M.KUMARASAMY
COLLEGE OF ENGINEERING
NAAC Accredited Autonomous Institution
Approved by AICTE & Affiliated to Anna University
ISO 9001:2015 & ISO 14001:2015 Certified Institution
Thalavapalayam, Karur – 639 113.

A Minor project Report

on

CERVICAL CANCER PREDICTION

Submitted in partial fulfilment of requirement for the award of the Degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

Under the guidance of

Mrs. P. KAYALVIZHI M.E.,
Assistant Professor/CSE

Submitted By

ARVIND VENKAT (927621BCS013)
DEEPAN RAJ G (927621BCS018)
DURAI MURUGAN V (927621BCS027)
GOBALA KRISHNAN G (927621BCS032)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

M.KUMARASAMY COLLEGE OF ENGINEERING
(Autonomous)

KARUR – 639 113

May 2024

M.KUMARASAMY
COLLEGE OF ENGINEERING
NAAC Accredited Autonomous Institution
Approved by AICTE & Affiliated to Anna University
ISO 9001:2015 & ISO 14001:2015 Certified Institution
Thalavapalayam, Karur – 639 113.

M.KUMARASAMY COLLEGE OF ENGINEERING

(Autonomous Institution affiliated to Anna University, Chennai)

KARUR-639113

BONAFIDE CERTIFICATE

Certified that this minor project report **"CERVICAL CANCER PREDICTION"** is the bonafide work of **ARVIND VENKAT (927621BCS013), DEEPAN RAJ. G (927621BCS018), DURAI MURUGAN. V (927621BCS027), GOBALA KRISHNAN. G (927621BCS032)** who carried out the project work during the academic year 2023-2024 under my supervision.

Signature

Mrs. P. KAYALVIZHI M.E.,

SUPERVISOR,

Department of Computer Science and Engineering,

M. Kumarasamy College of Engineering

Thalavapalayam, Karur-639113.

Signature

Dr. M. MURUGESAN M.E., Ph.D.,

HEAD OF THE DEPARTMENT,

Department of Computer Science and Engineering,

M. Kumarasamy College of Engineering

Thalavapalayam, Karur-639113.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VISION OF THE INSTITUTION

- To emerge as a leader among the top institutions in the field of technical education

MISSION OF THE INSTITUTION

- Produce smart technocrats with empirical knowledge who can surmount the global challenges
- Create a diverse, fully engaged, learner-centric campus environment to provide quality education to the students
- Maintain mutually beneficial partnerships with our alumni, industry, and Professional associations

VISION OF THE DEPARTMENT

- To achieve education and research excellence in Computer Science and Engineering.

MISSION OF THE DEPARTMENT

- To excel in academic through effective teaching learning techniques
- To promote research in computer science and engineering with thefocus on innovation
- To transform students into technically competent professionals with societal and ethical and responsibilities

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

PEO1: Graduates will have a successful career in software industries and R&D divisions through continuous learning.

PEO2: Graduates will provide effective solutions for real-world problems in the key domain of computer science and engineering and engage in lifelong learning.

PEO3: Graduates will excel in their profession by being ethically and socially responsible.

PROGRAM OUTCOMES(POs)

Engineering students will be able to do:

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
5. **Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. **Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-long learning:** Recognize the need for and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM SPECIFIC OUTCOME (PSO's)

- **PSO1: Professional Skills:** Ability to apply the knowledge of computing techniques to design and develop computerized solutions for the problems.
- **PSO2: Successful career:** Ability to utilize the computing skills and ethical values in creating a successful career.

ABSTRACT

The project aims to predict cervical cancer risk using machine learning techniques applied to a dataset of risk factors. Through comprehensive data exploration, preprocessing, and modeling, various algorithms including Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, AdaBoost, and Artificial Neural Networks. Key predictors such as age, number of sexual partners, pregnancies, and smoking habits were identified. The develop models exhibited high accuracy in predicting cervical cancer risk, demonstrating potential for targeted screening and personalized interventions. Future research will focus on model refinement and validation for real-world deployment in clinical settings.

ABSTRACT WITH POs AND PSOs

ABSTRACT	POs MAPPED	PSOs MAPPED
The project aims to predict cervical cancer risk using machine learning techniques applied to a dataset of risk factors. Through comprehensive data exploration, preprocessing, and modeling, various algorithms including Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, AdaBoost, and Artificial Neural Networks. Key predictors such as age, number of sexual partners, pregnancies, and smoking habits were identified. The develop models exhibited high accuracy in predicting cervical cancer risk, demonstrating potential for targeted screening and personalized interventions. Future research will focus on model refinement and validation for real-world deployment in clinical settings.	**PO1(3)** **PO2(3)** **PO3(2)** **PO4(2)** **PO5(3)** **PO6(1)** **PO7(2)** **PO8(2)** **PO9(3)** **PO10(2)** **PO11(2)** **PO12(2)**	**PSO1(2)** **PSO2(2)**

