Samuel Aboagye
Data Mining – Final Project
Individual Final Report

1. Introduction. An overview of the project and an outline of the shared work.

Cervical cancer is normally diagnosed through the performance of Biopsy by a trained physician. Biopsy is known to be one of the best ways to confirm the presence of cervical cancer as being benign or malignant. Automating the process of diagnosing the disease could be helpful in the early screening and detection of cervical cancer. From the literature, certain features are stronger predictors of the presence of cervical cancer. Through feature selection using Random Forest, the most important features were confirmed by the research. The 10 most important features were Age, Number of Sexual partners, IUD, Number of Pregnancies, Smoking, first Sexual Intercourse, Hormonal Contraceptives.

2. Description of your individual work. Provide some background information on the development of the algorithm and include necessary equations and figures.

Data cleaning was done by removing columns with more than 50% of missing data, as the inference from this data will be less efficient. These features were 'STDs: Time since first diagnosis' and 'STDs: Time since last diagnosis'. There were 4 targets in total and due to time constraints, we decided to focus our modelling on just one of them which was Biopsy. The choice of this Target was due to several research papers pointing to it being one of the most reliable ways of screening for the presence of cervical cancer.

We had to Over sample the observations with cancer to increase the sampling rate of this class as to keep up with the heaviness of the 'without cancer' class.

3. Describe the portion of the work that you did on the project in detail. It can be figures, codes, explanation, pre-processing, training, etc.

My portion of the project was researching cervical cancer, data preprocessing and building the models.

Preprocessing and Model Building

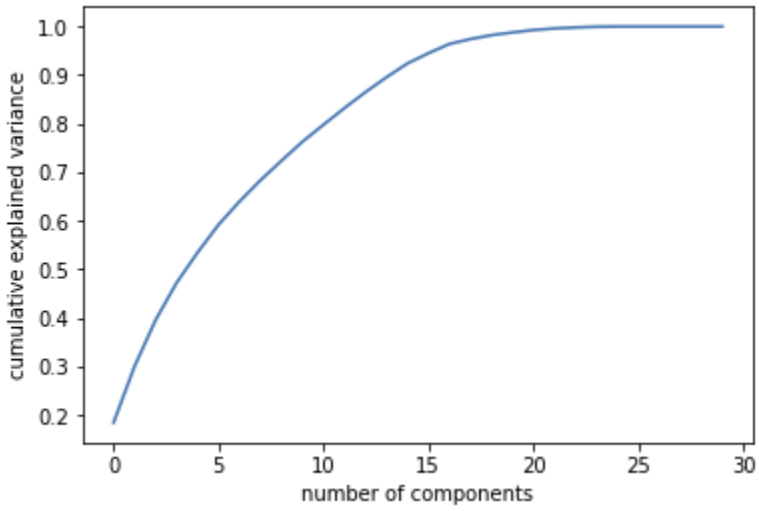
1. Missing Value imputation with a combination of dropping columns and mean replacement
2. Create Training and Validation data splits for training
3. Wrote the scripts for data cleaning and model evaluation

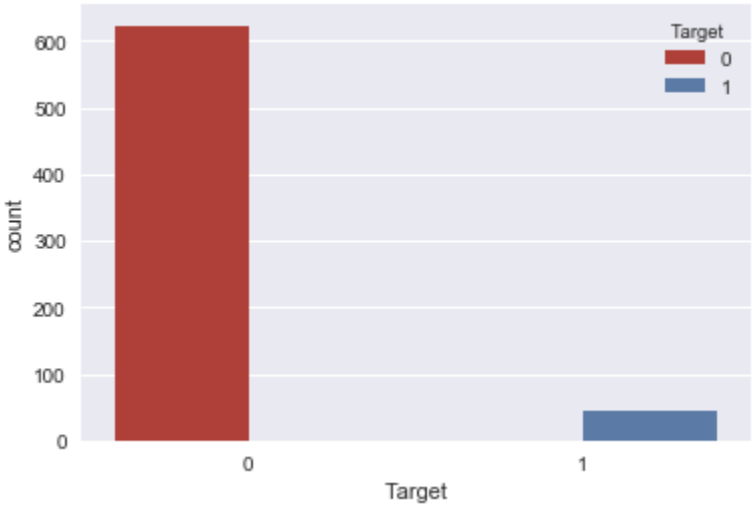
Model Building and Training

1. Built several models and trained them.
2. Tuned the models and evaluated them
3. Evaluate model before and after preprocessing

