CERVICAL CANCER PREDICTING SYSTEM USING MACHINE LEARNING

Akmeemana Pathira Kankanamge Chandi Prabodhani

209366B

Master of Science in Computer Science

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

August 2022

CERVICAL CANCER PREDICTING SYSTEM USING MACHINE LEARNING

Akmeemana Pathira Kankanamge Chandi Prabodhani

209366B

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Science in Computer Science

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

August 2022

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works.

Signature: Date:

The supervisor/s should certify the thesis/dissertation with the following declaration.

The above candidate has carried out research for the Masters thesis under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of the supervisor: Dr. Buddhika Karunarathne

Signature of the supervisor: Date:

ACKNOWLEDGEMENT

I would like to show my gratitude to Dr. Buddhika Karunarathne for guiding me to initiate and finding a better methodology to conduct research. His supervision greatly helped me in setting goals and engaging in research.

I'd like to convey our gratitude to Dr. Charith Chitraranjan and all of the lecturers at the University of Moratuwa's Department of Computer Science and Engineering for their assistance in overcoming this challenge.

Last, but not least, my heartfelt gratitude goes to my parents, husband, and friends for extending their kind hands in support of research's success.

ABSTRACT

Machine Learning has become a vital tool in everyday life, as well as a potent tool for automating most of the industries we want to automate. Machine Learning is a method of developing algorithms that learn from data, which might be labelled, unlabelled or learned from the environment. Machine Learning is employed in a variety of industries, including health care, where it provides much greater benefits through a proper decision and prediction processes. Because the machine learning in health care is scientific research, we must save, retrieve, and properly use information and data, as well as give knowledge about the difficulties that face the healthcare industry and proper decision-making.

Over the years, these technologies have resulted in significant advancements in the health-care sector. Medical experts employ the machine learning tools and techniques to analyse medical data in order to identify hazards and provide accurate diagnosis and treatment.

The paper aims to build a web application and put a trained machine learning model into production using Flask API. Here use cancer data to predict cervical cancer using machine learning. Therefore this project helps to use machine learning models for end-users or systems.

Keywords: Machine learning, Flask API, Python, Web application

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
RNN	Recurrent Neural Network
OCR	Optical Character Recognition
FRT	Facial Recognition Technology
RFE	Recursive Feature Elimination
LR	Logistic Regression
KNN	K-nearest neighbors

CHAPTER 1 – INTRODUCTION

1.1 Introduction

This chapter presents a research summary. It explains the history of Cervical Cancer Prediction Systems, the research topic, research goals, and the study's purpose. Finally, the final report's structure is discussed.

1.2 Background

In today’s world, One of the most powerful technologies is machine learning. These data will be meaningless unless we analyse them and uncover the hidden patterns. Machine learning techniques are used to locate valuable underlying patterns in complex data that would otherwise be difficult to find manually. Machine learning is a data-to-knowledge conversion technology. In the last 50 years, there has been a data explosion.

Each year, around 11,000 new instances of invasive cervical cancer are detected in the United States. However, the number of new cervical cancer cases has significantly decreased during the last few decades. There are several risk factors for cervical cancer leading to this biopsy examination. In this research, we use machine learning models and create a web application that is helpful for the health care sector.

To solve an issue, software engineers have always coupled human-created rules with data. Machine learning, on the other hand, employs data and answers to deduce the rules that govern a situation. The report tries to build a relationship between symptoms and the biopsy using supervised learning. In this project, new inputs can then be fed in of predefined symptoms and the Machine learning algorithm will then output a future prediction for the biopsy as to whether cancer exists or not.

This project aims to build a web application using Flask API. Essentially, APIs are used for many web applications. There are popular ML APIs as follows.

1. Machine Translation
2. The message Resonance
3. Question and Answers
4. User Modeling

Flask is one of the web service development frameworks in Python. The flask framework needs minimal configuration and light-weighted, and it can be controlled from within Python code. Therefore, it is a more popular framework.

Combining all these things and building a web application for users of the healthcare sector is the outcome of this research.

1.3 Research problem

Several research projects have been done in the past decades for Several research projects in the past decades for different diseases (Heart disease, Diabetes) using different machine learning models. (ANN, SVM, GNB, KNN)

The main drawback of these research studies is their models are low accurate and web applications are not user-friendly.

However, in today's hectic world, user-friendly applications are required. It is very crucial if researchers could develop user-friendly applications with high accuracy models.

1.4 Objectives

The main objective of this research is to build web applications from a machine learning model in Python using Flask for the healthcare sector. To achieve this there are sub-objectives to be accomplished,

1. Recognize the cervical cancer symptoms
2. Build a user-friendly web application

3. Determine the biopsy using this web application

1.5 Motivation

In most cases, the machine learning algorithms uncover the hidden pattern in a huge dataset and calculate the desired approximate final result. Use supervised learning methods to test the accuracy of certain prominent (ML) algorithms in this project. Supervised Learning systems strive to anticipate new results based on previous learning by learning the pattern from pre-existing data. Function-based, rule-based, rfunction-based, probability-based, and instance-based, are ML techniques for analysing existing data Various Machine Learning methods are introduced for supporting medical experts using various data mining strategies.

Cancer is a disease that is rapidly spreading throughout the population. Cancer is a disease of genes that causes cells to grow out of control. Despite the fact that there are numerous forms of cancer, they always begin with the uncontrolled proliferation of aberrant cells. Normal cells divide, grow, and die in a predictable pattern. Cancer cells outlive normal cells and create new aberrant cells because they continue to grow and divide.

Cancer cells form when DNA, which directs all activity in each cell, is damaged. When DNA is broken, the human body normally has the ability to repair it. Damaged DNA is not repaired in cancer cells, however. Damaged DNA can be passed down down the generations, and it is responsible for about 10% of all malignancies. Exposure to something in the environment or random cellular processes, on the other hand, is more likely to damage a person's DNA.

Cancer can start anywhere in the body and usually takes the form of a solid tumour. Liquid tumours are a term used to describe some diseases such as leukaemia and myeloma. Before spreading to other tissues and growing, cancer cells attack the blood and blood-forming organs (bone marrow).

The different types of cancer include:

Carcinomas are the most common malignancy, and they arise from cells that coat the body's internal and external surfaces. Breast, lung, colon, and prostate cancers are the most common tumours of this type in the United States.

Sarcomas are cancers that start in the body's supporting tissues, such as fat, bone, cartilage, muscle, and connective tissue

Lymphomas are cancers that develop in the immune system's lymph nodes and tissues.

Cancers of immature blood cells that form in the bone marrow and then aggregate in high quantities in the bloodstream are known as leukemias.

The primary site is the location where cancer begins. It might then spread to other places of the body (metastasize). Cancer is always named for the site where it first appeared, regardless of where it spreads. Metastatic breast cancer, rather than liver cancer, is the term used to describe breast cancer that has spread to the liver.

Different cancers behave in a variety of ways. For example, lung and breast cancers are two separate diseases react to different treatments and grow at different rates. As a result, cancer patients require treatment tailored to their specific type of cancer.

The accuracy of the decision support system, which is a subfield of AI, is acknowledged for its use. As a result, the most important goal in building a decision support system is to accurately foresee and identify a condition. We use a pre-existing data set, to train and evaluate this model in this system.

1.6 Thesis Structure

The following is how the thesis is structured: The project's goals, background, the problem, and motivation are all discussed in the first chapter. In chapter 2, go over some earlier work that has been done that is pertinent. It includes three parts. They create machine learning models, develop web applications with Flask and integrate

the model and deploy the web all on Heroku. The third chapter contains an in-depth overview of our methodology as well as many algorithms.

CHAPTER 2 - LITERATURE REVIEW

2.1 Overview

Research on the machine learning in production spread with a variety of applications. Many researchers have done research with various kinds of sectors. Machine learning can assist us in finding the best way back home, locating a product that meets our requirements, and even scheduling hair salon appointments. If we adopt a more hopeful view, we can see how the incorporating machine learning into projects might improve people's lives and perhaps propel society forward. Web applications are already an important part of people's lives, and integrating them with the power of machine learning can result in delightful and impressive user experiences.

When considering the previous related works, The author of Reference [2] suggested that the given data be pre-processed in order to isolate all of the metadata. The KNN algorithm is used to find the dataset's closest neighbors. Naive Bayes (NB) is an accurate machine learning method used to organize Arabic web documents, according to Reference [1. Economic situations have been investigated using the K-Nearest Neighbor classification. Economic distress and insolvency will maintain a certain distance, which can be predicted using the KNN technique.

2.2 Heart disease-related research studies

The Author in reference [11] proposed that they create a model to predict heart disease on physical and mental parameters. Then develop a web application that takes the input to predict heart disease. They used the below flow chart for their prediction system.

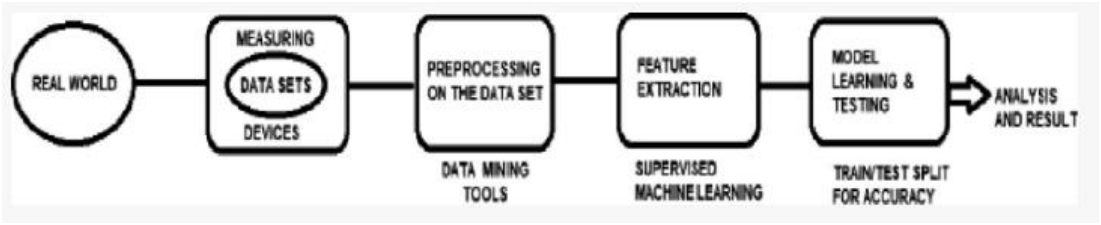


Fig 2.1 Flow chart of heart disease predicting the system

Below diagram shows how the website works.

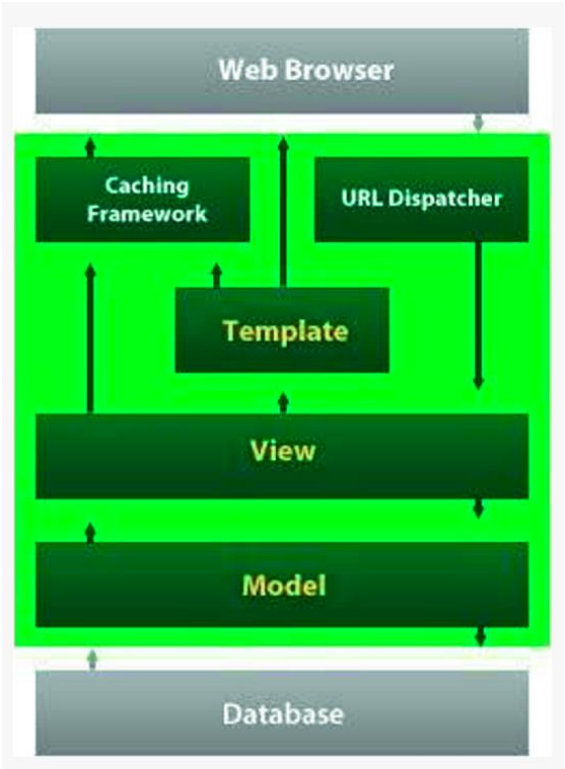


Fig 2.2 Experimental working of heart disease predictor.

