Cervical Cancer Risk: a Data Mining approach

Luca Gandolfi* Bruno Palazzi† Stefano Sacco‡

Abstract

Cervical Cancer is one of the most preventable type of cancer, despite this it kills 300 thousands of women every year. A lot of studies have been published in order to discover causes and to detect malicious behaviours that can affect people's health. Data Mining can be helpful in this situations, according to the use of a detailed dataset. This paper proposes different approaches in order to explore, analyze and predict Cervical Cancer and its causes. In addition, results of 4 different classification models are presented, including a deep learning network, and it is shown that results are highly affected by the feature reduction task.

Keywords: cervical cancer, data exploration, classification, imbalance, feature reduction, clustering

1 Introduction

In the past years, the study of serious diseases with a machine learning approach has become very popular, although it is a challenging task because serious diseases are rare and therefore, data isn't available at large scale.

Cervical Cancer is one of the most preventable type of cancer. Despite this, it kills 300 thousands of women every year. A lot of studies have been published in order to discover causes and to detect malicious behaviours that could cause this disease, but it's still an open problem.

This project's aims are to find hidden relations between detailed women information and the risk of cancer, and to develop a predictive model able to predict whenever a woman should submit herself to a biopsy analysis.

2 Dataset

The dataset used in this analysis consists of several information about a sample of 858 women. In particular, women are described by 36 attributes, visible in details in Table 1, along with the number of missing values. This problem is studied in depth in section 3.1, along with strategies adopted to handle it. Attributes concern about patients' age, sexual activity, smoke addiction, usage of contraceptives and sexual diseases. The dataset is highly unbalanced according to the target attribute Biopsy, with only the 6.4% of the women resulted positive to the biopsy test. This distribution is showed in Figure 1.

The majority of women are young, women past forties and older are outliers. Age distribution is visible in Figure 2. It is known that early age at first sexual intercourse and early pregnancy are risk factors for cervical cancer, as shown in Louie K. [3], and therefore the dataset contains only women with at least one pregnancy. In section 3.2, is shown that this known correlation is visible in this dataset too.

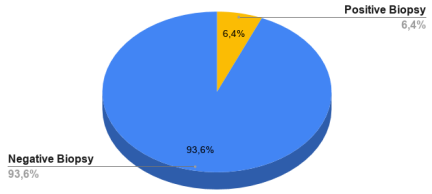

FIGURE 1: **Biopsy distribution: highly unbalanced.**

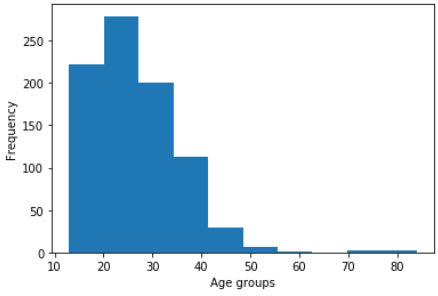

FIGURE 2: **Age distribution.**

3 The Methodological Approach

The methodological approach adopted in this project consists in 5 main tasks: (1) missing values analysis, to understand how to handle them, (2) correlation analysis among attributes, with a particular focus on the target attribute, (3) cluster analysis, (4) oversampling of the minority class to

DISCO, Università degli Studi di Milano Bicocca.
*Email: l.gandolfi3@campus.unimib.it ID Number: 807485
†Email: b.palazzi@campus.unimib.it ID Number: 806908
‡Email: s.sacco3@campus.unimib.it ID Number: 807532

TABLE 1: **Number of missing values for each attribute.**

Attribute	Missing values
Age	0
Number of sexual partners	26
First sexual intercourse	7
Num of pregnancies	56
Smokes	13
Smokes (years)	13
Smokes (packs/year)	13
Hormonal Contraceptives	108
Hormonal Contraceptives (years)	108
IUD	117
IUD (years)	117
STDs	105
STDs (number)	105
STDs:condylomatosis	105
STDs:cervical condylomatosis	105
STDs:vaginal condylomatosis	105
STDs:vulvo-perineal condylomatosis	105
STDs:syphilis	105
STDs:pelvic inflammatory disease	105
STDs:genital herpes	105
STDs:molluscum contagiosum	105
STDs:AIDS	105
STDs:HIV	105
STDs:Hepatitis B	105
STDs:HPV	105
STDs: Number of diagnosis	0
STDs: Time since first diagnosis	787
STDs: Time since last diagnosis	787
Dx:Cancer	0
Dx:CIN	0
Dx:HPV	0
Dx	0
Hinselmann	0
Schiller	0
Citology	0
Biopsy	0

better handle the imbalance problem and, finally, (5) classification, in order to develop a model with the lowest False Negatives rate. Each of this task is explained in details in the following subsections.

