Software Requirements Specification

for

BREAST CANCER PREDICTION SYSTEM

Prepared by :-

TEAM NAME: TEAM DECODER

SHUBHI GUPTA (E17CSE165)

MANMEET SINGH (E17CSE126)

SOUVIK MISHRA(E17CSE127)

MADHAV GINORIA(E17CSE186)

21/10/2018

**Table of Contents**

[1. Introduction 4](#_Toc496523408)

[1.1 Purpose 4](#_Toc496523409)

[2. Description 4](#_Toc496523411)

[2.1 Product Perspective 4](#_Toc496523412)

[2.2 Product Functions 4](#_Toc496523413)

[2.3 Operating Environment 5](#_Toc496523414)

[2.4 Assumptions and Dependencies 5](#_Toc496523415)

[3. External Interface Requirements 5](#_Toc496523416)

[3.1 User Interfaces 5](#_Toc496523417)

[3.2 Hardware Interfaces 5](#_Toc496523418)

[3.3 Software Interfaces 5](#_Toc496523419)

[3.4 Communications Interfaces 5](#_Toc496523420)

[4. System Features 6](#_Toc496523421)

[4.1 System Feature 1 6](#_Toc496523422)

[4.2 System Feature 2 (and so on) 6](#_Toc496523423)

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Date** | **Reason For Changes** | **Version** |
|  |  |  |  |
|  |  |  |  |

# Introduction

## Purpose

* Breast cancer represents one of the diseases that make a high number of deaths every year.
* It is the most common type of all cancers and the main cause of women's deaths worldwide.
* Classification and data mining methods are an effective way to classify data.
* Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions.

In this paper, a performance comparison between different machine learning algorithms:

* + - Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) datasets is conducted.
    - The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity.
    - Experimental results show that SVM gives the highest accuracy (97.13%) with lowest error rate. All experiments are executed within a simulation environment and conducted in WEKA data mining tool.

# Description

## Product Perspective

* Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today’s society.

## Product Functions

* **Function** is to use Machine Learning Techniques to diagnose breast cancer and database techniques to store the information for future reference

## Operating Environment

* The user can access the website using any web browser that supports html5 given that all are up-to-date.

## Assumptions and Dependencies

* That the software could accurately predict the breast cancer when user provides the information and also project completion in the given time.

# External Interface Requirements

## User Interfaces

* Machine learning using Python.
* Databases using Mysql.

## Hardware Interfaces

* Laptops having Core i3 and i5 processors.

## Software Interfaces

The application is developed using:

* GUI (software not yet decided).

## Communications Interfaces

* Mails will be sent to users for verifications and other important notifications which will be achieved through php mail(). A SMTP server will also be required to send mails.

# System Features

## Baseline algorithm checking

From the dataset, we will analysis and build a model to predict if a given set of symptoms lead to breast cancer. This is a binary classification problem, and a few algorithms are appropriate for use. Since we do not know which one will perform the best at the point, we will do a quick test on the few appropriate algorithms with default setting to get an early indication of how each of them perform. We will use 10 fold cross validation for each testing.

The following non-linear algorithms will be used, namely: **Classification and Regression Trees (CART)**, **Linear Support Vector Machines (SVM)**, **Gaussian Naive Bayes (NB)** and **k-Nearest Neighbors (KNN)**.

## Evaluation of algorithm on Standardised Data

The performance of the few machine learning algorithm could be improved if a standardised dataset is being used. The improvement is likely for all the models. I will use pipelines that standardize the data and build the model for each fold in the cross-validation test harness. That way we can get a fair estimation of how each model with standardized data might perform on unseen data.