Note:1-Low,2-Medium,3-High

SUPERVISOR **HEAD OF THE DEPARTMENT**

TABLE OF CONTENT

LIST OF FIGURES

ACRONYMS/LIST OF ABBREVIATIONS

ML - Data Frame

DTR - Decision Tree Algorithm

LR - Linear Regression Algorithm

SVM - Support Vector Machine

ANN - Artificial Neural Network

CHAPTER 1
INTRODUCTION

In this hands-on project, we embark on a journey to construct and train an XGBoost model with the purpose of predicting cervical cancer in a cohort of 858 patients. The dataset, meticulously curated at the 'Hospital Universitario de Caracas' in Caracas, Venezuela, encapsulates a rich trove of demographic information, behavioral habits, and historical medical records for each of these individuals.

Cervical cancer, a global health concern claiming the lives of approximately 300,000 women worldwide, underscores the urgency of continued research and proactive healthcare measures. Encouragingly, due to advances in medical screening, the cervical cancer mortality rate has witnessed a commendable 74% reduction from 1955 to 1992.

A pivotal revelation in cervical cancer studies has been the identification of High sexual activity Human papilloma virus as a primary contributing factor. This insidious virus significantly heightens the risk of cervical cancer. Additionally, certain lifestyle choices such as the use of oral contraceptives, multiparity, and smoking amplify the susceptibility, particularly among women harboring HPV. Notably, individuals with weakened immune systems, often associated with conditions like HIV, face an elevated risk of HPV infection.

1.1 OVERVIEW

Identify the most relevant features that could contribute to predicting cervical cancer risk. This may involve statistical analysis, domain knowledge, or feature selection algorithms. Additionally, create new features or transform existing ones to improve the predictive power of the model. Choose appropriate machine learning algorithms for the task of predicting cervical cancer risk. Commonly used algorithms include logistic regression, decision trees, random forests, support vector machines, and gradient boosting methods like XGBoost. The selection may depend on the size and complexity of the dataset, as well as the

desired interpretability of the model.

1.1.1 DOMAIN INTRODUCTION

Cervical cancer is a significant global health concern, being one of the leading causes of cancer-related mortality among women. Early detection and accurate diagnosis are critical in reducing mortality rates and improving patient outcomes. Traditional methods of screening and diagnosis, such as Pap smears and HPV testing, have limitations in terms of sensitivity, specificity, and accessibility. Recent advancements in machine learning (ML) offer promising avenues to enhance the detection, diagnosis, and management of cervical cancer through more precise and efficient analysis of medical data.

1.1.2 NPYTHON

Python is the most popular programming language for machine learning due to its extensive libraries and frameworks. Libraries like scikit-learn, TensorFlow, Keras and PyTorch provide tools for data preprocessing, model development, and deployment. Python's readability and versatility make it a preferred choice for data scientists and researchers in the agriculture domain.

1.1.3 JUPYTER NOTEBOOK

Jupyter Notebook is a web-based application used to create and share interactive notebook documents, which can contain live code, text, data visualization, video, and other computational outputs. Created by Project Jupyter, the application is open-source and supports the use of over 40 programming languages, including Python, R and Scala.

Jupyter Notebook showcases real-time code results and imagery and can execute cells in any order. This makes it a useful tool for quick code experimentation, designing code presentations or facilitating data science workflows.

1.2 OBJECTIVE

The objective of a cervical cancer prediction project is to develop a predictive model that can accurately assess the risk of cervical cancer in individuals. The goal is to create a tool that can aid in early detection and prevention of cervical cancer. Thereby improving patient outcomes through timely intervention, treatment, and education about preventive measures such as vaccination and regular screening.

CHAPTER 2
LITERATURE SURVEY

[1] Title: Prediction of Cervical Cancer Using Machine Learning Techniques

Author: Singh J, Sharma S

The authors conducted a survey-based study on cervical cancer detection, including performance analysis to determine the accuracy of various distinctive types of architecture in an artificial neural network (ANN), where the ANN was used for identifying cancerous, normal, and abnormal cells.

[2] Title: Cervical cancer prediction through different screening methods using data mining

Author: Alam T.M, Khan A, Iqbal A, Abdul W, Mushtaq M

Five different machine learning algorithms are used by authors, including random forest, KNN, C5.0, SVM, and RPart. After finishing the training and evaluating the performance of all the classifiers (C5.0, RF, RPART, SVM, and KNN), the best options in terms of accuracy were investigated, showing values of 88%, and 88%. Machine learning (ML) algorithms such as decision tree, random forest, and logistic regression were used in conjunction with the voting model.