3.1 Missing Values Replacement

As visible in Table 1, there are lots of missing values for several attributes. In order to obtain a more consistent dataset, it is highly important to manage this lack of data. The attributes *STDs: Time since first diagnosis* and *STDs: Time since last diagnosis* have been removed due to the high number of missing values, furthermore, it is impossible to impute them because they are Missing Completely at Random (MCAR) and it is impractical to reach data sources, in order to know their real values. For *Number of sexual partners* and *Num of pregnancies* attributes, it has been decided to replace them with the most frequent value among the patients with the same *Age* value. Attributes related to *Smokes* have been replaced with the most frequent value and, consequently, they have been set at zero. For what concerns Sexual Transmitted Diseases, attributes regarding *Time since diagnosis* have been removed due to massive lack of data, while *AIDS* and *cervical condylomatosis* have been removed because they were composed of only zeros. Finally, the other missing attributes have been imputed according to the most frequent values. After this pre-processing task, the dataset was composed of 32 attributes, including the target, without missing values.

3.2 Correlation Analysis

The aim of correlation analysis is to find hidden relations and hidden patterns among attributes. Since the number of attributes is high, not all attributes are representative and good descriptors for the target attribute Biopsy, so the correlation analysis could help to perform a first feature selection task.

Looking at the correlation plot in Figure 3, it can be seen that the most correlated attributes with the target class are Schiller, Hinselmann and Citology, that are other tests to perform in order to discover cervical cancer. A more difficult analysis would consider all these attributes and Biopsy together as target features, but in this project it has been studied only the latter mentioned.

Following what have been shown in [3], this correlation can be seen in this dataset too (Figure 4). In particular, it is clear that the majority of Biopsy-positive women are young, near twenties, with an age lower than twenty at first sexual intercourse.

Analysing the correlation between cervical cancer and sexual transmitted diseases, it can be seen that they are poorly correlated. In particular, the distribution in the dataset shows how it is easier to have a positive Biopsy result while not having STDs, according to the pie plot in Figure 5.

Dx is a Genomic test to discover Breast Cancer. Looking at Table 2 it is possible to see that the probability of resulting positive to a Dx test while having Cervical cancer is higher than the probability of being positive to a Dx test while not having Cervical cancer. However, the probability of

