Machine Learning Engineer Nanodegree

Capstone Project “Predicting causes of cervical cancer”

I. Definition

Project Overview

Cancer starts when cells in the body begin to grow out of control. Cells in nearly any part of the body can become cancer, and can spread to other areas of the body. Cervical cancer starts in the cells lining the cervix -- the lower part of the uterus (womb).

The American Cancer Society's estimates for cervical cancer in the United States for 2017 are:

- About 12,820 new cases of invasive cervical cancer will be diagnosed.
- About 4,210 women will die from cervical cancer.

Cervical pre-cancers are diagnosed far more often than invasive cervical cancer.

Cervical cancer was once one of the most common causes of cancer death for American women. But over the last 40 years, the cervical cancer death rate has gone down by more than 50%. The main reason for this change was the increased use of the Pap test. This screening procedure can find changes in the cervix before cancer develops. It can also find cervical cancer early – in its most curable stage.

Cervical cancer tends to occur in midlife. Most cases are found in women younger than 50. It rarely develops in women younger than 20. Many older women do not realize that the risk of developing cervical cancer is still present as they age. More than 15% of cases of cervical cancer are found in women over 65. **[1]**

The dataset used in this project was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset comprises demographic information, habits, and historic medical records of 858 patients. Several patients decided not to answer some of the questions because of privacy concerns (missing values). **[2]**

Problem Statement

We can see, that there is a big number of features in the dataset. We should understand which of them are not important for determining, if a woman has cervical cancer or not. It will help doctors to quickly analyze which patients are in the main risk group and it will help patients not to provide all personal information.

We will begin with analyzing all the features and trying to predict, if a woman has the decease or not, based on all of them. Then we will try to significantly reduce features number and predict the same thing, based on the reduced number. We should get the same or better results as with all features.

Metrics

We will use the following metrics:

- **Precision**

  $$P = \frac{T_p}{T_p + F_p}$$

  Precision (P) is defined as the number of true positives (T_p) over the number of true positives plus the number of false positives (F_p). **[3]**
- **Recall**

  $$R = \frac{T_p}{T_p + F_n}$$

  Recall (R) is defined as the number of true positives (T_p) over the number of true positives plus the number of false negatives (F_n). **[3]**
- **F1-score**

  $$F1 = 2\frac{P \times R}{P + R}$$

  F1-score, which is defined as the harmonic mean of precision and recall. **[3]**
- **Confusion matrix**
  By definition, entry i, j in a confusion matrix is the number of observations actually in group i, but predicted to be in group j. **[4]**
  In our case the matrix looks like this:

  $$\begin{bmatrix} true\ positive & false\ positive \\ false\ negative & true\ negative \end{bmatrix}$$

II. Analysis

Data Exploration

The dataset is a CSV file. It contains 858 data records (patients), 32 features and 4 target values. There are many '?' instead of unknown data, so we should convert them into N/A while reading the CSV.

```
import pandas as p
df = p.read_csv('risk_factors_cervical_cancer.csv', na_values='?')
```

Features:

Feature	Meaning
Age	Age of the patient
Number of sexual partners	Number of sexual partners of the patient
First sexual intercourse	Age of patient's first sexual intercourse
Num of pregnancies	Number of patient's pregnancies
Smokes	Shows if patient smokes (0 – No, 1 – Yes)
Hormonal Contraceptives	Shows if patient takes hormonal contraceptives (0 – No, 1 – Yes)
Hormonal Contraceptives (years)	Number of years of taking hormonal contraceptives
IUD	Shows if patient uses IUD (intrauterine device) (0 – No, 1 – Yes)
IUD (years)	Number of years of using IUD
STDs	Shows if patient has STDs (sexually transmitted diseases) (0 – No, 1 – Yes)
STDs: std_name	Shows if patient has std_name STD (0 – No, 1 – Yes)
Dx: dx_value	Shows if patient had diagnostic of dx_value

Target values:

'Hinselmann', 'Schiller', 'Citology', 'Biopsy' – if even one of these values shows 1, it means that patient has risk of having cervical cancer. So, we introduce new column 'AllCancerFlag': value = 0, when all 'Hinselmann', 'Schiller', 'Citology' and 'Biopsy' columns are 0, and value = 1, when at least one of them is 1.

