VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

Department of Computer Engineering

Project Report on

Predictive Analytics in Pregnancy Database

In partial fulfillment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai Academic Year 2017-2018

Submitted by

Deepanshu Garg (D17 - C , Roll no - 17)
Deepesh Garg (D17 - C , Roll no - 18)
Pratik Nichat (D17 - C , Roll no - 41)
Shruti Sindhi (D17 - C , Roll no - 57)

Project Mentor

Ms. Rupali Hande
Ms. Sunita Suralkar

(2017-18)

VIVEKANAND EDUCATION SOCIETY’S INSTITUTE OF TECHNOLOGY

Department of Computer Engineering

Certificate

This is to certify that ***Deepanshu Garg, Deepesh Garg, Pratik Nichat & Shruti Sindhi*** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on “***PREDICTIVE ANALYTICS IN PREGNANCY DATABASE***” as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor ***Ms. Rupali Hande, Ms. Sunita Suralkar and Dr. Raj Hingorani (Industry Mentor)*** in the year 2017-2018.

Programme Outcomes	Grade
PO1,PO2,PO3,PO4,PO5,PO6,PO7, PO8, PO9, PO10, PO11, PO12 PSO1, PSO2	

Date:

Project Guide:

Project Report Approval

For B. E (Computer Engineering)

This project report entitled ***Predictive Analytics in Pregnancy Database*** by ***Deepanshu Garg, Deepesh Garg, Pratik Nichat & Shruti Sindhi*** is approved for the degree of ***Bachelor of Engineering (B.E.) Degree in Computer Engineering.***

Internal Examiner

External Examiner

Head of the Department

Principal

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

--	--
(Signature)	(Signature)
--	--
(Name of student and Roll No.)	(Name of student and Roll No.)
--	--
(Signature)	(Signature)
--	--
(Name of student and Roll No.)	(Name of student and Roll No.)

Date:

जवाहरलाल नेहरू पोर्ट ट्रस्ट
JAWAHARLAL NEHRU PORT TRUST

जनेपन्यास
JNPT

ISO 9001 : 2008
ISO 14001:2004
ISO 27001:2013
OHSAS 18001: 2007

पत्तन कार्यालय : प्रशासन भवन, शेवा, नवी मुंबई - 400 707. Port Office : Administration Bldg., Sheva, Navi Mumbai - 400 707.
मुख्य सतर्कता अधिकारी Chief Vigilance Officer-(022) 2724 2292; मुख्य प्रबंधक (प्रशासन) एवं सचिव Chief Manager (Admn.) & Secy-(022) 2724 2233;
मुख्य प्रबंधक (यातायात) Chief Manager (Traffic)-(022) 2724 2377; मुख्य प्रबंधक (यां.एवं वि.अ.) Chief Manager (M&EE)-(022) 2724 2218;
मुख्य प्रबंधक (वित्त) Chief Manager (Fin)-(022) 2724 2241; मुख्य प्रबंधक (प. यो. वि.) Chief Manager (PP&D)-(022) 2724 2326;
उप-संरक्षक Dy. Conservator (022) 2724 2301; हार्बर मास्टर Harbour Master - (022) 2724 4173.
Website : www.jnport.gov.in E-mail : info@jnport.gov.in

No.JNP/A/TC/2018 Date:- 03/04/2018

CERTIFICATE

This is to certify that **Mr. Deepanshu Garg** a Computer Engineering student of Vivekanand Education Society's Institute of Technology, Chembur, Mumbai has undergone Industrial Exposure and imbibed technical medical guidance in **Jawaharlal Nehru Port Trust Hospital** from **22.08.2017** to **10.03.2018** and has satisfactorily completed the **BE project on Predictive Analytics in Pregnancy Database in JNPT Hospital** under the guidance of Dr. Raj Hingorani (Industry mentor).

(S.R.Aphale)
Asstt.Manager(HRD)
सहायक प्रबंधक (मा.सं.वि)
Assistant Manager (HR.)

मुंबई कार्यालय : 1107, रहेजा सेंटर, 214, फ्री प्रेस जर्नल मार्ग, नरिमन पॉईंट, मुंबई - 400 021. दूरभाष : 91-22-6616 5600 फैक्स : 91-22-6743 1116.
MUMBAI OFFICE : 1107, RAHEJA CENTRE, 214, FPJ MARG, NARIMAN POINT, MUMBAI - 400 021. TEL.: 91-22-6616 5600 FAX : 91-22-6743 1116.
हमेशा हिन्दी में पत्राचार करके देश का गौरव बढ़ाएँ ।

No.JNP/A/TC/2018

Date:- 03/04/2018

CERTIFICATE

This is to certify that **Mr. Deepesh Garg** a Computer Engineering student of Vivekanand Education Society's Institute of Technology, Chembur, Mumbai has undergone Industrial Exposure and imbibed technical medical guidance in **Jawaharlal Nehru Port Trust Hospital** from **22.08.2017** to **10.03.2018** and has satisfactorily completed the **BE project on Predictive Analytics in Pregnancy Database in JNPT Hospital** under the guidance of Dr. Raj Hingorani (Industry mentor).

(S.R.Aphale)
Asstt.Manager(HRD)
सहायक प्रबंधक (मा.सं.वि)
Assistant Manager (HRD)

जवाहरलाल नेहरू पोर्ट ट्रस्ट
JAWAHARLAL NEHRU PORT TRUST
पत्तन कार्यालय : प्रशासन भवन, शेवा, नवी मुंबई - 400 707. Port Office : Administration Bldg., Sheva, Navi Mumbai - 400 707.
मुख्य सतर्कता अधिकारी Chief Vigilance Officer-(022) 2724 2292; मुख्य प्रबंधक (प्रशासन) एवं सचिव Chief Manager (Admn.) & Secy-(022) 2724 2233;
मुख्य प्रबंधक (यातायात) Chief Manager (Traffic)-(022) 2724 2377; मुख्य प्रबंधक (यां.एवं वि.अ.) Chief Manager (M&EE)-(022) 2724 2218;
मुख्य प्रबंधक (वित्त) Chief Manager (Fin)-(022) 2724 2241; मुख्य प्रबंधक (प. यो. वि.) Chief Manager (PP&D)-(022) 2724 2326;
उप-संरक्षक Dy. Conservator (022) 2724 2301; हार्बर मास्टर Harbour Master - (022) 2724 4173.
Website : www.jnport.gov.in E-mail : info@jnport.gov.in

जनेपन्यास
JNPT

ISO 9001 : 2008
ISO 14001:2004
ISO 27001:2013
OHSAS 18001: 2007

No.JNP/A/TC/2018

Date:- 03/04/2018

CERTIFICATE

This is to certify that **Mr. Pratik Nichat** a Computer Engineering student of Vivekanand Education Society's Institute of Technology, Chembur, Mumbai has undergone Industrial Exposure and imbibed technical medical guidance in **Jawaharlal Nehru Port Trust Hospital** from **22.08.2017** to **10.03.2018** and has satisfactorily completed the **BE project on Predictive Analytics in Pregnancy Database in JNPT Hospital** under the guidance of Dr. Raj Hingorani (Industry mentor).

(S.R.Aphale)
Asstt.Manager(HRD)

सहायक प्रबंधक (मा. .)
Assistant Manager (HRD)

मुंबई कार्यालय : 1107, रहेजा सेंटर, 214, फ्री प्रेस जर्नल मार्ग, नरिमन पॉईंट, मुंबई - 400 021. दूरभाष : 91-22-6616 5600 फैक्स : 91-22-6743 1116.
MUMBAI OFFICE : 1107, RAHEJA CENTRE, 214, FPJ MARG, NARIMAN POINT, MUMBAI - 400 021. TEL.: 91-22-6616 5600 FAX : 91-22-6743 1116.
हमेशा हिन्दी में पत्राचार करके देश का गौरव बढ़ाएँ ।

जवाहरलाल नेहरू पोर्ट ट्रस्ट
JAWAHARLAL NEHRU PORT TRUST

जनेपन्यास
JNPT

ISO 9001 : 2008
ISO 14001:2004
ISO 27001:2013
OHSAS 18001: 2007

पत्तन कार्यालय : प्रशासन भवन, शेवा, नवी मुंबई - 400 707. Port Office : Administration Bldg., Sheva, Navi Mumbai - 400 707.
मुख्य सतर्कता अधिकारी Chief Vigilance Officer-(022) 2724 2292; मुख्य प्रबंधक (प्रशासन) एवं सचिव Chief Manager (Admn.) & Secy-(022) 2724 2233;
मुख्य प्रबंधक (यातायात) Chief Manager (Traffic)-(022) 2724 2377; मुख्य प्रबंधक (यां.एवं वि.अ.) Chief Manager (M&EE)-(022) 2724 2218;
मुख्य प्रबंधक (वित्त) Chief Manager (Fin)-(022) 2724 2241; मुख्य प्रबंधक (प. यो. वि.) Chief Manager (PP&D)-(022) 2724 2326;
उप-संरक्षक Dy. Conservator (022) 2724 2301; हार्बर मास्टर Harbour Master - (022) 2724 4173.
Website : www.jnport.gov.in E-mail : info@jnport.gov.in

No.JNP/A/TC/2018 Date:- 03/04/2018

CERTIFICATE

This is to certify that **Ms. Shruti Sindhi** a Computer Engineering student of Vivekanand Education Society's Institute of Technology, Chembur, Mumbai has undergone Industrial Exposure and imbibed technical medical guidance in **Jawaharlal Nehru Port Trust Hospital** from **22.08.2017** to **10.03.2018** and has satisfactorily completed the **BE project on Predictive Analytics in Pregnancy Database in JNPT Hospital** under the guidance of Dr. Raj Hingorani (Industry mentor).

(S.R.Aphale)
Asstt.Manager(HRD)

सहायक प्रबंधक (मा.सं.वि)
Assistant Manager (HRD)

मुंबई कार्यालय : 1107, रहेजा सेंटर, 214, फ्री प्रेस जर्नल मार्ग, नरिमन पॉईंट, मुंबई - 400 021. दूरभाष : 91-22-6616 5600 फैक्स : 91-22-6743 1116.
MUMBAI OFFICE : 1107, RAHEJA CENTRE, 214, FPJ MARG, NARIMAN POINT, MUMBAI - 400 021. TEL.: 91-22-6616 5600 FAX : 91-22-6743 1116.

हमेशा हिन्दी में पत्राचार करके देश का गौरव बढ़ाएँ ।

जवाहरलाल नेहरू पोर्ट ट्रस्ट
JAWAHARLAL NEHRU PORT TRUST

जनेप न्यास
JNPT

पत्तन कार्यालय : प्रशासन भवन, शेवा, नवी मुंबई - 400 707. Port Office : Administration Bldg., Sheva, Navi Mumbai - 400 707.
मुख्य सतर्कता अधिकारी Chief Vigilance Officer-(022) 2724 2292; मुख्य प्रबंधक (प्रशासन) एवं सचिव Chief Manager (Admn.) & Secy-(022) 2724 2233;
मुख्य प्रबंधक (यातायात) Chief Manager (Traffic)-(022) 2724 2377; मुख्य प्रबंधक (यां.एवं वि.अ.) Chief Manager (M&EE)-(022) 2724 2213;
मुख्य प्रबंधक (वित्त) Chief Manager (Fin)-(022) 2724 2241; मुख्य प्रबंधक (प. यो. वि.) Chief Manager (PP&D)-(022) 2724 2326;
उप-संरक्षक Dy. Conservator (022) 2724 2301; हार्बर मास्टर Harbour Master - (022) 2724 4173.
Website : www.jnport.gov.in E-mail : info@jnport.gov.in

JN/TC/A/2017

Date: 21.08.2017

To,
Vivekanand Education Society's Institute of Technology,
Chembur,
Mumbai.

Sub: Permission for B.E. Project.

Sir

Kindly refer to your letter on the above subject, it is to inform you that this organisation is mitting to provide technical medical guidance for B.E. Project on the topic Predictive Analytics in Pregnancy Database on following terms and condition.The said Project should be completed by March 2018.

Name of the Students allowed are as detailed below.

01. Mr. Pratik Nichat,
02. Ms. Shruti Sindhi,
03. Mr. Deepanshu Garg,
04. Mr. Deepesh Garg,

Terms & Conditions:

1. Student will be treated as unpaid trainee. No lodging, boarding and transport facilities will be provided to trainee student.
2. An undertaking will be obtained from trainee student for maintaining discipline during the training period.
3. JNPT will not bear any liability during the training period due to any accident or personal injury.
4. JNPT will not be liable for any payment or extending any facilities to the student.
5. The project work completed by the above student is used only for academic purpose and is never divulged in public without proper permission. A copy of project work prepared by the student will be submitted to JNPT & thereafter certificate will be issued.
6. In case any stundent nominated by your institute needs training in operational area or to visit Dock/Port area of the Port then such student shall arrage personal protactives such as helmet, safety shoes etc. of their own and no such protectives will be provided by the port in the absence of such proteatives,entry to these areas will be denied to student as a Part of safety measure which they should note.
7. The concerned student may please be informed to report to the Sr.Dy.Cheif Medical Officer JNPT. for further instructions please.

Thanking you,

Yours faithfully,

Aphale
(S.R.Aphale)
Asstt.Manager (HRD)

Sr.Dy.Chief. (Medical Officer)

मुंबई कार्यालय : 1107, रहेजा सेंटर, 214, फ्री प्रेस जर्नल मार्ग, नरिमन पॉईंट, मुंबई - 400 021. दूरभाष : 91-22-6616 5600 फैक्स : 91-22-6743 1116.
MUMBAI OFFICE : 1107, RAHEJA CENTRE, 214, FPJ MARG, NARIMAN POINT, MUMBAI - 400 021. TEL.: 91-22-6616 5600 FAX : 91-22-6743 1116.
हमेशा हिन्दी में पत्राचार करके देश का गौरव बढ़ाएँ ।

ACKNOWLEDGEMENT

We are thankful to JNPT Hospital, JNPT training center and our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to **Mr. S.R. Aphale, Dr. Raj Hingorani** (Industry Mentor), Assistant Professors **Ms. Rupali Hande and Ms. Sunita Suralkar** (Project Guide) for their kind help and valuable advice during the development of project synopsis and for their guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr.(Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair ,** for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement at several times.