2.3 Diabetes disease-related research studies

Implement a web application to anticipate diabetic disease as described in reference [6]. Among other diseases, diabetes is regarded as one of the most dangerous. To create this online application, they used an Artificial Neural Network with an

82.35% prediction rate. The diabetes prediction system's flow chart is shown below.

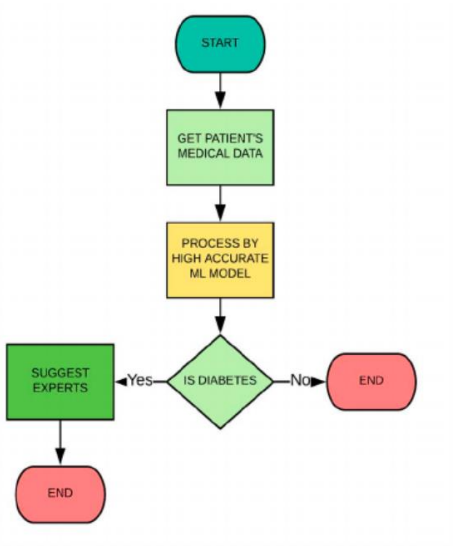


Fig 2.3 Flow chart of diabetes predicting a system

They used four algorithms to choose what is the best model. They used SVM, KNN, GNB, and ANN. The below table shows the different dataset sizes and accuracy of each method.

Training Dataset Size	SVM	KNN	GNB	ANN
368	63	64	76	63
468	63	68	77	63
568	63	68	76	63
668	63	66	76	63
Average	**63**	**66.5**	**76.25**	**63**

Fig 2.4 Comparison of four algorithms

They used the Min-Max scalar method to perform some data preprocessing to improve the detection accuracy. Using this strategy, they were able to achieve a higher level of accuracy than before. Based on their dataset, Artificial Neural Network attained a detection accuracy of 82.35 percent. Finally, they used the ANN model to create a web application. The web application's final appearance is depicted in the diagram below.

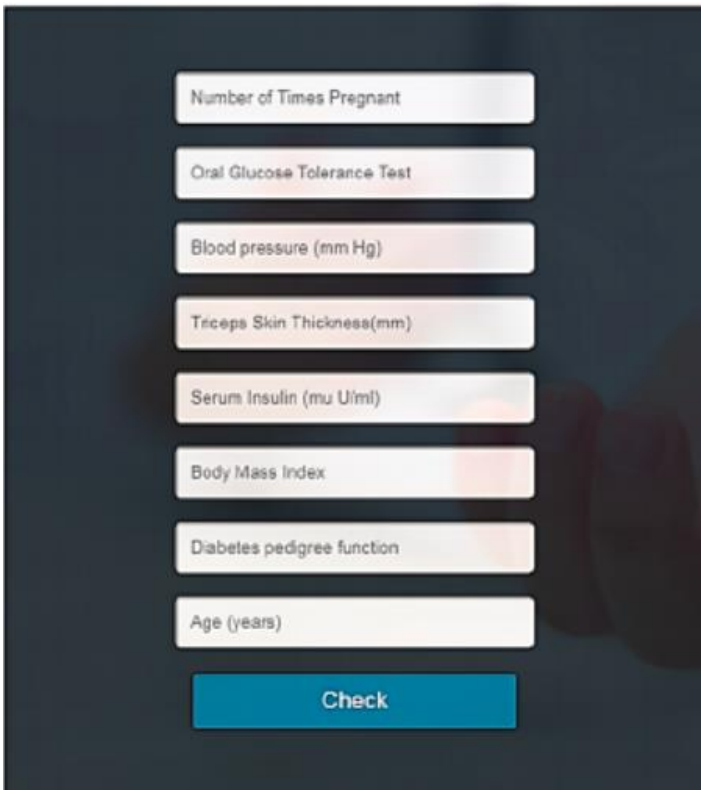


Fig 2.5 The proposed web application to predict diabetes

And they compare their research results with other research works. Among other research works their proposed method has the highest accuracy.

Model Name	Accuracy
Gaussian Naive Bayes (GNB) [13]	76.52%
General Regression Network (GRNN) [14]	80.21%
Backpropagation Genetic Algorithm (BGA) [15]	74.80%
Fuzzy Min Max (FMM) [16]	69.28%
Our Proposed Model (ANN with MMS)	**82.35%**

Fig 2.6 Comparison of the model with other models

Reference [12] shows another prediction system for diabetes disease. They used the methodology below to implement this ML model. Their dataset size is 769 records and 9 features.

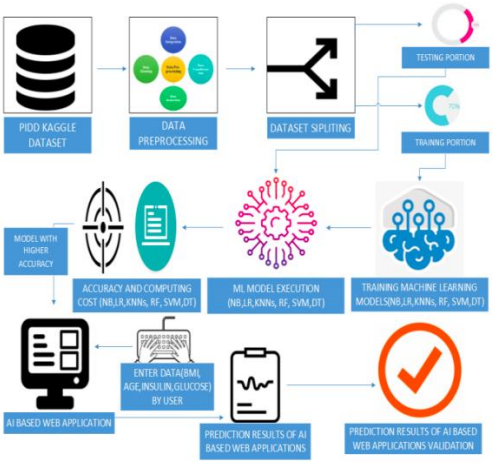


Fig 2.7 Methodology for predicting the model

They used the six machine learning algorithms and compared each algorithm.

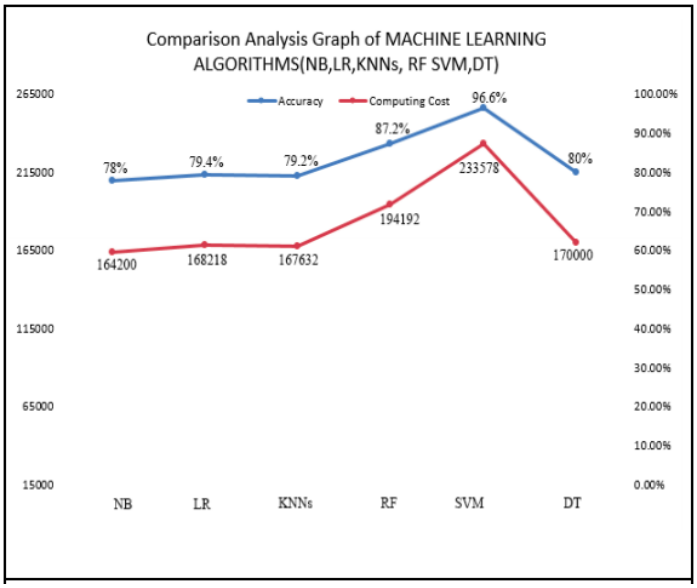


Fig 2.8 Comparative analysis

SVM had the highest accuracy of 96.6% of the algorithms tested, while having a somewhat high computing cost. They employed SVM to construct AI-based online applications to more reliably forecast diabetes status, according to these findings. The below image shows the artificial intelligence-based web application for diabetic disease predictions.

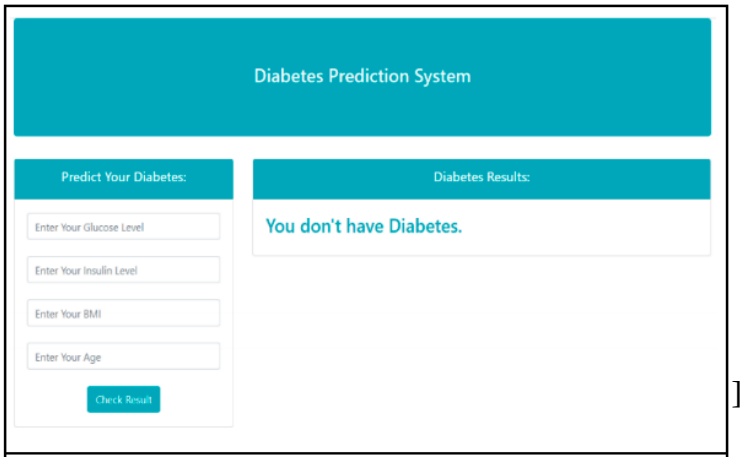


Fig 2.9 The Web application Overview

2.4 Cancer disease-related research studies

In reference [14] predict the cancer disease. They create hospital management systems including a user monitoring system, a doctor management system, an administration system, and a hospital management system. They used the below structure to build their management system.

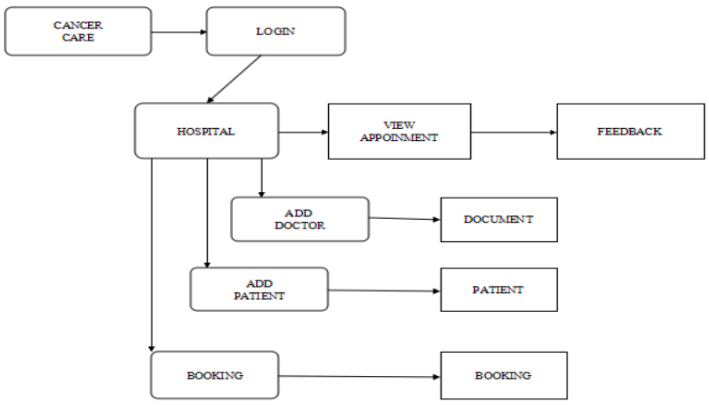


Fig 2.10 Hospital Management system

Cancer prediction is the only part of their research. They used the K- Means clustering algorithm for classifying cancer and non-cancer patients. The decision tree is used to identify the accurate status of illness that could be associated with the patient.

