Cervical Cancer Risk Assessment

A project by Soumyadip Kundu

Springboard Data Sciences Career Track

Background

- 2nd most frequent gynecological cancer
- 570,000 new cases in the world out of which 11000 in the US (WHO, 2018)
- 90% fatality in low- and middle-income countries
- Mortality rate can be reduced by early diagnosis, screening and treatment

Problem Statement

- Build a predictive machine learning model which can assess all the risk factors and will lead us to ascertain the major risk factors associated with cervical cancer

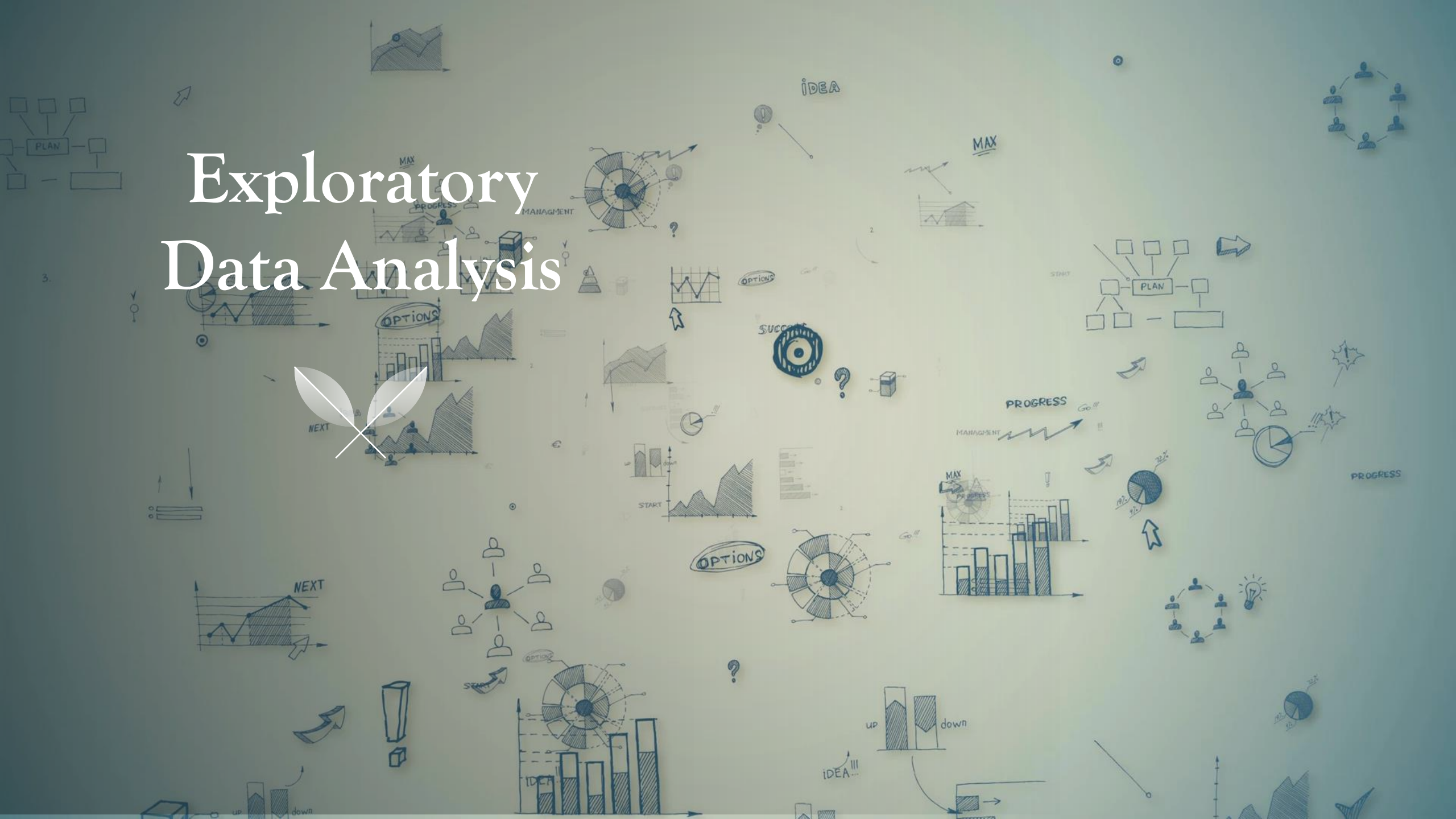
Dataset

- Obtained from the UCI data repository and downloaded through Kaggle.
- 857 rows, 33 columns
- Each feature contains a question from a survey that was asked to the women.