Computer Engineering Department

COURSE OUTCOMES FOR B.E PROJECT

Course Outcome	Description of the Course Outcome
CO 1	Able to apply the relevant engineering concepts, knowledge and skills towards the project.
CO2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solution for the problem.
CO 4	Able to interpret the data and datasets to be utilized.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.
CO 10	Able to apply the project domain knowledge to sharpen one's competency.
CO 11	Able to develop professional, presentational, balanced and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment and platforms for creating innovative solutions for the project.

ABSTRACT of the project

Pregnancy and the expectation of the birth of the child bring joy to the entire families. Not only the parents to be, but the grandparents and siblings of the expectant couple all wish to share the joy of this miraculous event. The thought of soon having a toddler in the family is a wondrous thing. Above all this, pregnancy is one of the most delicate stages in every woman's life requiring increased medical care and attention. It induces a variety of abnormalities which, may lead to flinty situations if unnoticed or neglected. Such abnormalities and complication can turn severe and initiate instances leading to death of both mother and infant. They should be protected from such complications. With immense growth of big data in biomedical and healthcare communities, precise analysis of medical data and early disease detection is viable. But the analysis accuracy is reduced when the quality of medical data is incomplete. Legion of algorithms are known but some of them can be greatly used in the biomedical field for varied analysis and predictions. The modified prediction models are experimented over real-life hospital data collected. The project works on effective performance in predicting risk levels during pregnancy from the collected, standardized and transformed data efficiently using supervised machine learning classification algorithms like decision trees or random forest

Index

Chapter 1

Introduction

1.1 INTRODUCTION

Pregnancy and the expectation of the birth of the child bring joy to the entire families. The thought of soon having a toddler in the family is a wondrous thing. We propose a system in which the healthcare providers should be to assess the high risk of conditions based on different estimable parameters like age, weight gain, blood pressure, blood group, haemoglobin, blood sugar, urine exam, HIV tests, thyroid profile etc. and necessary preventive action may be taken or help of the higher centres can be obtained. Also the treatment that should be ideally prescribed under the circumstances should be predicted along with the suggestion of mode of delivery that should be considered for delivering the baby. This newly developed system will help maintain up to date record of the patient and help us take better and fast decision on the available patient history. It will also help us analyse the areas where improvement is required. In dearth of medical expertise in emergency situations, this system can help in taking some initial decision. Due to compulsory registering process women delivering at private hospitals and homes won't remain out of the government's information net. If the women move in near places for delivery, the system will make sure the health care providers at the new place have complete and adequate information about the expectant mother. Information mostly reaches the district headquarters or the government only when it audits maternal death or premature death of the infant. If system is developed properly it will help sort all such problems and in near future can also be connected to various other systems to provide automated issuing of Birth and Death certificate. It is possible to generate different statistical data like birth rate, maternal mortality rate, and fetal mortality rate for different national programmes and also keep track of medication and vaccination of patients, high risk pregnancies and reporting incidence of normal deliveries, forceps or C-Sections conducted. The data in future can also be used in predicting the population in near future.

1.2 MOTIVATION OF THE PROJECT

A News article remarking," The personalized tiding that used to stay inside the four walls won't stay personalized anymore", forge the interest to know more about it. The article discussed that the women has to break the news of her pregnancy as soon as she is aware about it to the state government (Tamil Nadu). The concept of the government to make the pregnant women to register themselves with the health department to ensure that the pregnancy and childbirth happen safely developed a notion to apply our knowledge and skills in this sector.

Though the idea is ambitious or aspiring or one can even say self-seeking. But will take its roots if brain muscles are used appropriately.

1.3 DRAWBACK OF THE EXISTING SYSTEM

- In many of the existing healthcare setups in rural area, there is unavailability of specialist doctors. Hence the antenatal to postnatal facilities are provided by ANM (Auxiliary Nurse Midwifery) or equivalent.
- Early detection of complications are difficult in some situations due to lack of knowledge. Such situations can lead to acute emergencies which can complicate the health of the mother.
- The dearth of medical specialists in emergency situations in rural areas may result in losing the mother or child or both.
- The data collected if any is either incomplete or insufficient to be used for productive purposes to calculate statistical data to be used for various government programmes.

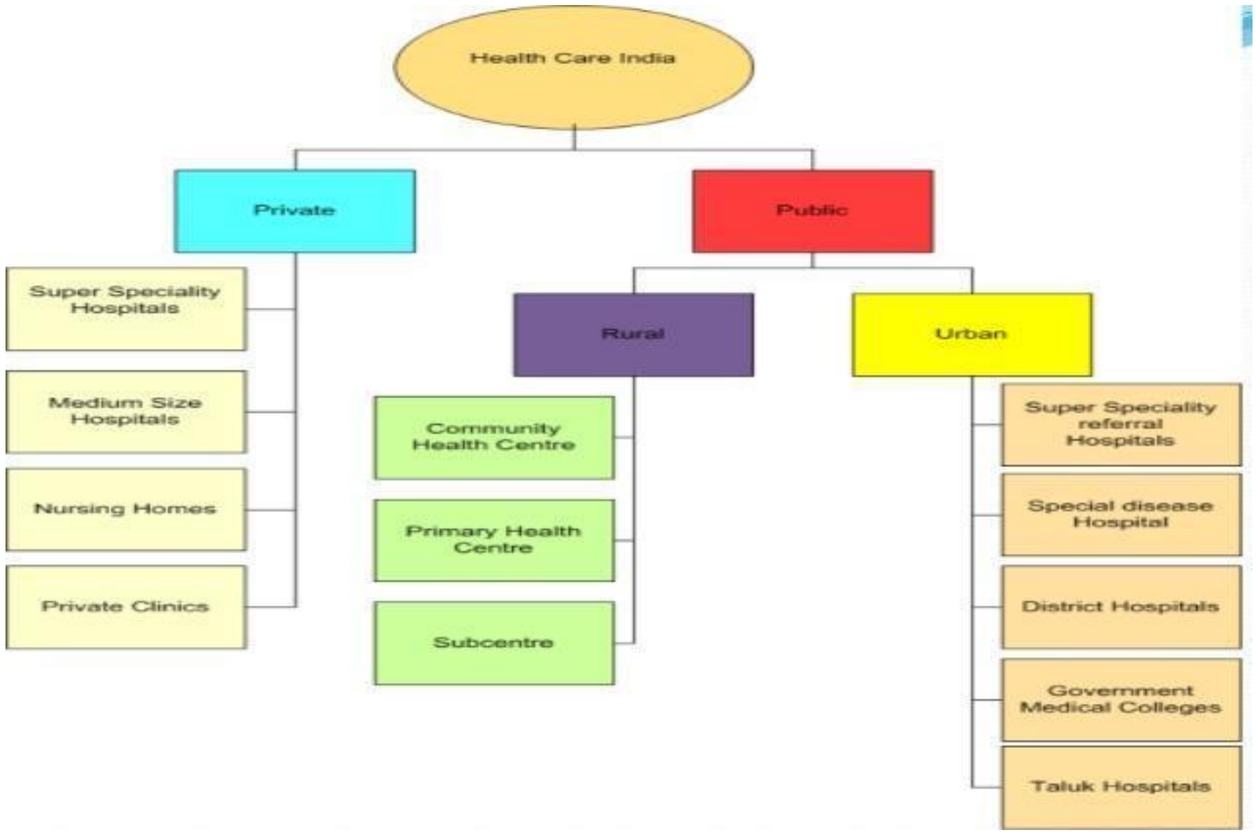


Figure 1.3.1 : Existing healthcare system in India

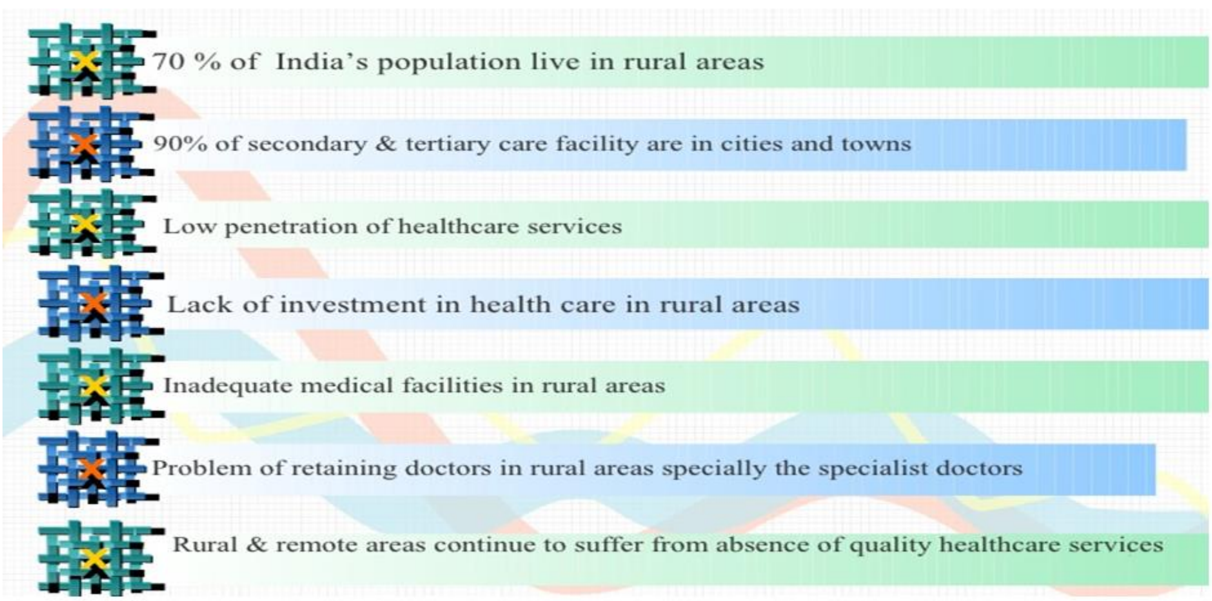


Figure 1.3.2 : Lacuna in the existing system

1.4 PROBLEM DEFINITION

- We propose a web based system in which the healthcare providers (specialised as well as non-specialised) should be to assess the high risk of conditions based on different estimable parameters like age, weight gain, blood pressure, blood group, haemoglobin, blood sugar, urine exam, HIV tests, thyroid profile etc. and necessary preventive action may be taken or help of the higher centres can be obtained.
- Also the treatment that should be ideally prescribed under the circumstances should be predicted along with the suggestion of mode of delivery that should be considered for delivering the baby.
- This newly developed system will help maintain up to date record of the patient and help us take better and fast decision on the available patient history.
- Baby weight predictions can also be made depending on the patient parameters.

1.5 RELEVANCE OF THE PROJECT

Every idea or thought is probed whether it is worthwhile as well as feasible or not. It also takes check into factors like market need, economical feasibility, legal feasibility and many more. This newly developed system will help maintain up to date record of the patient and help us take better and fast decision on the available patient history. It will also help us analyse the areas where improvement is required. In dearth of medical expertise in emergency situations, this system can help in taking some initial decision. Due to compulsory registering process women delivering at private hospitals and homes won't remain out of the government's information net. If the women move in near places for delivery, the system will make sure the health care providers at the new place have complete and adequate information about the expectant mother. Information most reaches the district headquarters or the government only when it audits maternal death or premature death of the infant. If system is developed properly it will help sort all such problems and in near future can also be connected to various other systems to provide automated issuing of Birth and Death certificate. The data can future also be used in predicting the population in near future.

1.6 METHODOLOGY

There are many methods, technique or procedure to achieve a target. But at times the sequence or the upshot use doesn't matter. All that matters is how well they are implemented. The methods and materials used to conduct work are shown in following steps given below:

1. Create pregnancy data set. Following attributes have been collected for predicting risk prediction in patients:

Attributes	**Data type**
age	numeric
trimester	numeric
weight	numeric
obstretic_history (number of pregnancies, number of abortions, previous mode of delivery, contracted pelvis , History of hypertension, History of gestational diabetes)	numeric/string
complaints (Amenorrhea, vomiting, weakness, pain in abdomen, oedema feet, Lethargy, headache, spotting pelvis, leaking pelvis, leg cramps, burning in micturition, pedal and abdominal wall oedema, Fever with cold and cough)	string
temperature	numeric

pulse	numeric
blood_pressure	numeric
blood_group	string
haemoglobin	numeric
blood_sugar	numeric
thyroid_stimulating_hormone	numeric
urine_profile	string
amniotic fluid	string
usg	numeric
Lie of the baby	string
excessive weight gain	string

Table 1.6.1 : Attributes used in Analytics for mode of delivery, complications and treatment

2. Standardize and transform data.

Transform data in required form i.e converting ordinal values of attributes in numeric form and maybe using dimensionality reduction algorithm like pca to get rid of redundant attributes if any.

3. Divide pregnancy dataset into training and testing dataset.

Training dataset is used for construction of the learning model and test dataset for evaluating generated model for determining the accuracy.

4. Apply machine learning classification algorithm like decision trees or k-nearest neighbours on training data set and build (train) the model.

5. Analyse the testing dataset with the trained model and obtain the results and accuracy.

Chapter 2

Literature Survey

2.1 Research Papers

2.1. a. ABSTRACT OF THE RESEARCH PAPER

PAPER 1: A COMPARATIVE STUDY OF ALGORITHMS FOR RISK PREDICTION IN PREGNANCY

Pregnancy is one of the most delicate stages in every woman's life requiring increased medical care and attention. Pregnancy induces a variety of abnormalities which may lead to severe complications if unnoticed or neglected. Pregnancy complications are health problems that are caused due to changes in physiological parameters during the period of gestation. Pregnancy complications can lead to severe maternal illness by women during pregnancy, at delivery and after delivery. The aim of this paper is to predict the present complications in the health of a pregnant woman using two classification algorithms namely C4.5 decision tree classification algorithm and Naive Bayes Classification Algorithm. The selected algorithms are powerful and popular tools used for the tasks of classification and prediction in Data mining. These two algorithms use pregnancy data collected from pregnant women in different stages of pregnancy to predict their present health state and the associated health complications. This study focuses on identifying the best algorithm among the two classification algorithms to predict the health status of each pregnant woman and its associated complication. Applying these classification techniques on pregnancy related data the status of risk in the health of any pregnant women can be identified and maternal and foetal mortality rate can be reduced.