[3] Title: Women's knowledge and attitudes towards cervical cancer prevention

Author: Mukama T, Ndejjo R, Musabyimana A, Halage A, Musoke D

Cervical cancer was detected using a dataset containing four target parameters (biopsy, cytology, Schiller, and Hinselmann), as well as 32 risk factors, collected from the University of California (UCI). Machine learning (ML) algorithms were applied, including the decision tree and decision jungle approaches. The study observed that the decision tree algorithm showed a higher value. In another study using the Microsoft Azure ML tool, an appropriate data mining technique was developed from the boosted decision tree, decision forest, and decision jungle algorithms to detect cervical cancer.

[4] Title: Study of adaboost and gradient boosting algorithms for predictive analytics

Author: Bahad P, Saxena P

Bahad and Saxena presented a survey-based study on cervical cancer prevention from the perspective of women in Bug, IRI, and Mayuge in Eastern Uganda, using a questionnaire to collect data from 900 women aged 25 to 49 years. After measuring and scoring the women's knowledge and statements about cervical cancer treatment, the data was analyzed using Stata 12.0 software. After doing bivariate and multivariate analysis, the authors discovered that 794 women, or roughly 88.2%, had heard of the condition. A majority of 557 women (70.2%) acquired their information from the radio, while a minority of 120 women (15.1%) got their information from health care organizations.

[5] Title: Using machine learning for predicting cervical cancer

Author: Weegar R, Sundström K

The authors analyzed various machine learning approaches used from 2006 to 2017 to diagnose cervical cancer. In this research, a comparison was made using existing relevant works based on cervical cancer medical data, to determine the benefits and drawbacks of different approaches. Most studies had used unbalanced medical image datasets. The survey also mentioned employing deep learning to predict cervical cancer. Furthermore, the goal was to see how well the Cox proportional hazard regression model and the deep learning neural network model predicted survival in cervical cancer patients. A dataset from the University of California, Irvine, was used in the study, which included age, number of pregnancies, contraceptive use, smoking habits, and chronological records of sexually transmitted infections (STDs). The study's essential purpose was to use Hinslemann screening methods to predict cervical cancer. With 10-fold validation, a data mining strategy was used with the boosted decision tree, decision forest, and decision jungle approaches.

CHAPTER 3
FEASABILITY STUDY

Feasibility study is carried out when there is a complex problem or opportunity. It is considered as the primary investigation which emphasizes on “Look before You Loop” approach to any project. A Feasibility study is undertaken to determine the possibility of either improving the existing system or developing a completely new system. We are going to develop the new system which is feasible as our application is very user friendly and easy to understand.

3.1.1 ECONOMIC FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the organization can pour into the work and development of the system is limited. The expenditures must be justified. Thus, the developed system as Well within the budget and this was achieved because most of the technologies used are freely.

3.1.2 TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes for the implementing this system.

3.1.3 OPERATIONAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also, able to make some constructive criticism, as he is the final user ofthe system.

3.1.4 EXISTING SYSTEM

To assess the efficacy of the proposed approach, a comparative analysis of its performance is conducted against state-of-the-art models that specifically focus on cervical cancer detection. This evaluation involves considering a selection of recent studies from the literature, which serve as benchmarks for comparison. The reasons behind the superior performance of the proposed method compared to existing approaches lie in two factors: handling missing values and ensemble voting classifier. The unique combination of techniques, addressing missing values, ensemble learning, and class imbalance handling, are the key factors contributing to the observed improvements in accuracy. Unlike some of the previous methods that may not explicitly address the issue of missing values, this study incorporated a KNN imputation technique coupled with SMOTE up-sampled features. Furthermore, the proposed method employs a stacked ensemble voting classifier that integrates the predictions of three individual classifiers. This ensemble approach often proves beneficial by reducing overfitting, leveraging the strengths of multiple classifiers, and providing more robust predictions.

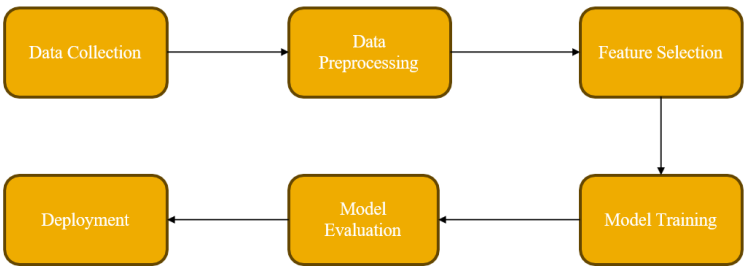

Figure 3.1 EXISTING SYSTEM

3.1.5 PROPOSED SYSTEM

The study utilized a dataset obtained from Kaggle, a reputable source of publicly available datasets. To address missing values and improve the performance of learning

models, preprocessing steps were conducted. The KNN imputer was employed to handle missing values. Subsequently, the data was split into a 70:30 ratio, with 70% allocated for model training and 30% for testing.