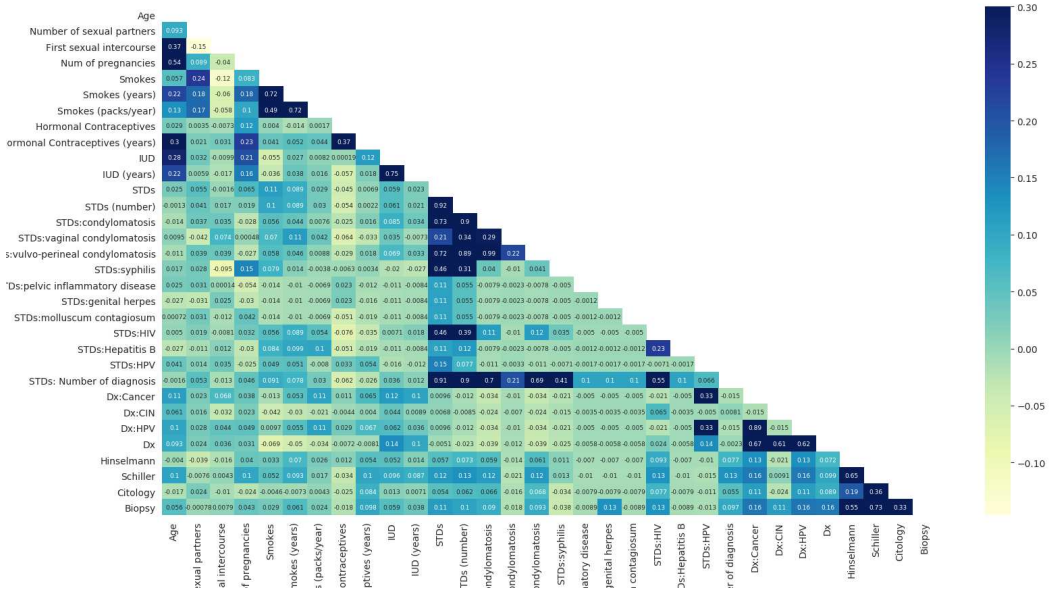

FIGURE 3: **Correlation matrix after Data Manipulation.**

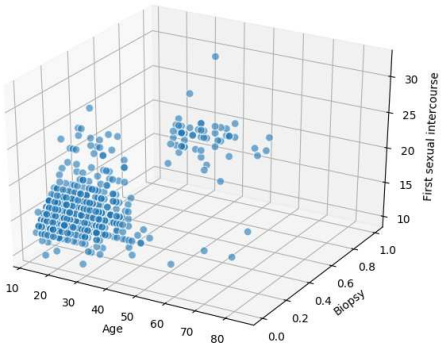

FIGURE 4: **Correlation between age and first sexual intercourse according to Biopsy.**

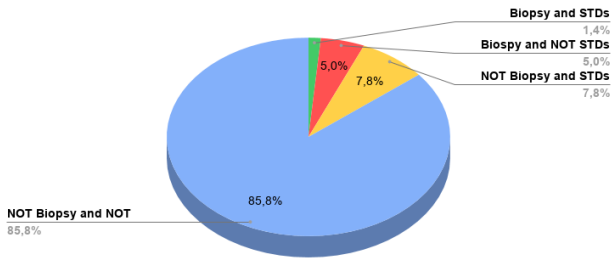

FIGURE 5: **Correlation between Biopsy and Sexual Transmitted Diseases.**

women having a Breast cancer is very low in this dataset, and therefore it can't be demonstrated a real strong correlation.

TABLE 2: **Relation between Biopsy and Dx tests.**

	Positive Biopsy	Negative Biopsy	Freq
Positive Dx	7	17	0.1458
Negative Dx	48	786	0.0216

Finally, it can be seen that pregnancies improve the probability of Cervical cancer. In particular, Figure 6 shows how the number of Biopsy-positive women stays on the same level even if the total number of women decreases while the number of pregnancies grows.

Despite what emerged in recent studies, like [4], in this dataset smoking isn't a factor of Cervical Cancer and the correlation is very low. However, there are not enough information to make the decision to remove Smokes columns and this dropping task will be performed during the Feature Selection step, described in section 3.5.1.

3.3 Cluster Analysis

Clustering algorithms permit to obtain groups of objects that share common characteristics, and to find out hidden relations among data. Furthermore they provide an easy visual inspection of these groups, in order to have a high level representation of them. Cluster analysis helps to better

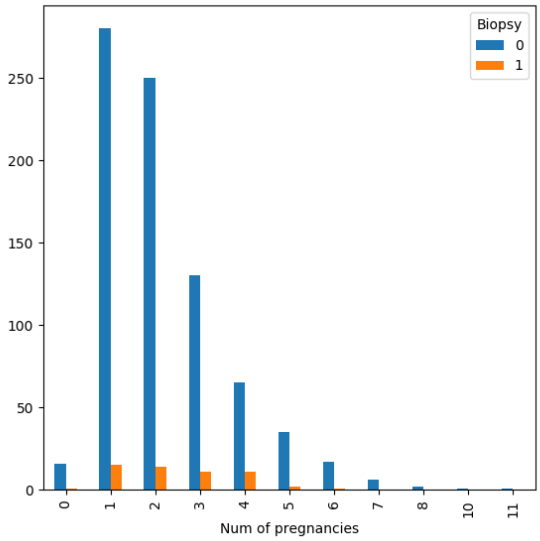

FIGURE 6: **Number of child vs Biopsy.**

understand which type of group are present in this sample of women and which properties are shared among people inside a group. It has been decided to find 2, 3 and 4 different clusters applying a K-means algorithm, computing a maximum of 350 iterations. It has been decided to exploit only attributes related to Biopsy to obtain clusters, in this way only 9 attributes have been selected: attributes related to women's age and sexual activities (4), attributes related to smoke addiction (2), hormonal contraceptives usage (2) and Dx presence (1).