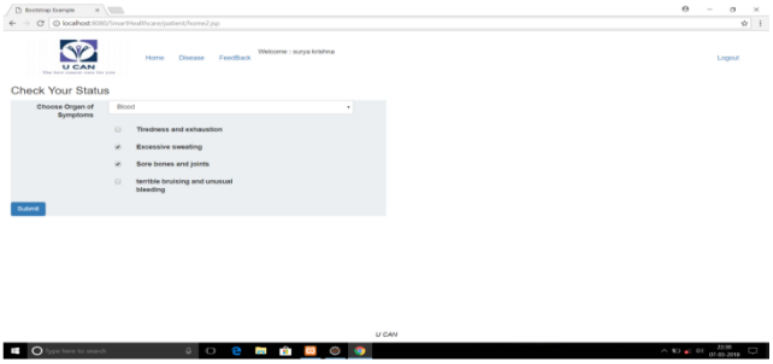


Fig 2.11 Cancer Predicting system

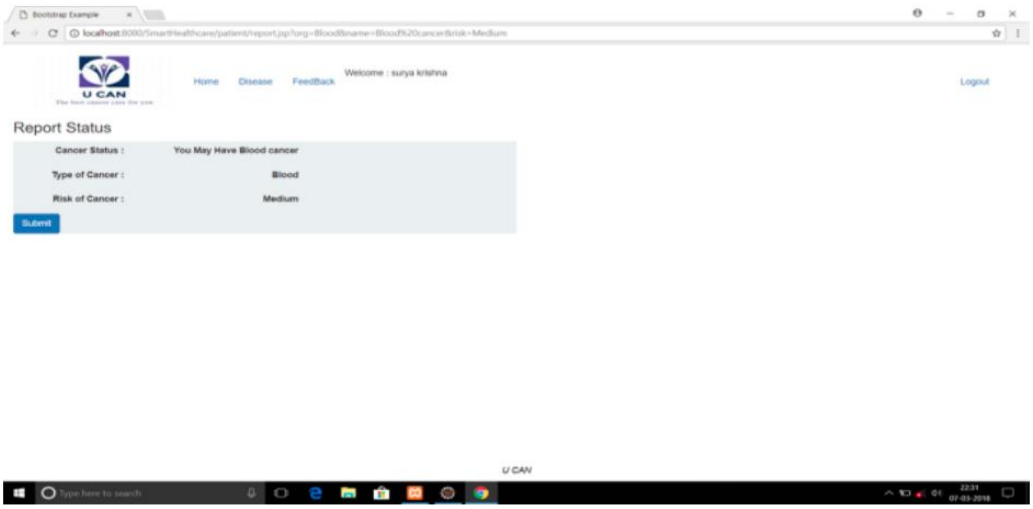


Fig 2.12 Cancer Prediction Risk Level

The above two figures show the appearance of the cancer prediction system, and they identify what is the risk level of the disease.

Deep learning in cancer prediction is another research according to this topic[15]. The accuracy of cancer prediction will help clinic management of cancer patients significantly. Deep learning was suggested because it produces more accurate predictions when dealing with massive amounts of data. They discovered seven

major hurdles in using a deep learning algorithm to analysis cancer prognosis and achieve high performance.

Completely interconnected NN and CNN models have been used to predict cancer prognosis in various studies and have performed well.
According to this paper, cancer prediction has improved in recent years as additional data, kinds have been available to better evaluate disease states. Other studies have used the principal component analysis and clustering of an auto encoder to classify cancer kinds. Support vector machines, Bayesian networks, semi-supervised learning, and decision trees have all been used to predict cancer and have shown some promise.

Another study is utilizing the machine learning approaches to predict breast cancer. [16] The second-worst of the malignancies that have been discovered so far is Breast cancer. They compare all supervised machine learning techniques which are (SVM), K-nearest neighbors, random forests, (ANNs), and LR. The accuracy, sensitivity, specificity, precision, a false-negative rate, false-positive rate, F1 score, and Matthews Correlation Coefficient, are used to evaluate the research's performance. Their findings show that ANNs have the highest precision, accuracy, and F1 scores. SVM has the highest accuracy and precision. There are 699 cases in this dataset that are either benign or harmful.

Jasvinder Singh and Sandeep Sharma [17] present a model for determining the appropriate stage of infection. Their proposed model takes into account ten characteristics of cervical cancer that are divided into four stages. Wireless sensors are used to collect data from the patient in the suggested model for predicting the stage of cervical cancer. The acquired analogue data is updated in the data repository before being transformed to digital using an Arduino device. Six machine learning techniques are trained with the updated repository.

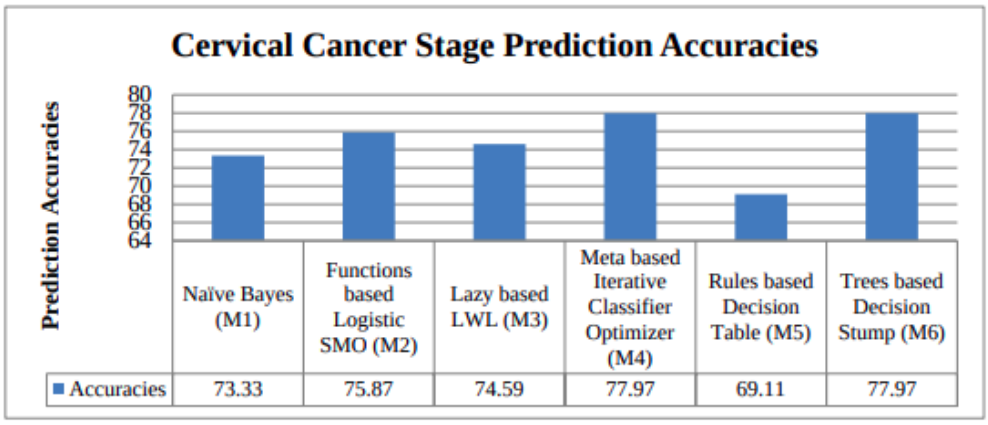


Fig 2.13 Accuracies Comparison for Cervical Cancer Stage Prediction

F. Asadi, C. Salehnasab, and L. Ajori devised a strategy for predicting cervical cancer [18]. Personal health status, marital status, social standing, contraceptive dose, level of education, and the number of cesarean deliveries were revealed to be meaningful predictors in all of the algorithms. In This study decision tree algorithms were used to find the most important outcomes. Because the number of essential predictors for analysis is reduced in the proposed models, the computing cost is minimized. The disease can be anticipated more accurately with the use of machine learning. Furthermore, Strengthening a patient's health and socio-cultural level looks to be beneficial.

This literature review is based on three parts. The first, part is based on how to create a machine learning a model. This section uses supervised learning and tests, various algorithms to take the best machine learning a model. The second part is based on how to develop a web application with the flask and integrate your model and the third part is deploying the web application on Heroku.

CHAPTER 3 - IDENTIFICATION OF RELEVANT TECHNOLOGIES

3.1 Introduction

Different techniques for predicting cancer disease, as well as the limitations and challenges in previously developed models, were discussed in the previous chapter. In this research, using Machine Learning Techniques, develop a cancer disease prediction model and developing a user-friendly web application to get a hint for the patient if he has a risk for cervical cancer.

3.2 Machine Learning

Machine learning refers to the process of creating computer algorithms that can imitate human intelligence. Computer science, probability and statistics, artificial intelligence, information theory, control theory, psychology, and philosophy are all concepts that are incorporated. Machine learning algorithms are used in a wide range of applications including speech, recognition, medicine, computer vision, email filtering, and many more when traditional algorithms are incapable of accomplishing the required tasks.

Another sort of artificial intelligence is machine learning, which seeks to understand the structure of data and fit it into models. Machine learning is an area of computer science that differs from traditional methods of computation. Machine learning, trains on data, and uses statistical analysis to produce values. These algorithms look for patterns in data and use them to make crucial choices and forecasts about the dataset. Furthermore, machine learning is a very iterative process in which the algorithms learn on their own by analysing previous data.

In today’s world, most technologies use machine learning. OCR technology transforms text images into movable types. FRT allows social media to identify

tag, users, and share among friends. Some recommended engines, show which movies, videos, and telegrams to watch next based on user preferences.

Machine learning algorithms use historical data as input and develop ML models using training data to forecast output values for the future. To predict new output values, machine learning algorithms are used historical data as input and create ML models using training data. There are several steps needed to be followed when developing an ML model to predict values. The ML model development process is explained in the following.

Machine Learning model development process

There are 6 key steps to building a machine learning a model.

- Data Collection
- Data preprocessing
- Model creation
- Evaluation
- Parameter tuning
- Prediction

It describes each step below one by one.

Data Collection

For a given problem, It's necessary to gather data and use this data to build the machine learning a model. The quality and quantity of these data are very important because they will directly affect the accuracy of the model. The dataset may be an existing one or created using scratch.

Data Preprocessing

Duplicate, missing, and noisy data make up the majority of the real-world dataset. To improve data quality, data preparation is required. Quality data must be used to build the quality model. The duplicate or missing values distorted overall statistics.

The outliers and inconsistent data points are caused to faulty prediction. These quality data are needed for this. Data preprocessing is done in several steps.

- Acquire the dataset
- Handling null values and duplicate values
- Handling categorical variables
- Scaling the features
- Split the dataset into two parts: training and testing.

The Model Creation

After training a machine learning algorithm on the collected data, a machine learning model produces the result. It is critical to select a model that is appropriate for the task. The most crucial phase in machine learning is training. To detect patterns and make predictions have to provide prepared data to the machine learning a model during training. As a result, the model learns from the data and can accomplish the task assigned. As it is trained, the model gets better at predicting.

The test data is used to check whether a model has learned efficiently and assess the model. So, in testing, the phase model is evaluated using different evaluation metrics.

Evaluation

After creating the model, it is required to check how the model is performing. It can be done by testing the model's performance using previously unseen data. The testing dataset is used in this stage to evaluate the model. If we test on the same data that were used for training, we will most likely not get the expected good result because the algorithm is already familiar with the data and recognizes similar patterns that it recognized previously. This will provide a high level of precision disproportionately. Following are some of the evaluation metrics that are used in model evaluation.

Parameter Tuning

We assess whether the accuracy of a model can be enhanced in any way once it has been built and tested in this step. The parameters of the model are fine-tuning in this step. The programmer sets the variables are called parameters. At a given value of the parameter, the accuracy will be at its highest. Parameter tuning is the process of determining these settings.

Prediction

Finally, The machine learning a model can predict values for data that have not been seen.

Machine Learning Approaches

Supervised Learning

In this approach, create a machine learning a model using input features and labelled output data. The model can detect relationships between input data and output data. Supervised getting to know excels at regression and classification troubles, which include figuring out the category of an information article or forecasting the quantity of income for a destiny date. The goal is to recognize statistics within the context of a particular hassle.

Un Supervised Learning

In this strategy, the algorithm is intended to discover similarities or patterns on its own for provided unlabelled data. The algorithms are trained on unlabelled, uncategorized, unclassified, and test data. Unsupervised learning algorithms, rather than responding to input, search for patterns in data and react in response to the presence or absence of such patterns in each new set of data.

Semi-supervised Learning

This method combines supervised and unsupervised learning components. In this method, tagged or unlabelled datasets can be used as inputs. Researchers in machine learning have demonstrated that unlabelled data with a few labelled data can improve learning performance significantly, despite the lack of training labels in some cases.

Reinforcement Learning

The study of how software agents should act in a given environment to maximize reward. When the best possible action/decision should be taken by a computer utilizing its own experience, this method is used. The model is given an answer key to training itself in supervised learning. In Reinforcement Learning, however, no solution key is provided, and the agent must select how to accomplish the assignment on their own. Because a training dataset is not available, it will have to learn on its own through experience or trial and error.

CHAPTER 4 – METHODOLOGY

4.1 Introduction

The design of the research project is described in Chapter 4. This chapter is broken down into four sections. In the first section, the details of the training dataset are described. Data preprocessing is described in the second section. The discussion of developing a machine learning algorithms takes up the third section. Finally, the chapter's summary is discussed.