Now we will calculate some statictics values.

Primarily, we gather statistics about values in columns:

```
print df.describe().T
```

We can see, that two columns ('STDs:cervical condylomatosis' and 'STDs:AIDS') have equal minimum and maximum values; so, these columns are not representative for our research. We remove them from our dataset:

```
df.drop(['STDs:cervical condylomatosis', 'STDs:AIDS'], axis=1, inplace=True)
```

Now we gather general information about the dataset.

Number of patients:

```
n_patients = df.shape[0]
```

Number of features:

```
n_features = df.shape[1] - 5
```

Number of patients, having cervical cancer:

```
n_has_cancer = len(df[df['AllCancerFlag'] == 1])
```

Number of patients without cervical cancer:

```
n_has_no_cancer = len(df[df['AllCancerFlag'] == 0])
```

Cancer rate:

```
cancer_rate = n_has_cancer / float(n_has_no_cancer) * 100
```

Total number of patients: 858

Number of features: 30

Number of patients, having cervical cancer: 102

Number of patients, not having cervical cancer: 756

Cancer rate: 13.49%

As we can see we have small dataset with many features. One of the main difficulties is cancer rate: we have very few cases of cancer to learn on.

The second difficulty is large number of N/A values in the dataset.

```
df_wo_null = df.dropna(axis=0, how='any')
print df_wo_null.shape[0]
```

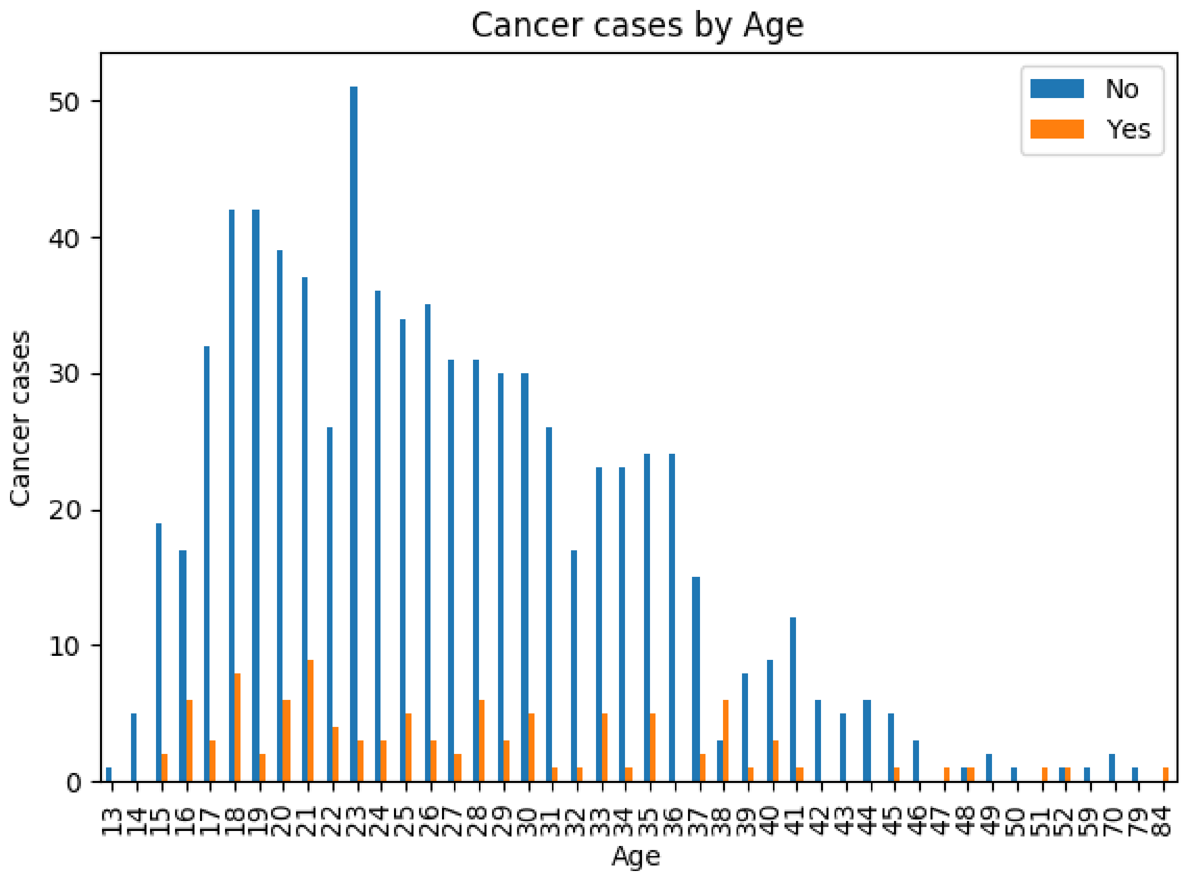
Only 59 rows are without N/A values. We cannot sacrifice so many data, so we need to replace N/A with adequate data: for numeric features we replace N/A with the median value and for categorical features we replace N/A with 1.

