CANCER CELL PREDICTION

USING MACHINE LEARNING FOR PYTHON

SUBMITTED BY:

Moinak Ghosh (13000218083)
Koustav Banerjee (13000218090)
Dipanjan Guchait (13000218096)
Arju Singh (13000218110)

PROJECT GUIDE : SOFIKUL MULLICK

Start Date : 18/08/2020	End Date : 19/09/2020

TECHNO MAIN SALT LAKE

ARDENT COMPUTECH PVT. LTD.
High-End Technology Training and Project

CERTIFICATE

This is to certify that Moinak Ghosh, Koustav Banerjee, Dipanjan Guchait and Arju Singh successfully completed the project titled "Cancer Cell Prediction" under my supervision during the period from 18th August 2020 to 19th September 2020 which is in fulfilment of their training in Data Science and Machine learning.

Signature of the Supervisor

Date: September 19, 2020

Sofikul Mullick

CONTENTS :-

ACKNOWLEDGEMENT

The achievement that is associated with the successful completion of any task would be incomplete without mentioning the names of those people whose endless cooperation made it possible. Their constant guidance and encouragement made all our efforts successful

We take this opportunity to express our deep gratitude towards our project mentor ***Sofikul Mullick***, for giving such valuable suggestions, guidance and encouragement during the development of this project work.

ABSTRACT

Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods. Therefore, these techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance. A variety of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. Even though it is evident that the use of ML methods can improve our understanding of cancer progression, an appropriate level of validation is needed in order for these methods to be considered in the everyday clinical practice. In this work, we present a review of recent ML approaches employed in the modelling of cancer progression. The predictive models discussed here are based on various supervised ML techniques as well as on different input features and data samples.

INTRODUCTION

Machine Learning (ML) is one of the core branches of Artificial Intelligence. It's a system which takes in data, finds patterns, trains itself using the data and **outputs an outcome.**

Machine learning is not new to cancer research. Artificial neural networks (ANNs) and decision trees (DTs) have been used in cancer detection and diagnosis for nearly 20 years (Simes 1985; Maclin et al. 1991; Ciccheti 1992). Today machine learning methods are being used in a wide range of applications ranging from detecting and classifying tumors via X-ray and CRT images (Petricoin and Liotta 2004; Bocchi et al. 2004) to the classification of malignancies from proteomic and genomic (microarray) assays (Zhou et al. 2004; Dettling 2004; Wang et al. 2005). According to the latest PubMed statistics, more than 1500 papers have been published on the subject of machine learning and cancer. However, the vast majority of these papers are concerned with using machine learning methods to identify, classify, detect, or distinguish tumors and other malignancies. In other words machine learning has been used primarily as an aid to cancer diagnosis and detection (McCarthy et al. 2004). It has only been relatively recently that cancer researchers have attempted to apply machine learning towards cancer prediction and prognosis. As a consequence, the body of literature in the field of machine learning and cancer prediction/prognosis is relatively small.

PROBLEM STATEMENT

Every year, Pathologists diagnose 14 million new patients with cancer around the world. That's millions of people who'll face years of uncertainty.

Pathologists have been performing cancer diagnoses and prognoses for decades. Most pathologists have a 96–98% success rate for diagnosing cancer. They're pretty good at that part.