For cervical cancer detection, the proposed system utilized an ensemble approach called XGB + RF + ETC. Ensemble models are powerful techniques that combine the predictions of multiple models to enhance accuracy and robustness. Each model in the ensemble has its own strengths and weaknesses, and their combination leads to improved overall performance. The proposed approach for cervical cancer detection combines three popular algorithms: XGB, RF, etc. The workflow diagram of the proposed approach is depicted in Figure 3.1.

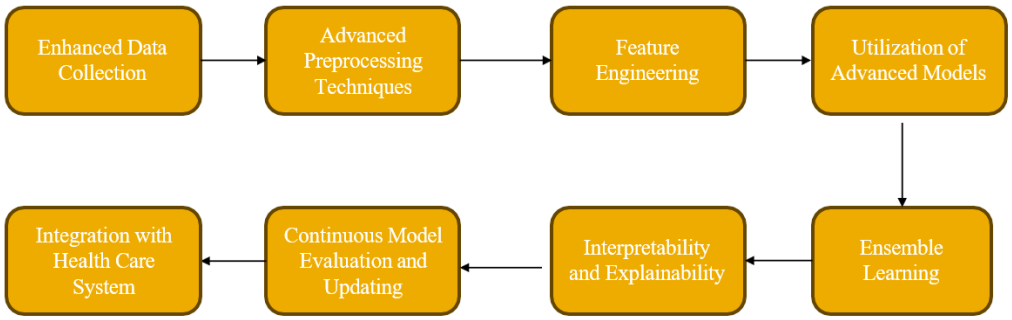

Figure 3.2 Proposed System

3.2 ALGORITHMS USED

3.2.1 LOGISTIC REGRESSION ALGORITHM

Enables learning enhance model performance by using multiple learners. RF is also a kind of enable learning. Following the RF bagging method reduces the chances of results being affected by outliers. This works well for both categorical and continuous data. Datasets do not need to be scaled, and the higher the number of learners, the more computational resources are required for complex models. In this algorithm, the decision is made by voting. Such an algorithm is called ensemble learning. Random forests are made up

of many trees or shrubs. Just as there are many trees in the forest, random forests also have many decision trees. The decision that most trees make is considered the final decision.

3.2.2 SUPPORT VECTOR MACHINE ALGORITHM

The support vector machine algorithm can be used for classification and regression problems. However, SVMs are quite popular for relatively complex types of small or medium classification datasets. In this algorithm, data points are separated by a hyperplane, and the kernel determines what the hyperplane will look like. If we plot multiple variables in a normal scatter plot, in many cases, that plot cannot separate two or more data classes. The kernel of an SVM is a significant element, which can convert lower-dimensional data into higher-dimensional space, and thus differentiate between types.

3.2.3 K-NEAREST NEIGHBORS ALGORITHM

KNN is a non-parametric algorithm that does not assume anything about the distribution of the data at its core. However, a new data point is classified using KNN by looking at its k nearest neighbors in the training set and being assigned to the class that is most prevalent among them. Further, the user selects the value of k, which establishes how many neighbors should be considered when producing a prediction. KNN can be used for regression problems by taking the average or median of the k nearest neighbors, as well as for binary and multi-class classification tasks which makes it beneficial for being simple to use and requiring no training time. However, when working with sizable datasets, it might be computationally expensive and is regarded as a lazy learning algorithm because it does not create a model from the training set of data. Instead, it only maintains the training data and generates predictions based on how far the new data point is from the training set's previous data points.

3.2.4 ADAPTIVE BOOSTING ALGORITHM

The adaptive boosting technique creates a powerful learner by combining the knowledge of several weak learners. In this scenario, every single weak learner utilizes the exact same input, often known as a training set. Every initial input or piece of training data

is given the same amount of importance. The responsibility for correcting the incorrect predictions made by the first weak learner is passed on to the next weak learner, who is given greater weight on the predictions made by the first weak learner and is turned over to the next weak learner. As a result, the errors that the second weak learner made in its predictions are passed on to the following weak learner in the same fashion, but with increased weight. The same process is continued until the number of inaccurate forecasts is reduced to a manageable level. In the end, a powerful learner is developed via the combined efforts. In this way, the amount of inaccuracy in the forecast is reduced.

3.2.5 DECISION TREE ALGORITHM

Both classification and regression problems can be solved with the classification and regression tree or CART algorithm, which is also called the DT. The DT looks a lot like the branches of a tree, which is why the word ‘tree’ is included in its name. The decision tree starts from the ‘root node’ just as the tree starts from the root. From the root node, the branches of this tree spread through different decision conditions; such nodes are called decision nodes (and called leaf nodes after making a final decision).

