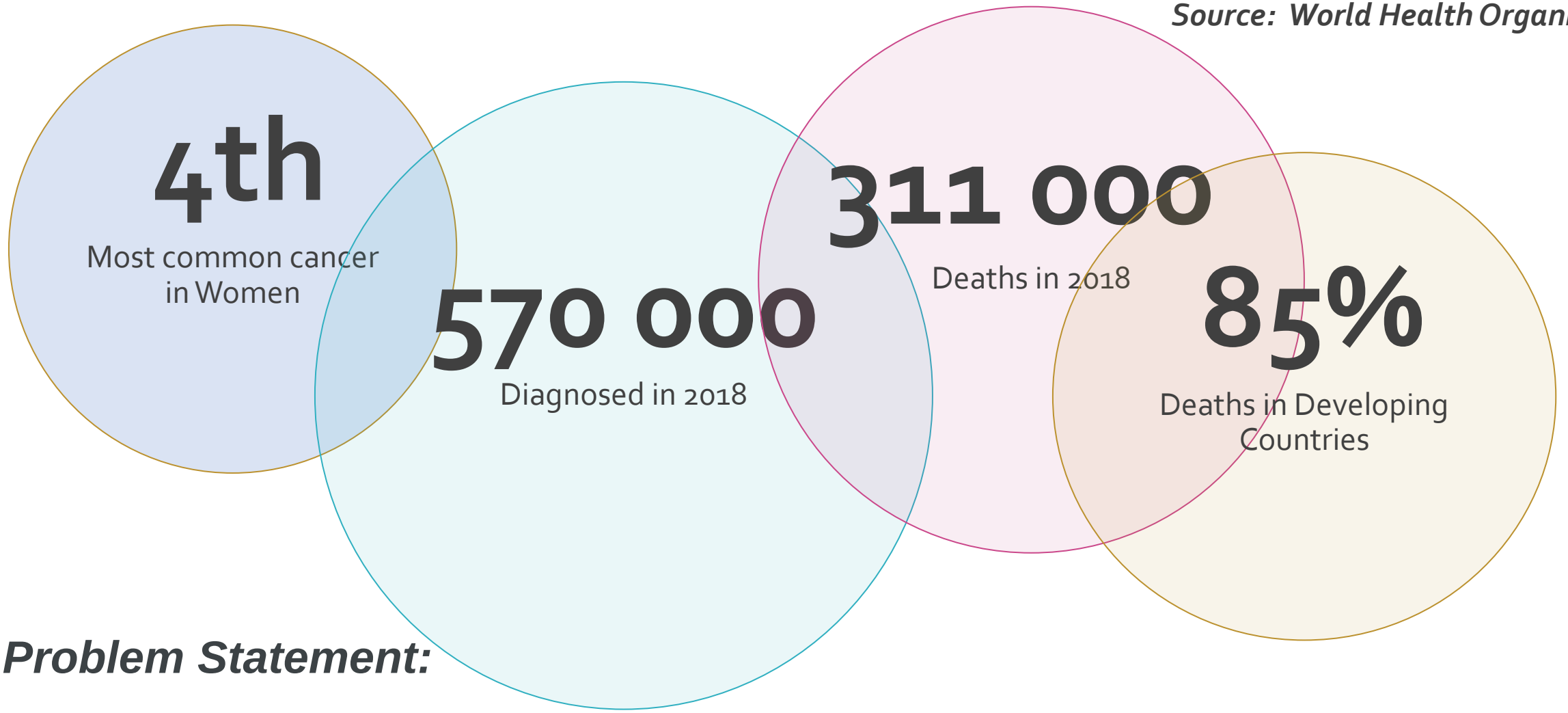


Image of a Pap Smear specimen under a microscope

Global Prevalence of Cervical Cancer

Source: World Health Organization



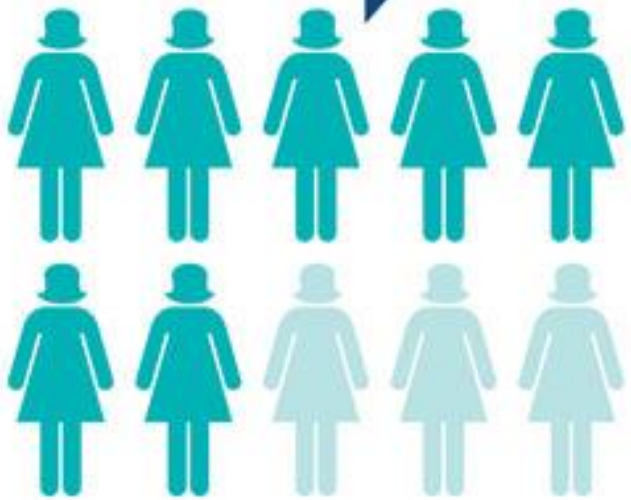
Problem Statement:

Towards A prediction model of cervical cancer to identify high risk groups to focus scarce screening & testing capacity in developing countries

Explanatory Side: Source US CDC at https://www.cdc.gov/vitalsigns/cervical-cancer/infographic.html

Missed opportunities for cervical cancer screening

In 2012, **8 million women** were not screened in the last **5 years.**



7 out of 10 women who were not screened had a regular doctor and health insurance.