An important task while performing Cluster Analysis is Cluster Validity which permits to evaluate results obtained by a Clustering Algorithm. In order to do that it has been used silhouette coefficient measure on clusters found by K-means algorithm. Results obtained has been shown in Table 3. Optimal number of clusters has been individuated in 3, achieving the highest overall mean Silhouette score. In addition, it allowed to obtain more well-defined clusters than other configurations. In Figure 7 it is showed a pie chart representing the percentage of data belonging to each of three different clusters.

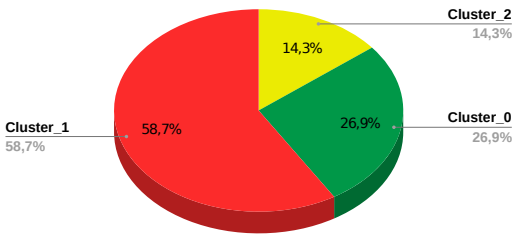

FIGURE 7: **Pie Chart representing distribution of women per cluster**

TABLE 3: **Mean Overall Silhouette coefficient for different clusters' size and Silhouette coefficient scores for size 3.**

	2 Clusters	3 Clusters	4 Clusters
Overall	0.450	**0.588**	0.404
	Green cluster	Red cluster	Yellow cluster
Scores	0.643	0.628	0.323

To better understand how these 3 Clusters are composed, it has been provided a t-Distributed Stochastic Neighbor Embedding (t-SNE) [2] plot visible in Figure 8, representing groups in a two-dimensional space. Clusters are mostly separated and well defined, but slightly scattered.

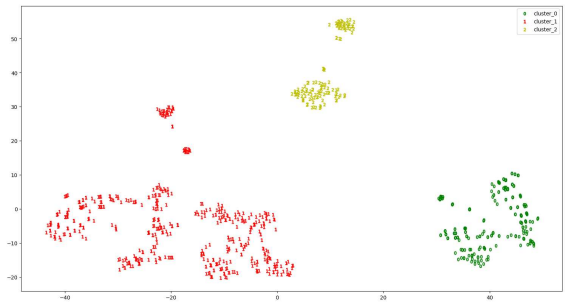

FIGURE 8: **t-SNE plot of Clusters in two dimensional space.**

Analyzing Figure 8 and Table 3, it's possible to notice that the Yellow cluster is the less cohesive: graphically it's composed of 2 different dense areas, separated with a low density area. To improve the quality of the cluster, it has been decided to split it further into 2 sub clusters, which are actually more cohesive. After a new computation of Silhouette Coefficient, higher results (shown in Table 4) have been obtained, while graphically in Figure 9 it's possible to see that the old yellow cluster is now split in 2 different clusters (yellow and blue) without low-density areas. In addition, the overall Silhouette Coefficient has increased compared with the overall score achieved looking for 4 clusters from scratch (Table 3).

TABLE 4: **Silhouette coefficient scores after improving cluster quality.**

	Green	Red	Yellow	Blue	Overall
Scores	0.643	0.628	0.588	0.479	0.621

Every cluster has been investigated, in order to understand which characteristics are shared among women belonging to the same group. Clusters are well defined by 2 attributes: Smokes and Hormonal Contraceptives. In particular, two

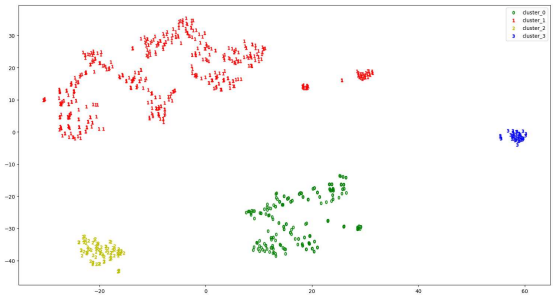

FIGURE 9: **t-SNE plot of Clusters after splitting the yellow cluster.**

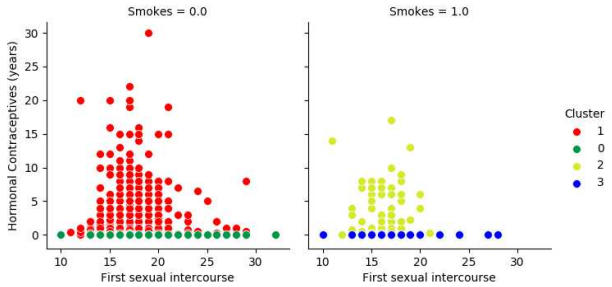

FIGURE 10: **Cluster Inspection on First Sexual Intercourse, Hormonal Contraceptives (years) and Smokes attributes**

clusters are made up only of women addicted to smoking, while the other two are made up only of women who are not addicted to smoking. Considering this first division, both cluster pairs differ in the use of hormonal contraceptives. Summarizing:

- *Red cluster*: composed of 504 women who have never smoked but have used hormonal contraceptives.
- *Green cluster*: composed of 231 women who neither smokes nor use hormonal contraceptives.
- *Yellow cluster*: composed of 85 women with smoking addiction and hormonal contraceptives usage.
- *Blue cluster*: composed of 38 women with smoking addiction and never used hormonal contraceptives.

Following this relevance, a visual analysis has been carried out on these 2 attributes considering also the age at first sexual intercourse, because it presents a vastly variety of values. Hormonal contraceptives has been taken in account as its years attribute, to provide a better visualization. It is obvious that this choice does not affect cluster distributions, since women who have never used hormonal contraceptives presented zero years of hormonal contraceptives usage, while some women who have used hormonal contraceptives for an extremely low time, presented values near zero on hormonal contraceptives (years) attribute. This situation lead to overlapping points near zero on the y-axis, as shown in Figure 10, for both Red and Yellow cluster.

3.4 Oversampling

The imbalance problem is always hard to handle in situations where the minority class has an high importance and thus it requires to be correctly classified. Due to the low number of data, applying undersampling to the majority class was unfeasible. As oversampling technique, SMOTE [1] has been used, which creates synthetic data near to already existing minority class data. This approach allows to create dense areas in the space. A problem of SMOTE is the original sparsity of the data classes: considering the nearest neighbours could confuse the learner instead of help to improve recognition. Classes have been re-balanced to a close 50%-50% distribution. It is worth while to mention that synthetic data has not been used during test phase in the classification task: only original data has been used to evaluate models.

3.5 Classification

In order to develop a model able to predict whenever a woman should perform a Biopsy test, four different models have been tested, each of these with two different feature reduction techniques. In the following sections, the choices taken are described in detail, while results are analyzed in section 4.

3.5.1 Feature Selection

In order to avoid the *curse of dimensionality* problem, two different filters have been implemented to select a subset of attributes to impute a classification model. In this way, *Uni-variate* and *Multi-variate* filters have been developed. Uni-variate filter considers every feature in an isolated way and it makes the assumption of indipendecy among features. It finds the subset of attributes by choosing an association measure between candidate attributes and target attribute, sorting them according to their score. It has been decided to choose a gain ratio as association measure, exploiting the Information Gain value associated to every feature. This filter allows to identify irrelevant attributes, in order to discard them, and redundant attributes that need to be kept. Once attributes have been sorted, the first 12 out of 31 have been selected, to learn a classifier. Multi-variate filter jointly identifies both irrelevant and redundant attributes according to the target variable: a good subset of features must contains attributes which are strongly correlated with the target attribute but uncorrelated among them. It has been decided to choose a Multi-variate filter based on Correlation Feature Selection, in order to find a subsest of 12 attributes.

Once feature selection process ended, two different predictor have been used to compare Uni-variate and Multi-variate

subset of attributes, in order to find out which one fits better. Both models have pretty good performances considering accuracy score, but the univariate one obtains a 97% score and improves the multi-variate by a single percentage point.

Due to the best results achieved by the first one, it has been decided to use the 12 features selected by the Uni-variate model because they would allow our classifier to achieve better results. Selected features are:

- Schiller
- Hinselmann
- Citology
- Hormonal Contraceptives (years)
- STDs:HIV
- Dx
- Dx:HPV
- Dx:Cancer
- DxCIN
- Smokes (packs/year)
- IUD
- Hormonal Contraceptives

In Figure 11 it is showed a 2-dimensional representation of the 12 features selected, using t-SNE. Dense areas of Biopsy-negative women are clearly visible, while Biopsy-positive women are more sparse.