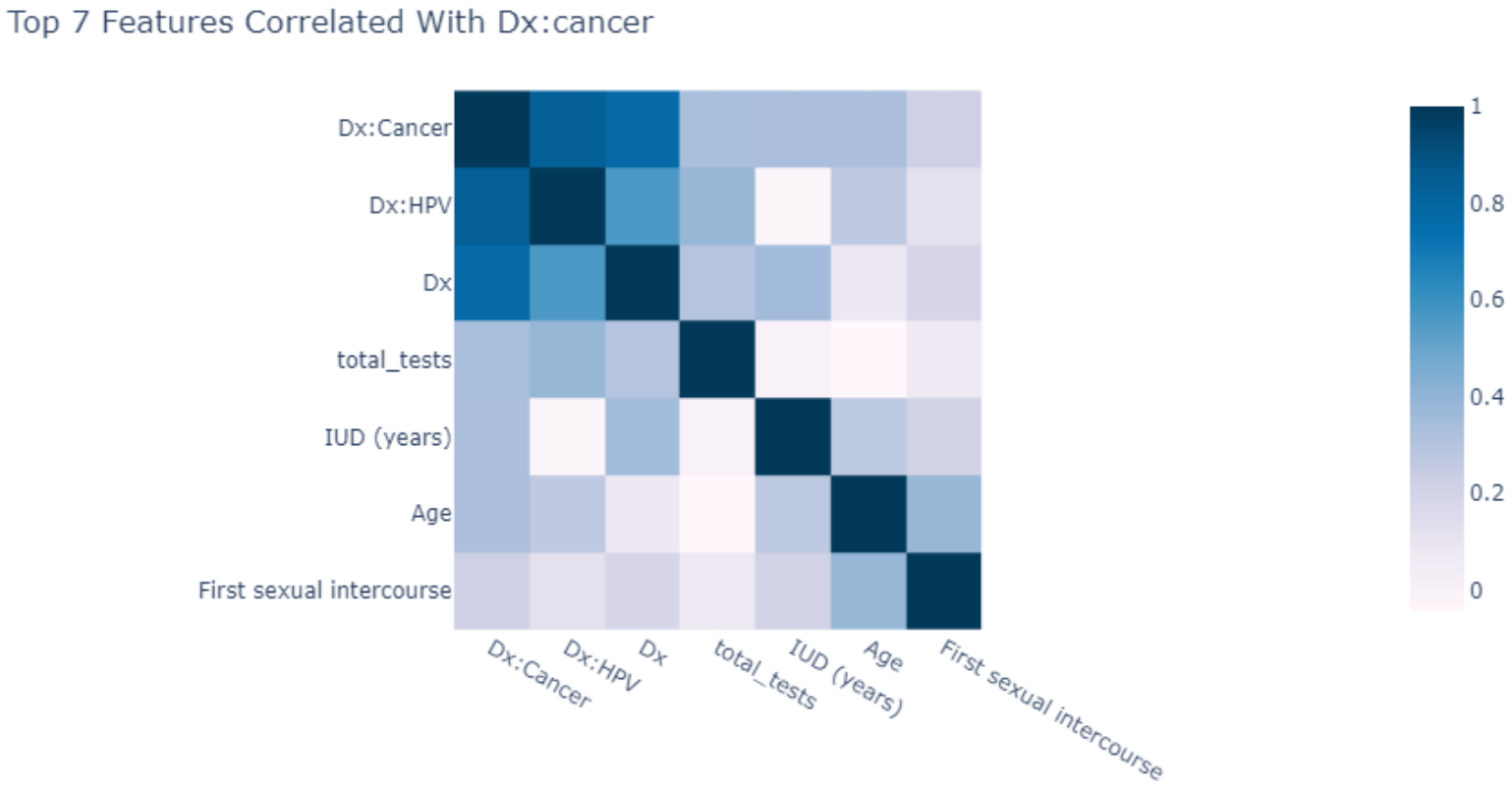
Data Cleaning and Preprocessing

- 2 out of 35 columns had more than 90% missing values
- Simple Imputer used to impute missing values with the median
- ADASYNC used to solve class imbalance problem of the target variable, “Dx:Cancer”