A decision tree is a type of supervised learning algorithm that is commonly used in machine learning to model and predict outcomes based on input data. It is a tree-like structure where each internal node tests on attribute, each branch corresponds to attribute value and each leaf node represents the final decision or prediction.

CHAPTER 4

PROJECT METHODOLOGY

The proposed research methodology is classified into several segments: research dataset, data preprocessing, predictive model selection (PMS), and training method. Figure 1 depicts an architectural diagram of the proposed research; by looking at Figure 1, it can be clearly observed that the architectural diagram has been separated into four phases, because the model presented in this research performs some essential tasks in each stage. Details on research data collection are described in the Research Dataset section. The Data Preprocessing section mentions how to remove noise from the dataset and make it useful for feeding in machine learning. The type of predictive model selected to predict cervical cancer in this research is shown in the PMS portion. The requisites for model training are shown in the Training Methods section.

The training data will be fed to the system at the beginning of the model training. Then, ML algorithms are adopted. After that, model input data and new input data are applied to the scheme to train the architecture properly. Finally, prediction is performed on the newly accumulated data.

4.2 MODULE DESCRIPTION

4.2.1 META DATA

The UCI repository contributed to the dataset "Cervical Cancer Risk Factors for Biopsy". The collection contains information about 858 people's activities, demographics, and medical history. Multiple missing values occur in this dataset for hospital patients because of several patients declining to answer questions due to privacy concerns. The collection has 858 instances, each with 32 properties. The dataset includes 32 variables and the histories of 858 female patients. The dataset includes 32 variables and the histories of 858 female patients, including factors such as age, IUD, smokes, STDs, and so on.

4.2.2 DATA PREPROCESSING

- Data preprocessing is divided into three sections, which are as follows: data cleaning, data transformation, and data reduction.

- Data preprocessing is critical since it directly impacts project success. Data impurity occurs when attributes or attribute values contain noise or outliers, and redundant or missing data [30]. We have removed the missing values and outliers from this dataset.

- The data transformation stage is kept in place to change the data into suitable forms for the mining process. This research combines normalization, attribute selection, discretization, and concept hierarchy generation. When dealing with a huge amount of data, analysis becomes more difficult when the data dimension is large.

- The data reduction approach is employed in this research to overcome this. It seeks to improve storage efficiency, while lowering the cost of data storage and processing. We have applied the dimension reduction technique because it is another useful technique that can be used to mitigate overfitting in machine learning models. For that, we have applied the principal component analysis (PCA) technique.

4.2.3 PREDICTIVE MODEL SELECTION

Several machine learning classification algorithms have been used in the PMS, namely support vector machine (SVM), decision tree classifier (DTC), random forest (RF), logistic regression (LR), gradient boosting (GB), XGBoost, adaptive boosting (AB), and K-nearest neighbor (KNN). This section has highlighted some of the algorithms that have achieved a satisfactory level of accuracy on the adopted research dataset.

CHAPTER 5
RESULT AND DISCUSSION

Machine Learning algorithm is trained using a training data set to create a model. When new input data is introduced to the ML algorithm, it makes a prediction based on the model. The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine Learning algorithm is deployed. If the accuracy is not acceptable, the Machine Learning algorithm is trained again and again with an augmented training data set.

The Machine Learning process involves building a Predictive model that can be used to find a solution for a Problem Statement. To understand the Machine Learning process let us assume that you have been given a problem that needs to be solved by using Machine Learning.

Based on the findings of this research, it can be stated that the objectives of this paper have been achieved. Its research methodology was enriched with a set of algorithms including decision tree (DT), logistic regression (LR), support vector machine (SVM), K-nearest neighbors (KNN), adaptive boosting, gradient boosting, random forest (RF), and XGBoost. The research has reached a satisfactory result for both predictions and classification. This investigation also observed that the DT and RF algorithms were used in conjunction with the Microsoft Azure machine learning (ML) method to achieve a proper data mining technique for predicting cervical cancer.

CHAPTER 6
CONCLUSION AND FUTURE SCOPE

In conclusion, Early detection increases the likelihood of successful treatment in the pre-cancer and cancer stages. Being aware of any signs and symptoms of cervical cancer can also aid in avoiding diagnostic delays. This research has focused on cervical cancer using conventional machine learning (ML) principles and several traditional machine learning algorithms, such as decision tree (DT), logistic regression (LR), support vector machine (SVM), and K-nearest neighbors (KNN). In terms of cervical cancer prediction, the highest classification has been achieved with the random forest (RF), decision tree (DT), adaptive boosting, and gradient boosting algorithms. The results of these algorithms are applied to identify the most relevant predictors. We have received satisfactory accuracy compared to the support vector machine algorithm. The findings of this study revealed that the SVM model could be used to find the most important predictors. As the number of essential predictors for analysis decreases, the computational cost of the proposed model decreases. The disease can be predicated more accurately with the use of machine learning. Furthermore, boosting patients' personal health and socio-cultural status can lead to cervical cancer prevention.

