CANCER PREDICTION USING CARET AND KNN

[1] [2] [3] [4]

Dheeraj.R , Hariprasath.R , Akshay Kannan.V , Nishanth Kumar.S

[1] *Under graduate, computer science & engineering, srm institute of science and technology, Tamil Nadu, INDIA*

[2] *Under graduate, computer science & engineering, srm institute of science and technology, Tamil Nadu, INDIA*

[3] *Under graduate, computer science & engineering, srm institute of science and technology, Tamil Nadu, INDIA*

[4] *Under graduate, computer science & engineering srm institute of science and technology, Tamil Nadu, INDIA*

ABSTRACT

Abstract: Cancer has been characterised 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. Predicting the cancer disease using caret and kNN algorithm helps to classify cancer patients into higher or lower risk. CARET (short for Classification And REgression Training) is a set of functions that attempt to streamline the process for creating predictive models. kNN - K-Nearest Neighbour. It is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbours. This algorithms segregates unlabelled data points into well defined groups.

Keywords: - *CARET, kNN Algorithm, Prediction, and R language.*

1. INTRODUCTION

Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. It is a dreadful disease. This project is designed to develop a cancer prediction system that allows the users to classify cancer patients into higher or lower risk. It is aimed at developing a software model that predicts the cancer using CARET (short for Classification And REgression Training) and kNN algorithm.This project using R language for coding the application. It incorporates all of the standard statistical tests, models, and analyses, as well as providing a comprehensive language for managing and manipulating data. Here on, the data is analyzed using various packages in R. The caret package is used to form the kNN algorithm. kNN modeling and prediction is a simple algorithm that regresses over k nearest neighbour variables and uses the collected data to analyze the obtained data. The pander package is used to represent the analyzed data in the form of tables for easy recognition and readability. Thus the prediction is used to classify patients into higher or lower risk.

1.1.Existing system

- Cancer prediction by oncologists
- Support vector machine(SVM)
- Decision Trees
- Learning vector quant
- Class package for functional classification

1.2.Issues in Existing System

- Outdated methods are used for prediction.
- Regression Principle has not been used in existing system.
- It cannot form various number of model.
- Accuracy as a parameter is not used for testing discrepancies.

1.3.Proposed System

- The proposed system is to build a cancer prediction system using R language as a base.
- It reduces the error caused by human intervention in cancer prediction and increases the accuracy of prediction and diagnosis of the disease.
- CARET (short for Classification And REgression Training) and kNN - K-Nearest Neighbour algorithms are used which provide high levels of accuracy in prediction.

2. SYSTEM ARCHITECTURE

In this chapter we are going to discuss about the system architecture of the cancer prediction system. There are 32 variables that contribute to the tumor's initiation and progression, which are recorded and stored in the dataset; the variables include radius, texture, volume, size, etc. of the cancer cells is uploaded as a training set into R language and the kNN algorithm is applied upon them to get the predicted outcome. This predicted outcome is taken as input for Tableau and image magick software as a combination, then the predicted analysis outcomes are presented to the user in the form of graphical representation. The architecture of this system is kept as simple as possible to make it accessible to a wide range of consumers and to maintain a simple user interface.