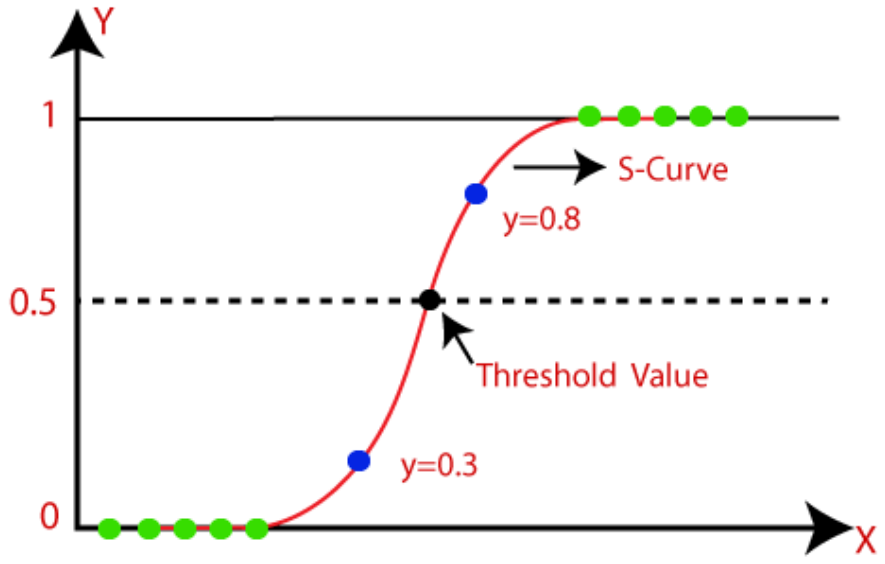
The problem comes in the next part. According to the Oslo University Hospital, the accuracy of prognoses is only 60% for pathologists. A prognosis is the part of a biopsy that comes after cancer has been diagnosed, it is predicting the development of the disease.

GOAL

Here our goal is to create & implement a Machine Learning Algorithm which can be used to solve the above mentioned problem and can be used as a tool to help the Medical Sector.

REQUIRED STATISTICAL CONCEPTS FOR THIS PROJECT (LOGISTIC REGRESSION)

In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. It is a statistical **model** that in its basic form uses a **logistic** function to **model** a binary dependent variable, although many more complex extensions exist.
In **regression** analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a **logistic model** (a form of binary **regression**).



APPROACH FOR THE MODEL BUILDING

Data Gathering:

We need to gather data from various pathology labs, testing clinic to get a proper dataset which can be further used to develop our prediction model.

Data Analysis:

After gathering the data we need to analyze it to remove null values, redundancies and misleading data values to create a perfect dataset which can be easily fit into a Machine Learning Model.

Model Creation:

We are using Logistic Regression for the prediction. We need to create a Machine Learning Model using Logistic Regression which can be used for the perfect prediction from our analyzed data.

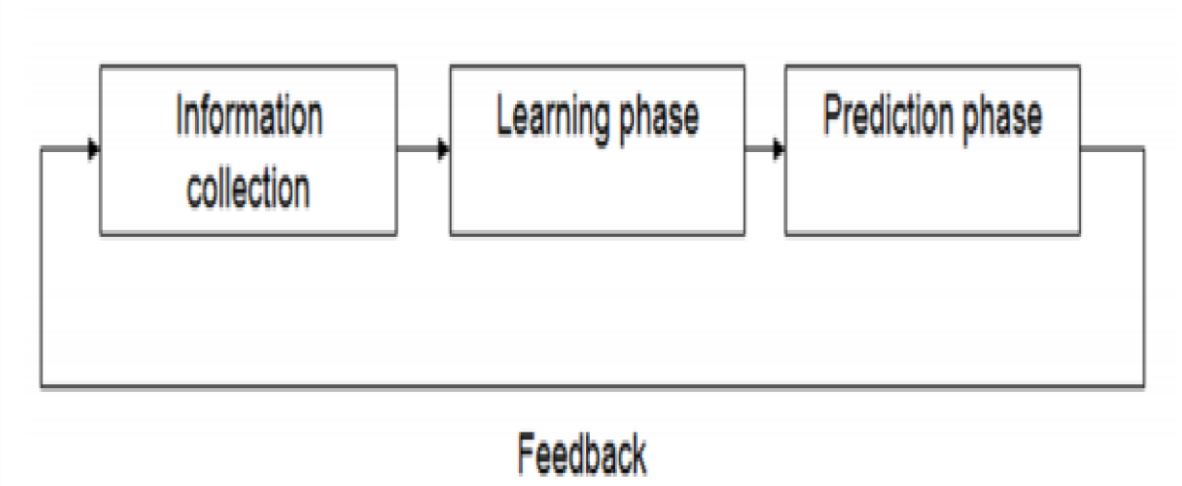
Training the Model:

We will be going to train our dataset with our Machine Learning Model & Proceed for the final step of the model building

Testing & Implementation:

We will test our model and check the accuracy. Based on the accuracy we will be going for further improvements. Also we will be testing our model with new datasets to make it more predictable and accurate.