APPENDIX

IMPORTING OF LIBRARIES

```
import numpy as np
import pandas as pd
import matplotlib. pyplot as
plt
import seaborn as sns
```

IMPORT DATASET

```
data = pd.
read_csv("C:/Users/gk143/OneDrive/Desktop/kag_risk_factors_cervical_cancer.csv")
print (data. Shape)
data.head()
```

SPLITTING DATASETS FOR TRAINING AND TESTING

```
# determining the null values in each column

data = data.replace("?", np.nan)
data = data.convert_objects(convert_numeric = True)

data.isnull().sum()

data.info()

# determining the null values in each column

data = data.replace("?", np.nan)
data = data.convert_objects(convert_numeric = True)

data.isnull().sum()

# 0 means not cancer affected and 1 means cancer affected cell

data['Biopsy'].value_counts()
```

```
# correlation plot

f, ax = plt.subplots(figsize = (10, 8))

corr = data.corr()
sns.heatmap(corr, mask = np.zeros_like(corr, dtype = np.bool),
            cmap = sns.diverging_palette(10, 10, as_cmap = True), square = True, ax = ax)
# Biopsy vs no. of sexual partners

#categorical to categorical
fig, (ax1,ax2) = plt.subplots(2, 1, figsize = (15, 8))
sns.countplot(x = 'Number of sexual partners', data = data, ax=ax1)
sns.barplot(x = 'Number of sexual partners', y = 'Biopsy', data = data, ax=ax2)

#continuous to categorical
facet = sns.FacetGrid(data, hue='Biopsy',aspect=4)
facet.map(sns.kdeplot,'Number of sexual partners',shade= True)
facet.set(xlim=(0, data['Number of sexual partners'].max()))
facet.add_legend()
```

INDEX

```
<!DOCTYPE html>
<html lang="en">

<head>
  <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
  <meta name="viewport" content="width=device-width, initial-scale=1, maximum-
scale=1.0" />
  <title>Cervical Cancer Prediction</title>

  <!-- CSS  -->
  <link href="https://fonts.googleapis.com/icon?family=Material+Icons" rel="stylesheet">
  <link href="../static/css/materialize.css" type="text/css" rel="stylesheet"
media="screen,projection" />
</head>

<body>
  <nav class="purple darken-4 accent-3" role="navigation">
    <div class="nav-wrapper container"><img
       src="C:\Users\kalas\PycharmProjects\MSC projects\A-Prediction-Project-in-Python-
Flask-main\logo">
```

```
    <ul class="right hide-on-med-and-down">
      <li><img src="C:\Users\kalas\PycharmProjects\MSC projects\A-Prediction-Project-in-Python-Flask-main\cclogo">
      </li>
    </ul>

    <ul id="nav-mobile" class="sidenav">
      <li><a href="#">Navbar Link</a></li>
    </ul>
    <a href="#" data-target="nav-mobile" class="sidenav-trigger"><i class="material-icons">menu</i></a>
   </div>
  </nav>


  <div class="section no-pad-bot" id="index-banner">
   <div class="container">
    <br><br>
    <h1 class="header center blue-text">Cervical Cancer Prediction</h1>
    <div class="row center">
     <h5 class="header col s12 light black-text">Using Machine Learning Technique!
      <br>
     </h5>
    </div>

    <div class="row">
     <form action='/predict' method="post" class="col s12">
      <div class="row">
       <div class="input-field col s4">
        <label for="first_name" style="font-size: 20px"><b>Age</b></label>
        <br>
        <input placeholder="Age" name="age" id="age" type="text" class="validate">
       </div>

       <div class="input-field col s4">
        <label for="_name" style="font-size: 20px"><b>Marital_status</b></label>
        <br>
        <input id="_name" name="marital_status" placeholder="1 - Single/2 - Married/3 - Divorced/4 - Widowed"
          type="text" class="validate">
       </div>