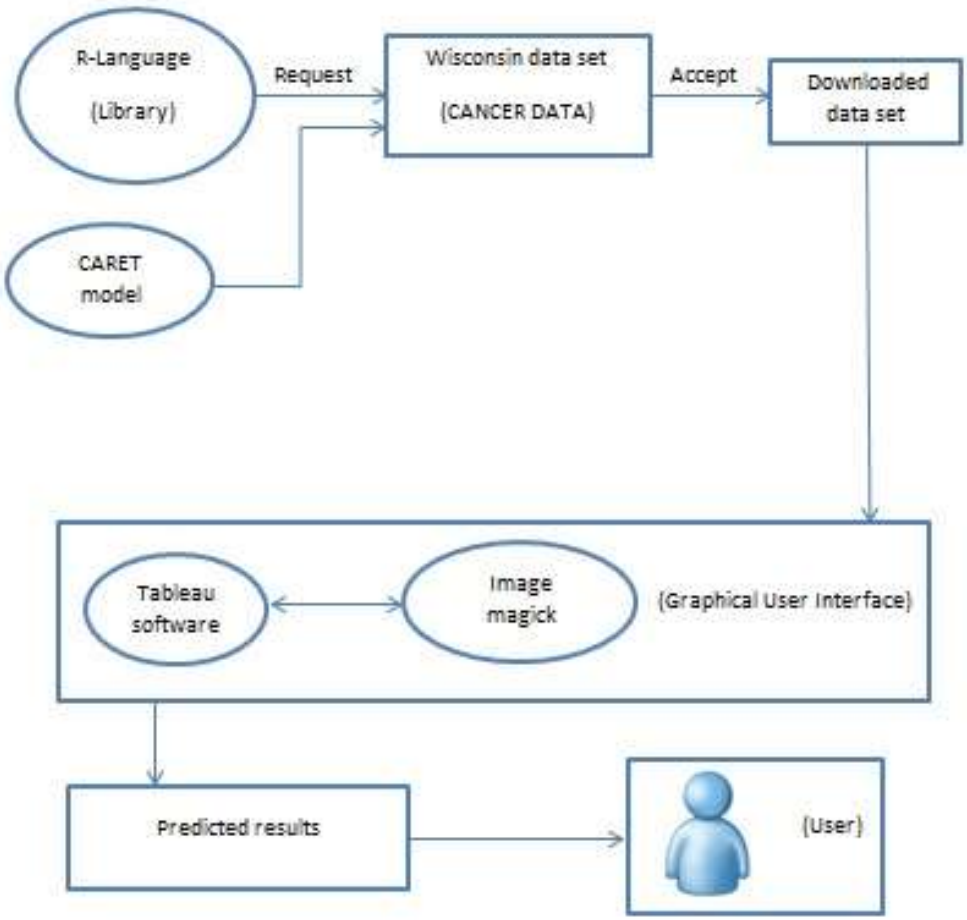

Fig -1 System Architecture Diagram

3. IMPLEMENTATION

The project named cancer prediction using kNN algorithm and CARET has been built using R language in R studio. System has satisfied all the proposed work. This project is implemented successfully. This platform will give an excellent result in terms of user interaction accuracy of prediction of cancer. These are the following modules that have been implemented in this project:

1. Input cancer variable module
2. Data aggregation
3. Analysis of data
4. Prediction of stock
5. Graph Plotting

3.1.INPUT CANCER VARIABLE MODULE

There are various factors and variable that define cancer cells. The genomic data is collected with biological knowledge and stored in a database which is collective called the dataset. This module's purpose is to connect the dataset to R language so that it can be processed to predict cancer. There are 32 variables that contribute to the tumor's initiation and progression, which are recorded and stored in the dataset; the variables include radius, texture, volume, size, etc. of the cancer cells.

3.2.Data aggregation

In this module all the variables are sorted into a super variable. The data from input cancer variable module gathered and expressed in a summary form, for statistical analysis.

3.3.Analysis of data

In this module the aggregated data is analyzed using the different mathematical algorithms such as KNN are used in finding recurring patterns.

3.4.Prediction of stock

In this module the data that has been Processed so far is subjected to CARET (Classification And REgression Training).

3.5.Graph plotting

In this module all the analyzed data is converted into form of graphical representations with the help of Tableau software and Imagemagicks platform.The plotting is based upon Bollinger's band. Bollinger Bands are volatility bands placed above and below a moving average. Volatility is based on the standard deviation, which changes as volatility increases and decreases.The bands automatically widen when volatility increases and narrow when volatility decreases.

4. CONCLUSIONS

Thus, hereby we conclude that the proposed system removes all the drawbacks of the existing system, decreases human intervention greatly and increases the efficiency and accuracy of the cancer prediction system. R Language and kNN algorithm has been implemented upon raw data and prediction is carried out. The system provides an user friendly interface to provide ease of access to a wide range of users.

5. REFERENCES

1. Hiba Asri,Hajar Mousannif,Hassan Al Moatassim,Thomas Noel , " Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016).

2. Mohammad R. Mohebian Hamid R. Marateb MarjanMansourian , Miguel Angel Mañanas ,Fariborz Mokarian , "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning", Computational and Structural Biotechnology Journal 15 (2017) 75–85 .

3. N. Suguna, and Dr. K. Thanushkodi, "An Improved k-Nearest Neighbor Classification Using Genetic Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 2, July 2010.

4. Venkat Reddy Korupally,Subba Rao Pinnamaneni, "Big data analytics for Diagnosis and Prognosis of Cancer using Genetic Algorithm", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (3) , 2016, 1251-1253.

5. Ritu Parna Panda, Prakalpa Prakash Barik and P. Alok Kumar Prusty, "A Review Paper on Big Data in Lung Cancer Big Data Analytics in Lung Cancer", International Journal of Trend in Research and Development, Volume 3(5), ISSN: 2394-9333.

6. Carmela Dantas Barbosa , "Challenges with Big Data in Oncology", Barbosa, Journal Orthop Oncol 2016, 2:2.