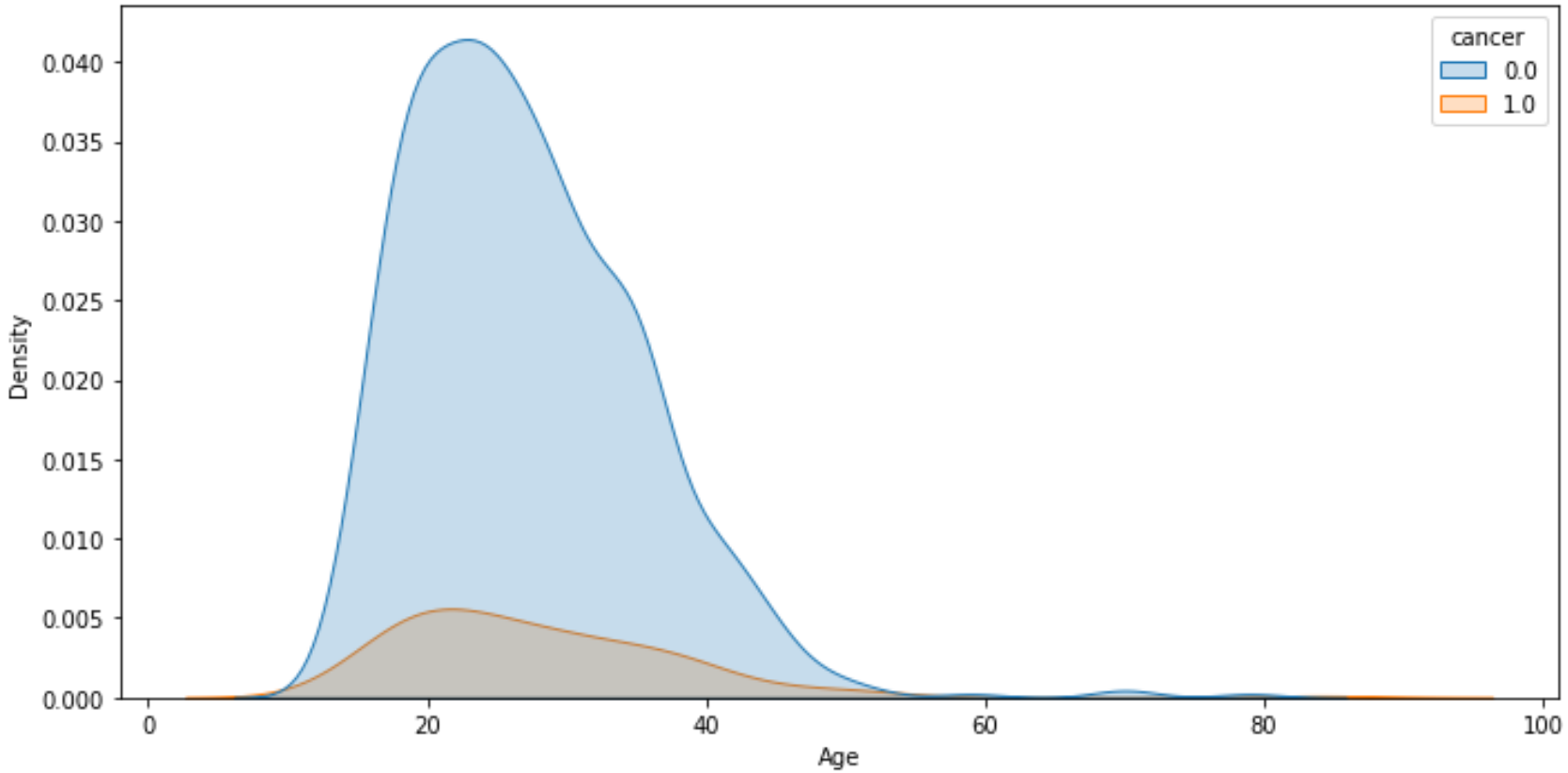
Exploratory Data Analysis



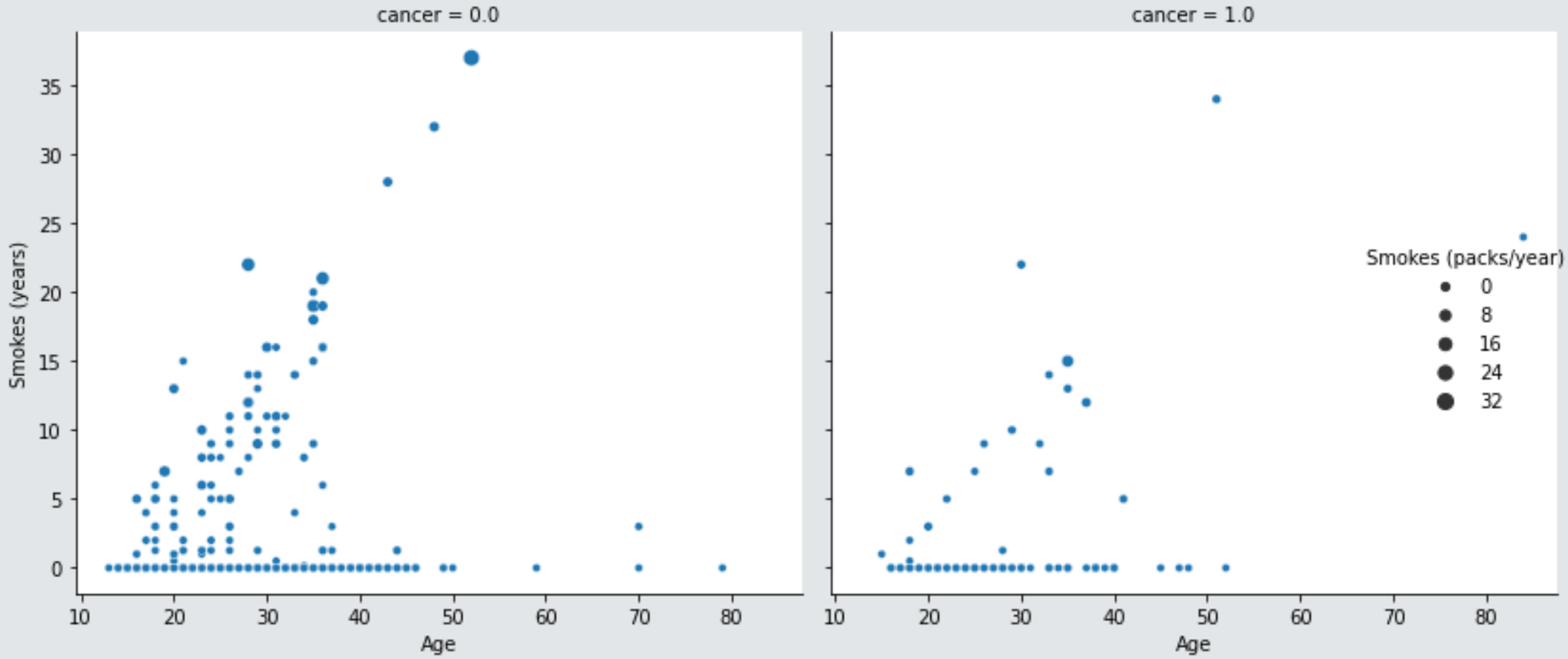
Heatmap of the top 7 features associated with Cervical Cancer



Age vs Cancer (Distribution)