4.2 Problem

The proliferation of abnormal cells in the cervix's lining is characterized as cervical cancer. [20] The most important thing is there is more chance of detection in the early stages.

Five-Year Survival Rates by Percentage	
Localized	92%
Regional	58%
Distant	17%
All Stages combined	66%

Fig 4.1 Stage Rates by Precentage

Localized means that the cancer has not spread beyond the cervix. Regional cancer has spread past the cervix and distant indicates that the cancer has spread to nearby organs.

The goal is to create a web application using a machine learning a model which predicts the outcome of a biopsy and certify the presence or absence of cervical cancer.

4.3 Data Collection

To predict cervical cancer disease, The data was gathered at 'Hospital Universitario de Caracas' in Caracas, Venezuela. It contains many risk factors leading to a biopsy. 858 records and 36 columns are contained in this dataset. This dataset has 35 features and biopsy target variables.

In the US 11,000 new cases are diagnosed each year. Each year cervical cancer kills 300,000 women worldwide.

The features cover habits, demographic information, and historic medical records.

Some of the attributes of the dataset are,

- Age
- Sexually Transmitted Diseases
- Number of sexual patterns
- Number of pregnancies
- First sexual intercourse
- Smokes
- IUD
- Hormonal Contraceptives

4.4 Analyzing data

Using the data analysis It can be identified are the features more affected by the cervical cancer biopsy. As well as use these more-imported features to do model prediction. They provide insights into current data and are helpful to understand patterns and other information in the dataset. Using analysis results, we can come up with better mechanisms to deal with future data.

Analysis data

The missing Values Imputation

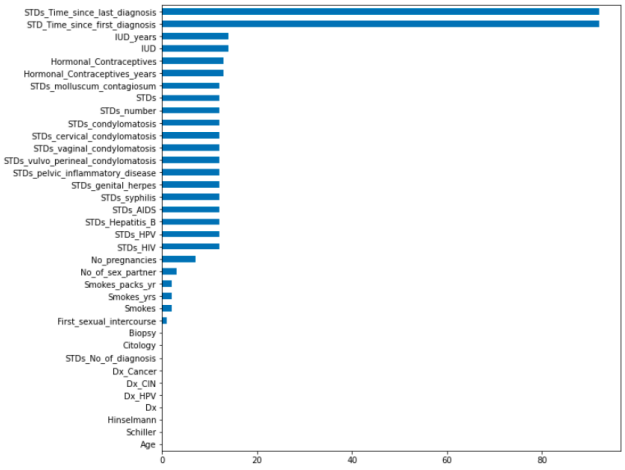


Fig 4.2 Missing value percentage of the dataset

STD time since the last diagnosis and STD time since the first diagnosis are having more than 80% null values. So dropping it off these two columns. Smokes and first sexual intercourse have a very little amount of missing values. So remove these missing records of these two columns.

This dataset is sensitive and medical. Therefore not use the method of imputation using mean, median, and mode. Here used machine learning models to fill the missing values. There are a few steps to follow to fill the missing values.

1. First, remove the columns, that need to be imputed from the independent column list.
2. Fill the null values with mode or median depending on their data type.
3. Choose the column that needs to be imputed as ‘Y’ and all other features as ‘X’
4. Build an ML model and train it and predict the missing values of ‘Y’.
5. DecisionTreeRegressor is used for numerical columns and used DecisionTreeClassifier for categorical columns.

Partitioning the features into categorical and numerical

Exploratory data analysis

Uni Variate analysis

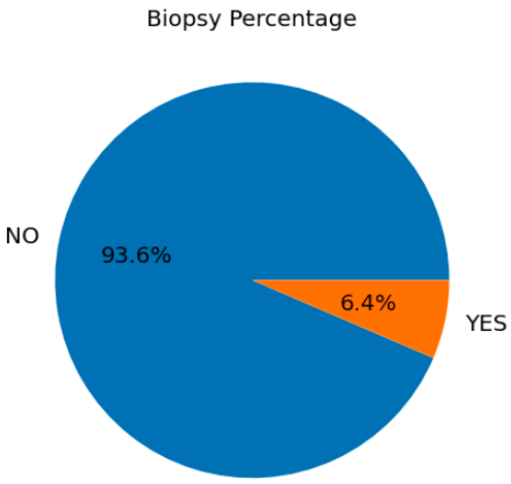


Fig 4.3 Biopsy Percentage

According to the dataset, 6.4% of people have cervical cancer. There is an imbalance in the data which cares about the model building section.

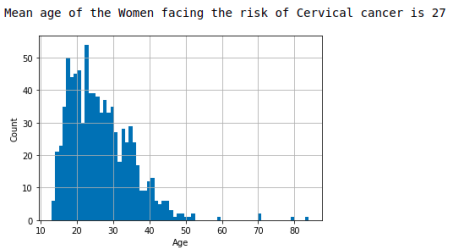
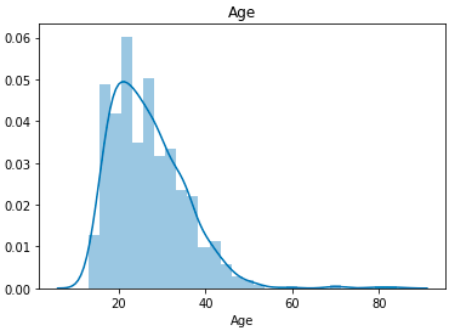


Fig4.4 Age analysis



Most of the patients are in the age group 20 - 40 and the Mean age of the women facing the risk of cervical cancer is 27.

All these features are grouped into four parts.

- Smoking habit attributes
- Sexual habit attributes
- Birth control attributes
- STDs

- Smoking habits

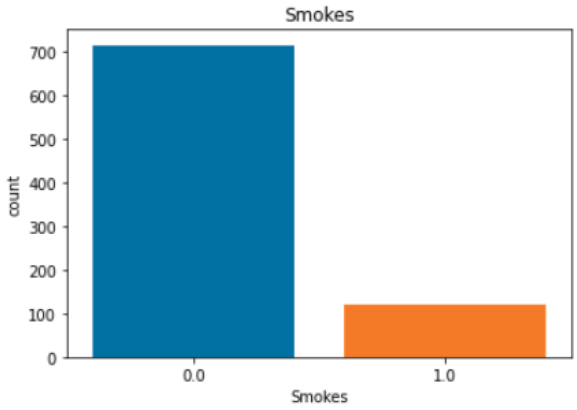


Fig 4.5 Smoking habit analysis

A relatively larger proportion (700) of the patients are non-smokers and a small amount (100) of patients are smokers.

- Sexual Habits

Most of the patients had 0 - 5 sexual partners.

The larger group of patients have first sexual intercourse between 15 - 20 years

Most of the patients have 0 -3 pregnancies in their life.

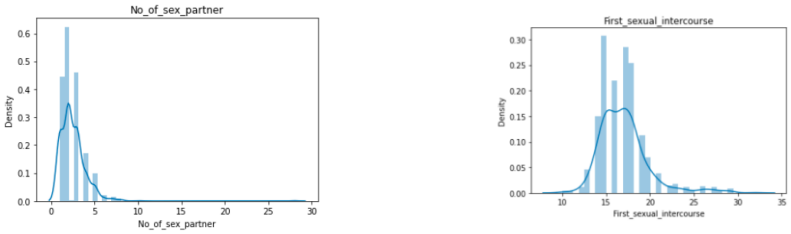


Fig 4.6 Sexual habit analysis

Birth Control Habits

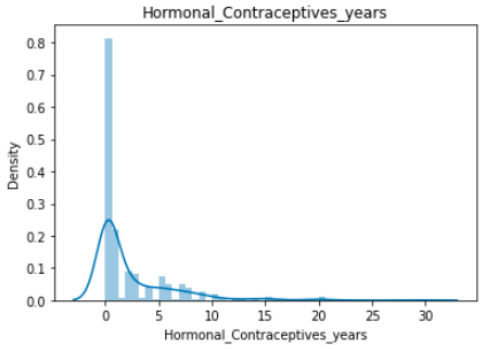


Fig 4.7 Birth control analysis

Most patients use birth control methods for less than 2 years and few patients use birth control methods for more than 2 years.

- STDs (Sexually Transmit Diseases)

All of the STDs effect a very small number of people.

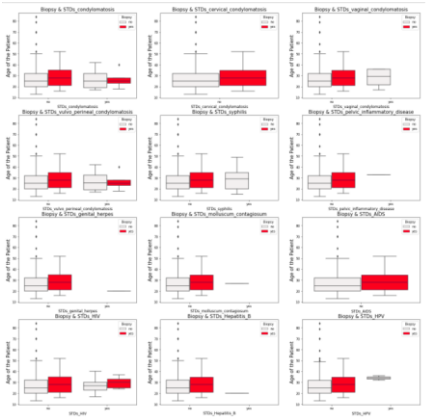


Fig 4.8 STD analysis

Multivariate Analysis

- Age and Sexual Habits Vs Biopsy

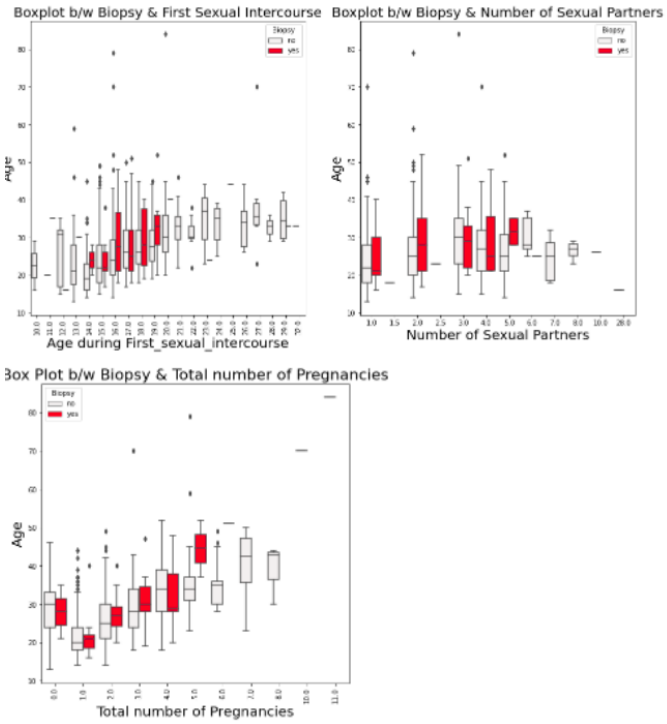


Fig 4.9 Age and Sexual habits vs Biopsy

First Sexual Intercourse means the age when the patient had their first sexual intercourse. The number of sexual partners means the total number of sexual partners the patient had and the total number of pregnancies means the patient had a total number of pregnancies.

Most patients aged 15 -18 had their first sexual intercourse. Most of the patients with cancer risk are aged group 20 - 35.

The persons who have sexual partners between 1 & 3 are more prone to be tested as positive in Biopsy tests and they are predominantly in the age group of 20 to 35.

Higher the number of pregnancies higher the chance of getting a positive biopsy.