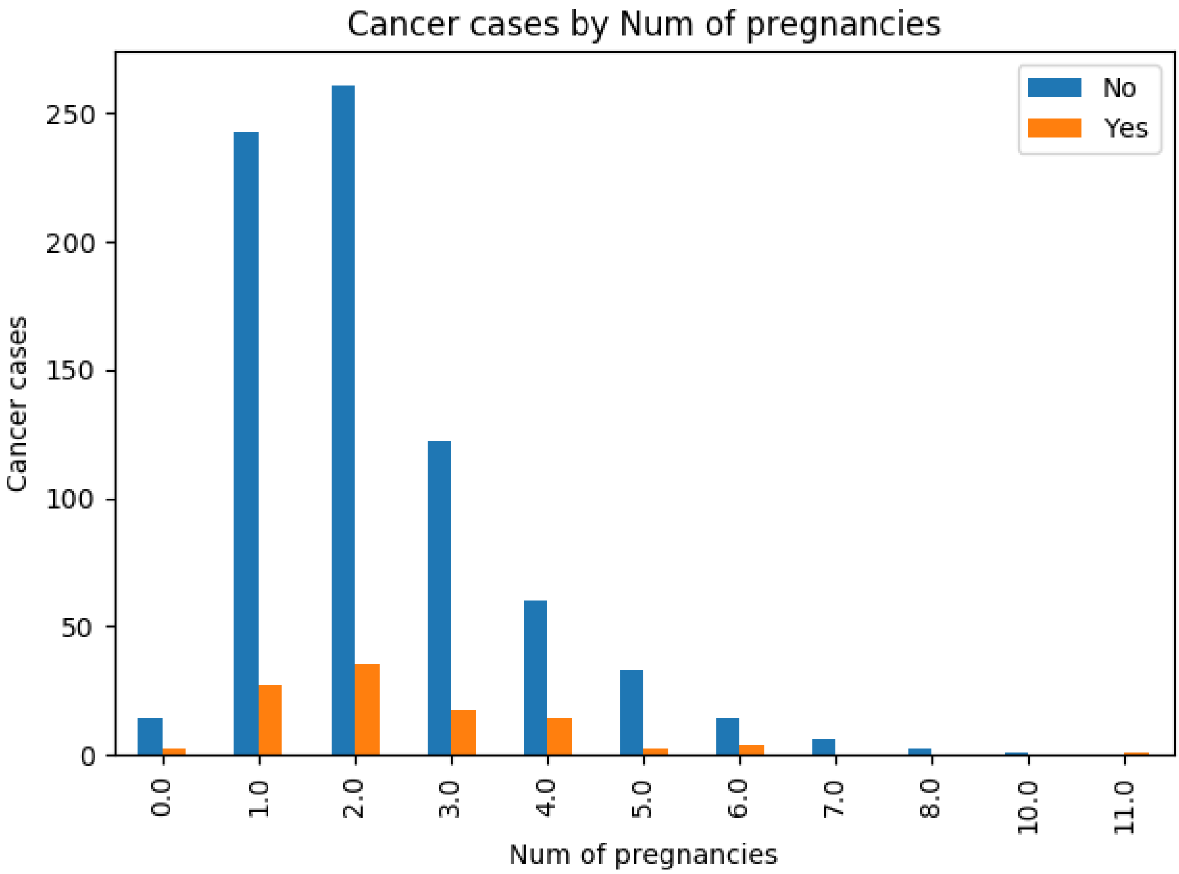
Exploratory Visualization

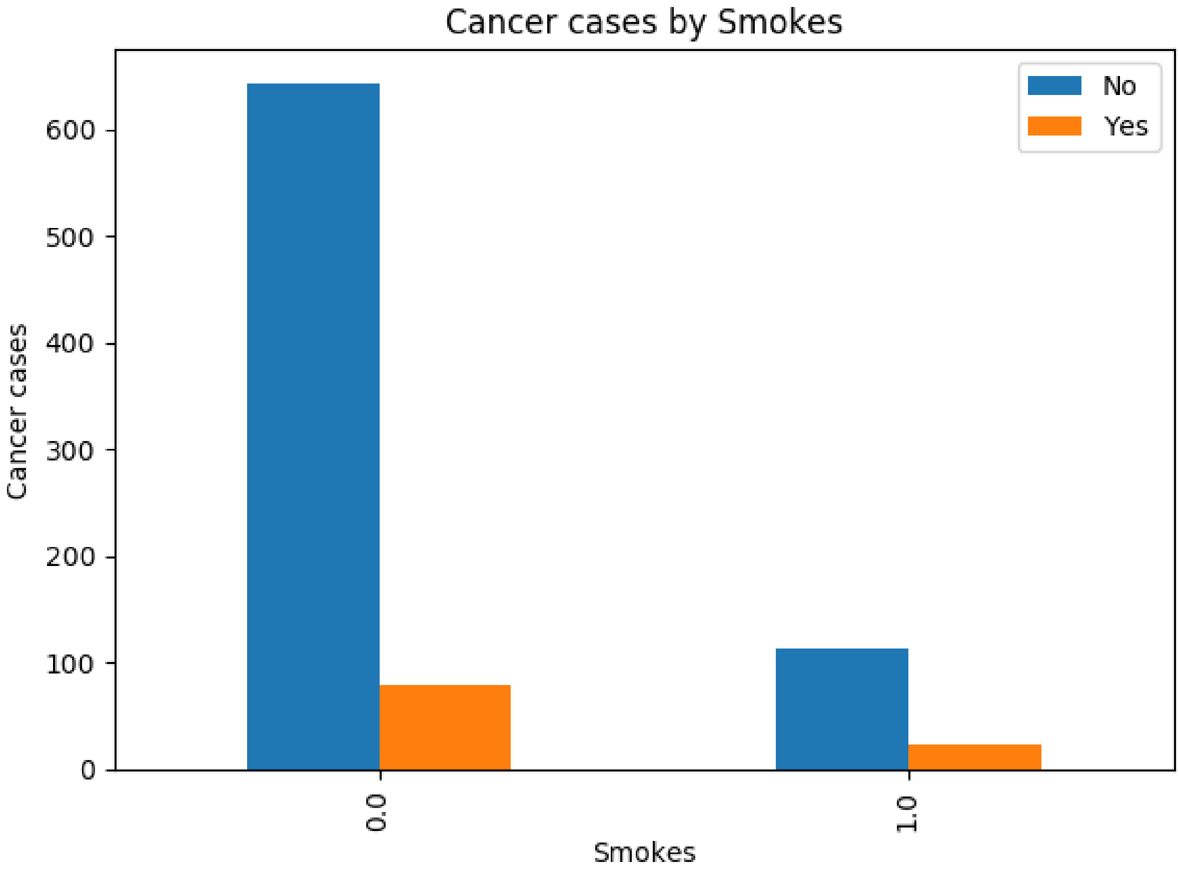
It is very unlikely, that visualization will help us in this case.

There are 30 features and we need to determine which of them are most important. It is not likely, that our result will be one or two features, so without proper algorithm it will take very long time to determine main features visually.

Let's build some plots to try to find some clues about our dataset.







