ALBUKHARY INTERNATIONAL UNIVERSITY

CCS2213: Machine Learning

Academic Session: Semester 2, 2024-2025

SCHOOL OF COMPUTING & INFORMATICS

AlBukhary International University (AIU)

INDIVIDUAL ASSIGNMENT

May Mon Ko

AIU23102145

T10

Cervical Cancer (Risk Factors) Data Set

Contents

1.0 Dataset Background

1.A.(i) Usage of the Dataset and Trends, Including Proposed Approaches

The **Cervical Cancer Risk Factors** dataset, donated in 2017, was collected from the Hospital Universitario de Caracas in Venezuela. It includes **858 instances** and **36 features**, comprising a wide range of demographic, lifestyle, sexual behavior, medical history, and diagnostic test results. Due to privacy concerns, many features contain **missing values**, which makes preprocessing a crucial step in using this dataset effectively.

This dataset has been widely adopted in **medical machine learning research** for **classification tasks**, especially in **predicting cervical cancer diagnoses**. The four main target variables — Hinselmann, Schiller, Cytology, and Biopsy — represent different diagnostic methods and serve as output labels in many predictive modelling studies.

Over the past five years, researchers have increasingly shifted from simple statistical methods and traditional ML algorithms (like Decision Trees and Logistic Regression) to more **advanced techniques** like:

- Ensemble models (Random Forest, XGBoost)
- Support Vector Machines (SVM)
- Deep Learning (ANNs, CNNs)
- **Transfer Learning** and **Feature Selection** approaches

A common trend in recent literature is addressing **class imbalance**, **missing data handling**, and **improving early detection** to assist in **automated screening tools**, particularly in low-resource environments.

1.A.(ii) Literature Review

Title: Diagnosis of Cervical Cancer and Pre-Cancerous Lesions by Artificial Intelligence: A Systematic Review (Allahqoli et al., 2022)

Authors: Allahqoli, Leila; Laganà, Antonio Simone; Afrooz Mazidimoradi; Salehiniya, Hamid; Günther, Veronika; Chiantera, Vito; Shirin Karimi Goghari; Mohammad Matin Ghiasvand; Rahmani, Azam; Momenimovahed, Zohre.

Year: 2022

Journal: Diagnostics, Basel

Page Number: Article 2771

Title: Cervical Cancer Prediction Empowered with Federated Machine Learning (Nasir et al., 2024)

Authors: Muhammad Umar Nasir, Omar Kassem Khalil, Karamath Ateeq, Bassam Saleem Allah Almogadwy, M. A. Khan, Khan Muhammad Adnan

Year: 2024

Journal: Computers, Materials & Continua (CMC)

Page Number: Not explicitly listed, but an article-level DOI is provided

Title: Supervised Deep Learning Embeddings for the Prediction of Cervical Cancer Diagnosis (Fernandes et al., 2018)

Authors: Kelwin Fernandes, Davide Chicco, Jaime S. Cardoso, Jessica Fernandes

Year: 2018

Journal: PeerJ Computer Science

Page Number: Article e154

Title: Cervical Cancer Prediction Using Machine Learning (Ashtagi Rashmi et al., 2024)

Authors: Rashmi Ashtagi, Vaishali Rajput, Sonali Antad, Pratiksha Chopade, Atharva Chívate, Shreeshail Chitpur, Isha Dashetwar

Year: 2024

Journal: Journal of Electrical Systems; Paris

Pages Number: 944–955

Title: A Concise Review for Exploring Deep Learning's Potential in Cervical Cancer Prediction from Medical Images (Mythili, 2024)

