Assessment 2

Applied Artificial Intelligence

CS5079

Yassin Dinana

52094480

15/12/2020

Introduction:

This paper focuses on two different tasks, both tasks cover important areas in Artificial Intelligence, the first task is not presented in this PDF report, it is submitted as a Python Notebook where both the report and the code are included, the first task is mainly technical where it includes technical implementation to predict whether a female patient will be diagnosed with cervical cancer. The second task does not include technical implementation as mainly focuses on the ethical considerations for data scientists and A.I engineers when working with sensitive datasets.

In the first task, a dataset is given which include many different indicators and factors for predicting if a woman will get cervical cancer in the future, the dataset includes many factors such as demographic data and medical history for different patients. Many technical methods are approached to pre-process the data, train the data using the specific Gradient Boosted Decision Tree algorithm, use the Shapley technique to visualise the output of the implemented machine learning model, and prove which features in the dataset had high probability of influencing the patients.

It is very sensitive when dealing with specific data, when working with medical data for many different patients, it is important to consider all the ethics and challenges related to this dataset. Task 2 mainly focuses on the ethical considerations that are examined when working with the cervical cancer dataset. The task is also focusing on applying fairness and unbiased results for different types of patients when working with such dataset.

Task 2

1. Ethical considerations that data scientist have to take into account when dealing with medical data sets

Data scientists always have deal with different types of datasets that can range from many applications and environments. The collection of personal data is a contentious issue and is prone to public scrutiny as it has it has been placed under a microscope in recent years. Private data needs to be dealt with carefully as it includes private information such as race, gender, health, and biometrics, as a result, they must be used while keeping their sensitivity in consideration.

The dataset used in this paper is a medical dataset that includes the risk factors that can cause cervical cancer based on the personal data of 859 patients. When dealing with a medical dataset such as the one given, the data scientist must take some ethical considerations into account.

The first step that a data scientist must take when gathering the data is to inform consent and take the patient's acceptance before collecting the data. It is important that the patients accept to give their personal data to science at the earliest possible stage and the patients must have the right to withdraw their data at any point of the research and testing process [4].

Medical data is very private and confidential as it contains personal information, therefore, another ethical approach that must be taken in consideration is the security, safety, and confidentiality of the data, meaning that the data scientist must not share the data with any third parties or use the data for other purposes than the ones mentioned to the patients. Data sharing and misuse without consent is unethical and must not occur; patients and customers must have full knowledge of how their private data is used.

When dealing with medical data, the data scientist should make sure not to standardize unfair biases, the means that the developed results from the research should not be biased such as racist or sexist and no different results should be obtained depending on the patient's ethnicity, skin colour, religion. When training different samples on different amounts of data, the algorithms can absorb unconscious biases [5].

2. What steps are necessary to make sure that the trained model is unbiased and uphold fairness requirements

It is mandatory to make sure the trained machine learning model develops fair and transparent results; some steps are required to achieve statistically accurate results that avoid discrimination. In order to develop the steps that will avoid bias, it is important to understand what might cause the biased results when training a model.

When dealing with sensitive datasets, the data scientist must do the appropriate pre-processing. The main reasons that might cause bias and discrimination in the results are:

- Imbalanced Training Data: When the training data is imbalanced this means that the number of observations in the training class is not equally distributed, this where there will be a majority class and a minority class which might cause bias, an example could be more women than men in a cancer detection dataset, therefore, the model output will be biased. [6]

- Reflecting Past Discrimination: This could be a main reason why bias happens in the results after training a dataset, reflecting past discrimination means that the model is trained on previous experiences where discrimination has been an issue. An example could be that more women where rejected from a company when applying than men. The model learns from that and therefore is biased in favour of men's applications.

Now after understanding where the bias and unfairness resulted from trained models might occur, there are different steps and approaches that can be implemented to avoid it.

There are many different ways to deal with bias after training ML model such as:

a. Setting guidelines and rules to eliminate bias.
b. Data cleaning and sharing
c. Keep close monitor to models while training
d. Training different datasets.