- Smoking and sexual habits vs biopsy

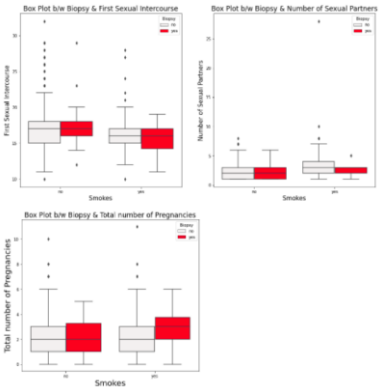


Fig 4.10 Smoking and Sexual habits vs Biopsy

Those who smoke and have first sexual intercourse at a younger age between 14 to 18 get a more chance to have a positive biopsy and no clear relationship between smoking and the number of sexual patterns. Smoking and non-smoking patients have nearly the same number of sexual partners.

The person who smokes and has a higher number of pregnancies gets the more chance of positive biopsy results.

- Age and smoking habits vs Biopsy

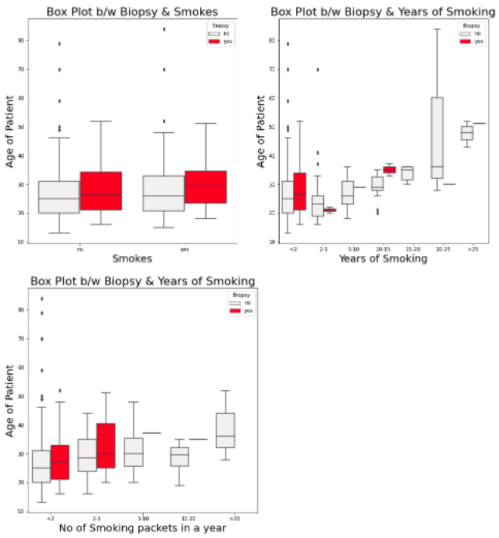


Fig 4.11 Age and Smoking habits vs Biopsy

Smoking people who aged high are more chance to be positive. A person who has smoked for at least one year is more chance to be positive. A person who aged high and smokes more packets a year gets a high chance of being a positive biopsy.

- Birth Control Attributes & age vs biopsy

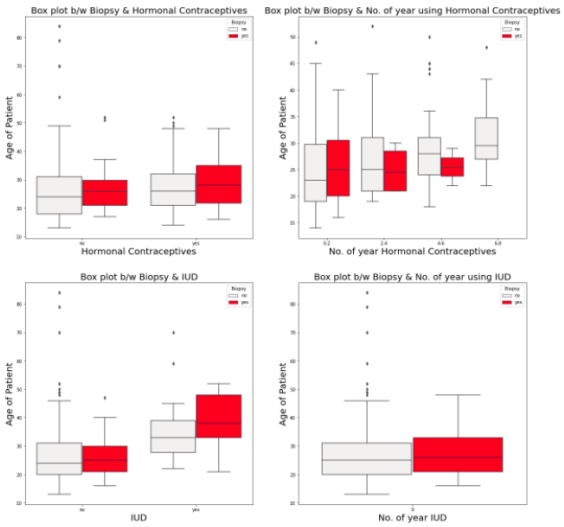


Fig 4.12 Age and Birth Control Attributes vs Biopsy

High age and no use of birth control methods, show positive results in biopsy test. The patients with 0-4 years of usage in hormonal contraceptives and the average age between 20 years & 30 years show positive results Biopsy test.

The persons who did not use the IUD and with lesser age between 25 years & 35 years show positive for Biopsy test, whereas, among those who used the IUD, the high-aged people (around 40) are more prone to cancer.

4.5 Feature Engineering

Outlier treatment

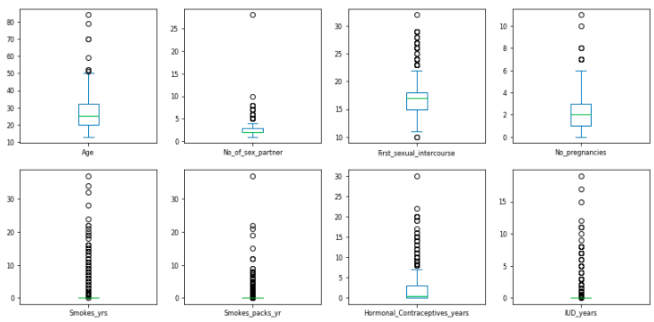


Fig 4.13 Data containing outliers

The above graph implies that the data contains outliers.

Below shows the outlier treatment graph.

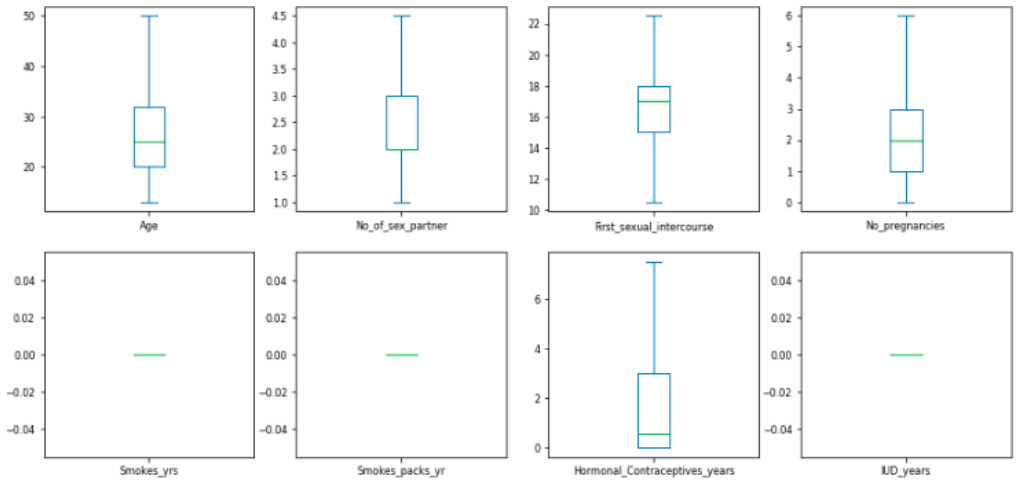


Fig 4.14 Data without outliers

CHAPTER 5 – EVALUATION

5.1 Introduction

This chapter discusses the details of the results obtained from the evaluations done for a model prediction of cervical cancer.

5.2 Basics

In the evaluation, divide this dataset into two parts: training and testing datasets, with 70% used for testing and 30% used for training. Use five models that are Logistic Regression, Decision Tree, Random Forest, GaussianNB, and KNN.

Logistic Regression

Logistic Regression is used in many social science applications. Specially Logistic regression is used when the target variable is categorical.The sigmoid function is used in this model. Output is 1 or 0 [21]

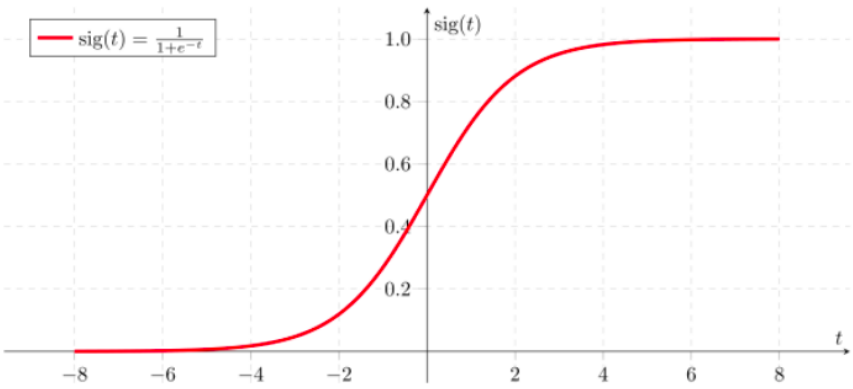


Figure 2: Sigmoid Activation Function

Fig 5.1 Sigmoid Function

The logistic function is defined as follows.

logistic() = 11 + exp(-)

This model not only the classification model but also give the probabilities also. This is an advantage of this model. It has many logistic regression types. When used multinomial logistic regression model it can be used binary classification to multi-class classification [21]

Decision Tree

The Decision Tree approach is a supervised learning method for classification and regression that uses decision trees. This is straightforward and simple to comprehend, and trees may be pictured. Branches represent decision rules, internal nodes represent dataset properties, and leaf nodes represent the decision tree's conclusion. Because it has a tree-like structure, the rationale behind the decision tree is simple to comprehend.

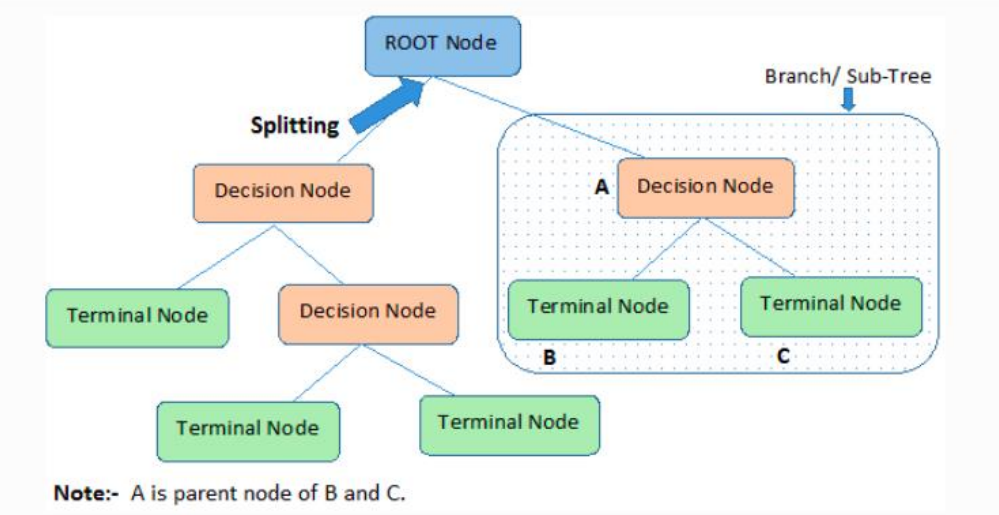


Fig 5.2 Decision tree structure

5.3 Evaluation

Using the above model classification finally get the summary of the results in the below table

	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
0	LogisticRegression	0.974403	0.952381	0.571429	0.533333	0.615385	0.756118
1	Decision Tree	1.000000	0.952381	0.647059	0.733333	0.578947	0.849789
2	Random Forest	1.000000	0.936508	0.529412	0.600000	0.473684	0.778903
3	GaussianNB	0.146758	0.095238	0.116279	1.000000	0.061728	0.518987
4	KNN	0.950512	0.936508	0.333333	0.266667	0.444444	0.622785

Fig 5.3 Five model evaluation results

This is very sensitive medical data, recall score is given higher importance because recall means the how many correctly predicted positive values. Among these five models choose the Decision tree and Random Forest models because of higher recall and high roc_auc scores. Recall should be given higher importance because it's necessary to predict actual cancer patients as cancer patients accurately.