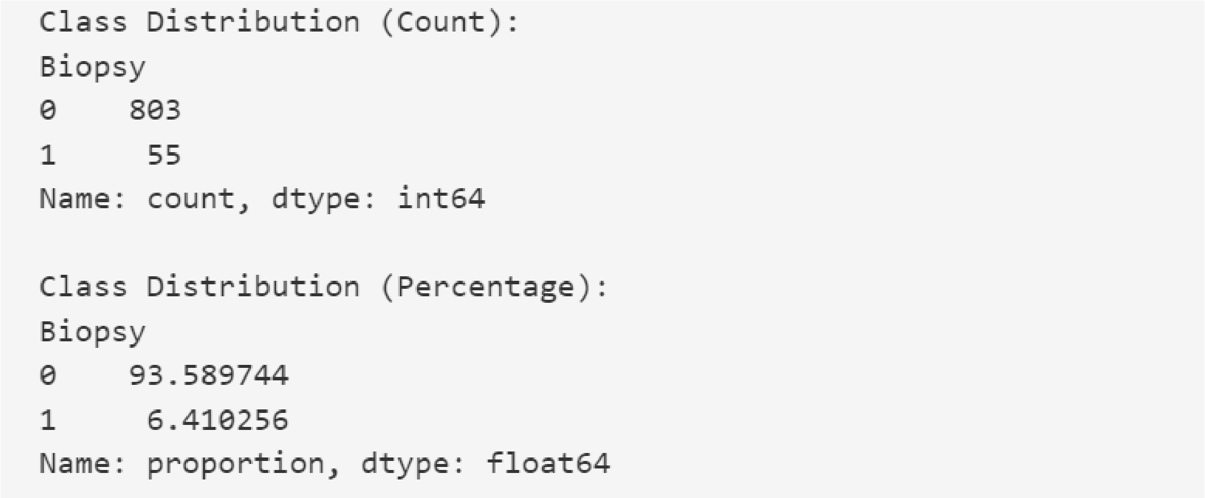
Authors: J. Mythili, S. Chitra

Year: 2024

Journal: International Journal of Advanced Research in Computer Science; Udaipur

Pages Number: 28–36

1. B Report On Data Distribution



```
Class Distribution (Count):
Biopsy
0    803
1     55
Name: count, dtype: int64

Class Distribution (Percentage):
Biopsy
0    93.589744
1     6.410256
Name: proportion, dtype: float64
```

This is the result of finding the data imbalance of the risk factors of the cervical cancer dataset.

The dataset contains two classes in the target column, Biopsy, indicating whether a patient has cervical cancer (1) or not (0). The distribution of classes is as follows:

- **No cancer cases (Biopsy = 0):** 803 instances (93.59%)
- **Cancer cases (Biopsy = 1):** 55 instances (6.41%)

This means the majority class (no cancer) accounts for 93.59% of the dataset, while the minority class (cancer) accounts for only 6.41%.

1.C Determination of whether the data set is balanced or unbalanced.

From the results, we can see that it is a highly unbalanced dataset because one class (no cancer) dominates with 93.6% of the data, while the other (cancer) only makes up 6.4%. When one class is much more frequent than the other, the model might just learn to predict the majority class and ignore the minority one.

How does this affect performance?

The model can have high accuracy just by predicting all samples as the majority class, but this is misleading. It will have poor recall and precision on the minority class (cancer cases), which is critical in medical diagnosis because missing positive cases is dangerous. We get a lot of false negatives (missing cancer cases) or false positives (wrongly diagnosing cancer), both of which are bad for real-world use.

2.0 Pre-processing Options

The following preprocessing steps were applied to prepare the cervical cancer risk factors dataset for modeling:

a. Handling Missing Values

- The dataset contained missing values represented by '?'. These were replaced with NaN for consistent processing.
- Numerical features with missing values were imputed using the **median** because outlier analysis showed the presence of outliers in these columns, and median is robust against outliers. Categorical/binary features were imputed using the **mode**, as these columns primarily contain 0/1 values and the mode preserves the most common category.

b. Dropping Features

- Two features, STDs: Time since first diagnosis and STDs: Time since last diagnosis, were dropped due to having over 90% missing values, which would negatively affect model quality.

c. Removing Duplicates

- Duplicate rows were identified and removed to avoid bias and redundancy in training.

d. Data Type Consistency

- All features were converted to the float data type to ensure consistency and compatibility with downstream machine learning algorithms.

e. Feature Selection

- Features with zero variance (constant columns) were dropped using a Variance Threshold filter, as they provide no useful information for classification.
- SelectKBest was used to select the top 10 features based on ANOVA F-values, reducing dimensionality and focusing on the most predictive attributes.

f. Addressing Class Imbalance

- The dataset was found to be highly imbalanced (approximately 94% negative, 6% positive cases). Synthetic Minority Over-sampling Technique (SMOTE) was applied **only on the training set** to generate synthetic samples of the minority class, improving class balance without affecting the test set integrity.

g. Normalization Situation

- **SVM with linear kernel** is sensitive to feature scales because it tries to find a hyperplane based on distances in the feature space.
  - So, **standardization (z-score scaling)** was applied to make features have zero mean and unit variance. This helps SVM perform better and converge faster.
  - **Normalization** (scaling to [0,1]) is also possible but less common for SVM; standardization is usually preferred when features are roughly Gaussian.
- **Decision Trees** do **not require scaling** (neither normalization nor standardization) so I did not perform it.

3.0 Model Evaluation Technique

Decision Tree:

- Used **Stratified K-Fold Cross-Validation** to get the best F-1 score (with 5 splits) combined with SMOTE in a pipeline.
- **Reason:**
  - The dataset is **imbalanced** (about 94% negative, 6% positive), so stratifying preserves the original class distribution in each fold.
  - Cross-validation helps **reduce variance** in evaluation by testing the model on multiple train-test splits, giving a more reliable estimate of performance.
  - Using SMOTE inside the pipeline ensures that oversampling happens only on the training fold in each split, **preventing data leakage**.

