Individual Report

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

Cervical Cancer is a leading cause of death in women. In the United States, there are over eleven thousand new diagnoses of cervical cancer each year. . The stated purpose of this project is to ultimately find a link between certain medical conditions that we use as variables, and the increased risk of being diagnosed with cervical cancer. This project aims to evaluate and compare different classifiers on the Cervical Cancer dataset. There are 4 targets, but for the purposes of this project only one of them, which is the Biopsy, will be classified. The target classes are Boolean but have been encoded. Any target with a value of (1) is encoded for True, which represents malignant cervical cancer, and any target with a value of (0) represents benign cervical cancer. A Logistic Regression, SVM, Decision Tree and Random Forest were trained on the Data. Logistic Regression was our choice of classifier because it had the best combination of Classification Metrics.

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

In the previous publications, people have predicted the risk assessment of cervical cancer using various supervised machine-learning models. Our project was more oriented towards addressing the potential issues that could rise without proper pre-processing. To learn the effect of pre-processing, we had to run various models on the raw data, and see as to what preprocessing methods could be used to address the underlying issues.

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.

A trained physician normally diagnoses cervical cancer through the performance of Biopsy. 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, and Hormonal Contraceptives.

The dataset was collected from the Hospital Universitario de Caracas in Caracas, Venezuela. The data set we utilized had compiled data based on the medical records and demographics of 858 patients, although there is some missing data points since some patients refused to disclosure parts of their medical history [3]. The data set covers a range of information from the patient's age, whether or not they smoke, to the number of pregnancies and sexual partners they've had in the past. There's also an extensive list covering a range of sexually transmitted diseases that a patient may have had been diagnosed with in the past. All these attributes are believed to be possible factors in increasing a person's risk of having invasive cervical cancer.

The performance of different machine learning algorithms on the cervical cancer dataset before and after the preprocessing is explored and compared, to study the effect of preprocessing the data, and to use the best machine learning algorithm for the risk assessment of cervical cancer. Performance measurement of the supervised algorithms. Here, five supervised learning classifiers are trained on the dataset and their prediction accuracy for the test data were measured. This process is mainly divided into two stages where the algorithms were run:

3.3.1 Prior to the preprocessing of the data

3.3.2 After the preprocessing of the data

3.3.1 Models before preprocessing

Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is used to curb these imperfections. But prior to that, we ran several machine learning algorithms on the raw data to study and compare the effects of preprocessing, and as to how much information the data is providing, with regards to the biopsy results, in its entirety.

Given that over 90% of the data had one or more missing values, we had to get rid of observations that are in short supply of data, since running the models with the missing values will be very difficult to deal with. As a result of this, the data has vastly shrunk from over 850 observations initially to 59 in total, of which

only 9 patients are diagnosed with cervical cancer. The training test and the test set were split in such a way that the test set consisted of 30 percent of the observations (=17.7 ≈ 18), of which only 3 women were diagnosed with cancer. The various models that were run on this set of data were mentioned below.
The accuracy of the models run below are dependent on the quality of the raw data, and how much useful information can be gained directly from it. The table below demonstrates the comparison of accuracies of all the models run prior to any cleaning methodology:

| **Model** | **Accuracy** |
|---|---|
| Decision Tree | 94.44 |
| Random Forest | 100 |
| Support Vector Machine | 88.88 |
| *K-Nearest Neighbors | 83.33 |
| Naïve Bayes | 83.33 |

*The accuracy mentioned for the KNN model was the accuracy for the model when n = √59 ≈ 7

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

Given that we've only used one target variable (Biopsy), we would like to further delve into this project and see how other targets would perform on similar machine learning models, and if there is a way we can produce an resembling model that could much lower the false negative rate.

6. Calculate the percentage of the code that you found or copied from the Internet. For example, if you used 50 lines of code from the internet and then you modified 10 of lines and added another 15 lines of your own code, the percentage will be 50 − 10 50 + 15 × 100.

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 inesctec dot pt) - INESC 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

[4]Analytics Vidhya:

https://www.analyticsvidhya.com/blog/category/machine-learning/