       <div class="input-field col s4">
        <label for="_name" style="font-size: 20px"><b>Educational_status</b></label>
        <br>
```

```
            <input id="_name" name="educational_status"
              placeholder="1 - Ilitrate/2 - Can Read and Write/3 - Elementary/4 - HIgh
School/5 - Tertiary School"
              type="text" class="validate">
            </div>

            <div class="input-field col s4">
              <label for="last_name" style="font-size:
20px"><b>Number_of_Birth</b></label>
              <br>
              <input id="number_of_Birth" name="number_of_Birth"
placeholder="Number_of_Birth" type="text"
              class="validate">
            </div>
            <div class="input-field col s4">
              <label for="last_name" style="font-size: 20px"><b>History_of_STI Self or
Partner</b></label>
              <br>
              <input id="_name" name="history_of_STI" placeholder="1 - yes/2 - no/3 - UN"
type="text" class="validate">
            </div>
            <div class="input-field col s4">
              <label for="smoker" style="font-size: 20px"><b>VIA result</b></label>
              <br>
              <input id="VIA_result" name="VIA result" placeholder="1 - VIA positive/0 -
VIA negative" type="text"
              class="validate">
            </div>
            <div class="input-field col s4">

              <label for="last_name" style="font-size: 20px"><b>Occupation</b></label>
              <br>
              <input id="target_population_category" name="target_population_category"
              placeholder="1 - female commercial sex worker/2 - long distance drivers/3 -
mobile or daily laborers/4 - prisoners/5 - general population/6 - other  MARPS/ 7- other"
              type="text" class="validate">

            </div>

            <div class="input-field col s4">
              <label for="smoker" style="font-size: 20px"><b>Hiv Positive Linked with
ART</b></label>
              <br>
```

```
          <input id="hiv_positive_linked_wit_art" name="hiv_positive_linked_wit_art"
placeholder="0 - Yes/1 - No"
            type="text" class="validate">
          </div>

          <div class="input-field col s4">
            <label for="region" style="font-size: 20px"><b>Screened With Via</b></label>
            <br>
            <input id="screened_with_via" name="screened_with_via" placeholder="1 -
VIA/2 - HPV DNA and VIA"
            type="text" class="validate">
          </div>
        </div>

        <div class="row center">

          <button type="submit" class="btn-large waves-effect waves-light lamber lighten-
1"><strong>Predict</strong>
            <h4 style="color:Tomato;">cancer(1)</h4>
            <h4 style="color:DodgerBlue;">non cancer(0)</h4>
          </button>
        </div>
      </form>
    </div>

  </div>
 </div>
 <div>
  <div>
    <div class="dark-text">
      <br>
      <h4>{{pred}}</h4><br>
    </div>
    <br><br>
  </div>
 </div>

 <footer class="page-footer light-blue lighten-1">
  <div class="container">
    <div class="row">

      <h5 class="white-text">Objective of The Project</h5>
      <p class="grey-text text-lighten-4">The general objective of this project is to
develop a predictive model for
```

```
        cervical cancer by using ensemble machine-learning techniques.</p>
      </div>
    </div>
  </footer>


  <!--  Scripts-->
  <script src="https://code.jquery.com/jquery-2.1.1.min.js"></script>
  <script src="../static/js/materialize.js"></script>
  <script src="js/init.js"></script>

</body>

</html>
```

