A picture containing logo

Description automatically generated

Binary Classification

By,  
Rajendra Prasad Patil  
Data Mining – CS 634105  
Mid-term project report  
Email – ID : [rp98@njit.edu](mailto:rp98@njit.edu)

[rajendra7406@gmail.com](mailto:rajendra7406@gmail.com)

Source : <https://github.com/rp98njit/data-mining-final-project>

# Contents

[Contents 2](#_Toc89550118)

[Introduction 2](#_Toc89550119)

[SVM Algorithm Overview 2](#_Toc89550120)

[K Neighbors Algorithm Overview 2](#_Toc89550121)

[LSTM Algorithm Overview 2](#_Toc89550122)

[Implementation Overview 2](#_Toc89550123)

[Assumptions 3](#_Toc89550124)

[Requirement 3](#_Toc89550125)

[Software 3](#_Toc89550126)

[Hardware 3](#_Toc89550127)

[Implementation List of Source Code Files 4](#_Toc89550128)

[Patil\_RajendraPrasad\_finaltermproj.ipynb 4](#_Toc89550129)

[How to Run the Application 4](#_Toc89550130)

[GUI Mode: 4](#_Toc89550131)

[Data Set 4](#_Toc89550132)

# Introduction

## SVM Algorithm Overview

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. – Source Wikipedia

## K Neighbors Algorithm Overview

k-NN is a type of classification where the function is only approximated locally, and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically. – Source Wikipedia

## LSTM Algorithm Overview

Long short-term memory is an artificial recurrent neural network architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single data points, but also entire sequences of data. – Source Wikipedia

## Implementation Overview

Python is used as the programming language, and all algorithms are implemented using libraries. The implementation uses in-built memory for storing variables and intermediary data.

The program is implemented in Jupyter Notebook, and one has to restart the kernel and run all the cells.

The program will load the breast cancer dataset and dataset is loaded, fitted, and run in to the algorithms. The program is running into K-Fold datasets, and it outputs the different performance metrics listed below. Then the average of each metric is calculated and presented using ‘pandas data-frame’.

The performance metrics calculated:

1. True Negative
2. False Positive
3. False Negative
4. True Positivity
5. Sensitivity
6. Specificity
7. Precision
8. Accuracy
9. F1 Score
10. Error Rate
11. Negative Predicted Value
12. False Positive Rate
13. False Discovery Rate
14. False Negative Rate
15. Balanced Accuracy
16. True Skill Statistics
17. Heidke Skill Score

# Assumptions

Several assumptions were considered for successful run of the application. They are listed below:

1. The evaluator has an idea about machine learning and its dependent libraries.
2. The evaluator knows how to run the Jupyter notebook and has installed libraries successfully.

# Requirement

## Software

Operating System: Linux / Windows

Programming language: Python 3.5 +

IDE: Jupyter Notebook / Jupyter Lab

## Hardware

Hardware Configuration: 8GB Ram, 2.2 GHz

# Implementation List of Source Code Files

Python is used as the programming language, and it is implemented from scratch. The implementation uses in-built memory for storing variables and intermediary data.

The program consists of one python file listed below

## Patil\_RajendraPrasad\_finaltermproj.ipynb

When the program is run, it imports libraries, loads the dataset, performs operations on the dataset, the data is fed into different machine learning algorithms, several performance metrics are calculated, and it is displayed.

NOTE: Source code is present in the Appendix

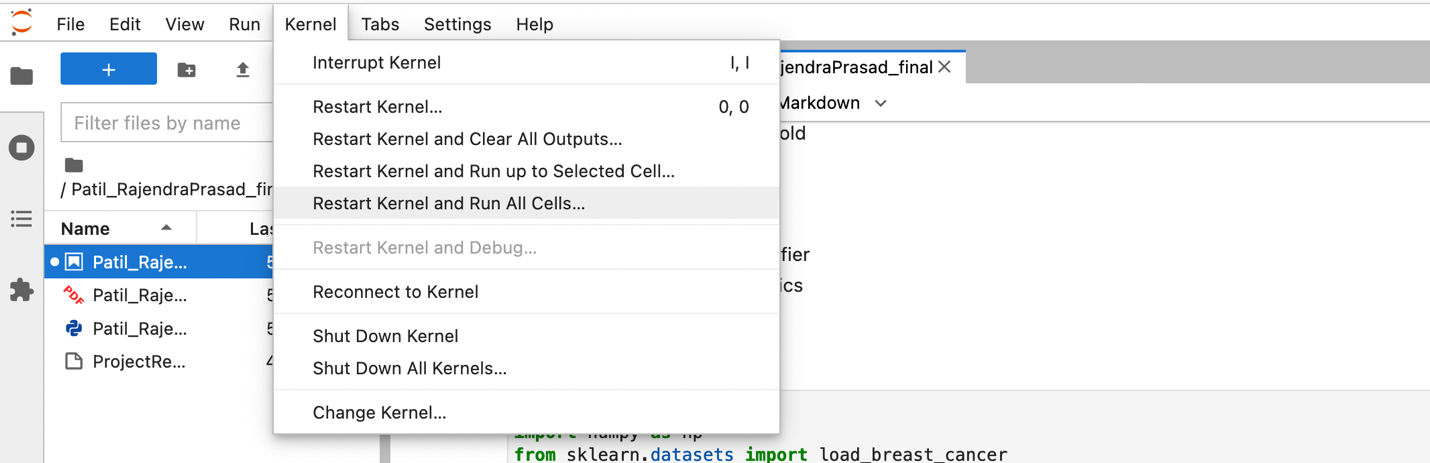
# How to Run the Application

Prerequisites:

1. Python 3.5 or higher (with pip preinstalled)

## GUI Mode:

Load the Jupyter Notebook into the Jupyter Lab. Go to Kernel menu and execute ‘Clear Kernel and Run all the cells’.



# Data Set

Breast Cancer Dataset provided by ‘Sklearn datasets’ library