As we can see, there is no visible correlation between cancer cases and features.

Algorithms and Techniques

The main algorithm we use in this assignment is ExtraTreesClassifier. We have the unbalanced dataset and ExtraTreesClassifier is known as one of the best algorithms to deal with such datasets. We compare this method with DecisionTreeClassifier to prove its advantage.

Both algorithms will be used with random_state=1.

Default value of the ExtraTreesClassifier parameter n_estimators is 10. If we rise this number, we will get better result, but execution time will also rise. As we have very small dataset, execution time will rise insignificantly, so we set n_estimators=100 for better performance.

Also we will use feature_importances_ to determine importance of the features.

Benchmark

Our goal is to minimize features number and get same (approximately) or better results, than we have with current features number. Because of it we do not have static benchmark, we have metrics values and with our algorithm of minimizing features number we should get better metrics results.

III. Methodology

Data Preprocessing

All necessary data preprocessing steps were described in Analysis.Data exploration section.

Implementation

Firstly, we train and fit DecisionTreeClassifier and ExtraTreesClassifier. We print out metrics values; these results are be our start point for improving results, because next comparisons will be made with these values.

Then we print values of features importances with feature_importances_ statement. It will help us to determine threshold values of features importance (greater value -> more important feature; we want to use only features, whose importance values are above some threshold value).

The first threshold value is 0.1, second – 0.01, third – 0.005. We can see, that the results in the case of third threshold begin to decrease, so we stop reducing the threshold. As we can see, the threshold value 0.01 is the best in this case.

Refinement

Now we will try to improve our results.

For that we will remove unnecessary features from the dataset. Firstly, we will remove features 'Smokes', 'Hormonal Contraceptives' and 'IUD'. We can remove these features, because, actually, we have more precise features 'Smokes (years)', 'Smokes (packs/year)' for 'Smokes' (if 'Smokes (years)' and 'Smokes (packs/year)' are greater, than 0, then, obviously, the woman smokes); 'Hormonal Contraceptives (years)' for 'Hormonal Contraceptives' and 'IUD (years)' for 'IUD'.

Secondly, we can remove features 'STDs: Time since first diagnosis' and 'STDs: Time since last diagnosis', because they do not represent very useful information (it is unknown, which STD was diagnosed), their influence is the least among other features, which values are greater, than 0.01 and, most importantly, it cannot cause diseases; it is just a statistical measure.

IV. Results

We run several iterations of training and fitting DecisionTreeClassifier and ExtraTreesClassifier models with datasets, determined by different thresholds of feature importance in feature_importances_ statement.

1. In first iteration we train and fit our DecisionTreeClassifier and ExtraTreesClassifier models with original dataset. We get results:

   ```
   DecisionTreeClassifier original:
   Training set:
                precision    recall  f1-score   support

             0       0.99      1.00      1.00       568
             1       1.00      0.96      0.98        77
   avg / total       1.00      1.00      1.00       645
   [[568   0]
    [  3  74]]
   Testing set:
                precision    recall  f1-score   support

             0       0.90      0.88      0.89       188
             1       0.21      0.24      0.23        25
   avg / total       0.82      0.81      0.81       213
   ```

```
[[166  22]
 [ 19   6]]

ExtraTreesClassifier original:
Training set:
             precision    recall  f1-score   support

          0       0.99      1.00      1.00       568
          1       1.00      0.96      0.98        77

avg / total       1.00      1.00      1.00       645
[[568   0]
 [  3  74]]
Testing set:
             precision    recall  f1-score   support

          0       0.89      0.97      0.93       188
          1       0.25      0.08      0.12        25
avg / total       0.81      0.86      0.83       213
[[182   6]
 [ 23   2]]
```

As we can see ExtraTreesClassifier gets much better results in testing set in correct determining absence of cancer (left upper corner), but gets worse results in correct determining of cancer presence (right bottom corner), which is crucial. Our goal is to improve these results.