PAPER 2: PREDICTION BASED HEALTH MONITORING OF PREGNANT WOMEN

Gestation or pregnancy a stage where women undergo several physiological changes, sometimes inducing complications turning severe and initiating instances leading to death of both mother and foetus. Pregnant women must thus be protected from complications arising during gestation period. Several classification algorithms are successfully implemented in several fields. Decision Tree Classification Method is one efficient method best suitable for medical diagnosis. A popular algorithm C4.5 Decision Tree classification algorithm is appropriate for classifying the pregnancy data. The algorithm constructs a learning model from the training data and later risks in pregnancy are predicted for unseen pregnancy data. The main aim of this paper is to optimise performance of C4.5 classification algorithm by applying on standardized and appropriate format of data. The paper highlights the effective performance achieved by C4.5 classifier in accurately predicting risk levels during pregnancy from the collected, standardized and transformed data efficiently.

PAPER 3: DISEASE PREDICTION BY MACHINE LEARNING OVER BIG DATA FROM HEALTHCARE COMMUNITIES

With big data growth in biomedical and healthcare communities, accurate analysis of medical data benefits early disease detection, patient care and community services. However, the analysis accuracy is reduced when the quality of medical data is incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. In this paper, we streamline machine learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities. We experiment the modified prediction models over real-life hospital data collected from central China in 2013-2015. To overcome the difficulty of incomplete data, we use a latent factor model to reconstruct the missing data. We experiment on a regional chronic disease of cerebral infarction. We propose a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm.

2.1.b INFERENCE DRAWN FROM PAPER

The learning model generated by C4.5 OR Naïve Bayes or k-nearest algorithm are easy to understand and require no domain experts. It can be inferred from the initial projects that these techniques are efficient, powerful and popular in classifying data and making desired predictions or predicting risk in pregnancy. From the studies one can easily wrap up why there is a lean towards these algorithms. These algorithms are powerful and most importantly efficient than other known algorithms. C4.5 algorithm is used over Naïve Bayes algorithm because it presents better performance and advantages of this algorithm can be pinpointed easily.

Currently, Hadoop technology is taking wind in HealthCare and HealthCare Intelligence. It is well known that changes are taking place continuously. But if they technology is not evolved with time it can be unremitting too. Since the data and the outcomes are dynamic and ever changing in healthcare sector, Hadoop technology is finding its roots. Hadoop is an infrastructure software for storing and processing large dataset. Medical datasets are far more big than what it is imagined or assumed, which at times cannot be stored on a single node(computer), in such case Hadoop is an intelligent option as it has a HDFS (HADOOP DISTRIBUTED FILE SYSTEM). In spite of being aware that Hadoop also works on structured and unstructured data, it is not preferred because the aim is not just storing and processing data but also solving the hard queries instantly. So using python is a better option since hard queries can be resolved in a better manner which using Hadoop is not possible as it is yet not capable of responding to the hard queries fired on traditional database in microseconds or instant of time. Also using Hadoop can create also lots of network traffic because batch processing environment as the processor is taken to the data.

2.2 Interaction with domain expert

We procured guidance from our industry mentor and understood the different risk factors that have a significant impact on the mother’s as well as child’s health. We also constructed the data sets keeping in mind the privacy of the patients. The different combinations of risk attributes that lead to complications were understood.

Chapter 3

Requirements of Proposed System

3.1 Functional Requirements:

1) Predicting high risk patients on the basis of age, weight gain, blood pressure.
2) Investigation based prediction (Blood group, haemoglobin, blood sugar, urine exam, HIV tests and thyroid profile.)

4) Risk of gestational diabetes and gestational hypertension.
5) Risk of pre-term or IUGR (Intra Uterine Growth Retardation) /IUD (Intra Uterine Death) of fetus.
6) Incidences of normal delivery, forceps assisted or Caesarian Sections and associated complications during delivery.
8) Prediction of weight of of new born baby.

3.2 Non-Functional Requirements

1) System should be user-friendly so that even a basic health provider should be able to use it.
2) Confidentiality of data should be maintained.
3) High sensitivity and specificity of prediction.
4) Database as well as the front end should be responsive.

3.3 Constraints

1) Privacy of the patient data needs to be maintained.

2) Proper authentication of the user is required.

3) Internet connection is mandatory for the system to be operable.

4) Basic computer knowledge (Data entry and retrieval) is must for user.

3.4 Hardware & Software Requirements

3.4.1 Technologies to be used

The backend database has been implemented using MongoDB, a NoSQL database.

We selected NoSQL to develop our system as it is able to handle:

- Large volumes of structured, semi-structured, and unstructured data
- Agile sprints, quick iteration, and frequent code pushes
- Object-oriented programming that is easy to use and flexible
- Efficient, scale-out architecture instead of expensive, monolithic architecture

Node.js has been used for server side scripting.

We selected Node.js to use as server side scripting as:

- Node.js is an open source server framework
- Node.js is free
- Node.js runs on various platforms (Windows, Linux, Unix, Mac OS X, etc.)
- Node.js uses JavaScript on the server

Here is how Node.js handles a file request:

Node.js eliminates the waiting, and simply continues with the next request.

Node.js runs single-threaded, non-blocking, asynchronous programming, which is very memory efficient.

Frontend of the system has been developed using HTML and JavaScript.

Chapter 4

Proposed Design

4.1 System Design

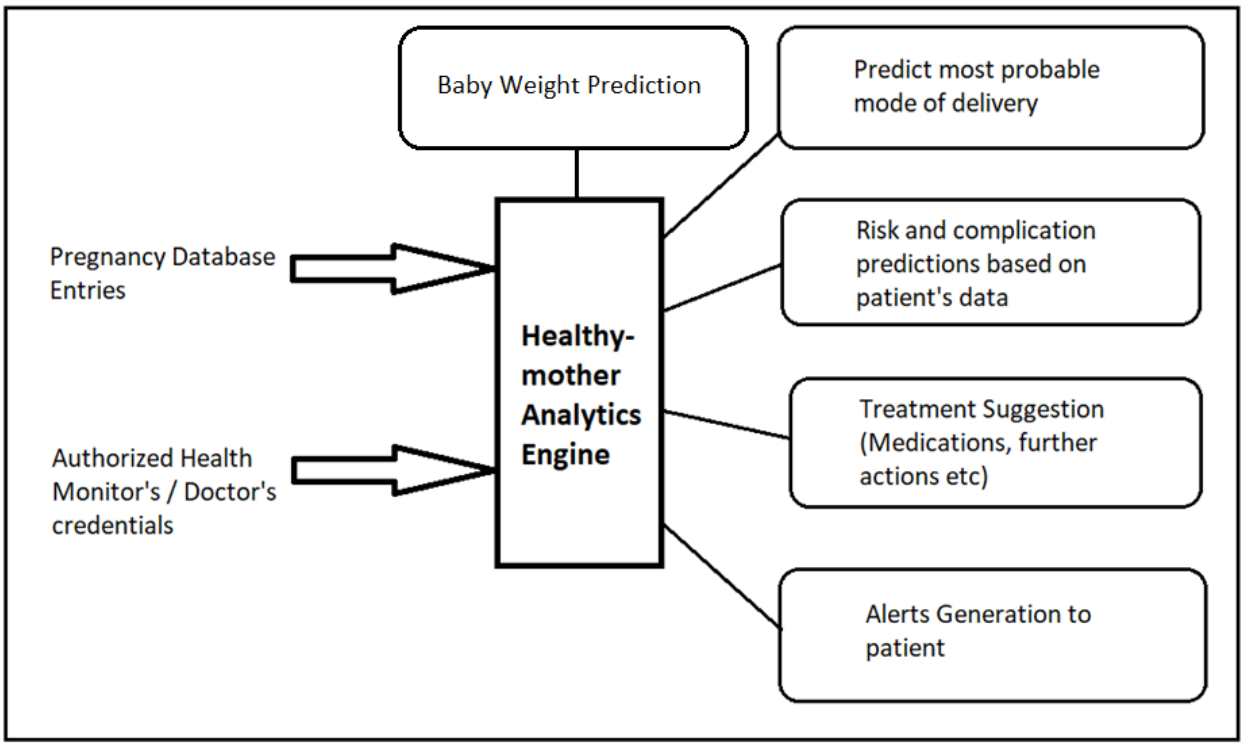


Figure 4.1.1 : System Design

The proposed system will take in pregnant female’s data as input which will be put in by authorised health monitor. The trained data set will also be referred to analyse the patient data and provide risk predictions, postnatal analysis and also provide statistical information to help in census calculation and other national programmes. The system also will provide alerts for appointments along with other predictions as shown.

4.2 Design of the proposed system

4.2.a. Data Flow Diagram

Level 0

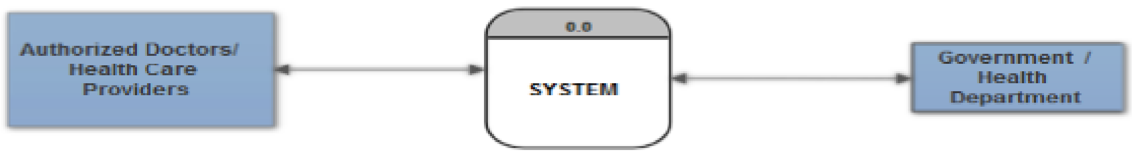


Figure 4.2.a.1 : Level 0 DFD

Level 0 shows a generalised data flow of the system wherein the system is viewed as an overall entity and authorised doctors or health care providers as well as the health dept uses the system.

Level 1

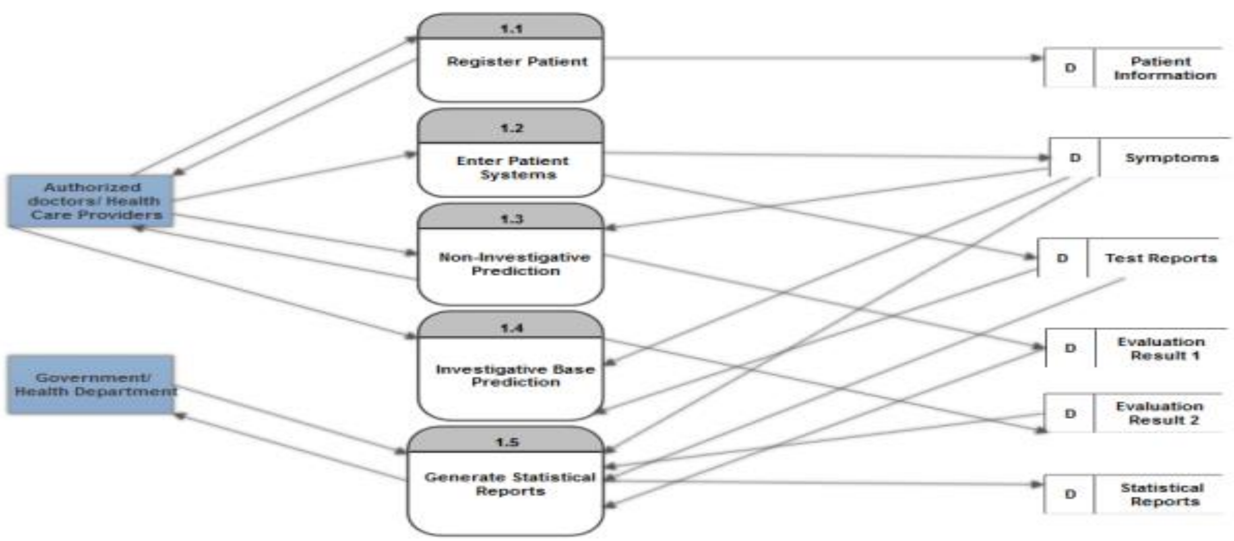


Figure 4.2.a.2 : Level 1 DFD

The level 1 diagram shows the expanded view of the system processes and the necessary databases required to store the data

Level 2

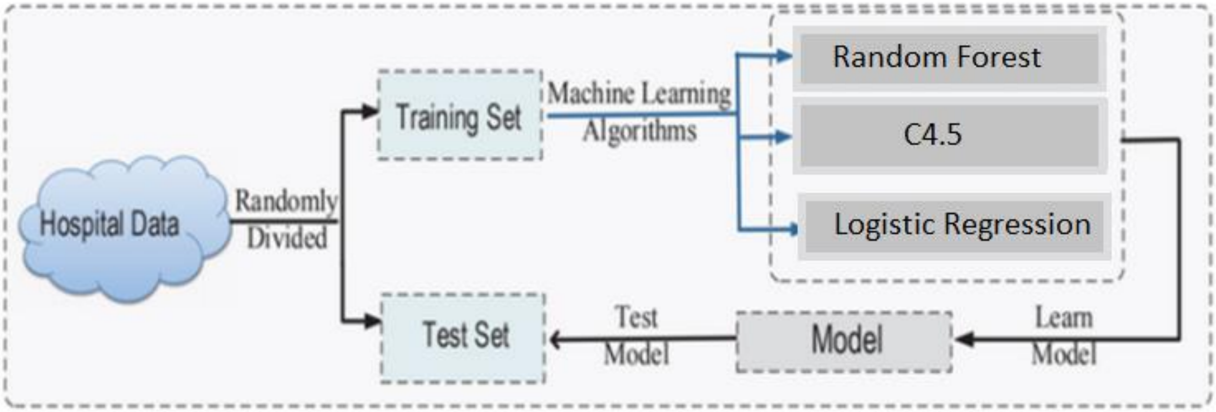


Figure 4.2.a.3 : Level 2 DFD

The level 2 DFD explains how the data is divided in training and testing data sets and the algorithms that we use to analyse the dataset and predict accordingly.

