

**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 Karunaratne

Signature of the supervisor:

Date:

ACKNOWLEDGEMENT

I would like to show my gratitude to Dr. Buddhika Karunaratne 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

Declaration.....	i
Acknowledgement.....	ii
Abstract.....	iii
Table of the contents.....	iv
List of figures.....	vi
List of tables.....	viii
List of Abbreviations.....	ix
1. Introduction.....	1
1.1. Introduction.....	1
1.2. Background.....	1
1.3. Research problem.....	2
1.4. Objectives.....	2
1.5. Motivation.....	3
1.6. Thesis Structure.....	4
2. Literature Review.....	6
2.1. Overview.....	6
2.2. Heart disease-related research studies.....	6
2.3. Diabetes disease-related research studies.....	7
2.4. Cancer disease-related research studies.....	11
3. Identification of relevant studies.....	15
3.1. Introduction.....	15
3.2. Machine Learning.....	15
4. Methodology.....	20
4.1. Introduction.....	20
4.2. Problem.....	20
4.3. Data Collection.....	21
4.4. Analysis data.....	21
4.5. Feature Engineering.....	28
5. Evaluation.....	32
5.1. Introduction.....	30
5.2. Basics.....	30
5.3. Evaluation.....	32

5.3.1 Feature selection.....	33
5.3.2 Hyper Parameter tuning.....	34
5.3.3 Ensembling methods.....	35
5.4. Implementation.....	36
6. Conclusion.....	40
6.1 Introduction.....	40
6.2. Overview of research.....	40
6.3. Limitations.....	41
6.4. Future works.....	41
6.5. Summary.....	41
REFERENCES.....	42
Appendix A – Web application with positive result.....	48
Appendix B - Web application with negative result.....	48
Appendix C – models.py File.....	48
Appendix D – app.py File.....	49

LIST OF FIGURES

Figure	Description
Figure 2. 1	Flow chart of heart disease predicting system
Figure 2.2	Experimental working of heart disease predictor
Figure 2. 3	Flow chart of diabetes predicting a system
Figure 2. 4	Comparison of four algorithms
Figure 2. 5	Proposed web application to predict diabetes
Figure 2. 6	Comparison of the model with other models
Figure 2. 7	Methodology for predicting the model
Figure 2.8	Comparative analysis
Figure 2.9	The web application Overview
Figure 2.10	Hospital Management system
Figure 2.11	Cancer Predicting system
Figure 2.12	Cancer Prediction - Risk level
Figure 2.13	Accuracies for Cervical Cancer Prediction
Figure 4.1	Stage Rates by Percentage
Figure 4.2	Missing value percentage of the dataset
Figure 4.3	Biopsy Percentage
Figure 4.4	Age analysis
Figure 4.5	Smoking habit analysis
Figure 4.6	Sexual habit analysis
Figure 4.7	Birth Control analysis
Figure 4.8	STD analysis
Figure 4.9	Age and Sexual habits vs Biopsy
Figure 4.10	Smoking and Sexual habits vs Biopsy
Figure 4.11	Age and Smoking habits vs Biopsy
Figure 4.12	Age and Birth Control Attributes vs Biopsy
Figure 4.13	Data with outliers
Figure 4.14	Data without outliers
Figure 5.1	Sigmoid Function
Figure 5.2	Decision Tree structure

Figure 5.3	Five model evaluation results
Figure 5.4	Target Variable imbalance
Figure 5.5	Results after sampling
Figure 5.6	Feature Selection Techniques
Figure 5.7	Feature selection results
Figure 5.8	Hyper parameter tuning results
Figure 5.9	Final results
Figure 5.10	Overview of the web application

LIST OF TABLES

Table	Description
Table 5.1	Comparison table with benchmark models

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