SOURCE: Behavioral Risk Factor Surveillance System, 2012.

How HPV infection can lead to cervical cancer

It could take years to decades

Cervical Cancer Risk Factors Dataset

Check out the beta version of the new UCI Machine Learning Repository we are currently testing! Contact us if you have any issues, questions, or concerns. Click here to try out the new site.

Cervical cancer (Risk Factors) Data Set

Download: Data Folder, Data Set Description

Abstract: This dataset focuses on the prediction of indicators/diagnosis of cervical cancer. The features cover demographic information, habits, and historic medical records.

Data Set Characteristics:	Multivariate	**Number of Instances:**	858	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	36	**Date Donated**	2017-03-03
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	161532

Dataset from University of California Irvine UCL:

https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29#

Risk Factors Dataset

(int) Age
(int) Number of sexual partners
(int) First sexual intercourse (age)
(int) Num of pregnancies
(bool) Smokes
(bool) Smokes (years)
(bool) Smokes (packs/year)
(bool) Hormonal Contraceptives
(int) Hormonal Contraceptives (years)
(bool) IUD
(int) IUD (years)
(bool) STDs
(int) STDs (number)
(bool) STDs:condylomatosis
(bool) STDs:cervical condylomatosis
(bool) STDs:vaginal condylomatosis
(bool) STDs:vulvo-perineal condylomatosis
(bool) STDs:syphilis
(bool) STDs:pelvic inflammatory disease
(bool) STDs:genital herpes
(bool) STDs:molluscum contagiosum
(bool) STDs:AIDS
(bool) STDs:HIV
(bool) STDs:Hepatitis B
(bool) STDs:HPV

Highly Correlated

(int) STDs: Number of diagnosis
(int) STDs: Time since first diagnosis
(int) STDs: Time since last diagnosis

90% Missing

(bool) Dx:Cancer
(bool) Dx:CIN
(bool) Dx:HPV
(bool) Dx

No meaningful documentation

(bool) Hinselmann: target variable
(bool) Schiller: target variable
(bool) Cytology: target variable
(bool) Biopsy: target variable

←Gold Standard Target

- Dataset has 35 attributes of demographic information, habits, and historic medical records of **858 patients collected at Venezuela hospital**
- Missing values as patients decline answering due to privacy especially for sensitive questions on sexual practices & STDs
- 80% of attributes have missing values, especially, STDs diagnosis & time have 90% missing data. Numerical attributes missing values replaced by median, STD Boolean values replaced with True & 3 attributes with 90% missing values dropped. Feature Dx dropped as no meaningful data explanation available
- Cervical cancer gold standard testing on **Biopsy** used as Target variable, the other three test results dropped

- Significant **correlation** exists between the attributes as expected since sexual practices are linked to incidence of STDs and among STDs positive correlation exist as well
- This can be visualised through a heatmap and the table of variance inflation factors. The group of STDs attributes are particularly highly correlated
- Min-max **Normalisation** also performed as some attributes like Age and number of years smoking have different ranges
- **Feature extraction** performed using cumulative explained variance to determine 10 principal components (accounting for about 95% of variance) and PCA technique to obtain these components
- Feature extraction as against Feature selection (dropping attributes) as no available information whether correlation due to just common surrogate for sexual orientation or if certain STDs might be better indicators of cervical cancer. In other words, certain STDs might predispose the patient to higher and persistent risk of HPV infection
- Target attribute of Biopsy heavily skewed towards negative (95%). To prevent a high accuracy but low precision model, dataset is rebalanced using **SMOTE** technique

Algorithms & Improvements

- Baseline outcome using a set of different algorithms with K-fold cross validation
- K-Nearest Neighbors, Decision Tree and Support Vector Machine perform best in the group of algorithms with accuracy of about 0.8
- Closer examination of these three algorithms also reveals that the harmonic mean (f1 value) of precision and recall of presence of cervical cancer at around 0.85. This is visualized using the confusion matrix
- This means that overall, the respective models predict the correct biopsy outcome 85% (accuracy) of the time, and 85% of positive predictions have positive biopsy (precision) and 85% of those with cervical cancer get a positive prediction (recall)

Algorithms & Improvements

- Improvements performed using ensemble method of Random Forest and gradient boosting algorithms. These methods combine many smaller models and should arrive at an overall higher performance model
- Through adjustments in hyperparameters like the number of estimators, the learning rate (gradient boosting) and depth, both Random Forest and Gradient Boosting deliver better performance with f1 harmonic mean of 0.9