3.5.2 Feature Extraction

One of the most famous technique for feature extraction is Principal Component Analysis (PCA). In order to maintain an high percentage of information, 15 principal components have been extracted from the 31 attributes obtained after the pre-processing (section 3.1). Extracting 15 PC means to ensure a 99.9% of original information in the dataset. In Figure 12 it is shown a 2-dimensional representation of the 15 components, using t-SNE. Different clusters are visible in the plot, but there isn't a cluster containing mainly Biopsy-positive women. For this reason, developing a model able to classify Biopsy will be a difficult task. Different number of principal components have been tested, but none of them could improve vectors separation. The sparsity of these vectors is also higher than the feature selection results vectors.

3.5.3 Models

Different approaches have been tested in order to find the best-performing model. In particular, 4 models have been defined:

- 2 Support Vector Machines
- 1 Multi Layer Perceptron
- 1 Keras Deep Learning Neural Network

All models have been applied to original *features selected* data and *features extracted* data, and to the same *oversampled* data. A 3-Fold Stratified Cross Validation learning method has been used. The low fold's number is necessary because original data is highly unbalanced and, therefore, an higher number of folds would lead to a lower number of minority class samples in test set. In Table 5 are reported results about the Feature Selection approach, while in Table 6 are reported results about the PCA approach. To provide a general overview of the performances, 3 metrics are compared: Pecision, Recall and F1-score. An higher focus has been given to results on the Biopsy-positive class. In particular, tables show in bold characters the higher results achieved in terms of f1-score on the positive class.

4 Discussion

In this section, models performances are described and analyzed. A general overview of results in terms of Precision, Recall and F1-score is summarized in Tables 5 and 6.

4.1 Models performances

Due to the low number of women positive to Biopsy test in this sample, it has been decided to focus the attention on Recall and F-measure of minority class. Recall and F-measure are very useful measures in case of imbalanced dataset, because they give important information about the model and its ability to discriminate data belonging to the minority class. In particular, Recall measure allows to select best model when there is a high cost associated with False Negative while F-measure is a measure to use if you need to seek a balance between Precision and Recall and there is an uneven class distribution (large number of Actual Negatives).

Considering the higher importance of correctly classify Biopsy-positive women, a micro average F1-score has been computed, which allows to consider in the same way both classes. The highest result has been obtained with both SVM SPegasos and SVM SMO models on *features selected* original data, while the lowest has been obtained with SVM SPegasos on *features selected* oversampled data. A general overview of results is shown in Figure 13, while micro average f1-score values are reported in Table 7.

In terms of Learning Time, Keras DNN is the most time consuming model: increasing the number of hidden layers leads to an higher number of parameters to be computed during the learning phase. This issue could be reduced increasing the batch size, but better results have been achieved using a small dimension, which allows to perform more iterations during the training epochs. The other classifiers are characterized by a very fast learning phase, despite they have been trained for more epochs.

