Justin Oakley and Micah Roberts

CSC 526-01

Prof. Manda

April 1st, 2019

CSC 526 Assignment 4 – Machine Learning Report

In this assignment, the dataset that was utilized in the machine learning experiment was called the “Cervical cancer (Risk Factors) Data Set”. The data came from 'Hospital Universitario de Caracas' in Caracas, Venezuela and was initially constructed for predicting indicators and diagnoses of cervical cancer in hospital patients. The original file contained thirty-six numerical features and eight hundred fifty-eight initial observations about demographic information, habits, and medical records about patients. Using this information, the first step taken in the machine learning process was to rid the data of any null values. After some analyzation, it was determined that two of the columns in the dataset, ‘STDs: Time since first diagnosis’ and ‘STDs: Time since last diagnosis’, each contained a vast amount of null values and therefore were dropped since they were not going to be useful in machine learning. Once this was done, then the leftover data could properly have all the remaining null values dropped, which meant that the edited data had thirty-four columns and six hundred sixty-four observations left. The next step was to drop any columns with mean or standard deviation values less than zero, split the data into the training and testing sets using a 0.75:0.25 ratio, and then standardize each column feature.

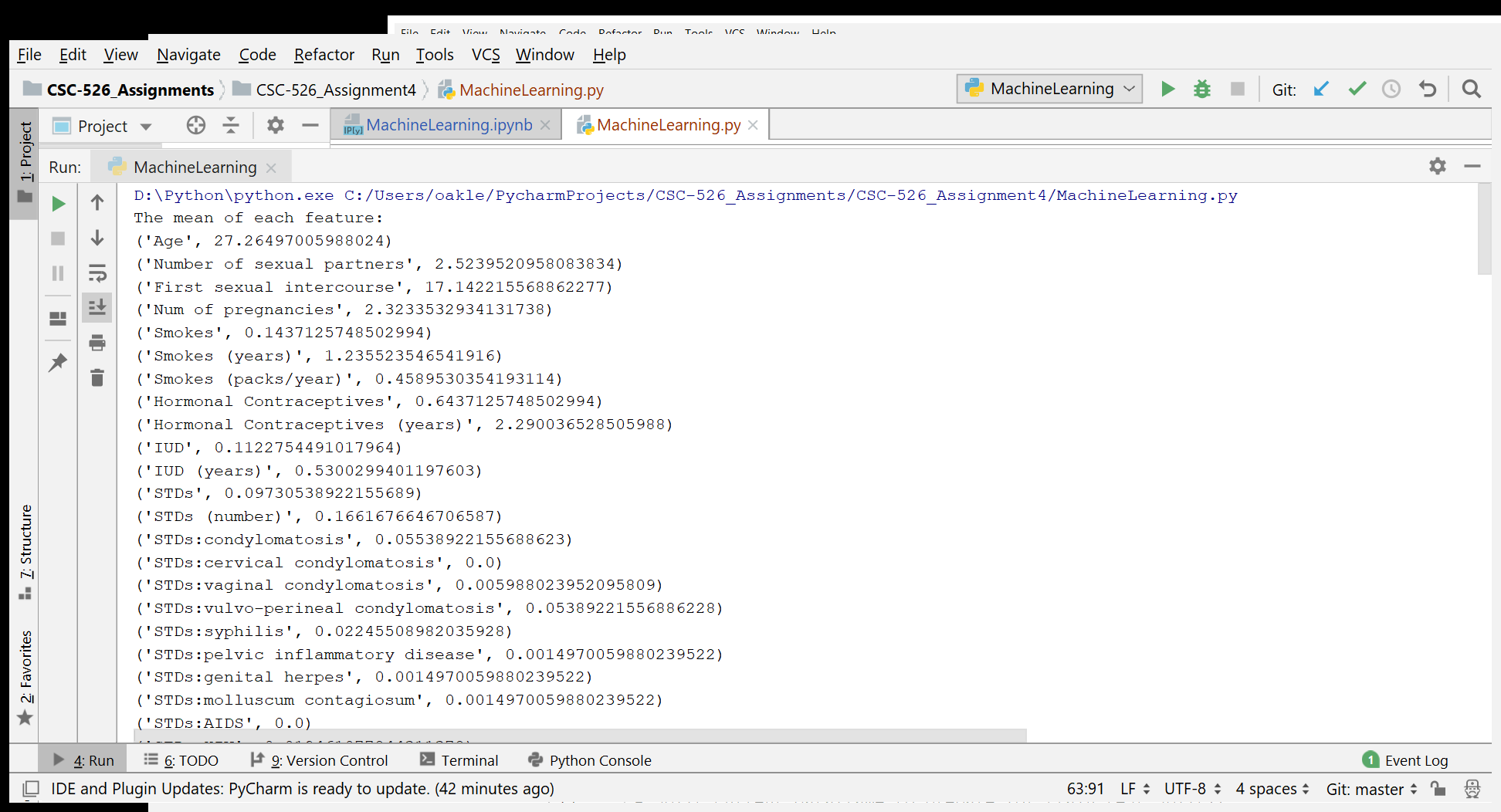
After preprocessing the data, a machine learning technique had to be chosen; for this assignment, classification was selected, and the chosen model type was Random Forest. Using this supervised learning technique, a model was fitted using the training data and each feature’s importance was calculated. With this knowledge, the following features were determined to be important in cervical cancer classification:

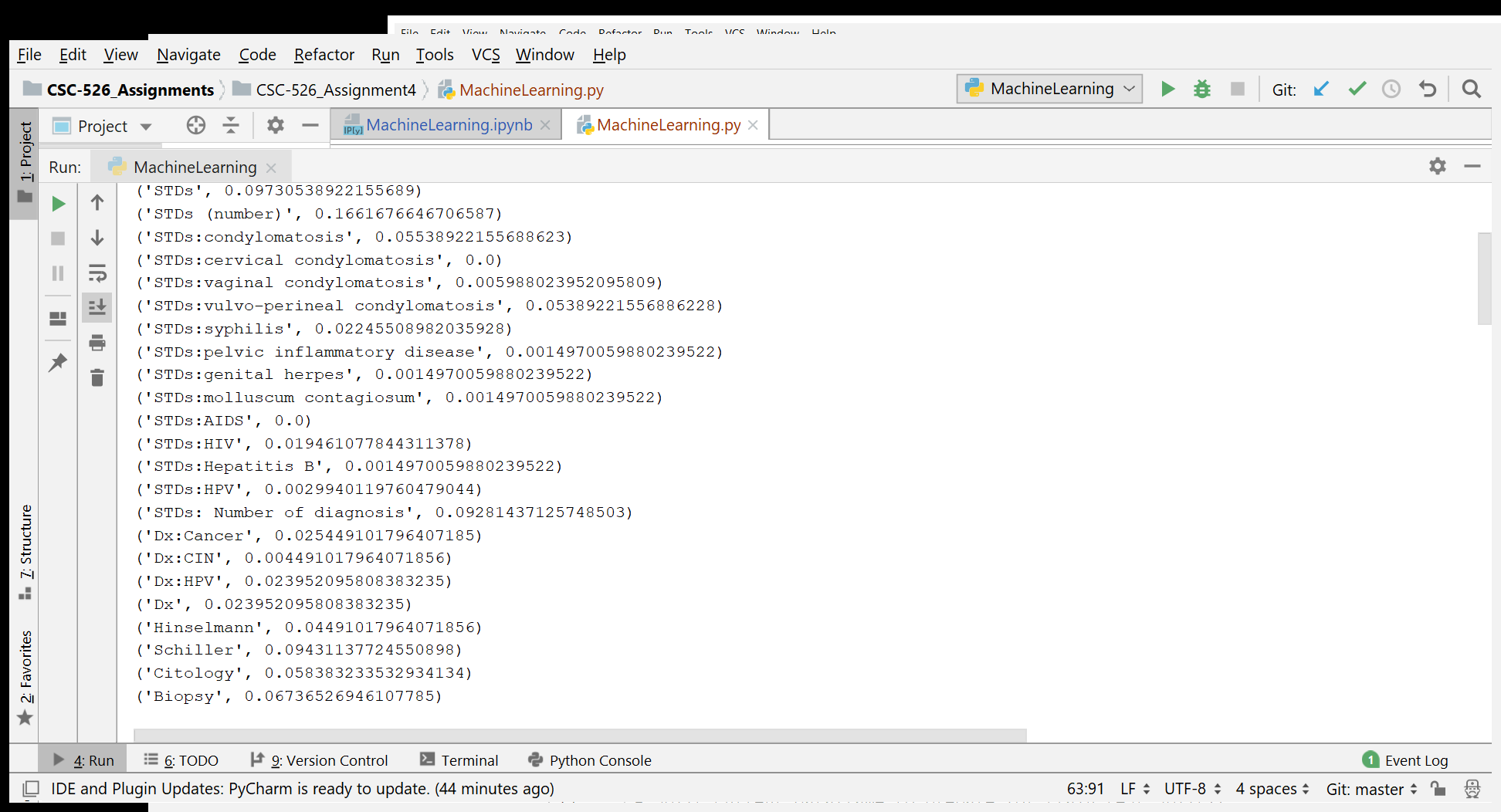
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 'Age' | 'Number of sexual partners' | 'First sexual intercourse' | 'Num of pregnancies' | 'Smokes' | 'Smokes (years)' |
| 'Smokes (packs/year)' | 'Hormonal Contraceptives' | 'Hormonal Contraceptives (years)' | 'IUD' | 'IUD (years)' | 'STDs' |
| 'STDs (number)' | 'STDs: condylomatosis' | 'STDs:  vulvo-perineal condylomatosis' | 'STDs: HPV' | 'STDs: Number of diagnosis' | 'Dx: CIN' |
| 'Dx: HPV' | 'Dx' | 'Hinselmann' | 'Schiller' | 'Citology' | 'Biopsy' |