4.2.b. Activity Diagram

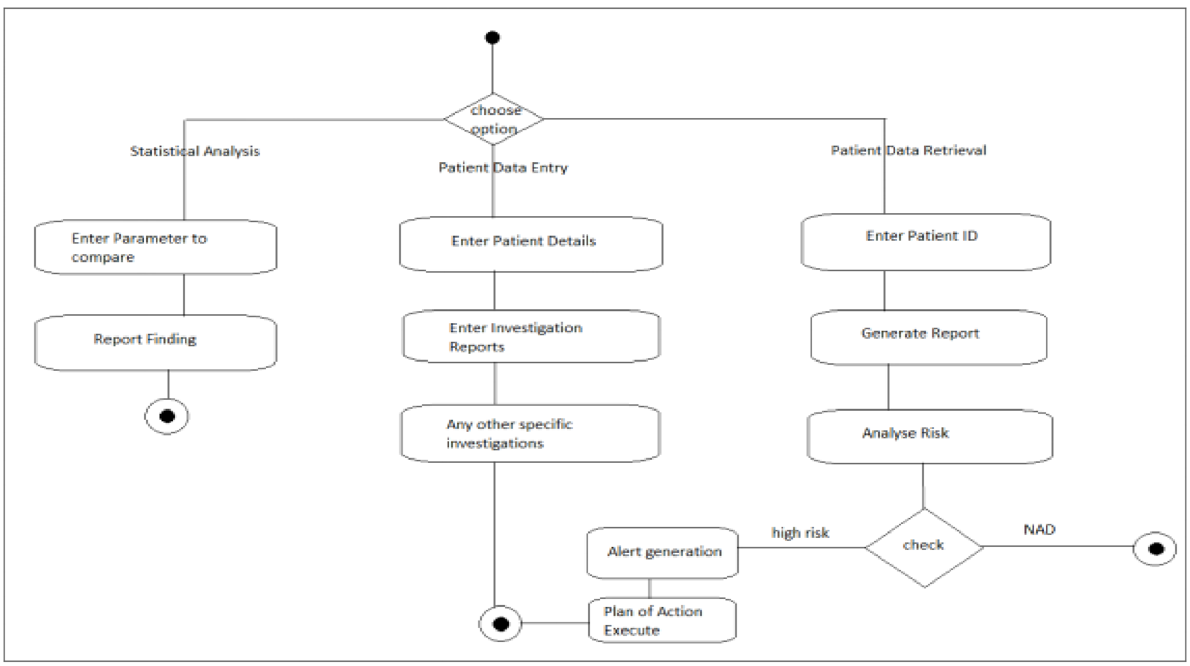


Figure 4.2.b.1 : Activity Diagram

The activity diagram shows the activities performed by the user of the system and it gives options to the user to do statistical analysis, patient data entry or patient data retrieval.

4.2.c. ER Diagram

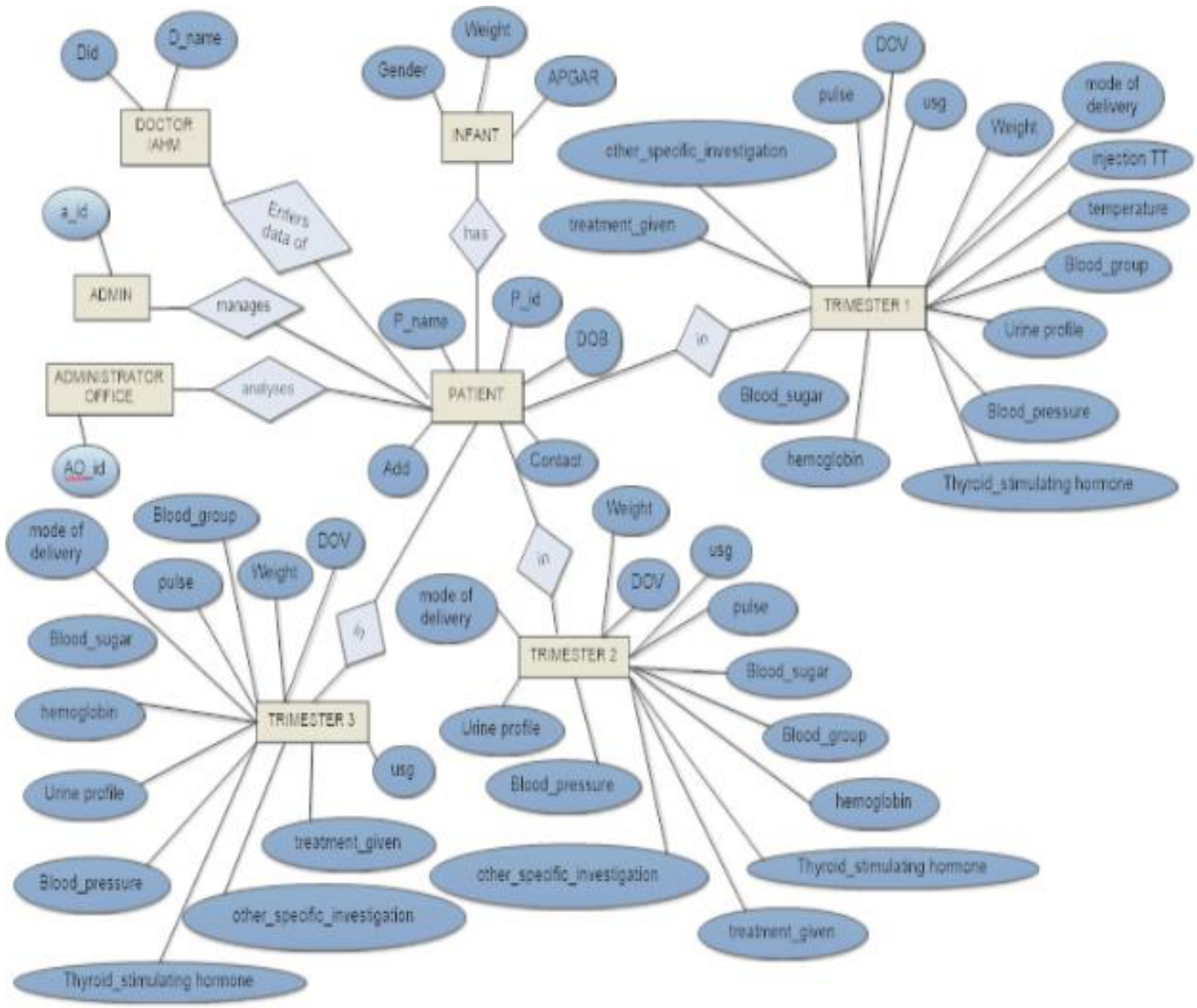


Figure 4.2.c.1 : ER Diagram

We have created the ER diagram showing the most relevant attributes of a patient required for the predictive analysis of risks in pregnancy.

4.3. Project Scheduling and tracking using Gantt chart

Task Name	Start Date	End Date	Status
Requirement Gatheri	08/21/17	09/26/17	Completed
Analysis	09/27/17	10/07/17	Completed
Design	10/08/17	10/31/17	Completed
Implementation	11/01/17	01/31/18	Completed
Testing	02/01/18	02/28/18	Completed
Deployment	03/01/18	03/19/18	Completed

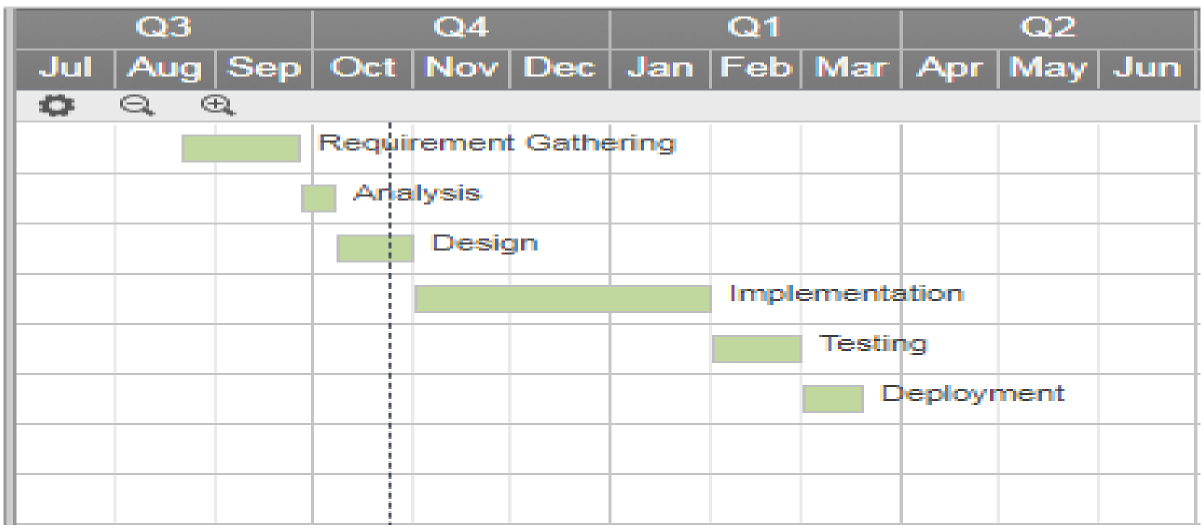


Figure 4.3.1 : Gantt Chart

5. Implementation Details

5.1. Algorithms utilized for the developed system

Supervised Learning

How it works: This algorithm consist of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning: Decision Tree, Random Forest, KNN etc.

Decision tree algorithm

A Decision Tree is an algorithm used for supervised learning problems such as classification . A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

Decision tree algorithm is one of the most powerful and well suited model for medical parameters.It doesn't require tuning of parameters for creation of model, it automatically creates subset of relevant feature needed for classification, as choices of splits are not on absolute magnitudes of attribute values, it is robust to outliers, it can easily handle missing data values. So this algorithm is best suited for solving prediction problems related to medical diagnosis.

How decision trees work:

- The decision tree is structured as a flowchart diagram. It has a single, starting root node, and can contain one or more leaf nodes, and internal nodes.
- Each node of the decision tree represents the test of single feature value.
- Each branch of the decision tree connects two nodes, an originating node, and a

destination node. A branch represents an outcome of the test of the branch's originating node.

- Each leaf node of the decision tree represents a classification. The end goal of the tree is to label your input samples, so each leaf fulfills that goal.

Algorithm :

1) Read the data set

>>X=pd.read_csv('Datasets/pregnancy.data',index_col=0)

2) Split data into training data and testing data

>>X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 7)

3) Create the model

>>dtree = tree.DecisionTreeClassifier(max_depth = 5, random_state = 2)

4) Train the decision tree model

>> dtree.fit(X_train_bag, y_train)

5) Test the accuracy of the model

>> dtree.score(X_test, y_test)

Random forests algorithm:

A single decision tree, tasked to learn a dataset might not be able to perform well due to the outliers, and the breadth and depth complexity of the data.

So instead of relying on a single tree, an advanced version of decision tree called random forests rely on a forest of cleverly grown decision trees. Each tree within the forest is allowed to become highly specialized in a specific area, but still retains some general knowledge about most areas. When a random forest classifier, it is actually each tree in the forest working together to cast votes on what label they think a specific sample should be assigned.

How random forest works:

Training:

Random forests make use of two techniques when training, one occurs at the forest level, and the other at the individual tree level. First, like any supervised classifier, you'll pass in a training set of samples along with "truth" labels when you create an instance of the class. Instead of sharing the entire dataset with each decision tree, the forest performs an operation which is essential a train / test split of the *training* data. Each decision tree in the forest randomly samples from the overall training data set. Through doing so, each tree exist in an independent subspace and the variation between trees is controlled. This technique is known as *tree bagging*, or bootstrap aggregating.

Tree bagging increases the accuracy of decision trees because while an individual decision tree might become hypersensitive to outliers and localized features, once all the results are averaged, the fringe results get blurred out. Therefore using random forests over a single tree decrease the variance of your classification results, without increasing the bias the way KNeighbor does when K is set too high. This technique only works if the individual trees are not correlated. If they were all trained on the same training set, they would only reinforce each other's decision. The bootstrapping or randomization of samples each tree is trained upon takes care of that.

Random forests also use one more trick. In addition to the tree bagging of training samples at the forest level, each individual decision tree further 'feature bags' at each node-branch split. This is helpful because some datasets contain a feature that is much correlated to the target. By selecting a random sampling of features every split, if such a feature were to exist, it wouldn't show up on as many branches of the tree and there would be more diversity of the features examined

Predicting

After fitting the forest, a prediction can be made for unseen samples by using the majority vote label assigned from each tree.

Error Testing

Since each tree within the forest is only trained using a subset of the overall training set, the forest ensemble has the ability to error test itself. It does this by scoring each tree's predictions against that tree's out-of-bag samples. A tree's out of bag samples are those forest training samples that were withheld from a specific tree during training. There's nothing unique about splitting your data between training and testing sets except that you have an independent set of unseen samples to validate the accuracy of your training. Part of the random forest algorithm is the creation of independent sets for the training of each tree, so an overall *out-of-bag error* metric can be calculated for the forest ensemble. This error value is defined as the mean prediction error for each training samples using only those trees that didn't have the sample in their bootstrap.

Algorithm :

1) Read the data set
>>X=pd.read_csv('Datasets/pregnancy.data',index_col=0)
2) Split data into training data and testing data
>>X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 7)
3) Create the model
>>forest = RandomForestClassifier(n_estimators = 30, max_depth = 10, oob_score = True, random_state = 0)
4) Train the decision tree model
>> forest.fit(X_train_bag, y_train)
5) Test the accuracy of the model
>> dtree.score(X_test, y_test)

Logistic Regression:

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. You can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function.

The fundamental equation of generalized linear model is:

$$g(E(y)) = \alpha + \beta x1 + \gamma x2$$

Here, g() is the link function, E(y) is the expectation of target variable and $\alpha + \beta x1 + \gamma x2$ is the linear predictor (α,β,γ to be predicted). The role of link function is to 'link' the expectation of y to linear predictor.

5.2. Evaluation of the developed system

After training the decision tree model on training data consisting of attributes mentioned in section III and testing on testing data the accuracy obtained while predicting mode of delivery in pregnancy is 95.65%, predicting complications in pregnancy is 91.3% and predicting treatment is 90.9%.

And after training the Random forests model on training data consisting of above attributes mentioned in section III and testing on testing data accuracy obtained while predicting mode of delivery in pregnancy is 97.63%, predicting complications in pregnancy is 95.65% and predicting treatment is 94%.

Since the Random forest algorithm performed better, we implemented the system using the random forest algorithm.

6. Testing

The developed system has been extensively tested at industry taking various test cases such as:

1. Case of anemia.
2. Case Pregnancy Induced Hypertension.
3. Case of contracted pelvis.
4. Case of Impending eclampsia.
5. Case of IUGR.
6. Case of threatened abortion.