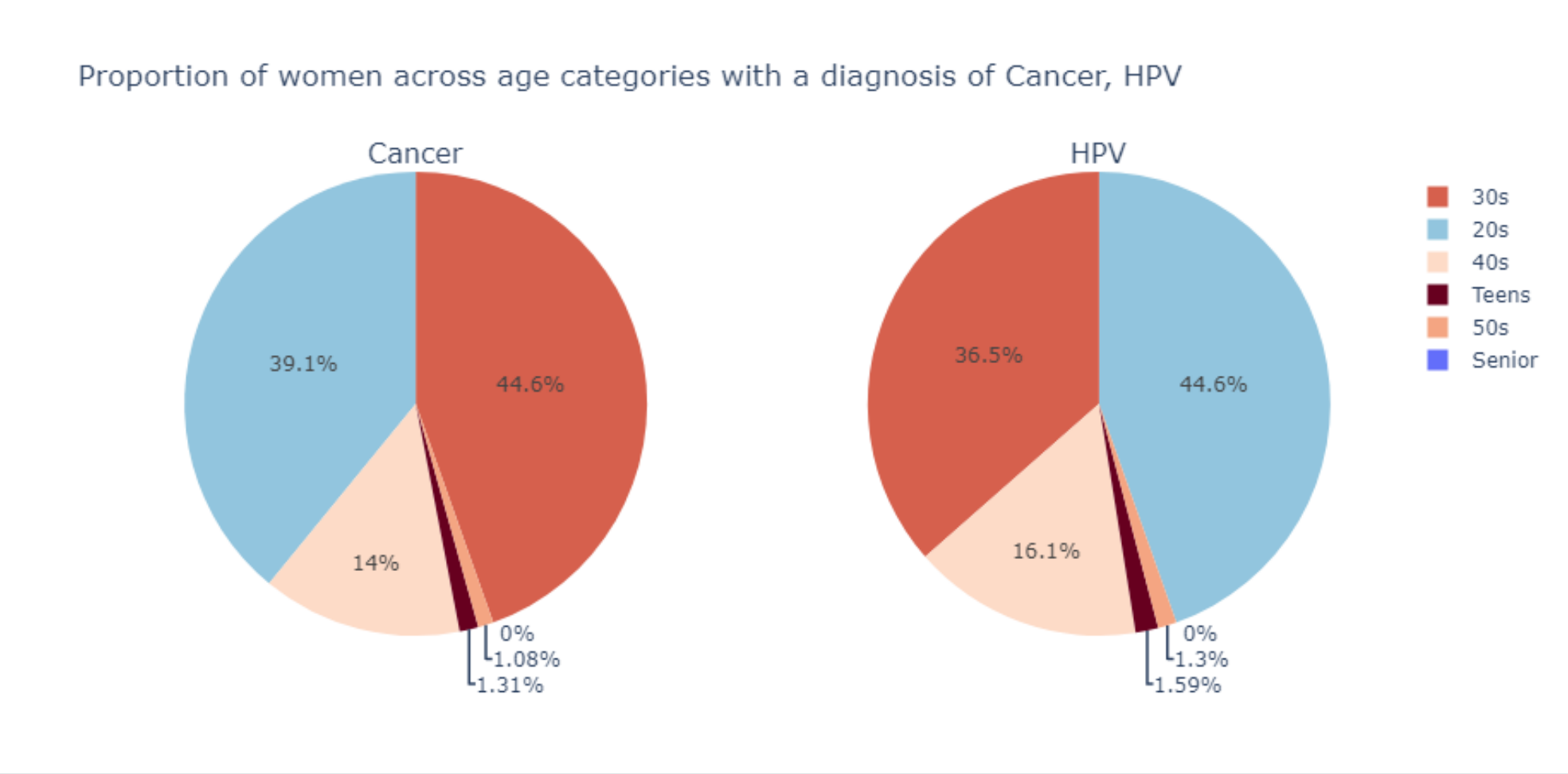


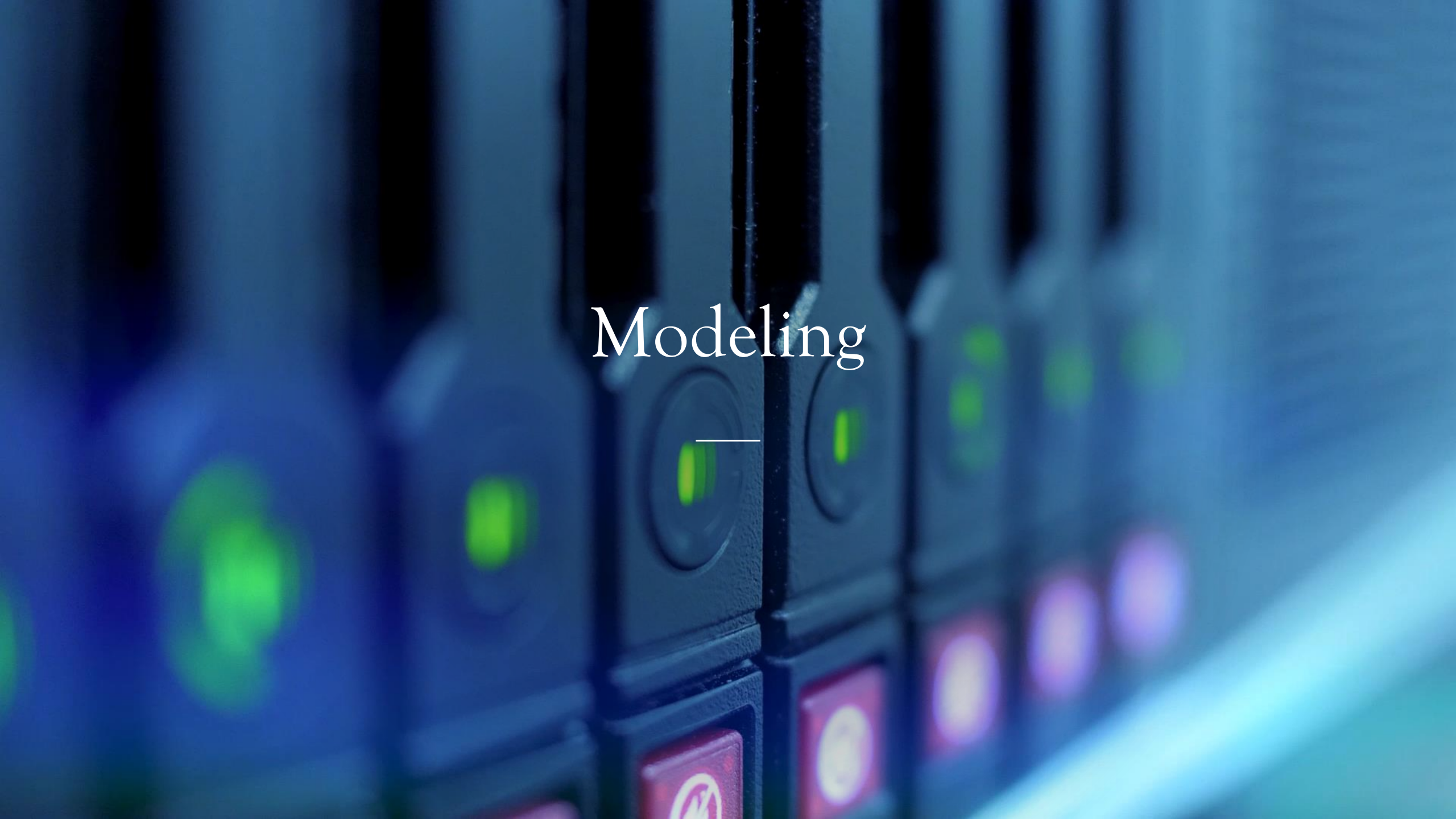
Correlation between Smoking, Age and Cervical Cancer