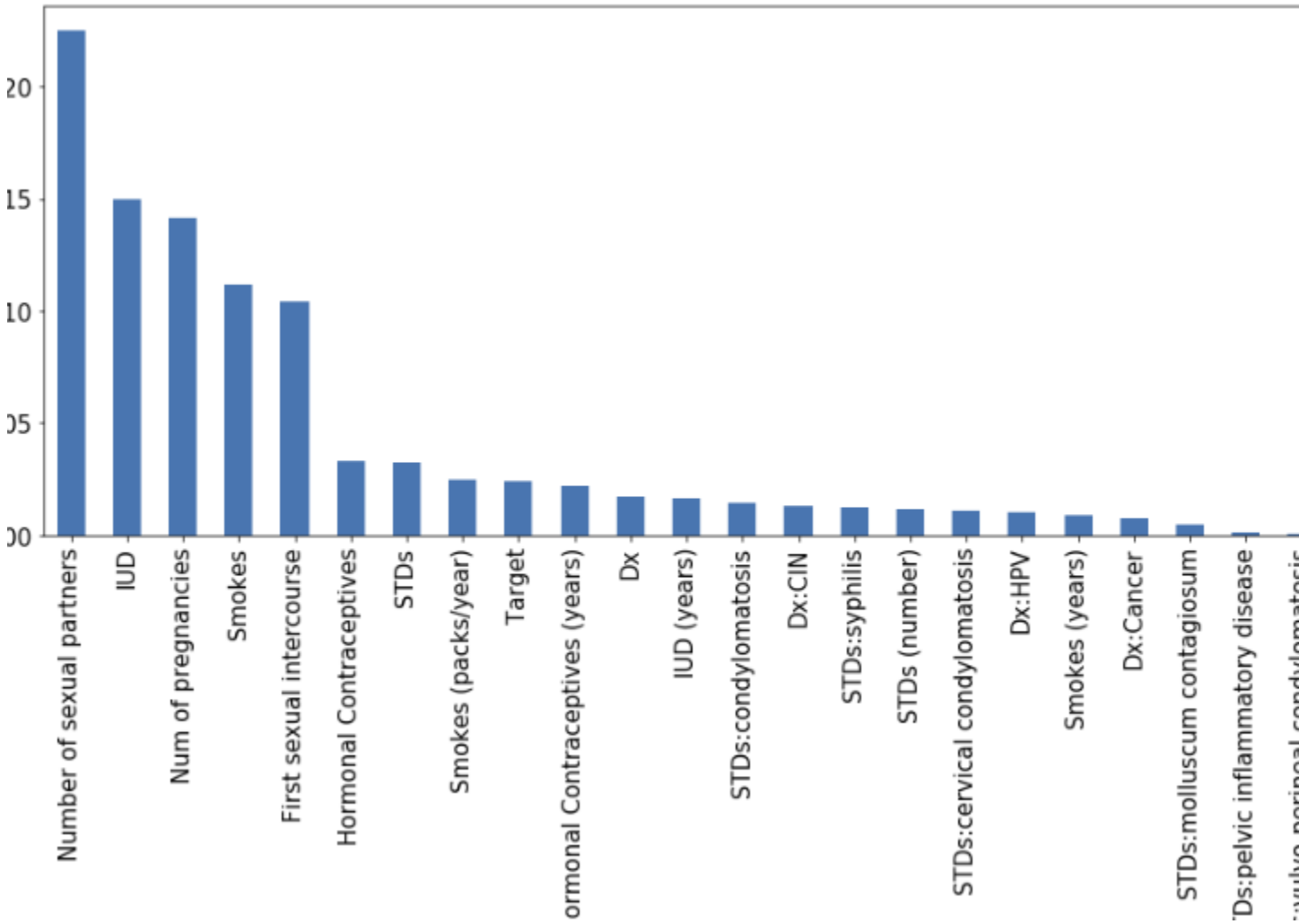
4. Results. Describe the results of your experiments, using figures and tables wherever possible. Include all results (including all figures and tables) in the main body of the report, not in appendices. Provide an explanation of each figure and table that you include. Your discussions in this section will be the most important part of the report.

1.Preprocessing affected the performance of our models
2. The important features identified during feature selection were confirmed by the literature as being vital for the diagnosis cervical cancer

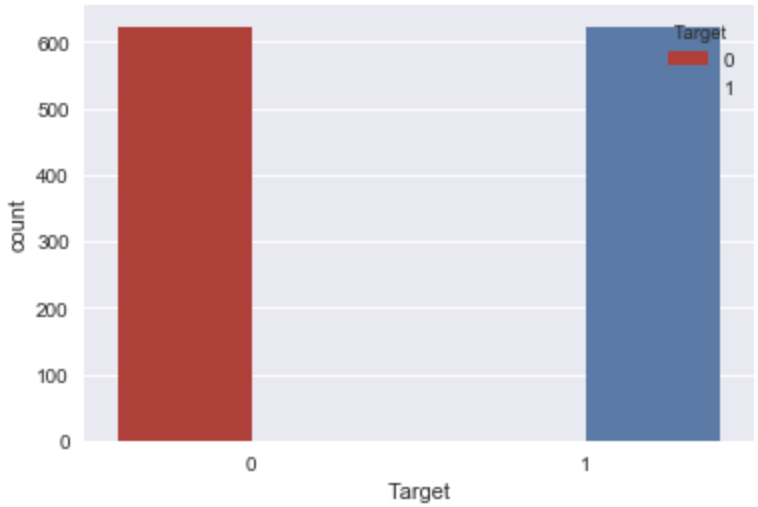




0	Age
1	Number of sexual partners
2	First sexual intercourse
3	Num of pregnancies
4	Smokes (packs/year)
5	Hormonal Contraceptives (years)
6	IUD (years)
7	STDs (number)



```
Random over-sampling:
1    623
0    623
Name: Target, dtype: int64
```



5. Summary and conclusions. Summarize the results you obtained, explain what you have learned, and suggest improvements that could be made in the future.

1.SVC performed the best on the training data.
2. I have learned that preprocessing is time consuming and its essentially for proper training of the models.
3. I will like to work more on preprocessing techniques and model tuning in the future

6. Calculate the percentage of the code that you found or copied from the internet.

795 total lines of code
Internet = 321 lines of code
Internet total = 40%

7. References.

[1]Center for disease Control and Prevention (CDC). https://www.cdc.gov/cancer/cervical/basic_info/risk_factors.htm

[2]Source: Kelwin Fernandes (kafc at inINESC TEC & FEUP, Porto, Portugal. Jaime S. Cardoso - INESC TEC & FEUP, Porto, Portugal. Jessica Fernandes - Universidad Central de Venezuela, Caracas, Venezuela.

[3] American Cancer Society: https://www.cancer.org/content/dam/CRC/PDF/Public/8599.00.pdf