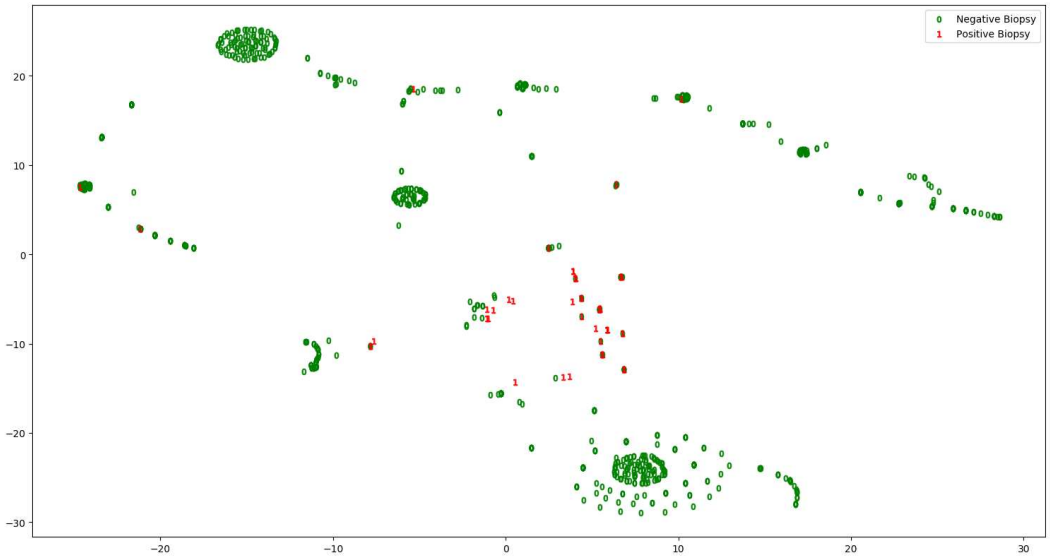

FIGURE 11: **t-SNE bidimensional representation of the 12 features selected.**

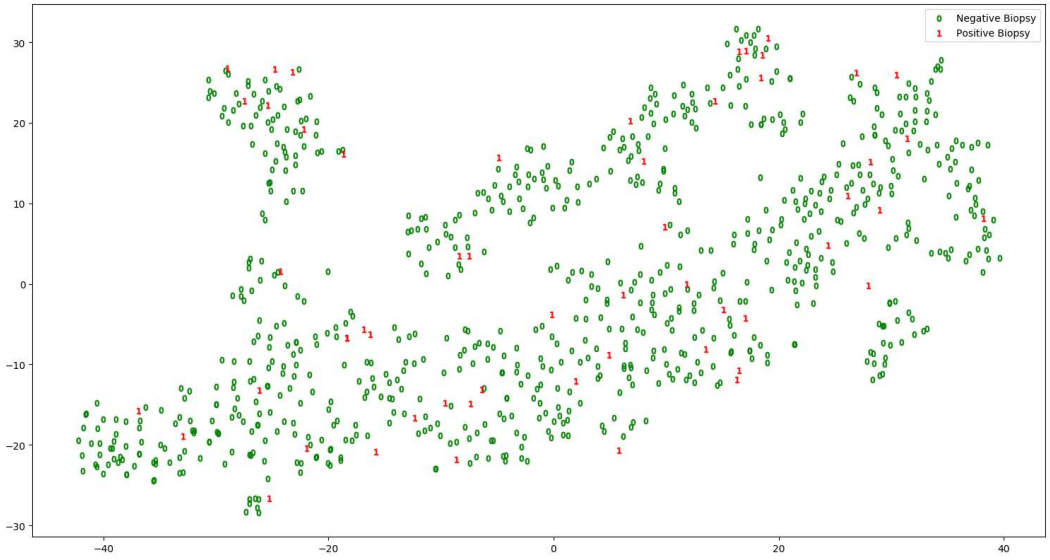

FIGURE 12: **t-SNE bidimensional representation of the 15 Principal Components extracted.**

TABLE 5: **Feature Selection approach's results achieved during experiments.**

		Original			**Oversampled**		
		Precision	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
SVM SPegasos	Negative Biopsy	0.9911	0.9676	0.9792	0.9933	0.5529	0.7104
	Positive Biopsy	0.6486	0.8727	**0.7442**	0.1265	0.9454	0.2231
SVM SMO	Negative Biopsy	0.9911	0.9676	0.9792	0.9935	0.9601	0.9765
	Positive Biopsy	0.6486	0.8727	**0.7442**	0.6097	0.9091	0.7299
MLP RProp	Negative Biopsy	0.9691	0.9763	0.9727	0.9686	0.9613	0.9650
	Positive Biopsy	0.6122	0.5454	0.5769	0.4918	0.5454	0.5172
Keras DNN	Negative Biopsy	0.9668	0.9801	0.9734	0.9847	0.9614	0.9729
	Positive Biopsy	0.6363	0.5091	0.5656	0.5811	0.7818	0.6666

Note. All results have been achieved computing a 3-Fold Cross Validation.

TABLE 6: **PCA approach's results achieved during experiments.**

		Original			**Oversampled**		
		Precision	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
SVM SPegasos	Negative Biopsy	0.9621	0.9801	0.9710	0.9753	0.9327	0.9535
	Positive Biopsy	0.6000	0.4364	0.5052	0.4000	0.6545	0.4965
SVM SMO	Negative Biopsy	0.9823	0.9701	0.9762	0.9897	0.9639	0.9767
	Positive Biopsy	0.6307	0.7454	0.6833	0.6184	0.8545	**0.7175**
MLP RProp	Negative Biopsy	0.9607	0.9738	0.9672	0.9711	0.9639	0.9675
	Positive Biopsy	0.5227	0.4182	0.4646	0.5246	0.5818	0.5517
Keras DNN	Negative Biopsy	0.9444	0.9938	0.9684	0.9744	0.9489	0.9615
	Positive Biopsy	0.6154	0.1454	0.2353	0.4605	0.6363	0.5343

Note. All results have been achieved computing a 3-Fold Cross Validation.