2. In second iteration we use only features with values more than 0.1.
   List of features:
   ['Age', 'Number of sexual partners', 'First sexual intercourse', 'Num of pregnancies', 'Hormonal Contraceptives (years)']
   Results:

```
DecisionTreeClassifier first features adjustment:
Training set:
             precision    recall  f1-score   support

          0       0.99      1.00      0.99       568
          1       1.00      0.92      0.96        77

avg / total       0.99      0.99      0.99       645

[[568   0]
 [  6  71]]
Testing set:
             precision    recall  f1-score   support
```

```
         0       0.88      0.91      0.90       188
         1       0.11      0.08      0.09        25

avg / total       0.79      0.82      0.80       213

[[172  16]
 [ 23   2]]
```

ExtraTreesClassifier first features adjustment:

Training set:

```
             precision    recall  f1-score   support

         0       0.99      1.00      0.99       568
         1       1.00      0.92      0.96        77

avg / total       0.99      0.99      0.99       645

[[568   0]
 [  6  71]]
```

Testing set:

```
             precision    recall  f1-score   support

         0       0.88      0.97      0.92       188
         1       0.14      0.04      0.06        25

avg / total       0.80      0.86      0.82       213

[[182   6]
 [ 24   1]]
```

As we can see we got much worse results, than in previous case. We removed too many features and encountered underfitting: we have too few features to create proper model. Then we should leave more features.

3. In third iteration we leave features with values more than 0.01.
   List of features:
   ['Age', 'Number of sexual partners', 'First sexual intercourse', 'Num of pregnancies', 'Smokes', 'Smokes (years)', 'Smokes (packs/year)', 'Hormonal Contraceptives', 'Hormonal Contraceptives (years)', 'IUD', 'IUD (years)', 'STDs (number)', 'STDs: Time since first diagnosis', 'STDs: Time since last diagnosis', 'Dx:Cancer', 'Dx:HPV']
   Results:
   DecisionTreeClassifier second features adjustment:
   Training set:

```
             precision    recall  f1-score   support
```

```
          0       0.99      1.00      1.00       568
          1       1.00      0.96      0.98        77

avg / total       1.00      1.00      1.00       645

[[568   0]
 [  3  74]]
Testing set:
             precision    recall  f1-score   support

          0       0.89      0.90      0.90       188
          1       0.22      0.20      0.21        25

avg / total       0.82      0.82      0.82       213

[[170  18]
 [ 20   5]]
ExtraTreesClassifier second features adjustment:
Training set:
             precision    recall  f1-score   support

          0       0.99      1.00      1.00       568
          1       1.00      0.96      0.98        77

avg / total       1.00      1.00      1.00       645

[[568   0]
 [  3  74]]
Testing set:
             precision    recall  f1-score   support

          0       0.90      0.96      0.93       188
          1       0.36      0.16      0.22        25

avg / total       0.83      0.87      0.85       213

[[181   7]
 [ 21   4]]
```

We got better results in determining correct presence of cancer cases, than in previous cases and almost same results in determining correct absence of cancer as in original case.

4. In fourth iteration we leave features with values more than 0.005.
   List of features:

```
['Age', 'Number of sexual partners', 'First sexual intercourse', 'Num of
pregnancies', 'Smokes', 'Smokes (years)', 'Smokes (packs/year)', 'Hormonal
Contraceptives', 'Hormonal Contraceptives (years)', 'IUD', 'IUD (years)', 'STDs
(number)', 'STDs:vulvo-perineal condylomatosis', 'STDs:genital herpes', 'STDs:HIV',
'STDs: Number of diagnosis', 'STDs: Time since first diagnosis', 'STDs: Time since
last diagnosis', 'Dx:Cancer', 'Dx:HPV', 'Dx']
Results:
DecisionTreeClassifier third features adjustment:
Training set:
             precision    recall  f1-score   support

          0       0.99      1.00      1.00       568
          1       1.00      0.96      0.98        77

avg / total       1.00      1.00      1.00       645

[[568   0]
 [  3  74]]
Testing set:
             precision    recall  f1-score   support

          0       0.89      0.85      0.87       188
          1       0.18      0.24      0.20        25

avg / total       0.81      0.78      0.79       213

[[160  28]
 [ 19   6]]
ExtraTreesClassifier third features adjustment:
Training set:
             precision    recall  f1-score   support

          0       0.99      1.00      1.00       568
          1       1.00      0.96      0.98        77

avg / total       1.00      1.00      1.00       645

[[568   0]
 [  3  74]]
Testing set:
             precision    recall  f1-score   support

          0       0.89      0.96      0.92       188
          1       0.20      0.08      0.11        25
```

```
avg / total       0.81      0.85      0.83       213

[[180   8]
 [ 23   2]]
```

We can see, that results started getting worse: we return to large number of features. We should return to our best iteration, features with values more than 0.01.