METHODOLOGY

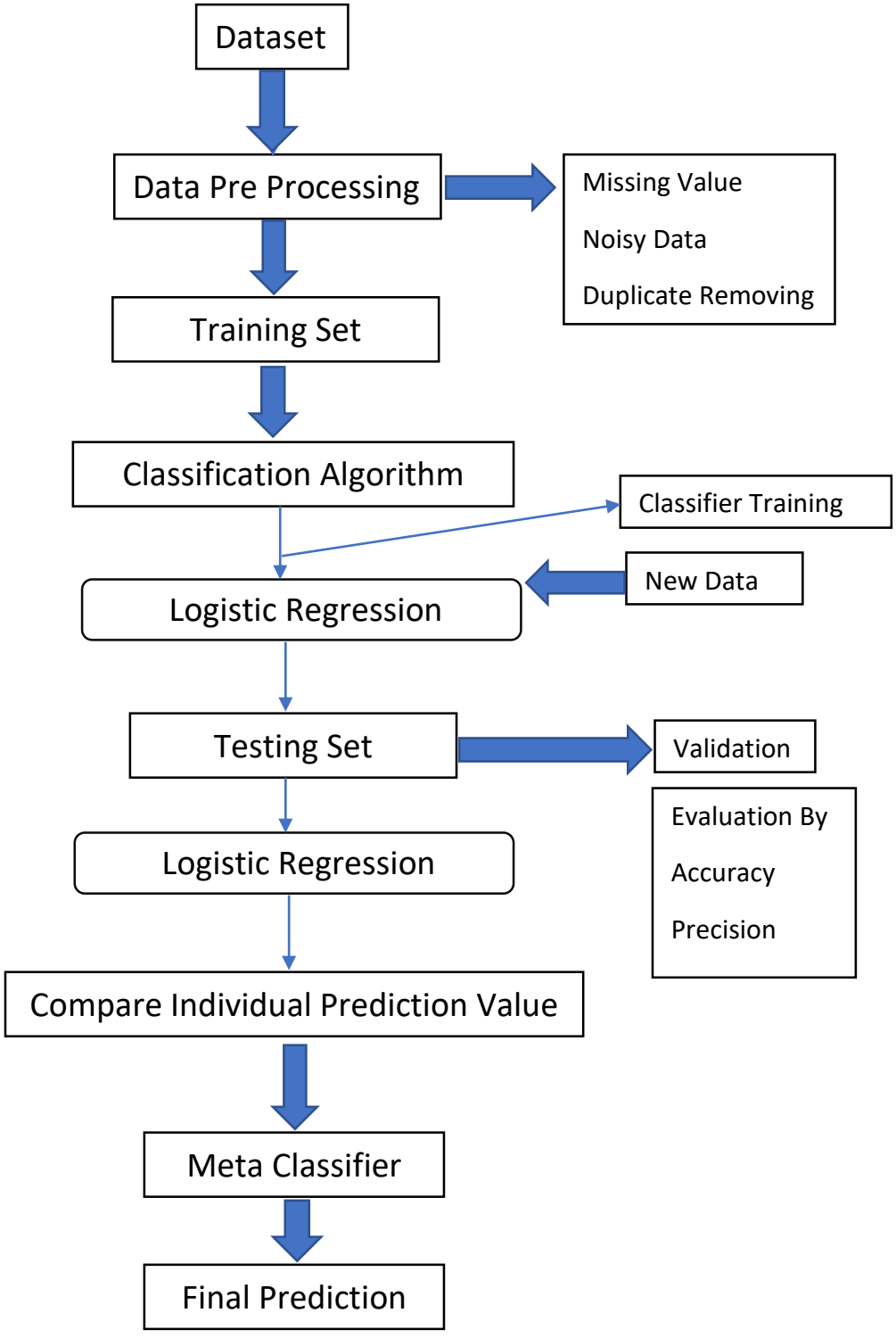


PROJECT OBJECTIVE

The objective of the project is:-

- To diagnostically predict whether or not a patient has cancer, based on certain diagnostic measurements included in the dataset.
- To study the symptoms of Cancer using Machine Learning.
- Extraction of important variable from the dataset useful for Cancer prediction.
- Data Visualization for the better prediction results