This means that out of the original thirty-six column features for cervical cancer dataset, only twenty-four of them contributed to classification. Finally, the accuracy of the classifier was computed, which gave a result of 0.982. In conclusion, the classification of whether a patient has cervical cancer can be determined by using the above twenty-four features and the Random Forest model.

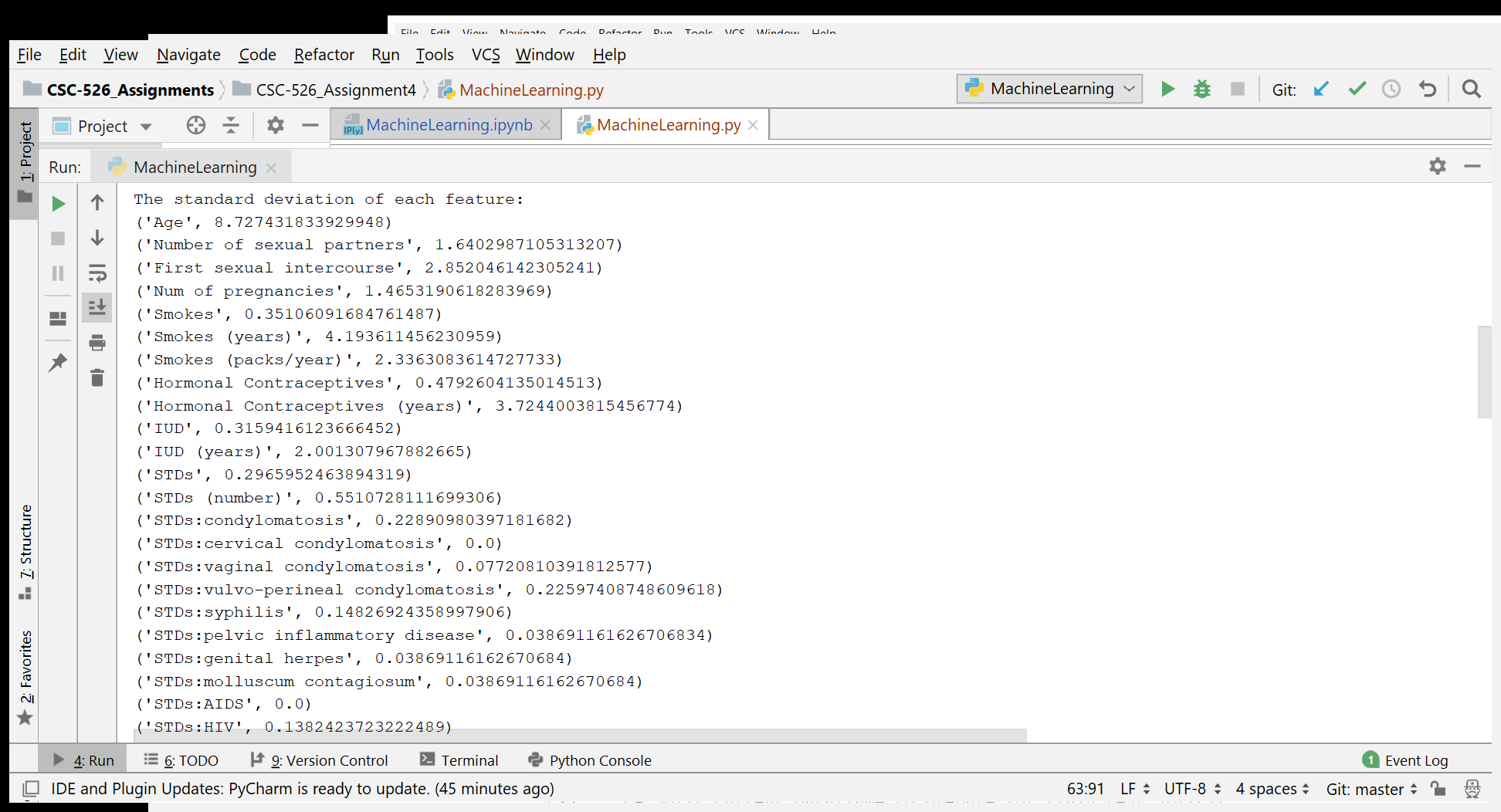
Python Code Screenshots

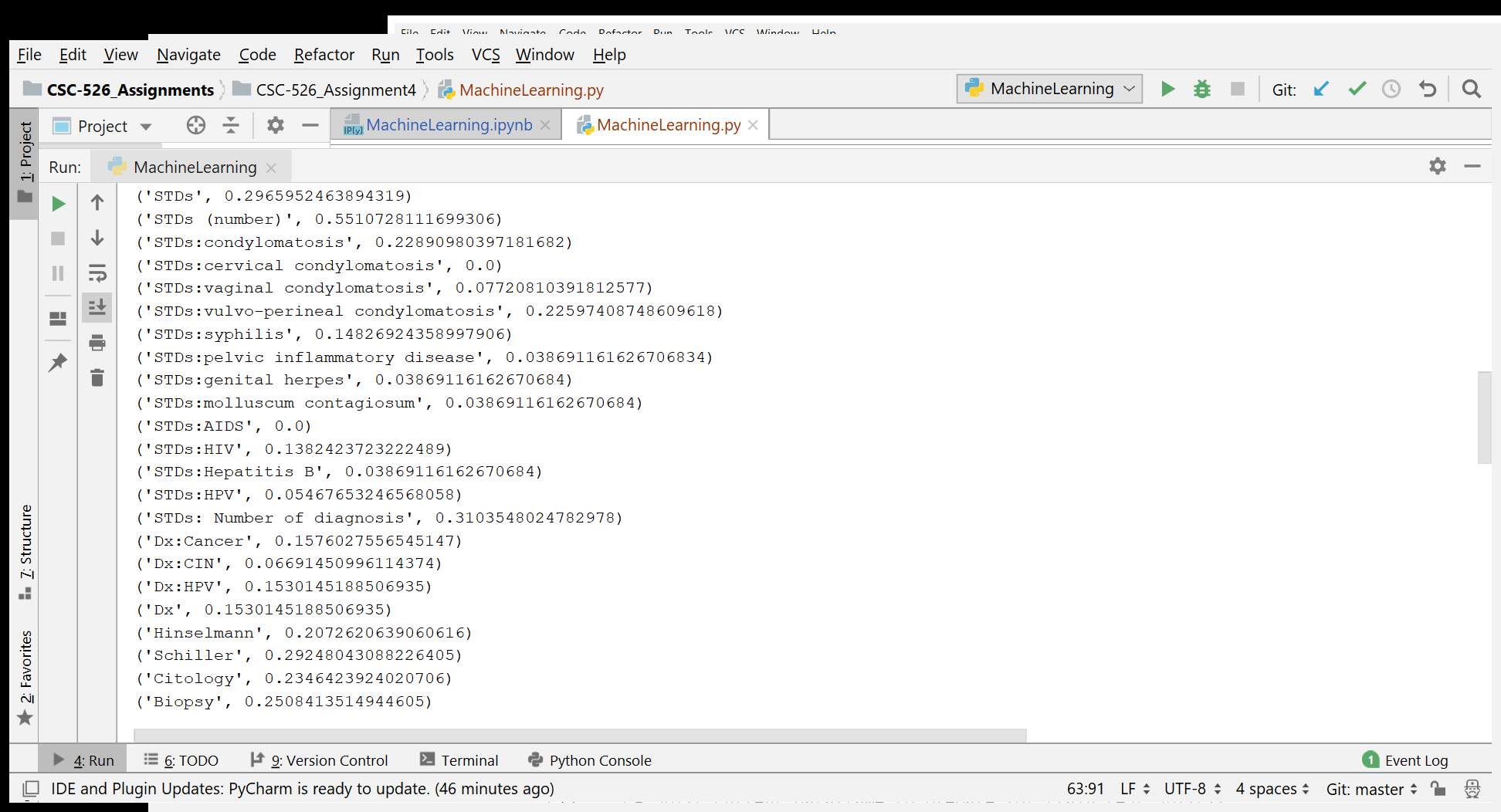
* The Mean of Each Feature



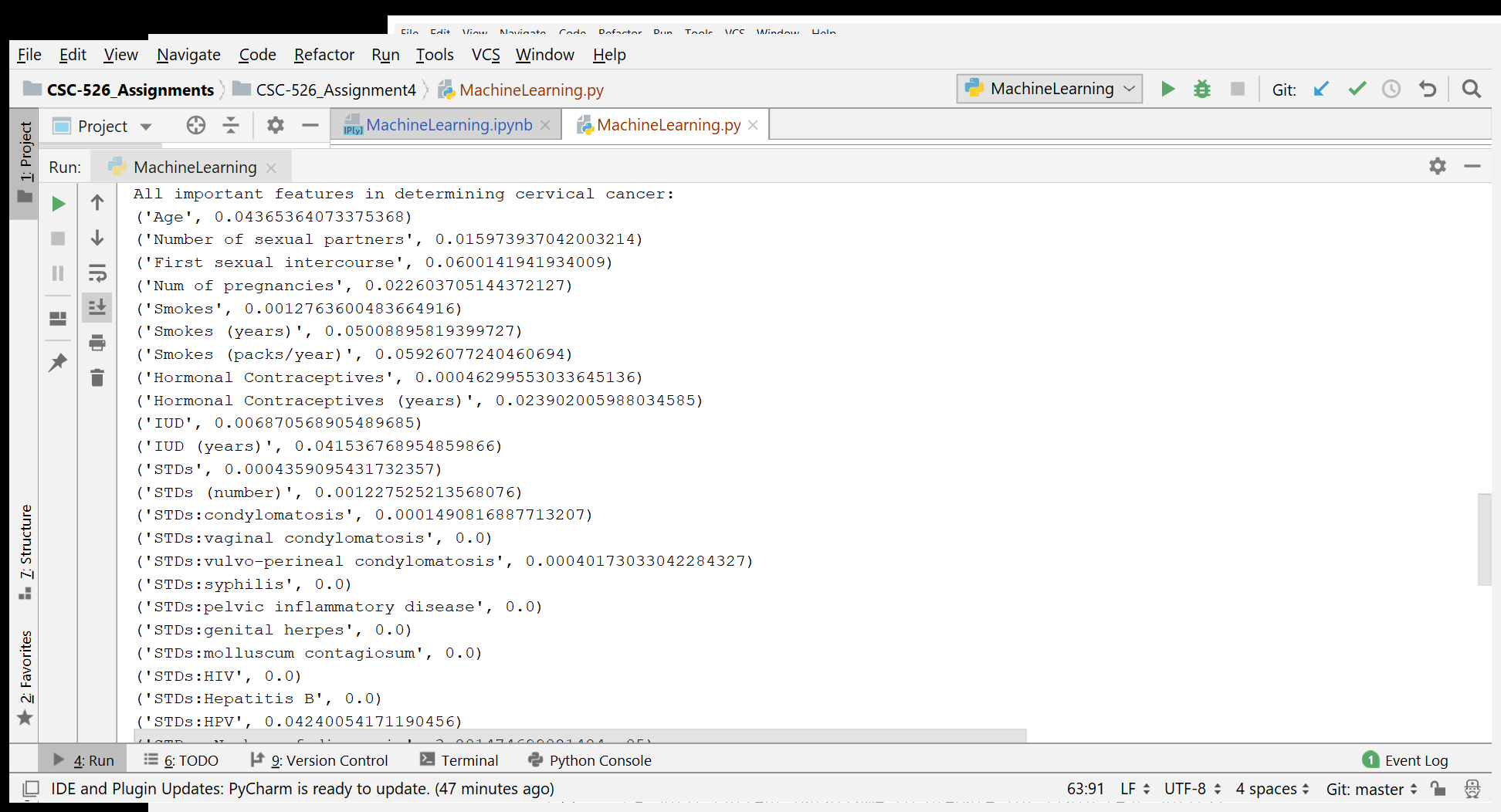


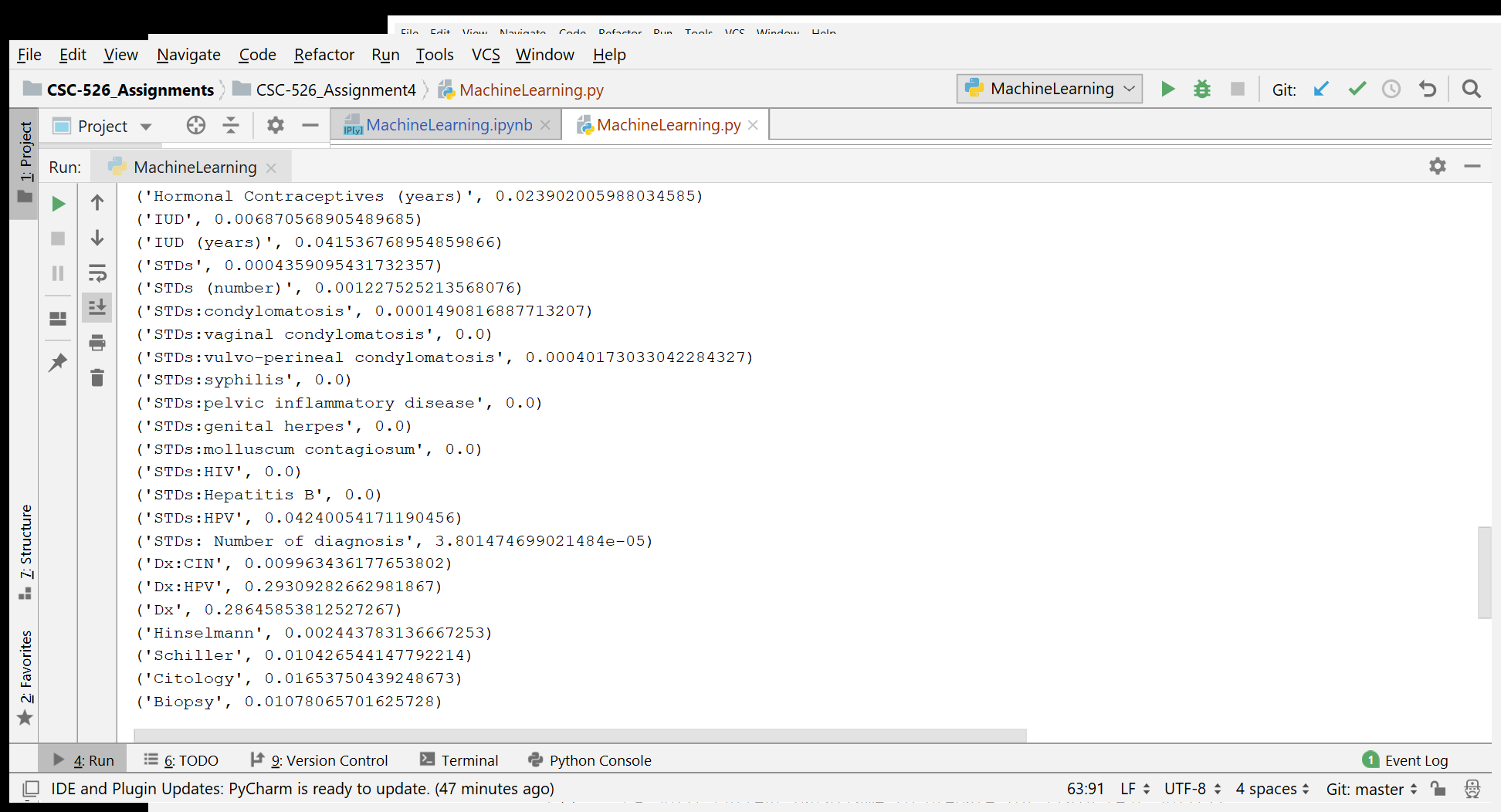
* The Standard Deviation of Each Feature





* All Important Features in Determining Cervical Cancer





* Most Important Features in Determining Cervical Cancer and Random Forest Model Accuracy