5. We remove some features as described in section Methodology.Refinement.
   List of features:
   ['Age', 'Number of sexual partners', 'First sexual intercourse', 'Num of pregnancies', 'Smokes (years)', 'Smokes (packs/year)', 'Hormonal Contraceptives (years)', 'IUD (years)', 'STDs (number)', 'Dx:Cancer', 'Dx:HPV']
   Results:
   DecisionTreeClassifier final features adjustment:
   Training set:

```
             precision    recall  f1-score   support

          0       0.99      1.00      1.00       568
          1       1.00      0.96      0.98        77

avg / total       1.00      1.00      1.00       645

[[568   0]
 [  3  74]]
```

Testing set:

```
             precision    recall  f1-score   support

          0       0.89      0.86      0.87       188
          1       0.16      0.20      0.18        25

avg / total       0.80      0.78      0.79       213

[[161  27]
 [ 20   5]]
```

ExtraTreesClassifier final features adjustment:
Training set:

```
             precision    recall  f1-score   support

          0       0.99      1.00      1.00       568
          1       1.00      0.96      0.98        77

avg / total       1.00      1.00      1.00       645

[[568   0]
 [  3  74]]
```

```
Testing set:
             precision    recall  f1-score   support

          0       0.90      0.98      0.94       188
          1       0.56      0.20      0.29        25

avg / total       0.86      0.89      0.86       213

[[184   4]
 [ 20   5]]
```

We got the best results of all we had seen before. As we can see we got maximum results for both absence and presence of cancer.

V. Conclusion

Reflection

In this project we had a goal to minimize features number and predict causes of cervical cancer.

We used DecisionTreeClassifier and ExtraTreesClassifier to train and test our models and feature_importances_ to determine the most important features in the dataset. The most important features we finally got from our research: 'Age', 'Number of sexual partners', 'First sexual intercourse', 'Num of pregnancies', 'Smokes (years)', 'Smokes (packs/year)', 'Hormonal Contraceptives (years)', 'IUD (years)', 'STDs (number)', 'Dx:Cancer', 'Dx:HPV'.

Main causes of cervical cancer according to other studies are: HPV ('Dx:HPV'), giving birth to many children ('Num of pregnancies'), smoking cigarettes ('Smokes (years)', 'Smokes (packs/year)'), using oral contraceptives ('Hormonal Contraceptives (years)'), being sexually active at a young age ('First sexual intercourse'), having many sexual partners ('Number of sexual partners'), HIV and age ('Age').**[5]**

As we can see we correctly determined most of main causes of cervical cancer and it allows us to say, that we chose correct model for describing it. Also we reduced the number of features from 32 to 11.

Main problems of this project were: small size of dataset and strongly unbalanced dataset (only 13.5% of women had cervical cancer). Small size of dataset did not allow us to remove any of the data, because any loss of data was crucial. Because of strongly unbalanced dataset we had great results in correct determining absence of cancer, but wrong determining presence of cancer.

Improvement

Main possible improvements are:

- Get significantly more data, especially with cancer cases.
- Remove as much unknown data as we can (gather more precise data).
- Make auto selection of the threshold of features importance.

Due to auto selection we can achieve more precise threshold and probably remove some extra features.

First two improvements will help us to predict more accurate results. Also we could try to test other models, because we would have large balanced dataset.

References

[1] https://www.cancer.org/cancer/cervical-cancer/about.html

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

[3] http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

[4] http://scikit-learn.org/stable/modules/model_evaluation.html#confusion-matrix

[5] https://www.cancer.gov/types/cervical/patient/cervical-treatment-pdq#section/all