DETAILED ARCHITECTURE OF WORK FLOW



PROJECT IMPLEMENTATION

Following Libraries are Used For the Project :-

Library (pandas):

pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, **real world** data analysis in Python. Additionally, it has the broader goal of becoming **the most powerful and flexible open source data analysis / manipulation tool available in any language**. It is already well on its way toward this goal.

Library (seaborn):

Seaborn is a graphic library built on top of Matplotlib. It allows to make your charts prettier, and facilitates some of the common data visualisation needs (like **mapping** a color to a variable or using **faceting**). This page gives

general tips concerning this awesome library. Visit individual chart sections if you need a specific type of plot. The Seaborn documentation is also very well done and help going further. Most of the customisations that work

for Matplotlib work for Seaborn, so do not hesitate to visit the Matplotlib page of the gallery. Last but not least, note that loading seaborn before a matplotlib plot allows you to benefit from its well looking style!

Library (matplotlib):

matplotlib.pyplot is a collection of command style functions that make matplotlib work like MATLAB. Each **pyplot** function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

In **matplotlib.pyplot** various states are preserved across function calls, so that it keeps track of things like the current figure and plotting area, and the plotting functions are directed to the current axes (please note that "axes" here and in most places in the documentation refers to the *axes* part of a figure and not the strict mathematical term for more than one axis).

Library (numpy):

NumPy (pronounced /ˈnʌmpaɪ/ (*NUM-py*) or sometimes /ˈnʌmpi/ (*NUM-pee*)) is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors.

LIMITATIONS OF THIS PROJECT

- Each narrow application needs to be specially trained.
- Require large amounts of *hand-crafted, structured* training data.
- Learning must generally be supervised: Training data must be tagged.
- Require lengthy offline/ batch training.
- Do not learn incrementally or interactively, in real time.
- Poor transfer learning ability, re-usability of modules, and integration.
- Systems are opaque, making them very hard to debug.
- Performance cannot be audited or guaranteed at the ‘long tail’.
- They encode correlation, not causation or ontological relationships.
- Do not encode entities, or spatial relationships between entities.
- Only handle very narrow aspects of natural language.
- Not well suited for high-level, symbolic reasoning or planning.

FUTURE-SCOPE OF THIS PROJECT

In the past, our dependency on macro-scale information (tumour, patient, population, and environmental data) generally kept the numbers of variables small enough so that standard statistical methods or even a physician's own intuition could be used to predict cancer risks and outcomes. However, with today's high-throughput diagnostic and imaging technologies we now find ourselves overwhelmed with dozens or even hundreds of molecular, cellular and clinical parameters. In these situations, human intuition and standard statistics don't generally work. Instead we must increasingly rely on non-traditional, intensively computational approaches such as machine learning. The use of computers (and machine learning) in disease prediction and prognosis is part of a growing trend towards personalized, predictive medicine (Weston and Hood 2004). This movement towards predictive medicine is important, not only for patients (in terms of lifestyle and quality-of-life decisions) but also for physicians (in making treatment decisions) as well as health economists and policy planners (in implementing large scale cancer prevention or cancer treatment policies).

SUMMERY

In this review, we discussed the concepts of ML while we outlined their application in cancer prediction/prognosis. Most of the studies that have been proposed the last years and focus on the development of predictive models using supervised ML methods and classification algorithms aiming to predict valid disease outcomes. Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain.

BIBLIOGRAPHY

- https://www.kaggle.com/learn/intro-to-machine-learning
- https://www.wikipedia.org/
- https://towardsdatascience.com/machine-learning-is-the-future-of-cancer-prediction-e4d28e7e6dfa#:~:text=Machine%20Learning%20is%20a%20branch,predicting%20the%20development%20of%20cancer.
- https://www.sciencedirect.com/science/article/pii/S2001037014000464