If a predicted healthy person is a cancer patient it is very dangerous and may be harmful to the patient life.

There is an imbalance in the target variable. Therefore we need to take care of it in this model evaluation section.

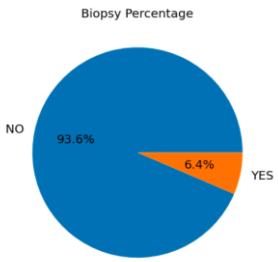


Fig 5.4 Target variable imbalance

This section used OVERSAMPLING TECHNIQUE - SMOTE to overcome this data balance. After using this technique both classes have the same proportions.

Below graph shows the Decision tree and Random forest model after sampling methods. It gets high recall and roc_auc than before.

	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
0	Decision Tree After Sampling	1.0	0.952381	0.647059	0.733333	0.578947	0.849789
1	Random Forest After Sampling	1.0	0.944444	0.611111	0.733333	0.523810	0.845570

Fig 5.5 - Results after sampling

5.3.1 Feature Selection

There are many feature selection techniques and the below image shows these types.

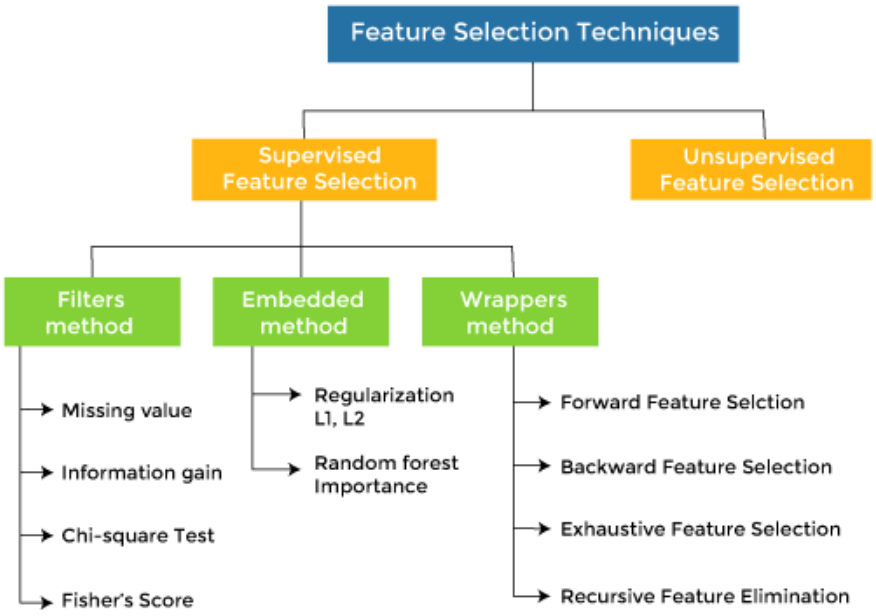


Fig 5.6 Feature Selection Techniques

RFE is a recursive greedy optimization approach, where features are selected taking a smaller and smaller subset of features. After trained each set of features determined the importance of each feature using accuracy.

Used RFE on Random Forest and Decision Tree separately and found the best features for both models separately. The features are also chosen based on recall score.

It can be getting the below table summary after feature selection.

	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
0	Decision Tree After Sampling	1.0	0.952381	0.647059	0.733333	0.578947	0.849789
1	Random Forest After Sampling	1.0	0.944444	0.611111	0.733333	0.523810	0.845570
0	Decision Tree after Feature Selection	1.0	0.988095	0.926829	0.904762	0.950000	0.950216
1	Random Forest after Feature Selection	1.0	0.992063	0.952381	0.952381	0.952381	0.974026

Fig 5.7 Feature selection results

The above shows that the recall score got better after feature selection. This score can be improved using hyper-parameter tuning.

5.3.2 Hyper Parameter tuning

The grid search is a popular way to perform hyper parameter tuning. This method attempts to choose a set of hyper parameters. Used Grid Search Cross-Validation for the decision tree.

Used Randomized search cross-validation for the random forest. There is a main difference between the RandomizedSearch CV and Grid search CV. This chooses hyper parameter sample combinations randomly from grid space.

Below shows the summary table after getting the hyper parameter tuning.

	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
0	Decision Tree After Sampling	1.00000	0.952381	0.647059	0.733333	0.578947	0.849789
1	Random Forest After Sampling	1.00000	0.944444	0.611111	0.733333	0.523810	0.845570
0	Decision Tree after Feature Selection	1.00000	0.988095	0.926829	0.904762	0.950000	0.950216
1	Random Forest after Feature Selection	1.00000	0.992063	0.952381	0.952381	0.952381	0.974026
0	Decision Tree after Hyperparameter Tuning	0.97989	0.960317	0.800000	0.952381	0.689655	0.956710
1	Random Forest After Hyperparameter Tuning	0.97989	0.968254	0.833333	0.952381	0.740741	0.961039

Fig 5.8 Hyper Parameter tuning results

The recall score is getting improved after hyper parameter tuning.

5.3.3 Ensembling Methods

Ensembling methods are approaches for creating many models and combining them to generate better outcomes. These strategies outperform a single model in terms of accuracy.

Here used three Ensembling methods. There are Bagging, Ada Boost, and Gradient boost techniques.

	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
1	Decision Tree After Sampling	1.000000	0.952381	0.647059	0.733333	0.578947	0.849789
2	Random Forest After Sampling	1.000000	0.944444	0.611111	0.733333	0.523810	0.845570
3	Decision Tree after Feature Selection	1.000000	0.988095	0.926829	0.904762	0.950000	0.950216
4	Random Forest after Feature Selection	1.000000	0.992063	0.952381	0.952381	0.952381	0.974026
5	Decision Tree after Hyperparameter Tuning	0.979890	0.960317	0.800000	0.952381	0.689655	0.956710
6	Random Forest After Hyperparameter Tuning	0.979890	0.968254	0.833333	0.952381	0.740741	0.961039
7	Bagged Decision Tree with Hyperparameter	0.979890	0.956349	0.765957	0.857143	0.692308	0.911255
8	Decision Tree ADA Boost with Hyperparameter	1.000000	0.988095	0.926829	0.904762	0.950000	0.950216
9	Gradient Boost	0.980804	0.976190	0.857143	0.857143	0.857143	0.922078

Fig 5.9 - Final results

There is no improvement in recall when using these methods. But it has improved the train score, but this is sensitive data we forced on the recall score. The final outcome is a Random forest after hyper parameter tuning.

Comparison tables with benchmark models

These values related to Random forest model

	Accuracy	**Precision**	**Recall**	**F1-score**
Benchmark 1 [24]	89%	90%	90%	89%
Benchmark 2 [25]	72.19%	NA	NA	NA
Benchmark 3 [26]	NA	NA	NA	91%
Proposed model	99.20%	95.23%	95.23%	95.23%

Table 5.1 - Comparison table with benchmark models

5.4 Implementation

In this section, I try to implement a user-friendly, attractive web application for users to get a hint that the person has cervical cancer or not. The above section found the best machine learning a model. Initially, the data set has 35 features and finally removes unnecessary features and filters them to 33 features. The removed attributes are time since the first diagnosis and time since the last diagnosis. There are so much missing data in these two features. The below figure shows the overview of the cancer prediction system web application. This is the part which has novelty in this project.

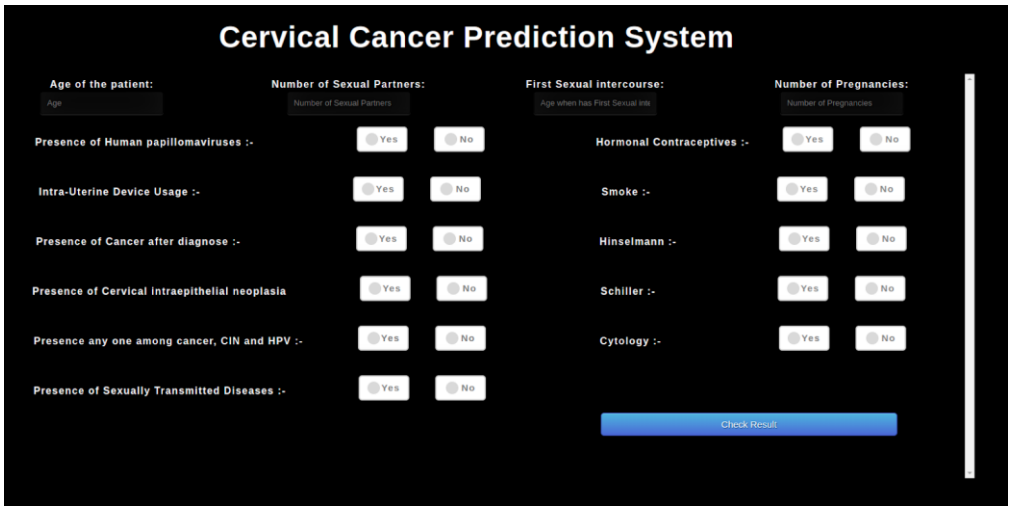


Fig 5.10 Overview of the web application

Below is a description of the features.

1. Age - The age of women is a numerical value.
2. The number of Sexual Partners -
3. First Sexual Intercourse - Her first sexual intercourse age.
4. Number of Pregnancies -
5. The presence of human papillomaviruses.
6. Hormonal Contraceptives - The woman uses hormonal contraceptives or not.
    1. Yes - Uses hormonal contraceptives
    2. No - Don’t use hormonal contraceptives

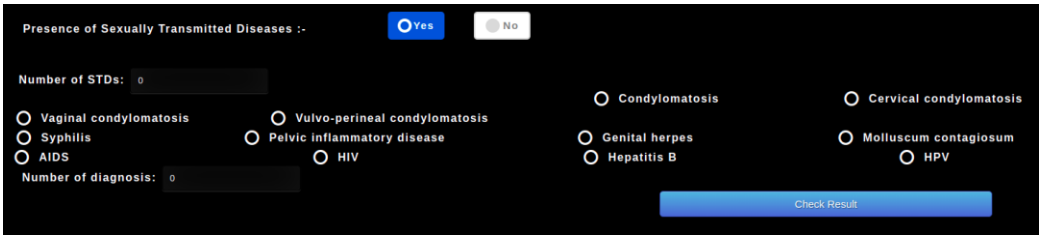
7. Intra-Uterine Device usage - It expresses whether the woman uses an intrauterine contraceptive device or not.
    1. Yes - Use IUD
    2. No - Don’t use IUD
8. Smoke - It expresses the woman smoke or not.
    1. Yes - Smoke
    2. No - Don’t smoke
9. Presence of cancer after diagnosis.
    1. Yes - Has cancer
    2. No - Hasn’t cancer
10. Hinselmann - It is also called colposcopy. It is a medical test that magnified the view of the cervix, vagina, and vulva.
11. Presence of Cervical intraepithelial neoplasia
12. Schiller - It is also a medical test in which iodine solution is applied to the cervix
    1. Yes - Positive result of the test
    2. No - Negative result of the test
13. Presence any one among cancer, CIN, and HPV.
14. Cytology - It is also known as the PaP smears test, and helps to identify abnormal cells in the cervix which has more chance of cervical cancer.
15. Hormonal Contraceptives (years) - If the hormonal contraceptive output is yes then show this field.