APP.PY MODULE

```
from flask import Flask, request, url_for, redirect, render_template
import pickle

# import xgboost

import numpy as np

app = Flask(__name__, template_folder="./templates", static_folder="./static")

Pkl_Filename = "model5.pkl"
with open(Pkl_Filename, "rb") as f:
    model = pickle.load(f)


@app.route("/")
def hello_world():
    return render_template("home.html")


@app.route("/predict", methods=["POST", "GET"])
def predict():
    features = [int(x) for x in request.form.values()]

    print(features)
    final = np.array(features).reshape((1, 9))
```

```
    print(final)
    pred = model.predict(final)[0]
    print(pred)

    if pred == 1:
        res_val = "** cervical cancer **"
    else:
        res_val = " no cervical cancer "

    return render_template("op.html", pred="Patient has {}".format(res_val))


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

5.1 Screenshot of interface module

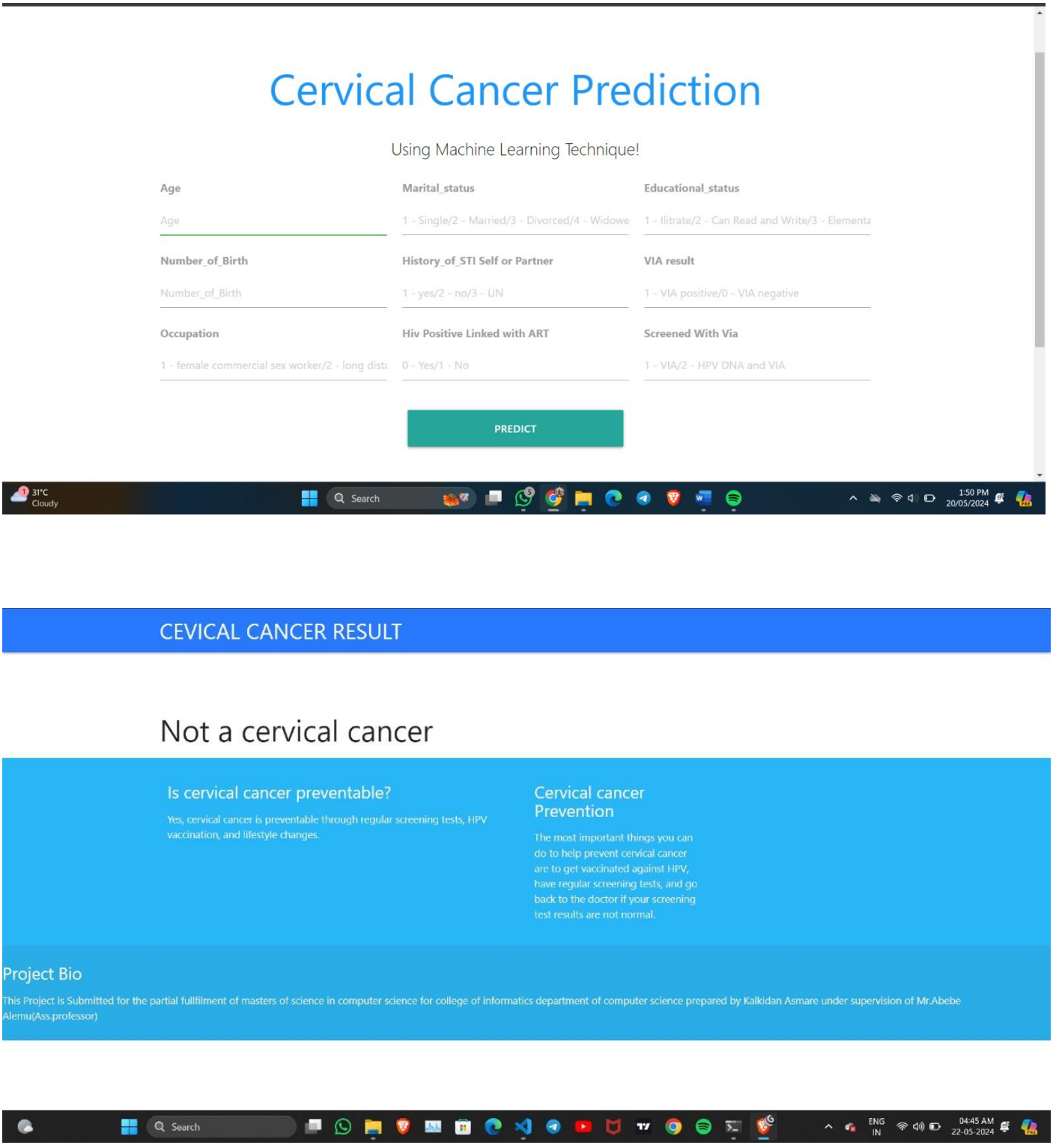

Figure 5.1 Screenshot of interface module

REFERENCES

[1] Martin, C.M.; Astbury, K.; McEvoy, L.; Toole, S.; Sheils, O.; Leary, J.J. Gene expression profiling in cervical cancer: Identification of novel markers for disease diagnosis and therapy. In Inflammation and Cancer; Springer: Berlin, Germany, 2009; Volume 511, pp. 333–359.

[2] Purnami, S.; Khasanah, P.; Sumartini, S.; Chosuvivatwong, V.; Sriplung, H. Cervical cancer survival prediction using hybrid of SMOTE, CART and smooth support vector machine. AIP Conf. Proc. 2016, 1723, 030017.

[3] Yang, X.; Da, M.; Zhang, W.; Qi, Q.; Zhang, C.; Han, S. Role of lactobacillus in cervical cancer. Cancer Manag. Res. 2018, 10, 1219–1229.

[4] Ghoneim, A.; Muhammad, G.; Hossain, M.S. Cervical cancer classification using convolutional neural networks and extreme learning machines. Future Gener. Comput. Syst. 2020, 102, 643–649.

[5] Rehman, O.; Zhuang, H.; Muhamed Ali, A.; Ibrahim, A.; Li, Z. Validation of miRNAs as breast cancer biomarkers with a machine learning approach. Cancers 2019, 11, 431

[6] Kamilaris, A., Prenafeta-Boldú, F. X., & Parascandolo, G. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70-90.

[7] Osuwa, A.; Öztoprak, H. Importance of Continuous Improvement of Machine Learning Algorithms From A Health Care Management and Management Information Systems Perspective. In Proceedings of the 2021 International Conference on Engineering and Emerging Technologies (ICEET), Istanbul, Turkey, 29–30 September 2021; pp. 1–5

[8] Prabhpreet, K.; Gurvinder, S.; Parminder, K. Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification. Inform. Med. Unlocked 2019, 16, 100151.

[9] Devi, M.A.; Ravi, S.; Vaishnavi, J.; Punitha, S. Classification of cervical cancer using artificial neural networks. Procedia Comput. Sci. 2016, 89, 465–472.

[10] Issah, F.; Maree, J.E.; Mwinituo, P.P. Expressions of cervical cancer-related signs and symptoms. Eur. J. Oncol. Nurs. 2011, 15, 67–72. [CrossRef].