The system has also been auto tested using test train split module in python in which the dataset is divided and 70% dataset is used for training and 30% is used for testing.

After training the decision tree model on training data consisting of attributes mentioned in section III and testing on testing data the accuracy obtained while predicting mode of delivery in pregnancy is 95.65%, predicting complications in pregnancy is 91.3% and predicting treatment is 90.9%.

And after training the Random forests model on training data consisting of above attributes mentioned in section III and testing on testing data accuracy obtained while predicting mode of delivery in pregnancy is 97.63%, predicting complications in pregnancy is 95.65% and predicting treatment is 94%.

Since the Random forest algorithm performed better, we implemented the system using the random forest algorithm.

7. Result Analysis

7.1. Screenshots of User Interface (UI)

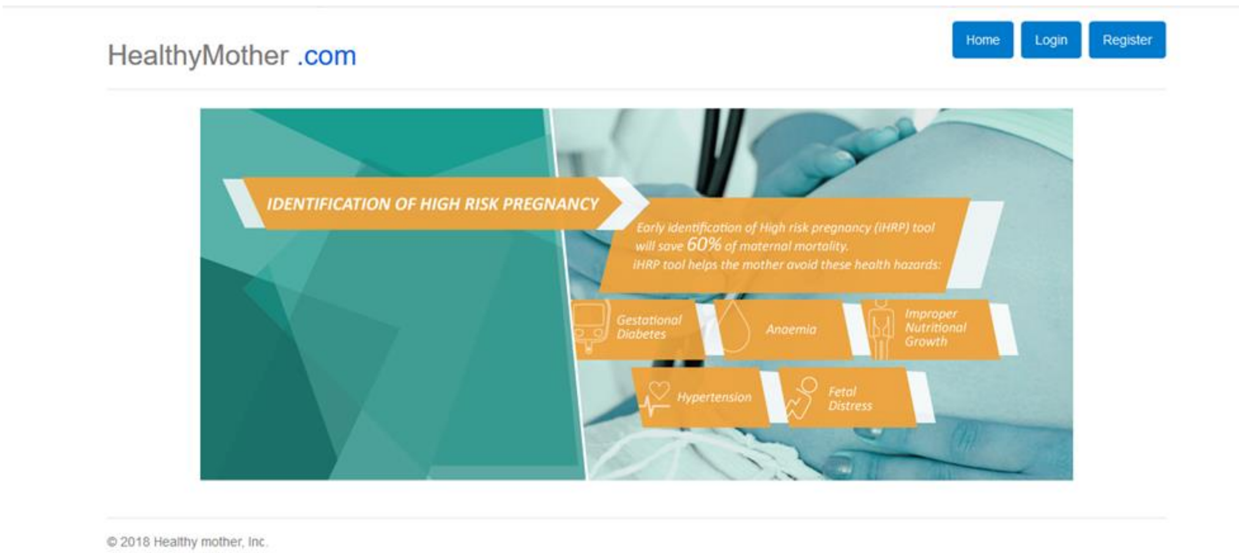


Figure 7.1.1 : Home Page

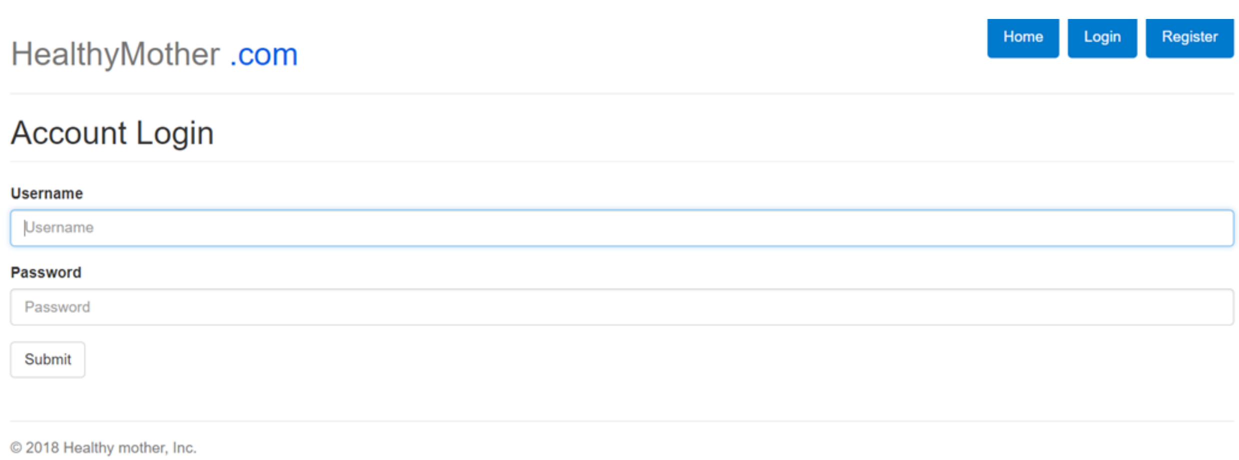


Figure 7.1.2 : Login Page

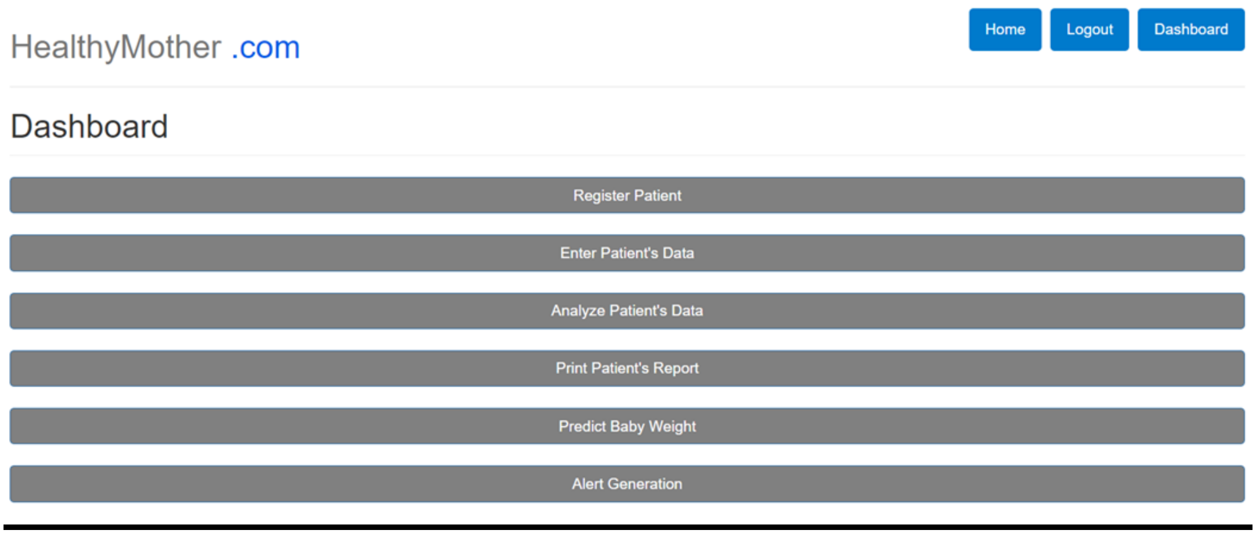


Figure 7.1.3 : Dashboard

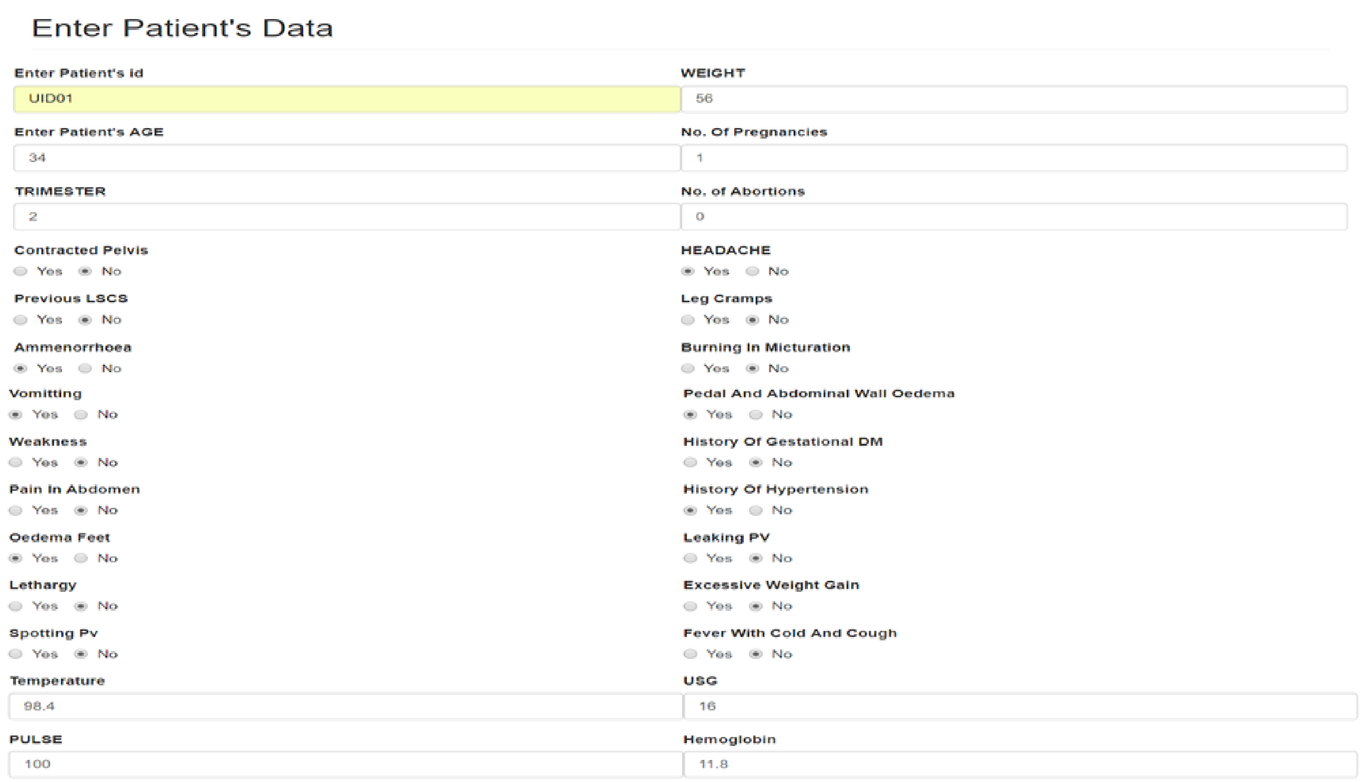


Figure 7.1.4 :Form for input data

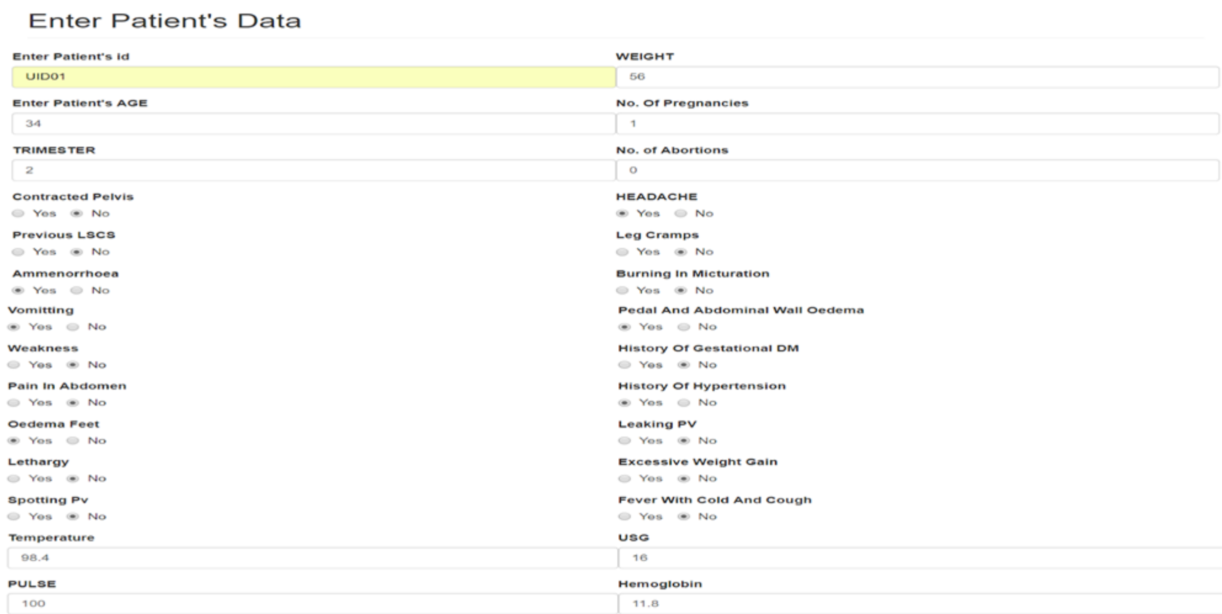


Figure 7.1.5 :Form for input data

7.2. Output Generated

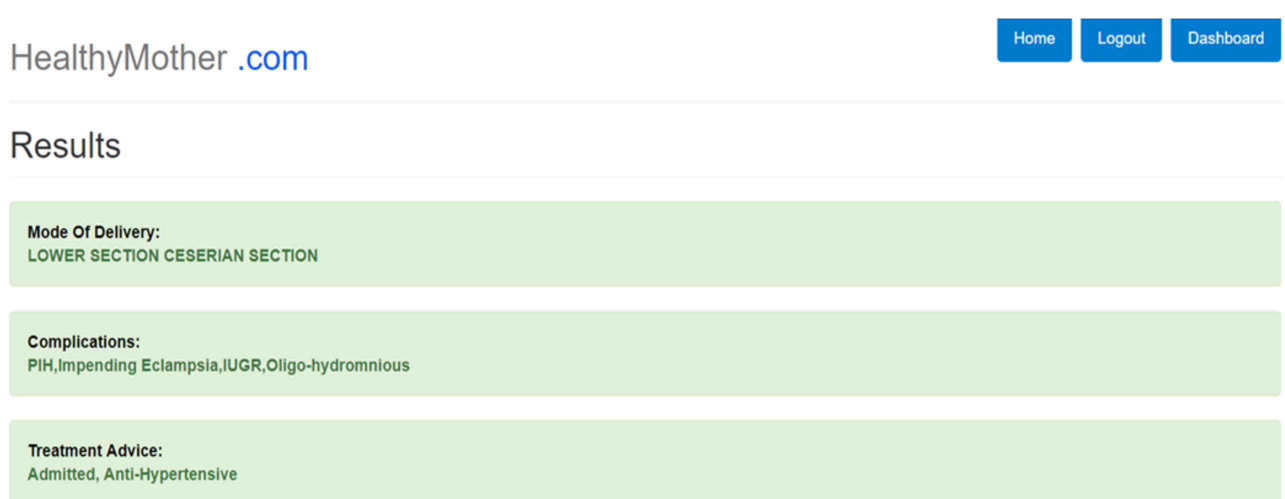


Figure 7.2.1 :Result Page

7.3. Report Generated

Healthy Mother   http://localhost:3000/reports/print

Home (/)   Logout (/users/logout)   Dashboard (/users/dashboard)