Correlation between Cervical Cancer, HPV and number of sexual partners

Cancer and HPV based on age categories





Modeling

Models used

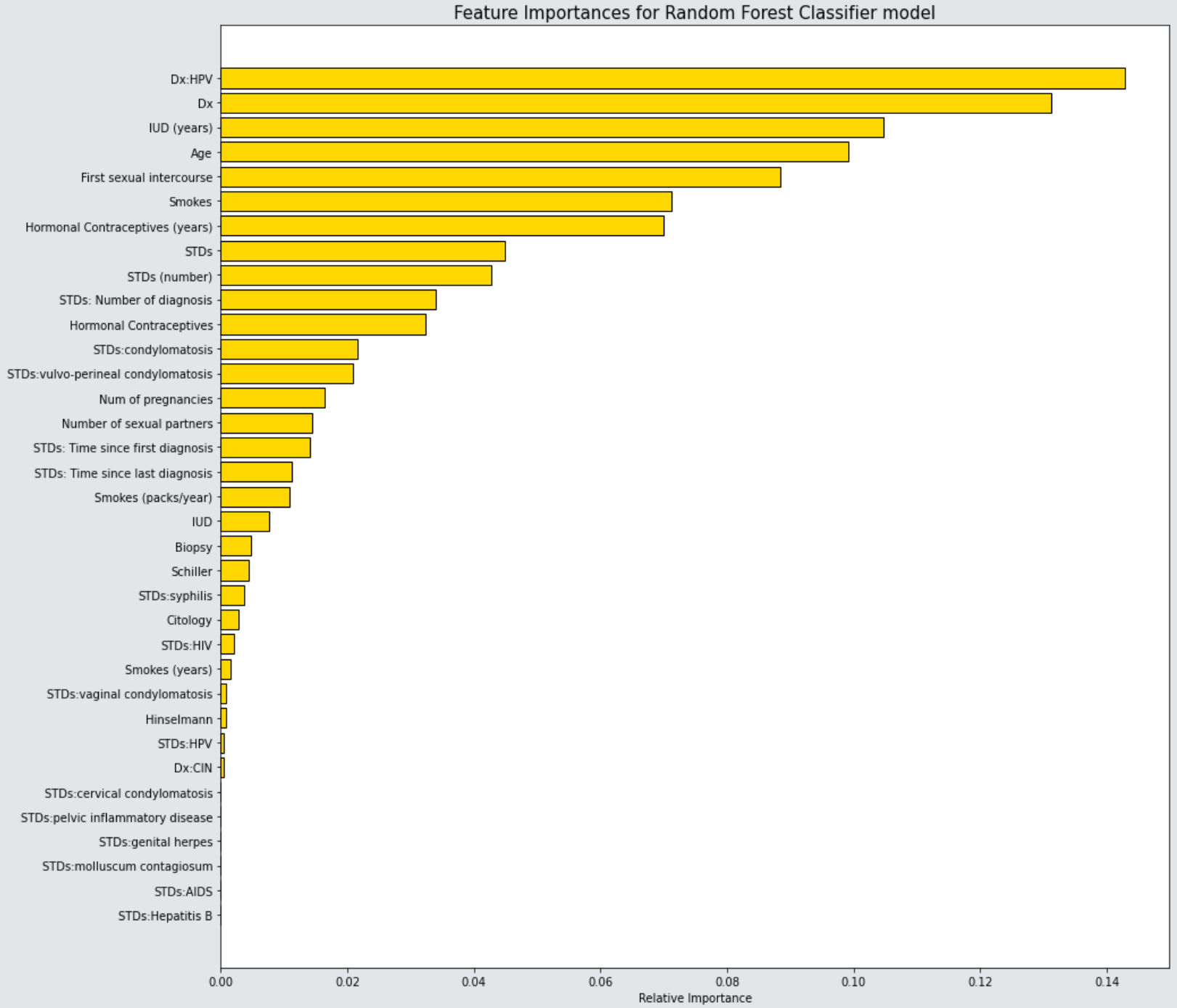
- Logistic Regression
- Random Forest Classifier
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)

Model Performance Evaluation

	Classifier Name	Accuracy Score	Precision Score	Recall Score	F1 Score
0	LogisticRegression	0.991071	0.991235	0.991071	0.991074
1	RandomForestClassifier	0.991071	0.991091	0.991071	0.991072
2	KNeighborsClassifier	0.961310	0.964200	0.961310	0.961314
3	SupportVectorClassifier	0.997024	0.997042	0.997024	0.997024

Hyperparameter Tuning

Hyperparameter tuning was done on Random Forest Classifier to show feature importance



Conclusion

Most important risk factors (in order):

- HPV diagnosis (most important)
- Diagnosis of other STDs
- Usage of IUDs (Intrauterine devices)
- Smoking
- First Sexual intercourse age
- Usage of Hormonal Contraceptives

Acknowledgements

- **Kenneth Gil-Pasquel (Mentor)**
- **Springboard team**
- **Kaggle**
- **Cover images – Google images**