In order tackle the first issue which is the "Imbalanced training data", the data scientist can balance it out by adding or removing data that do not have a strong effect on the predicted output until the data is balanced, in contemplation of choosing the most important data and removing the least important, this can be presented using a correlation heatmap that will help the data cleaning and make the balancing process easier. Another solution to deal with the imbalanced data set which might result in bias and unfairness is to train different models for each class, for example, one class can contain all the men, and another could contain all the women in the example mentioned above.

As stated before, the second issue that might cause result bias is "Reflecting Past Discrimination" and to deal with such problem, different techniques can be applied such as changing the learning technique and choosing the right learning model or use different algorithms or modifying the model after its being trained. Another technique to avoid biased results could be human centred solution, which means keeping human supervision on the output to improve fairness before releasing the output and moving on with the process [7].

3. Discuss whether for machine learning models deployed in the medical domain have to be unbiased and fair or in certain situation it is acceptable for them to have a bias. Please provide your rationale for supporting either or both points with examples.

When data scientists are dealing with sensitive data such as medical data, bias and unfairness could not be an option. All the information included in a medical dataset are very private for the patients such as the medical history, sexual preferences, and previous medical concerns. When biasing and unfairness occurs on such dataset the result might be inaccurate, and the prediction could affect the patient's health negatively [8].

A.I healthcare bias is dangerous and might affect the mental health of patients if predicted wrongly that they will be diagnosed with a specific disease. Six algorithms used on 60-100 million patients last year prioritised treatment coordination for white patients with the same degree of disease for black patients because of bias and reflecting on past discrimination [9].

The previous example states that the health risk for a black patient is less than the risk for a white patient which is incorrect and resulted from data bias, the output of the model can result in racial bias and discriminates different against patients because of the skin colour. Another issue that can arise from biased predictions when dealing with datasets is how the output might affect the patient. In the cervical cancer example, if the A.I model predicts that a woman will be diagnosed with cervical cancer in the future and this output is biased, this might affect the mental health and treatment of the patient and it might cause the treatment to be harder because of the mental effects and wrong predictions.

When dealing with medical datasets the accuracy and precision of the output should always be as accurate and no bias or unfairness is acceptable in such sensitive datasets as it might cause inaccuracies, which affect the mental health of patients and compromise the treatment. If the biased output of the model is used to train future models the result will be worse by causing more harm on future patients.

References:

[1] C.Molnar. 2020. “Interpretable Machine Learning”. Available at: https://christophm.github.io/interpretable-ml-book/ [Online]. Accessed 09 December 2020.

[2] D. Cortes. “Medical Fraud Detection Using SHAP values in features selections”. Available at: https://medium.com/@dan7cor/medical-fraud-detection-using-shap-values-in-feature-selection-4f98746da7a3 [Online]. Accessed 10 December 2020.

[3] S.Lundberg. “Implementation of SHAP”. Available at: https://github.com/slundberg/shap [Online]. Accessed 13 December 2020.

[4] M.Bak. “Big Data, Big Ethics: How to handle research data from medical emergency settings?”. Available at: https://blogs.biomedcentral.com/on-medicine/2018/09/13/big-data-big-ethics-handle-research-data-medical-emergency-settings/ [Online]. Accessed 12 December 2020

[5] P. Uria-Recio. 2018. “5 principles for big data ethics”. Available at: https://towardsdatascience.com/5-principles-for-big-data-ethics-b5df1d105cd3 [Online]. Accessed 12 December 2020.

[6] J.Jordan. 2018. “Learning from imbalanced data”. Available at: https://www.jeremyjordan.me/imbalanced-data/ [Online]. Accessed 13 December 2020

[7] N. Mehrabi, F.Morstatter, N.Saxena. 2019. “A survey on bias and fairness in machine learning”. Available at: https://arxiv.org/pdf/1908.09635.pdf [Online]. Accessed 13 December 2020.

[8] Anonymous Author(s). 2020. “Dataset Bias in Diagnostic AI systems: Guidelines for Dataset Collection and Usage”. Available at: http://web.mit.edu/juliev/www/CHIL_paper_bias.pdf [Online]. Accessed 14 December 2020.

[9] G.Nichols. 2020. “Artificial Intelligence in Healthcare is Racist”. Available at: https://www.zdnet.com/article/artificial-intelligence-in-healthcare-is-racist/ [Online]. Accessed 14 December 2020.