HealthyMother .com

Report

Patient Id:	p001
Age:	19
Trimester:	1
Weight:	52
Temperature:	98.3
Pulse:	80
Bp Systolic:	110
Bp Diastolic:	70
Hemoglobin:	10.5
Blood Sugar:	78
Thyroid Profile:	2.04
USG:	8
Predicted Mode of Delivery:	FULL TERM NORMAL DELIVERY
Predicted Complications:	No Complications So Far
Treatment Advised:	Folic Acid
Patient Id:	p001
Age:	19
Trimester:	1
Weight:	52
Temperature:	98.3
Pulse:	80
Bp Systolic:	110
Bp Diastolic:	70
Hemoglobin:	10.5
Blood Sugar:	78
Thyroid Profile:	2.04
USG:	8
Predicted Mode of Delivery:	FULL TERM NORMAL DELIVERY
Predicted Complications:	No Complications So Far
Treatment Advised:	Folic Acid

Print

© 2018 Healthy mother, Inc.

1 of 1   25-02-2018, 20:37

Figure 7.3.1 :Report Generated

8. Conclusion

8.1. Limitations

1. The developed system can’t be used in case of emergency.
2. The system operator should have minimum computer literacy.

8.2. Conclusion

It is perceived that the most of the system target only a particular section of the society. But the proposed system shot both rural and urban sector giving more value to the rural sector by providing alert system for regular appointments in the form of SMS. For better, smooth and well-bred analysis the traditional database is divided into three trimester. A neural network based pregnancy risk prediction system is being used for available structured and unstructured data from hospital,clinics or nursing homes..The learning model generated by k-nearest algorithm, decision tree algorithm or random forest algorithm are easy to understand and require no domain experts.

It can be inferred from the initial projects that these techniques are efficient, powerful and popular in classifying data and making desired predictions or predicting risk in pregnancy. From the studies one can easily wrap up why there is a lean towards these algorithms. These algorithms are powerful and most importantly efficient than other known algorithms. And providing up to 95 % of accuracy.

8.3. Future Scope

1. The system can be used for predictive analytics of the baby health too.
2. Analysis of database to find real time statistics (Census).
3. Implementation of the project can be done on a large scale so as to include patients from all parts of India.

9. References

[1] A Comparative Study of Classification Algorithms for Risk Prediction in Pregnancy Published in: TENCON 2015 - 2015 IEEE Region 10 Conference Electronic ISBN: 978-1- 4799-8641- 5

[2] Prediction based health monitoring in pregnant women Published in: Applied and Theoretical Computing and Communication Technology (iCATccT), 2015 International Conference Electronic ISBN: 978-1- 4673-9223- 5

[3] Disease Prediction by Machine Learning over Big Data from Healthcare Communities Published in: IEEE Access (Volume: 5) Electronic ISSN: 2169-3536

[4] Decision Tree Analysis using Weka, Machine Learning – Project II by Sam Drazin and Matt Montag, University of Miami

[5] C4.5 algorithm and Multivariate Decision Trees by Thales Sehn Korting Image Processing Division, National Institute for Space Research – INPE, S˜ao Jos´e dos Campos – SP, Brazil.

[6] Implementation of Decision Tree Algorithm to Analysis the Performance, Pooja Sharma[1], Asst. Prof. Rupali Bhartiya[2], Department Of Computer Science and Engineering,Shree Vaishnav Institute of Technology & Science, Indore, India[12] , International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 10, December 2012, ISSN (Print) : 2319-5940, ISSN (Online) : 2278-1021

[7] Improved C4.5 Algorithm For Rule Based Classification by Mohammed M Mazid, A B M Shawkat Ali, Kevin S Tickle, School Of Computing Science, Central Queensland University, Australia in Recent Advances in Artificial Intelligence, Knowledge Engineering And Data Bases, ISSN: 1790-5109, ISBN: 978-960-474-154-0

[8]United Nations Maternal Mortality Estimation Interagency Group, consisting of representatives from the World Health Organization (WHO), United Nations Children's Fund (UNICEF), the United Nations Population Fund (UNFPA), United Nations Population Division, the World Bank and Wikipedia.

[9] Predicting Students' Performance Using Id3 And C4.5 Classification Algorithms by Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, Department of Computer Engineering, Fr. C.R.I.T., Navi Mumbai, Maharashtra, India in International Journal

of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013, DOI : 10.5121/ijdkp.2013.3504.

[10] Improved C4.5 Algorithm For Rule Based Classification by Mohammed M Mazid, A B M Shawkat Ali, Kevin S Tickle, School Of Computing Science, Central Queensland University, Australia in Recent Advances in Artificial Intelligence, Knowledge Engineering And Data Bases, ISSN: 1790-5109, ISBN: 978-960-474-154-0

[11] Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification by Tina R. Patil, Mrs. S. S. Sherekar, Sant Gadgebaba Amravati University, Amravati, tnpatil2@gmail.com, ss_sherekar@rediffmail.com in International Journal Of Computer Science And Applications, Vol. 6, No.2, Apr 2013, ISSN: 09741011 (Open Access) NCAICN-2013, PRMITR,Badnera

[12] Analysis of Bayes, Neural Network and Tree Classifier of Classification Technique in Data Mining using WEKA by Yugal kumar1 and G. Sahoo2, 1Assistant Professor in CSE/IT Dept., Hindu College of Engineering, Industrial, Area, Sonepat, Haryana, India. yugalkumar.14@gmail.com, 2Professor in Dept. of Information Technology, Birla Institute of Technology, Mesra, Ranchi, Jhrakhand, India., gsahoo@bitmesra.ac.in, Computer Science & Information Technology (CS & IT), CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2236

[13]C4.5 algorithm and Multivariate Decision Trees by Thales Sehn Korting Image Processing Division, National Institute for Space Research – INPE, S˜ao Jos´e dos Campos – SP, Brazil.

[14] Human Talent Prediction in HRM using C4.5 Classification Algorithm by Hamidah Jantan, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Terengganu, 23000 Dungun, Terengganu, Malaysia Email: hamidahjtn@tganu.uitm.edu.my and Abdul Razak Hamdan and Zulaiha Ali Othman, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, Malaysia in Hamidah Jantan et al. / (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, 2526-2534, ISSN : 0975-3397

[15] Decision Tree Analysis using Weka, Machine Learning – Project II by Sam Drazin and Matt Montag, University of Miami

[16] Machine Learning Approach for Preoperative Anaesthetic Risk Prediction by Karpagavalli S1, Jamuna KS2, and Vijaya MS21,2 GR Govindarajulu School of Applied Computer Technology, PSGR Krishnammal College for Women, Coimbatore, India, Email:

karpagam@grgsact.com in International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009

[17] Implementation of Decision Tree Algorithm to Analysis the Performance, Pooja Sharma[1], Asst. Prof. Rupali Bhartiya[2], Department Of Computer Science and Engineering,Shree Vaishnav Institute of Technology & Science, Indore, India[12] , International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 10, December 2012, ISSN (Print) : 2319-5940, ISSN (Online) : 2278-1021

10. Project Review Sheet

Mentor at industry - Dr. Raj Hingorani

(Industry Project - JNPT Hospital)

Project Evaluation Sheet 2017 - 18

Class: D17 A/B/C

Group No: 31

Title of Project: Predictive analysis on Pregnancy Database

Group Members: Deepanshu Garg, Deepesh Garg, Pratik Sechpt, Shruti Sindhi

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life-long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage 1	4	4	3	2	3	2	2	2	2	2	2	2	4	4	38

Comments: Facility for personal account / Messaging. Overall Good idea.

Name & Signature Reviewer1

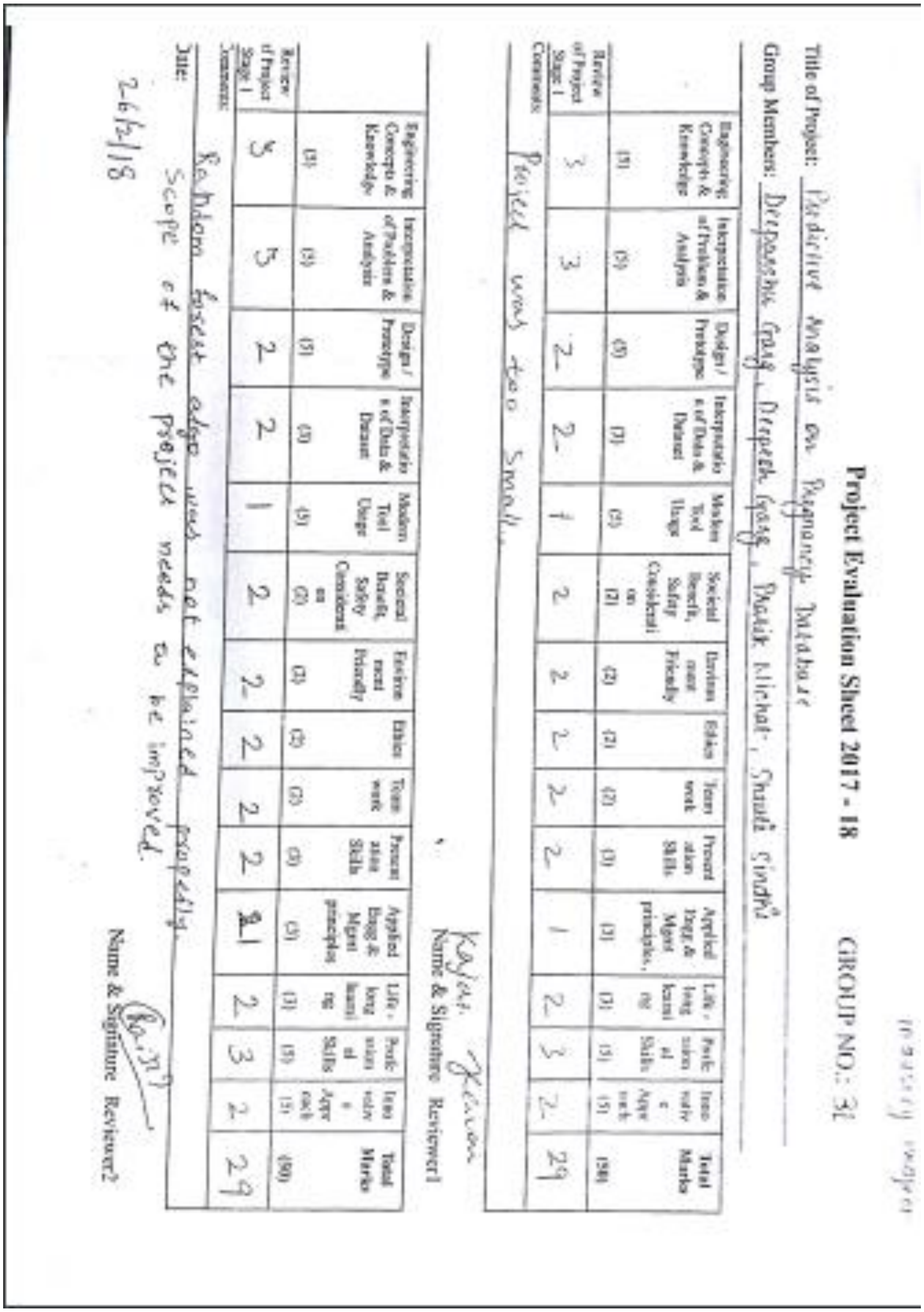
	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life-long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage 1	4	4	3	2	4	2	2	2	2	2	2	2	3	5	39

Comments:

Date: 26th September 2017

Name & Signature Reviewer2

Fig 10.1: Review Sem 7



Project Evaluation Sheet 2017 - 18 GROUP NO.: 31

Title of Project: Predictive Analysis on Pregnancy Database

Group Members: Deepanshu Garg, Deepesh Garg, Pratik Nichat, Shruti Sindhi

	Engineering Concepts & Knowledge	Interpretation of Problem & Analysis	Design / Prototype	Interpretation of Data & Dataset	Modern Tool Usage	Societal Benefit, Safety Consideration	Environment Friendly	Ethics	Team work	Presentation Skills	Applied Engg & Mgmt principles	Life-long learning	Professional Skills	Innovative Approach	Total Marks
	(5)	(5)	(5)	(3)	(2)	(2)	(2)	(2)	(2)	(3)	(3)	(3)	(5)	(5)	(50)
Review of Project Stage I	3	3	2	2	1	2	2	2	2	2	1	2	3	2	29

Comments: Project was too Small.