It indicates how many years the contraceptives used. It expresses the total number of years.

If click smoke yes then shows below two fields.

16. Smoke (Years) - How many years have women been smoking.
17. Smoke (Packs/Year) - The woman smokes the total number of packets of cigarettes per year.
18. IUD (Years) - It expresses how many years the woman was IUD used.

19. STD - It expresses the presence of Sexually Transmitted Diseases.
    1. Yes - Has Sexually transmitted disease
    2. No - Hasn't Sexually transmitted disease
20. STD (Number) - The woman has the total number of sexually transmitted diseases.
21. Condylomatosis
22. Cervical condylomatosis
23. Vaginal condylomatosis
24. Vulvo-perineal condylomatosis
25. Syphilis
26. Pelvic inflammatory disease
27. Genital herpes
28. Molluscum contagiosum
29. AIDS
30. HIV
31. Hepatitis B
32. HPV
33. The number of diagnoses - Total number of times, the STDs have been diagnosed.



This evaluation has four parts.

1. models.py
2. request.py
3. App.py
4. HTML/CSS

This file contains code that predicts the presence of cervical cancer based on the symptoms and habits. A random forest classifier is used as a model with the following parameters

- Criterion - gini
- Max_depth - 9
- Max_features - log2
- Max_leaf_nodes - 9
- N_estimators - 50

These values are obtained using the above evaluation part. This project used pickling to convert python objects into character streams.

Then create an API that receives cervical cancer information via a GUI and computer to predict the biopsy results based on our model. For this de-serialized pickle model to python object and used index.html to create the main page. When the user submits the form data via the POST method, the user can show the predicted biopsy result.

Finally used the request.py file to call APIs defined in app.py

CHAPTER 6 – CONCLUSION

6.1 Introduction

This section focuses on providing an overview of the overall study as well as future work for this research study. First, a summary of research has been provided and then the limitations of research as well as future works that have been determined to be carried out were discussed in this chapter.

6.2 Overview of research

Among several diseases, cancer is the most hazardous. Cancers come in a variety of forms. The second most common cancer is cervical cancer after breast cancer. There's a better possibility of catching this malignancy early on. Users are crucial to this project for this reason. There are 858 patient records in this dataset, and 35 characteristics with biopsy target variables. Accurate predictions of these facts will aid cervical cancer treatment and benefit patients. This has a lot of promise for use in hospitals and for improving cervical cancer management.

This thesis completes the first stage of the project. Machine Learning has a high capacity for processing large amounts of data, which is why it has been employed in medical fields.

That is using five techniques such as Random Forest, Logistic Regression, Gaussian NB, and KNN, Decision Tree methods to perform the highest accuracy model that is used for further research and various methods to improve the accuracy. Used sampling to overcome the imbalance of the target variable, Feature Selection, Hyper Parameter tuning, and Ensembling methods to improve accuracy.

First objective of this thesis is recognize the most important cervical cancer symptoms.

- Higher number of pregnancies
- Smokes
- High Age
- Did not use IUD with less Age

- Did not use hormonal contraceptives with high age

Second objective of this project is build user-friendly web application with high accuracy model and it is done and implement clearly in above chapter.

Final objective is Find the patient biopsy using this web application. It is also done in above chapter.
When it has positive biopsy then it shows “You hve risk of cervical cancer” and negative biopsy shows “You haven’t any risk of cervical cancer”.

Finally It Can achieve all objectives in thos project.

Most of the research studies done in this area, but most of them are not user-friendly.

6.3 Limitations

According to this study, the major limitations are the lack of data in the dataset, and some missing data. This dataset has only 858 records. There is a data imbalance of the target variable and it may affect to quality of the model. nuWhen prediction the model high quality model can be created that dataset has more data.

6.4 Future works

The next section of this research is to develop a user-friendly web application using a flask to perform the patient has cancer or not.

6.5 Summary

This chapter finalized the thesis by discussing the solution provided by the Random Forest approach for predicting cervical cancer and how it may be improved further by using various methods and converting all these data to a user-friendly web application. Users can input their STDs and other diseases through this web application then which gives the result of their risk of having cervical cancer or not.

REFERENCES

[1] Internet Society. 2021. *Artificial Intelligence & Machine Learning: Policy Paper | Internet Society*. [online] Available at: <https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/?gclid=CjwKCAiAkJKCBhAyEiwAKQBCku-9w5eyPXutXRbr5qQxr1dRHehOrlgjUsfjsQLWS4RQmfW_3E8r0RoCF7cQAvD_BwE>

[2] H. Shee, K. W Cheruiyot and S. Kimani, “Application of k-Nearest Neighbour Classification in Medical Data Mining,” International Journal of Information and Communication Technology Research, Volume 4, No. 4, April 2014.

[3] N. Lakhotia, "Integrating Machine Learning into Web Application with Flask", *Analytics Vidhya*, 2021.[Online]. Available: https://www.analyticsvidhya.com/blog/2020/09/integrating-machine-learning-into-web-applications-with-flask/.

[4] "Build and deploy your first machine learning web app - KDnuggets", *KDnuggets*, 2021. [Online]. Available: https://www.kdnuggets.com/2020/05/build-deploy-machine-learning-web-app.html.

[5] "Building a Web Application to Deploy Machine Learning Models", *Medium*, 2021. [Online]. Available: https://towardsdatascience.com/building-a-web-application-to-deploy-machine-learning-models-e224269c1331

[6] 2021. [Online]. Available: https://www.researchgate.net/publication/330880328_Implementation_of_a_Web_Application_to_Predict_Diabetes_Disease_An_Approach_Using_Machine_Learning_Algorithm.

[7] "Machine learning on mobile devices: 3 steps for deploying ML in your apps", *Medium*, 2021. [Online]. Available: https://heartbeat.fritz.ai/machine-learning-on-mobile-devices-3-steps-for-deploying-it-in-your-apps-48a0a24364a8.

[8] "Machine Learning Algorithm - an overview | ScienceDirect Topics", *Sciencedirect.com*, 2021. [Online]. Available: https://www.sciencedirect.com/topics/engineering/machine-learning-algorithm.

[9] "Types of cancer", *Cancer Research UK*, 2021. [Online]. Available: https://www.cancerresearchuk.org/what-is-cancer/how-cancer-starts/types-of-cancer.

[10] "Cancer Classification | SEER Training", *Training.seer.cancer.gov*, 2021. [Online]. Available: https://training.seer.cancer.gov/disease/categories/classification.html.

[11] R. Rastogi, D. Chaturvedi, S. Satya and N. Arora, "Intelligent Heart Disease Prediction on Physical and Mental Parameters: A ML Based IoT and Big Data Application and Analysis",

[12] H. Shahwani, "Machine Learning-based Web Application for Early Diagnosis of Diabetes", *Journal.buitems.edu.pk*, 2021. [Online]. Available: http://journal.buitms.edu.pk/j/index.php/bj/article/view/418/259.

[13] K. Kourou, T. Exarchos, K. Exarchos, M. Karamouzis and D. Fotiadis, "Machine learning applications in cancer prognosis and prediction".

[14] Ijitee.org, 2021. [Online]. Available: https://www.ijitee.org/wp-content/uploads/papers/v8i6s/F60600486S19.pdf.

[15]L. Qian et al., "Application of deep learning to predict underestimation in ductal carcinoma in situ of the breast with ultrasound",

[16] M. Islam, M. Haque, H. Iqbal, M. Hasan, M. Hasan and M. Kabir, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques",

[17] *Ripublication.com*, 2021. [Online]. Available: http://ripublication.com/ijaer19/ijaerv14n11_07.pdf.

[18] "Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer",

[19] Mayo Clinic. 2022. *Cervical cancer - Symptoms and causes*. [online] Available at: <https://www.mayoclinic.org/diseases-conditions/cervical-cancer/symptoms-causes/syc-20352501> [Accessed 10 May 2022].

[20] Cancer.org.au. 2022. *Cervical cancer | Causes, Symptoms & Treatments*. [online] Available at: <https://www.cancer.org.au/cancer-information/types-of-cancer/cervical-cancer> [Accessed 10 May 2022].

[21] Medium. 2022. *Logistic Regression — Detailed Overview*. [online] Available at: <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc> [Accessed 10 May 2022].

[22] Molnar, C., 2022. *5.2 Logistic Regression | Interpretable Machine Learning*. [online] Christophm.github.io. Available at: <https://christophm.github.io/interpretable-ml-book/logistic.html> [Accessed 10 May 2022].

[23] www.javatpoint.com. 2022. *Feature Selection Techniques in Machine Learning - Javatpoint*. [online] Available at: <https://www.javatpoint.com/feature-selection-techniques-in-machine-learning> [Accessed 10 May 2022].

[24] Kruczkowski, M., Drabik-Kruczkowska, A., Marciniak, A., Tarczewska, M., Kosowska, M. and Szczerska, M., 2022. *Predictions of cervical cancer identification by photonic method combined with machine learning*.

[25] Yin, Q, 2022. *The Application of Machine Learning in Cervical Cancer Prediction*. [online] Dl.acm.org. Available at: <https://dl.acm.org/doi/fullHtml/10.1145/3468891.3468894> [Accessed 11 July 2022].

[26] Nikookar, E., Naderi, E. and Rahnavard, A., 2022. *Cervical Cancer Prediction by Merging Features of Different Colposcopic Images and Using Ensemble Classifier*. [online] PubMed Central (PMC). Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8253312/> [Accessed 12 July 2022].

[27] Islam, A., Ripon, S. and Bhuiyan, M., 2022. [online] Available at: <https://www.researchgate.net/figure/Comparison-of-Accuracy-Recall-Precision-Kappa-and-F1-score-of-different-algorithms-on_tbl3_333595688> [Accessed 12 July 2022].

[28] Sun, L., Yang, L., Liu, X., Tang, L., Zeng, Q., Gao, Y., Chen, Q., Liu, Z. and Peng, B., 2022. *Optimization of Cervical Cancer Screening: A Stacking-Integrated Machine Learning Algorithm Based on Demographic, Behavioral, and Clinical Factors*.

[29] Kruczkowski, M., Drabik-Kruczkowska, A., Marciniak, A., Tarczewska, M., Kosowska, M. and Szczerska, M., 2022. *Predictions of cervical cancer identification by photonic method combined with machine learning*.

[30] Razali1, N., Mostafa1, S., Mustapha1, A. and Ibrahim, N., 2022. [online] Iopscience.iop.org. Available at: <https://iopscience.iop.org/article/10.1088/1742-6596/1529/2/022102/pdf> [Accessed 21 May 2022].