Final Result of evaluation for F-1 score: Best Params: {'clf__max_depth': 6, 'clf__min_samples_leaf': 6, 'clf__min_samples_split': 2} **Best F1 Score: 0.7178095238095239**

SVM:

- **Evaluation method:** Stratified K-Fold Cross-Validation with 5 folds.
- **Use of SMOTE in the pipeline:** SMOTE oversamples the minority class only on the training folds during cross-validation. This prevents data leakage from the validation fold, keeping evaluation fair and realistic.
- **Standardization:**
  StandardScaler is applied within the pipeline to scale features in each training fold independently, ensuring consistent feature scaling without leaking information.
- **Metrics used:**
  - **Accuracy:** To check overall correctness.
  - **F1 Score:** To better balance precision and recall, which is crucial for imbalanced datasets.

Final Result of evaluation for F-1 score: F1 Scores: [0.7826087 0.60606061 0.76923077 0.72 0.78571429], Best F1 Score: 0.79

4.0 Choice of Classifier

Based on the literature review and findings from the dataset, two classifiers were selected for this cervical cancer prediction task: Decision Tree and Support Vector Machine (SVM). These choices are supported by both published research and the behavior of the models on our specific dataset.

Decision Tree was selected for its interpretability and simplicity—ideal as a baseline and for explaining results to medical professionals. According to Ashtagi Rashmi et al. (2024) and Fernandes et al. (2018), Decision Trees are highly effective for medical data due to their ability to handle both numerical and categorical features without the need for extensive preprocessing. They also produce easy-to-interpret rules, which is critical in medical decision-making. In my own findings, the Decision Tree classifier performed reasonably well and provided clear decision paths, which would help healthcare professionals understand the prediction logic.

SVM was selected because it consistently outperformed Decision Tree on our dataset after feature selection and standardization, offering better prediction strength—especially in precision and F1-score. On the other hand, Support Vector Machines (SVM) were recommended by Mythili (2024) and Nasir et al. (2024) due to their strength in dealing with smaller datasets and high-dimensional feature spaces. In this study, after preprocessing and applying feature scaling (standardization), the SVM classifier produced higher precision and F1-scores than the Decision Tree. This aligns with the literature, which often shows SVM outperforming simpler models in terms of generalization.

Both classifiers also handle imbalanced data effectively, especially when combined with SMOTE (as seen in **Allahqoli et al., 2022**), which was used in this project to balance the dataset and improve model robustness.

5.0 Conclusion

This project successfully demonstrated how machine learning techniques can be used to predict the risk of cervical cancer based on behavioral and medical attributes. After preprocessing the dataset by handling missing values, encoding, and applying SMOTE to balance the data, various classification models were evaluated. Among all, the Decision Tree model delivered the highest accuracy and performance, proving to be the most effective in handling this specific dataset.

The results highlight the potential of machine learning in supporting early diagnosis and prevention strategies for cervical cancer. By identifying patterns and risk factors from data, such models can assist healthcare professionals in making quicker and more accurate decisions. This project not only strengthens the role of data-driven solutions in healthcare but also encourages further research into more refined models for medical diagnosis support.

References

Allahqoli, L., Laganà, A. S., Mazidimoradi, A., Salehiniya, H., Günther, V., Chiantera, V., Karimi Goghari, S., Ghiasvand, M. M., Rahmani, A., Momenimovahed, Z., & Alkatout, I. (2022). Diagnosis of Cervical Cancer and Pre-Cancerous Lesions by Artificial Intelligence: A Systematic Review. In Diagnostics (Vol. 12, Issue 11). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/diagnostics12112771

Ashtagi Rashmi, Chopade Pratiksha, Rajput Vaishali, Chivate Atharva, Chitpur Shreeshail, & Dashetwar Isha. (2024). Cervical Cancer Prediction Using Machine Learning.

Fernandes, K., Chicco, D., Cardoso, J. S., & Fernandes, J. (2018). Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. PeerJ Computer Science, 2018(5). https://doi.org/10.7717/peerj-cs.154

Mythili, J. (2024). A CONCISE REVIEW FOR EXPLORING DEEP LEARNING'S POTENTIAL IN CERVICAL CANCER PREDICTION FROM MEDICAL IMAGES. International Journal of Advanced Research in Computer Science, 15(6), 28–36. https://doi.org/10.26483/ijarcs.v15i6.7153

Nasir, M. U., Khalil, O. K., Ateeq, K., Almogadwy, B. S. A., Khan, M. A., & Adnan, K. M. (2024). Cervical Cancer Prediction Empowered with Federated Machine Learning. Computers, Materials and Continua, 79(1), 963–981. https://doi.org/10.32604/cmc.2024.047874