Kajal Keswani
Name & Signature Reviewer1

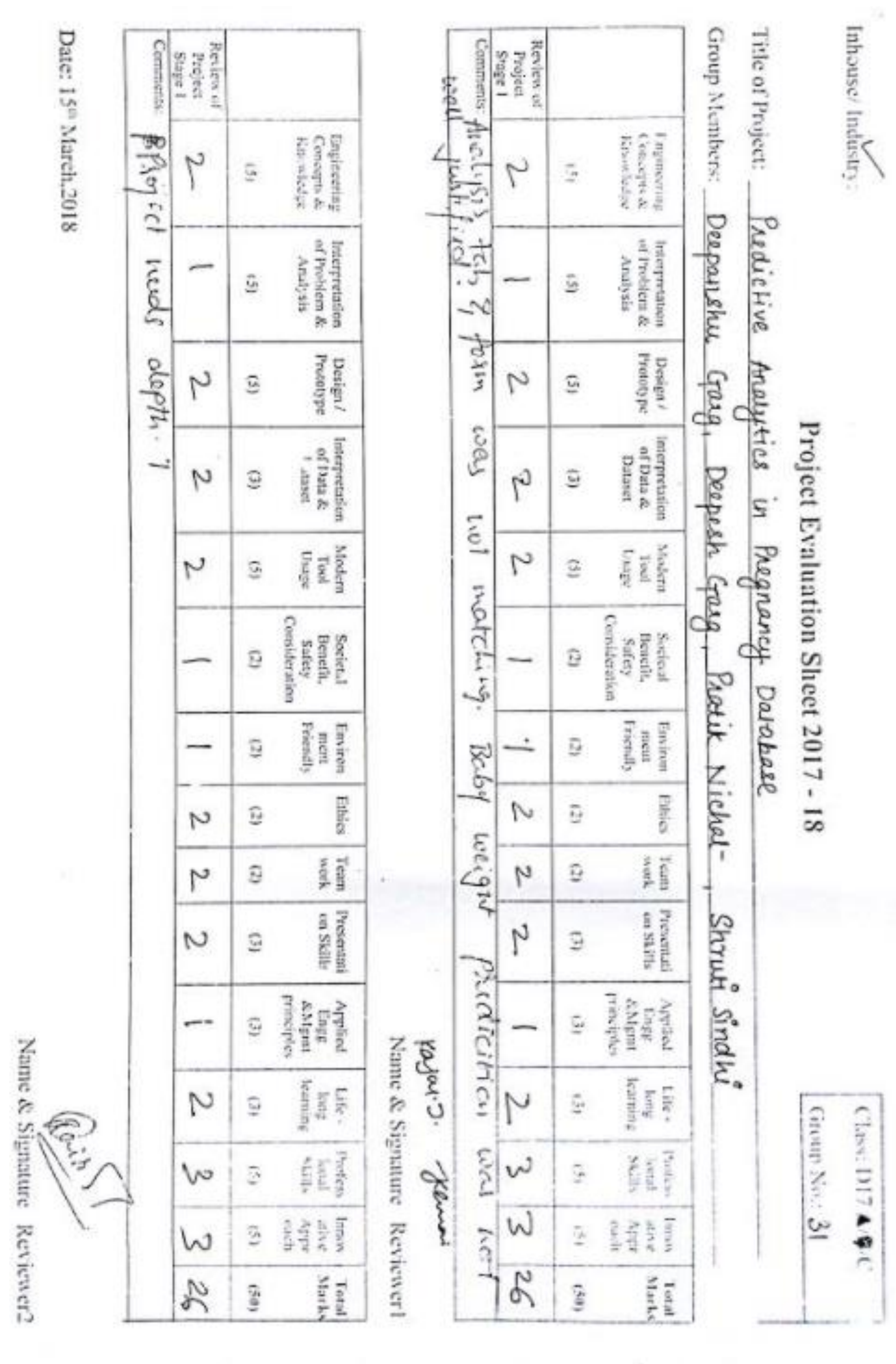
	Engineering Concepts & Knowledge	Interpretation of Problem & Analysis	Design / Prototype	Interpretation of Data & Dataset	Modern Tool Usage	Societal Benefit, Safety Consideration	Environment Friendly	Ethics	Team work	Presentation Skills	Applied Engg & Mgmt principles	Life-long learning	Professional Skills	Innovative Approach	Total Marks
	(5)	(5)	(5)	(3)	(5)	(2)	(2)	(2)	(2)	(3)	(5)	(3)	(5)	(5)	(50)
Review of Project Stage I	3	5	2	2	1	2	2	2	2	2	1	2	3	2	29

Comments: Random forest algo was not explained properly. Scope of the project needs to be improved.

Date: 26/2/18

Name & Signature Reviewer2

Fig 10.2: Review 3 Sem 8



Inhouse/ Industry: ✓

Class: D17 A/B/C

Group No.: 31

Project Evaluation Sheet 2017 - 18

Title of Project: Predictive Analytics in Pregnancy Database

Group Members: Deepanshu Garg, Deepesh Garg, Pratik Nichal, Shruti Sindhi

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life-long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage I	2	1	2	2	2	1	1	2	2	2	1	2	3	3	26

Comments: Analysis, tab & form was not matching. Baby weight prediction was not well justified?

Rajani D. Kamani

Name & Signature Reviewer1

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life-long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage I	2	1	2	2	2	1	1	2	2	2	1	2	3	3	26

Comments: Project needs depth ?

Date: 15th March,2018

Name & Signature Reviewer2

Fig 10.3: Review 4 Sem 8

11. Appendix

11.1 List of figures

11.2 List of tables

11.3 Paper Publications

ICECA 2018

2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018)

29,30,31 March, 2018 | http://iceca.org/index.html | smys375@gmail.com

28th February, 2018

To
PratikNichat, Deepesh Garg, Assistant Professor Rupali Hande, Deepanshu Garg,
Shruti Sindhi, Assistant Professor Sunita Suralkar.
Computer Engineering Department,
Vivekanand Education Society's Institute of Technology
Mumbai, India.

Paper ID: ICECA 2018–594

Subject: Acceptance Letter –2nd IEEE International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) – Reg.

Dear Author,
This is the notification to inform you that your Oral presentation proposal entitled Predictive Analytics in Pregnancy Database to the 2nd IEEE International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) organized by RVS Technical Campus, Coimbatore, India on 29, 30, 31 March, 2018 has been accepted as a result of blind reviews.

All registered papers will be published in IEEE Xplore
On behalf of the organization committee I would like to congratulate you.

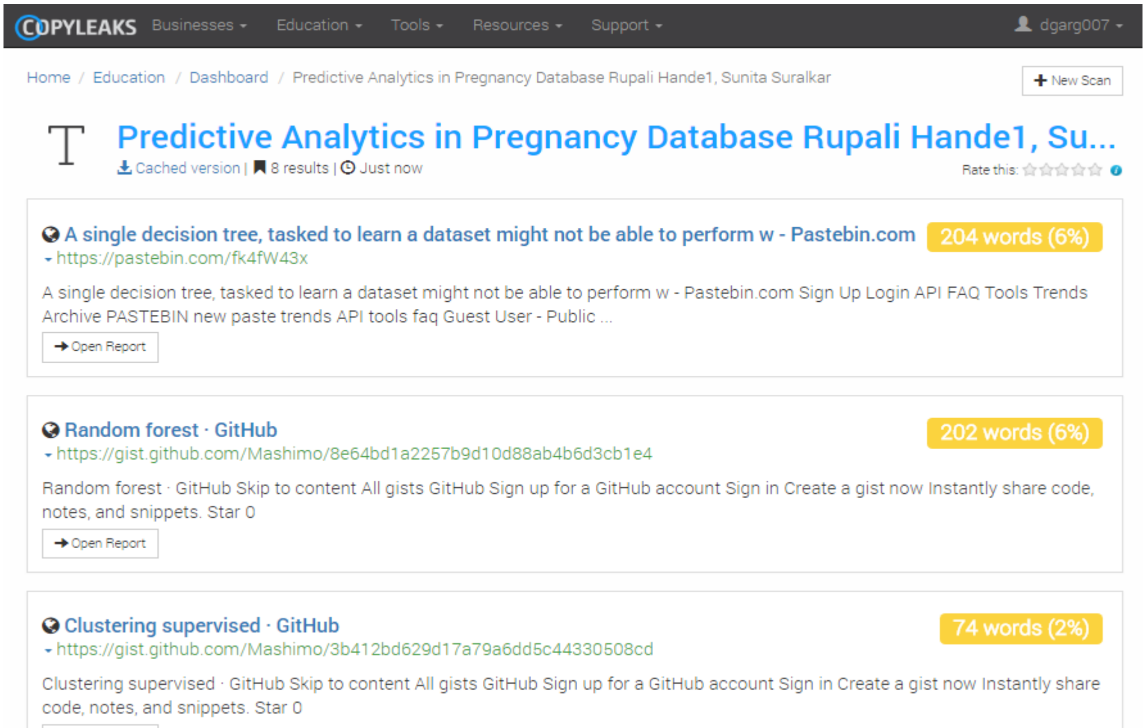
Yours sincerely,

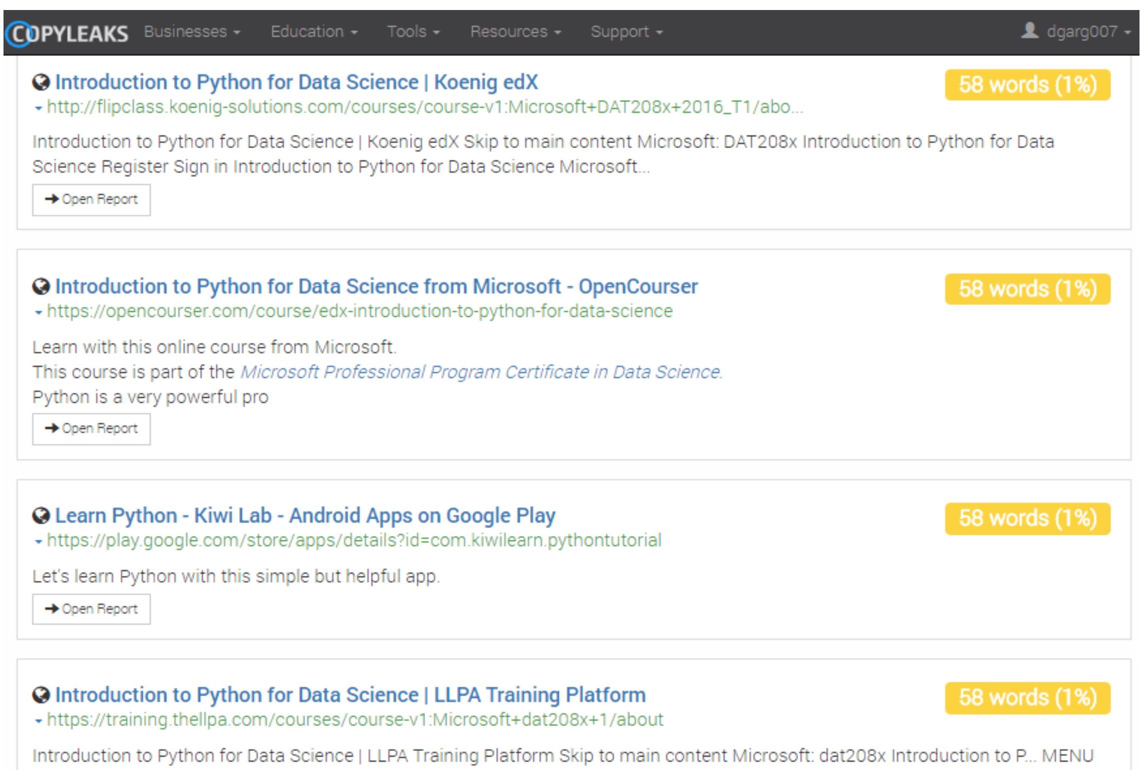
Dr. S. Smys,
Conference chair – ICECA 2018,
RVS Technical Campus,
Coimbatore, India.
smys375@gmail.com

TECHNICAL SPONSORS

IEEE

Plagiarism Report :





Paper published in 2nd IEEE International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) organized by RVS Technical Campus, Coimbatore, India

Predictive Analytics in Pregnancy Database

Pratik Nichat
Computer Engineering Department
Vivekanand Education Society's Institute of Technology
Mumbai, India
nichatpratik@gmail.com

Shruti Sindhi
Computer Engineering Department
Vivekanand Education Society's Institute of Technology
Mumbai, India
shrutisindhi1403@gmail.com

Deepesh Garg
Computer Engineering Department
Vivekanand Education Society's Institute of Technology
Mumbai, India
deepeshgarg6@gmail.com

Rupali Hande
Assistant Professor, Computer Engineering Department
Vivekanand Education Society's Institute of Technology
Mumbai, India
rupali.hande@ves.ac.in

Deepanshu Garg
Computer Engineering Department
Vivekanand Education Society's Institute of Technology
Mumbai, India
gargdeepanshu32@gmail.com

Sunita Suralkar
Assistant Professor, Computer Engineering Department
Vivekanand Education Society's Institute of Technology
Mumbai, India
sunita.suralkar@ves.ac.in

Abstract— **Pregnancy requires augmented medical care and attention. It induces a variety of abnormalities which, may lead to flinty situations if unnoticed or neglected. Such abnormalities and complications can turn severe and initiate instances leading to death of both mother and infant. With immense growth of big data in biomedical and healthcare communities, precise analysis of medical data and early disease detection is viable. But the analysis accuracy is reduced when the quality of medical data is incomplete. Legion of algorithms are known but some of them can be greatly used in the biomedical field for varied analysis and predictions. The modified prediction models are experimented over real-life hospital data collected. Our project works on effective performance in predicting risk levels (Complications), suggesting mode of delivery and treatment advice that needs to be prescribed during pregnancy from the collected, standardized and transformed data efficiently using supervised machine learning classification algorithms like Random Forest and Decision tree algorithm with a average accuracy of 85% and 80% respectively.**

Keywords- risk prediction; treatment prediction; pregnancy; decision tree classification algorithm; Random Forest algorithm; Decision tree; machine learning

I. INTRODUCTION

Pregnancy and the expectation of the birth of the child bring joy to the entire families. The thought of soon having a toddler in the family is a wondrous thing. We propose a system in which the healthcare providers should be to assess the high risk of conditions based on different estimable parameters like age, weight gain, blood pressure, blood group, hemoglobin, blood sugar, urine exam, HIV tests, thyroid profile etc. and necessary preventive action may be taken or help of the higher centers can be obtained. Also the treatment that should be ideally prescribed under the circumstances should be predicted along with the suggestion of mode of delivery that should be considered for delivering the baby. This newly developed system will help maintain up to date record of the patient and help us take better and fast decision on the available patient history. It will also help us analyze the areas where improvement is required. In dearth of medical expertise in emergency situations, this system can help in taking some initial decision. Due to compulsory registering process women delivering at private hospitals and homes won't remain out of the government's information net. If the women move in near places for delivery, the system will make sure the health care providers at the new place have complete and adequate information about the expectant mother. Information mostly reaches the district headquarters or the government only when it audits maternal death or premature death of the infant. If system is developed properly it will help sort all such problems and in near future can also be connected to various other systems to provide automated issuing of Birth and Death certificate. It is possible to generate different statistical data like birth rate, maternal mortality rate, and fetal mortality rate for different national programmes and also keep track of medication and vaccination of patients, high risk pregnancies and reporting incidence of normal deliveries, forceps or C-Sections conducted. The data in future can also be used in predicting the population in near future.