Results & Analysis

- Overall, Random Forest & Gradient Boosting are neck to neck with accuracy, precision & recall at 90%

Random Forest

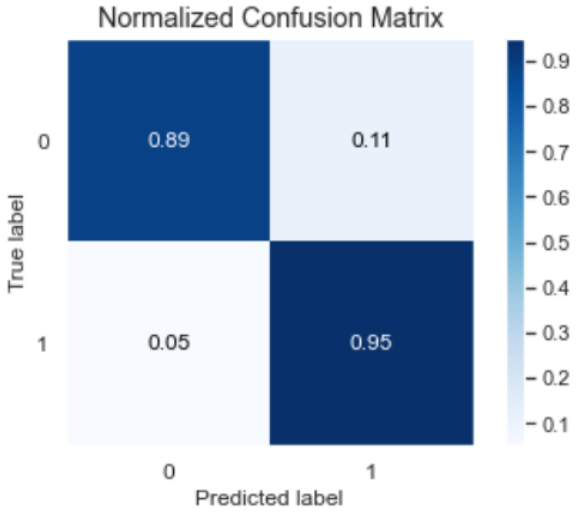
```
Accuracy:
0.922360248447205

Confusion matrix:
[[151  18]
 [  7 146]]

Classification report:
              precision    recall  f1-score   support

           0       0.96      0.89      0.92       169
           1       0.89      0.95      0.92       153

    accuracy                           0.92       322
   macro avg       0.92      0.92      0.92       322
weighted avg       0.92      0.92      0.92       322
```



Gradient Boosting

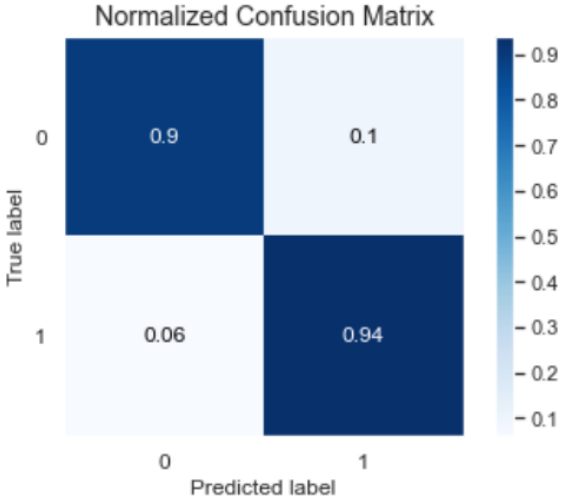
```
Accuracy:
0.9192546583850931

Confusion matrix:
[[152  17]
 [  9 144]]

Classification report:
              precision    recall  f1-score   support

           0       0.94      0.90      0.92       169
           1       0.89      0.94      0.92       153

    accuracy                           0.92       322
   macro avg       0.92      0.92      0.92       322
weighted avg       0.92      0.92      0.92       322
```



Results & Analysis

- Precision is “How many of those who we predict as positive are actually positive?”, or *Precision = TP/(TP+FP)*
- Recall answers, “Of all the people who are positive, how many of those we correctly predict as positive?”, or *Recall = TP/(TP+FN)*
- 90% of predicted positive by trained Random Tree Model indeed have cervical cancer (Precision) and Model successfully predicts positive 90% of the time for those who have cervical cancer (Recall)
- In other words, there is 10% error in that of those predicted as positive 10% are negative, and model misses 10% of those who indeed have cervical cancer. As such, the Model ***cannot be used in isolation*** to conclude as ultimate truth that a patient predicted as positive will go immediately for intrusive surgery. Conversely, a patient predicted as negative cannot be taken as free of cervical cancer for the rest of the life
- Instead, the **Model should be deployed alongside other screening tools** to complement and supplement with clear knowledge of the Model’s limitations such as to identify high risk groups and those flagged as positive to go for further tests such as pap smear. Those predicted as negative are exempted for this round of pap smear but if they get two consecutive negatives over say two to three years time, then it is mandatory for them to get the pap smear even if the Model predicts negative. This would take care of the False Negatives. Given that cervical cancer takes many years to develop (see explanatory slide), even if they miss the pap smear because of the False Negative, they will still be captured during the mandatory pap smear and intervention can take place

Conclusion

- Although the trained model has 90% accuracy, its predictability in real life is not certain due to small sample size. More data should be collected to test and refine model
- Due to sensitivity of sexual practices, STDs etc, less intrusive proxy indicators should preferably be devised
- Model can be deployed as a tool alongside others in cervical cancer screening. High risk groups can be identified and prioritised for other screening methods like pap smear and cytology.

*Prevention, **screening** & treatment*
can eliminate cervical cancer in one generation: WHO

Cervical screening testing vehicle in Taiwan, 2012