CANCER PREDICTIONSYSTEM USING MACHINE LEARNING PROBLEM STATEMANET

Group members

1. Mkhonta Thembinkosi 202003592
2. Msibi Vuyolwethu Samkelo 202003077
3. Nxumalo Neliswa 202004212
4. Mndzebele Mongi 202002370
5. Ngwenya Senanile 202002008

1. Background and Motivation

Cancer is a generic term for a large group of diseases that can affect any part of the body (WHO). According to WHO, cancer is the leading cause of death worldwide, accounting for nearly 10 million deaths in 2020 and most common cancer cases are breast, lung, prostate, skin and stomach etc. Each year, approximately 400 000 children develop cancer is rather a large number. Cancer is caused by transformation of normal cells into tumor cells in a multi-stage process that generally progresses from a precancerous lesion to a malignant tumor. These changes are a result of the interaction between a person's genetic factors and three categories of external agent including physical carcinogens such as ionizing radiation, chemical carcinogens such as tobacco smoke, alcohol etc and biological carcinogens such as infections from certain viruses or bacteria.

A number of people discover that they have cancer and there is less time for treatment as such we are proposing a web based machine learning cancer prediction system that will help predict a likelihood of someone getting cancer at a very early stage.

2. Problem Statement

Some cases, cancer is detected at a very late stage such that treatment and prevention no longer work, this weighs heavy on the patient as they have to accept that the chances of them facing death are very high. This also has an effect on the population as more people die, also the death toll increases thus reducing the life expectancy.

Here in Eswatini, statistical methods have been generally used for cancer classification. That is high risk/low risk however these are proven inefficient when it comes to handling high dimensional data. Also, cancer testing facilities are not easily accessible for the general population for reasons such as geographical location and some people don't even have the money at all. Our system will address such predicaments as it will be available for everyone to use thus people will be aware of their chances of getting cancer.

3. Methodology

Implementing this system will consist of a various step that range from model development to web system. The model creating will done on anaconda environment using Jupiter. The programming language that will be used is python version 3.8. The Cross Industry Standard Process for Data Mining (CRISP-DM) process model will be used to analyse the data and construct the prediction model. The CRISP-DM process model consists of the following steps:

- Business understanding: The phases focuses on clearly understanding the requirements and objectives of the developed model. It is where Determine business objectives, asses if resources are available for completing model, determining the data mining goals and creating a project plan.
- Data Understanding: This is the where data collection will occur and foe this project data collection will done online. The data exploration of where attributes will be determined, data description will occur and data quality confirmation will occur.

- Data Preparation: This is the stage where data cleaning and pre –processing will occur. This stage will ensure that the data will ensure that it is in the appropriate format to be used on the model. Data will also be split in to the training set and testing sets. This will have to be done according to the requirements of the proposed model.
- Modelling: Here a modelling technique will be chosen then built assessing the operation.
- Evaluation: Here the results of the model shall be evaluated and the whole process reviewed as to if it was done.
- Deployment: Based on the mined data a prediction system shall be built. A simple user interface shall be built to test in functionality and if its function according to requirements the model system shall the be transformed into a web page using tools such as flask. The web page programming shall be done using python for operations, HTML for page structure and cascading style sheets for better user experience and styling the web page.

4. Significance of the project

This project has a number of significances as it will help in a great deal. Firstly, since cancer is a deadly disease and is one of the leading causes of death, it has an effect on the population and increase in the death toll thus if we can predict the likelihood of someone having cancer at a very early stage, people can be aware of it and if it manifests, then they have the opportunity for early treatment and prevention. Secondly, because of geographical location, some people cannot access testing facilities and it is very costly for them to travel, since this is a web based system, everyone is able to access and use it

5. References

Cancer. (2022). Retrieved 3 November 2022, from https://www.who.int/news-room/fact-sheets/detail/cancer

Srinivas, P. R., Kramer, B. S., & Srivastava, S. (2001). Trends in biomarker research for cancer detection. *The lancet oncology*, *2*(11), 698-704.

Wulfkuhle, J. D., Liotta, L. A. and Petricoin, E. F.

Wulfkuhle, J., Liotta, L., & Petricoin, E. (2003). Proteomic applications for the early detection of cancer. *Nature Reviews Cancer*, *3*(4), 267-275. doi: 10.1038/nrc1043

.