[31] Alam1, T., Afzal Khan, M., Atif Iqbal, M., Wahab, A. and Mushtaq, M., 2022. [online] Thesai.org. Available at: <https://thesai.org/Downloads/Volume10No2/Paper_51-Cervical_Cancer_Prediction.pdf> [Accessed 11 May 2022].

[32] Tanimu, J., Hamada, M., Hassan, M. and Yusuf Ilu, S., 2022. [online] Shs-conferences.org. Available at: <https://www.shs-

conferences.org/articles/shsconf/pdf/2021/13/shsconf_etltc2021_04004.pdf> [Accessed 1 April 2022].

[33] Singh, J. and Sharma, S., 2022. [online] Ripublication.com. Available at: <https://www.ripublication.com/ijaer19/ijaerv14n11_07.pdf> [Accessed 2 May 2022].

[34] Khan, I., Aslam, N., Alshehri, R., Alzahrani, S., Alghamdi, M., Almalki, A. and Balabeed, M., 2022. [online] Available at: <https://www.hindawi.com/journals/sp/2021/5540024/> [Accessed 25 May 2022].

[35] Medicalnewstoday.com. 2022. *Cervical cancer: Symptoms, causes, stages, and treatment*. [online] Available at: <https://www.medicalnewstoday.com/articles/what-you-need-to-know-about-cervical-cancer>

[36] Ijitee.org. 2022. [online] Available at: <https://www.ijitee.org/wp-content/uploads/papers/v8i6s/F60600486S19.pdf>

[37] Nayan Kumar Sinha, Menuka Khulal, Manzil Gurung, and Arvind Lal, "Developing A Web based System for Breast Cancer Prediction using XGboost Classifier." [Online]. Available: https://www.researchgate.net/publication/342483667_Developing_A_Web_based_System_for_Breast_Cancer_Prediction_using_XGboost_Classifier. [Accessed: 05-Feb-2022]

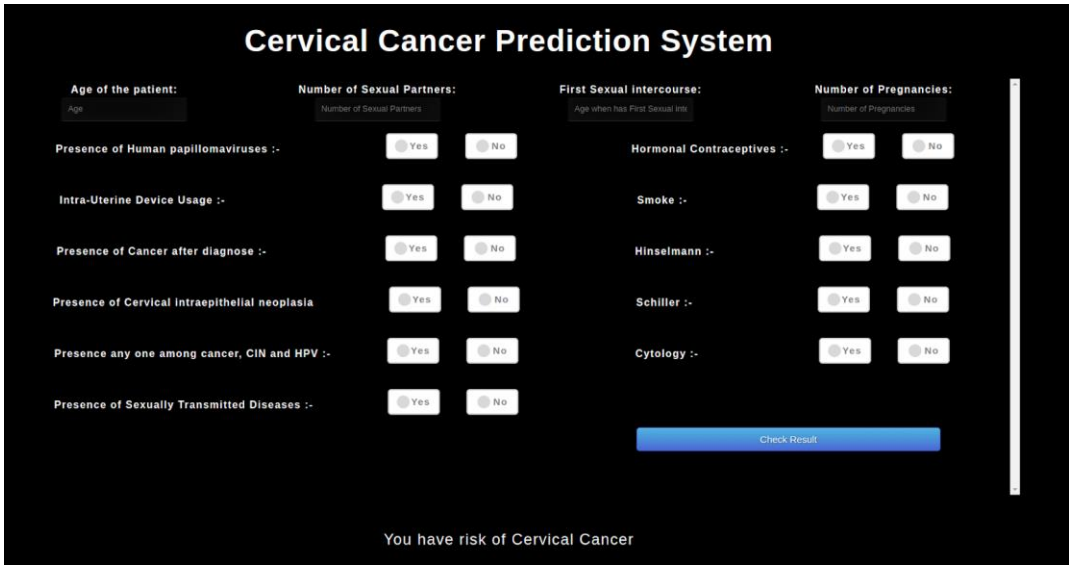
[38] T. Skillocity, "Breast cancer detection web app," *Skillocity*, 21-Jul-2021. [Online]. Available: https://skillocity.in/uncategorized/breast-cancer-detection-web-app/. [Accessed: 12-Jul-2022]

[39] M. A. Naji, S. E. Filali, K. Aarika, E. L. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine learning algorithms for breast cancer prediction and diagnosis," *Procedia Computer Science*, 08-Sep-2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921014629. [Accessed: 25-May-2022]

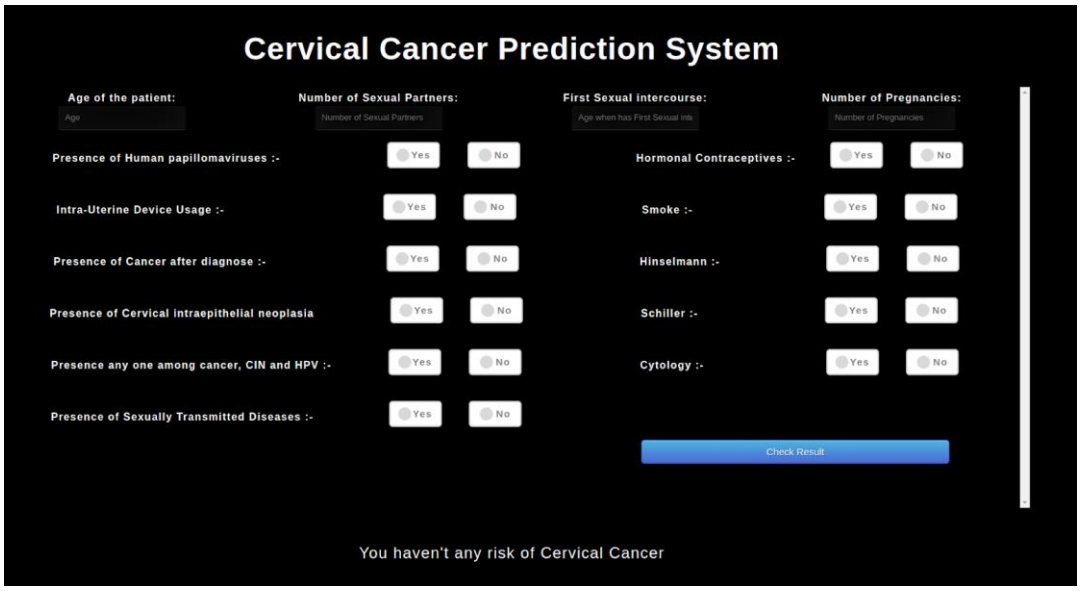
[40] H. Saleh, S. F. Abd-el ghany, H. Alyami, and W. Alosaimi, "Predicting breast cancer based on optimized deep learning approach," *Computational Intelligence and Neuroscience*, 19-Mar-2022.

[Online]. Available: https://www.hindawi.com/journals/cin/2022/1820777/.

Appendix A



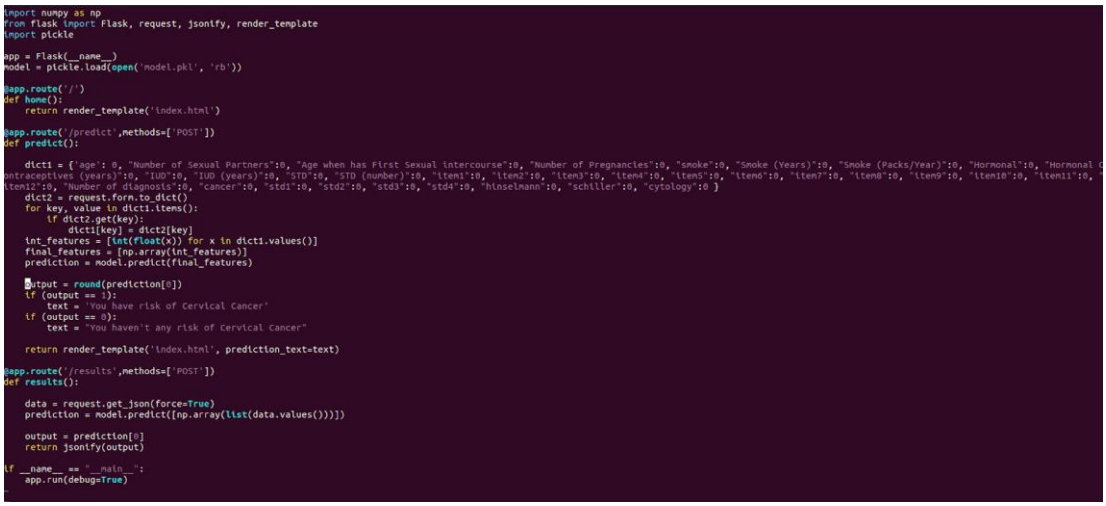
Appendix B



APPENDIX C

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import pickle

dataset = pd.read_csv('dataset.csv')

X = dataset.iloc[:, :33]

y = dataset.iloc[:, -1]

from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier(criterion= 'gini', max_depth = 9, max_features = 'log2', max_leaf_nodes=9, n_estimators=50).fit(X,y)


pickle.dump(model, open('model.pkl','wb'))

model = pickle.load(open('model.pkl','rb'))
~
```

APPENDIX D



```
import numpy as np
from flask import Flask, request, jsonify, render_template
import pickle

app = Flask(__name__)
model = pickle.load(open('model.pkl', 'rb'))

@app.route('/')
def home():
    return render_template('index.html')

@app.route('/predict',methods=['POST'])
def predict():

    dict1 = {'age': 0, "Number of Sexual Partners":0, "Age when has First Sexual intercourse":0, "Number of Pregnancies":0, "smoke":0, "Smoke (Years)":0, "Smoke (Packs/Year)":0, "Hormonal":0, "Hormonal C
ontraceptives (years)":0, "IUD":0, "IUD (years)":0, "STD":0, "STD (number)":0, "item1":0, "item2":0, "item3":0, "item4":0, "item5":0, "item6":0, "item7":0, "item8":0, "item9":0, "item10":0, "item11":0, "
item12":0, "Nunber of diagnosis":0, "cancer":0, "std1":0, "std2":0, "std3":0, "std4":0, "hinselmann":0, "schiller":0, "cytology":0 }
    dict2 = request.form.to_dict()
    for key, value in dict1.items():
        if dict2.get(key):
            dict1[key] = dict2[key]
    int_features = [int(float(x)) for x in dict1.values()]
    final_features = [np.array(int_features)]
    prediction = model.predict(final_features)

    output = round(prediction[0])
    if (output == 1):
        text = 'You have risk of Cervical Cancer'
    if (output == 0):
        text = "You haven't any risk of Cervical Cancer"

    return render_template('index.html', prediction_text=text)

@app.route('/results',methods=['POST'])
def results():

    data = request.get_json(force=True)
    prediction = model.predict([np.array(list(data.values()))])

    output = prediction[0]
    return jsonify(output)

if __name__ == "__main__":
    app.run(debug=True)
```

Git Hub Link

https://github.com/Chandi1994/PG-dip-Project