4.2 Results Analysis

Despite of what has been expected, an oversampling approach has not improved results, and has increased the difficulty in discrimination of Cervical Cancer. Recall has growth but, on the other hand, Precision has decreased. In particular, models were able to detect the majority of Biopsy-positive women but a lot of Biopsy-negative women were classified as positive. Looking at results it's clear that a lot of synthetic data doesn't help in the learning process: a better approach would have oversampled less data, for example achieving a 65%-35% balance, but KNIME platform doesn't support this kind of oversampling. It's unfeasible to apply an undersample on majority class and then to oversample, like proposed in [1], since the original number of samples is very low.

For what concerns results on original data, evaluating models in micro average has showed results higher than expected, but the real problem remains the correct classification of Biopsy-positive women, reducing the number of False Negatives, which is very more dangerous than classifying a woman as positive while she is not. That's because the proposed models should give an idea on when take the Biopsy exam, and suggest to a negative woman to take the exam is not as dangerous as not suggest it to a positive woman.

Focusing the attention on differences between Feature Selection approach and PCA approach, the first one has lead to higher performances. This result was already clear looking at t-SNE representations of both feature reduction approaches (Figure 11 and 12), since Feature Selection brought to vectors more grouped in dense areas than Principal Components vectors. The presence of dense areas facilitates learning a classification model.

Deep Learning is a powerful approach, but defining a

FIGURE 13: **Micro average F1 values for each classifier.**

TABLE 7: **F1-score micro avg results achieved during experiments.**

	Original		**Oversampled**	
	Feature Selection	*PCA*	*Feature Selection*	*PCA*
SVM SPegasos	**0.9615**	0.9452	0.5781	0.9149
SVM SMO	**0.9615**	0.9897	0.9557	0.9568
MLP RProp	0.9487	0.9382	0.9347	0.9394
Keras DNN	0.9499	0.9394	0.9499	0.9289

good Feed Forward Neural Network requires often a Hyperparameters Optimization task. That's the reason behind results of Keras Deep Learning NN, which are lower than Machine Learning techniques' results in most experiments.

5 Conclusions

Classification of serious diseases is a very challenging task, considering the high imbalance in most accessible datasets. Furthermore, it requires a background knowledge in medical domains, to better understand patterns in data.

Analyzing the dataset, it has emerged that young women and women having first sexual intercourse at early age are predisposed to Cervical Cancer. Early pregnancies also influences it, while Breast Cancer and Cervical Cancer are correlated. Clustering Analysis has detected groups of women described by similar characteristics, for example smoking addiction and hormonal contraceptives usage, but it is hard to extract additional knowledge with respect to the scientific literature.

In general, satisfying results have been achieved through the developed classification models. In particular, the best one was Support Vector Machines with SMO algorithm applied on the original features selected data. The Oversampling technique generates too many synthetic samples and therefore it gives a distorted representation of the minority class, which reduces discriminating capability of developed models.

An interesting future work could be considering to apply PCA on Feature Selections results, in order to obtain lower dimension vectors. PCA applied to original data could not individuate features that were able to discriminate classes. To handle this problem, more advanced techniques could be exploited, like Autoencoders.

References

[1] Bowyer, K. W., Chawla, N. V., Hall, L. O., & Kegelmeyer, W. P. (2011). SMOTE: Synthetic Minority Over-sampling Technique. *CoRR*, *abs/1106.1813* http://arxiv.org/abs/1106.1813.

[2] L.J.P. van der Maaten, G. H. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, (9(Nov)), 2579–2605.

[3] Louie K., de Sanjose S., D. M. (2009). Early age at first sexual intercourse and early pregnancy are risk factors for cervical cancer in developing countries. *British Journal of Cancer*, *100*, 1191–1197 https://doi.org/10.1038/sj.bjc.6604974.

[4] Roura, E., et al. (2014). Smoking as a major risk factor for cervical cancer and pre-cancer: Results from the EPIC cohort. *International Journal of Cancer*, *135*(2), 453–466, https://doi.org/10.1002/ijc.28666 https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.28666.