II. LITERATURE SURVEY

The aim of 'A comparative study of algorithms for risk prediction in pregnancy [1]' is to predict the present complications in the health of a pregnant woman using two classification algorithms namely C4.5 decision tree classification algorithm and Naive Bayes Classification Algorithm. The selected algorithms are powerful and popular tools used for the tasks of classification and prediction in Data mining. These two algorithms use pregnancy data collected from pregnant women in different stages of pregnancy to predict their present health state and the associated health complications.

The main aim of 'Prediction based health monitoring of pregnant women [2]' is to optimize performance of C4.5 classification algorithm by applying on standardized and appropriate format of data. The paper highlights the effective performance achieved by C4.5 classifier in accurately predicting risk levels during pregnancy from the collected, standardized and transformed data efficiently.

'Disease prediction by machine learning over big data from healthcare communities [3]' proposes a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital.

III. METHODOLOGY

There are many methods, technique or procedure to achieve a target. But at times the sequence or the upshot use doesn't matter. All that matters is how well they are implemented.

The methods and materials used to conduct work is shown in following steps given below:

1. Create pregnancy data set.

Following attributes have been collected for risk prediction in patients:

Attributes	**Data type**
age	numeric
trimester	numeric
weight	numeric
obstretic_history (number of pregnancies, number of abortions, previous mode of delivery, contracted pelvis , History of hypertension, History of gestational diabetes)	numeric/string
complaints (Amenorrhea, vomiting, weakness, pain in abdomen, oedema feet, Lethargy, headache, spotting pelvis, leaking pelvis, leg cramps, burning in micturition, pedal and abdominal wall oedema, Fever with cold and cough)	string
temperature	numeric
pulse	numeric
blood_pressure	numeric
blood_group	string
haemoglobin	numeric
blood_sugar	numeric
thyroid_stimulating_horm one	numeric
urine_profile	string
amniotic fluid	string
usg	numeric
Lie of the baby	string
excessive weight gain	string

Fig. 1. Different attributes and their datatypes.

2. Standardize and transform data. Transform data in required form i.e. converting ordinal values of attributes in numeric form and maybe using dimensionality reduction algorithm like pca to get rid of redundant attributes if any.

3. Divide pregnancy dataset into training and testing dataset. Training dataset is used for construction of the learning model and test dataset for evaluating generated model for determining the accuracy.

4. Apply machine learning classification algorithm like decision trees or random forest or k-nearest neighbors on training data set and build(train) the model.

5. Analyze the testing dataset with the trained model and obtain the results and accuracy.

IV. PROPOSED DESIGN

The proposed system will take in pregnant female's data as input which will be put in by authorized health monitor. The trained data set will also be referred to analyze the patient data and provide risk (complications like Pregnancy Induced Hypertension, anemia, Impending eclampsia, oligohydramnios etc.) predictions, predict most probable mode of delivery (Lower Section Caesarean Section / Full Term Normal Delivery/ Preterm Normal Delivery), ideal treatment to be provided(Folic acid/ iron/ calcium/ protein/ antihypertensive medications/ antiemetic and many other such prescriptions) and also visualize statistical information to help in census calculation and other national programmes. The system also will provide alerts to the patient for regular appointments.

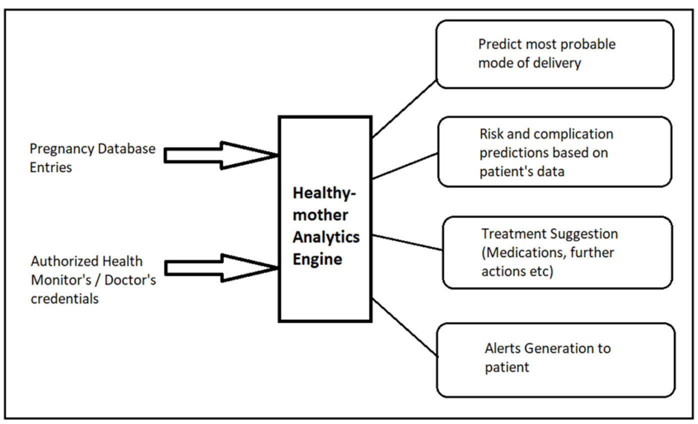


Fig. 2. Proposed design block diagram.

V. PROPOSED IMPLEMENTATION

We propose a solution using sklearn module in python which consists of various machine learning classifiers such as decision tree, random forests which can be applied to predict various risk conditions, ideal treatment for pregnant women.

The backend database has been implemented using MongoDB, a NoSQL database. Nodejs has been used for server side scripting. Frontend of the system has been developed using HTML and JavaScript.

Decision tree algorithm:

Decision tree algorithm is one of the most powerful and well suited supervised machine learning algorithms for medical parameters. It doesn't require tuning of parameters for creation of model, it automatically creates subset of relevant feature needed for classification, as choices of splits are not on absolute magnitudes of attribute values. It is robust to outliers. It can easily handle missing data values. So this algorithm is suited for solving prediction problems related to medical diagnosis.

How decision trees work:

- The decision tree is structured as a flowchart diagram. It has a single, starting root node, and can contain one or more leaf nodes, and internal nodes.
- Each node of the decision tree represents the test of single feature value.
- Each branch of the decision tree connects two nodes, an originating node, and a destination node. A branch represents an outcome of the test of the branch's originating node.
- Each leaf node of the decision tree represents a classification. The end goal of the tree is to label your input samples, so each leaf fulfills that goal.

Algorithm :
1) Read the data set
>>X=pd.read_csv('Datasets/pregnancy.data',index_col=0)
2) Split data into training data and testing data
>>X_training, X_testing, y_training, y_testing = train_test_split(X, y, test_size = 0.3, random_state = 7)
3) Create the model
>>dtree = tree.DecisionTreeClassifier(max_depth = 5, random_state = 2)
4) Train the decision tree model
>> dtree.fit(X_train_bag, y_train)
5) Test the accuracy of the model
>> dtree.score(X_test, y_test)

Random forest algorithm:

A single decision tree model, assigned to learn a dataset may not be able to perform well due to the noise, and the breadth and depth complexity of the data.

So rather depending on a single tree, an advanced version of decision tree called random forests depends on a forest of cleverly grown decision trees. Each tree preserves little general knowledge about entire data and becomes specialized in a specific area. In a random forest classifier, every tree casts vote that which label is most significant in deciding the result.

Training

Random forests uses two techniques while training, one occurring at the forest level and one at individual level. First, like any supervised classifier algorithm, we pass in a training set of samples along with "truth" labels when you create an instance of the class. Rather than sharing the entire dataset train/test split test is performed on the data. Each decision tree in the forest randomly samples from the overall training data set. By doing so, each tree exists in an independent subspace and the variation between trees is controlled. This technique is known as “tree bagging”, or “bootstrap aggregating”.

Tree bagging increases the accuracy of decision trees because while an individual decision tree might become hypersensitive to noise and localized attributes, once all the results are averaged, the border results get blurred out. Therefore using random forests over a single tree decrease the variance of your classification results, without increasing the bias. This technique only works if the individual trees are not correlated. If they were all trained on the same training set, they would only reinforce each other's decision. The bootstrapping or randomization of samples each tree is trained upon takes care of that.

Random forests also use one more trick. Along with tree bagging of training samples at the forest level, each individual decision tree also 'feature bags' at each node-branch split. This is useful because some datasets may contain a feature that is very correlated to the target. By selecting a random sampling of features every split, if such a feature were to exist, it wouldn't show up on as many branches of the tree and there would be more diversity of the features examined.

Predicting

After fitting (training) the forest, a prediction can be made for unseen samples by using the majority vote label assigned from each tree.

Error Testing

Since each tree within the forest is only trained using a subset of the overall training set, the forest group has the ability to error test itself. It does this by scoring each tree's predictions against that tree's out-of-bag samples. A tree's out of bag samples are those forest training samples that were withheld from a specific tree during training. There's nothing unique about splitting your data between training and testing sets except that you have an independent set of unseen samples to validate the accuracy of your training. Part of the random forest algorithm is the creation of independent sets for the training of each tree, so an overall *out-of-bag error* metric can be calculated for the forest ensemble. This error value is defined as the mean prediction error for each training samples using only those trees that didn't have the sample in their bootstrap.

Algorithm:

1) Read the data set
>>X=pd.read_csv('Datasets/pregnancy.data',index_col=0)
2) Split data into training data and testing data
>>X_training, X_testing, y_training, y_testing = train_test_split(X, y, test_size = 0.3, random_state = 7)
3) Create the model
>>forest = RandomForestClassifier(n_estimators = 30, max_depth = 10, oob_score = True, random_state = 0)
4) Train the decision tree model
>> forest.fit(X_train_bag, y_train)
5) Test the accuracy of the model
>> dtree.score(X_test, y_test)

VI. RESULTS

After training the decision tree model on training data consisting of attributes mentioned in section III and testing on testing data the accuracy obtained while predicting mode of delivery in pregnancy is 86.67%, predicting complications in pregnancy is 78.67% and predicting treatment is 66.67%.

And after training the Random forests model on training data consisting of above attributes mentioned in section III and testing on testing data accuracy obtained while predicting mode of delivery in pregnancy is 90.67%, predicting complications in pregnancy is 80% and predicting treatment is 76%.

Since the Random forest algorithm performed better, we implemented the system using the random forest prediction algorithm.

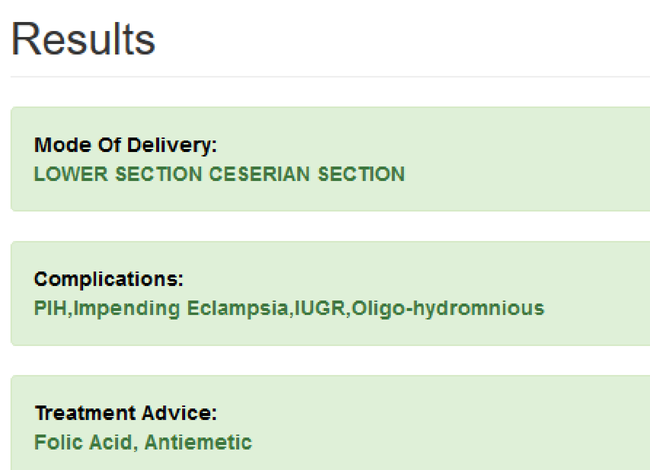


Fig 3. Sample output of system

VII. CONCLUSION

In this paper, we propose and implement a pregnancy risk prediction system using structured and unstructured data from hospital. The learning model generated by decision tree algorithm or random forest algorithms are easy to understand and require no domain experts. It can be inferred from the implementation of the system that these techniques are efficient, powerful and popular in classifying data and making desired predictions of mode of delivery suggestion, risks (complications) and treatment advice in pregnancy with good accuracy.

VIII. References

[1] A Comparative Study of Classification Algorithms for Risk Prediction in Pregnancy Published in: TENCON 2015 - 2015 IEEE Region 10 Conference Electronic ISBN: 978-1- 4799-8641- 5

[2] Prediction based health monitoring in pregnant women Published in: Applied and Theoretical Computing and Communication Technology (iCATccT), 2015 International Conference Electronic ISBN: 978-1- 4673-9223- 5

[3] Disease Prediction by Machine Learning over Big Data from Healthcare Communities Published in: IEEE Access (Volume: 5) Electronic ISSN: 2169-3536

[4] Decision Tree Analysis using Weka, Machine Learning – Project II by Sam Drazin and Matt Montag, University of Miami

[5] C4.5 algorithm and Multivariate Decision Trees by Thales Sehn Korting Image Processing Division, National Institute for Space Research – INPE, S˜ao Jos´e dos Campos – SP, Brazil.

[6] Implementation of Decision Tree Algorithm to Analysis the Performance, Pooja Sharma¹, Asst. Prof. Rupali Bhartiya², Department Of Computer Science and Engineering,Shree Vaishnav Institute of Technology & Science, Indore, India¹² , International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 10, December 2012, ISSN (Print) : 2319-5940, ISSN (Online) : 2278-1021

[7] Improved C4.5 Algorithm For Rule Based Classification by Mohammed M Mazid, A B M Shawkat Ali, Kevin S Tickle, School Of Computing Science, Central Queensland University, Australia in Recent Advances in Artificial Intelligence, Knowledge Engineering And Data Bases, ISSN: 1790-5109, ISBN: 978-960-474-154-0

[8]United Nations Maternal Mortality Estimation Interagency Group, consisting of representatives from the World Health Organization (WHO), United Nations Children's Fund (UNICEF), the United Nations Population Fund (UNFPA), United Nations Population Division, the World Bank and Wikipedia.

[9] Predicting Students' Performance Using Id3 And C4.5 Classification Algorithms by Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, Department of Computer Engineering, Fr. C.R.I.T., Navi Mumbai, Maharashtra, India in International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013, DOI : 10.5121/ijdkp.2013.3504.

[10] Improved C4.5 Algorithm For Rule Based Classification by Mohammed M Mazid, A B M Shawkat Ali, Kevin S Tickle, School Of Computing Science, Central Queensland University, Australia in Recent Advances in Artificial Intelligence, Knowledge Engineering And Data Bases, ISSN: 1790-5109, ISBN: 978-960-474-154-0

[11] Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification by Tina R. Patil, Mrs. S. S. Sherekar, Sant Gadgebaba Amravati University, Amravati, tnpatil2@gmail.com, ss_sherekar@rediffmail.com in International Journal Of Computer Science And Applications, Vol. 6, No.2, Apr 2013, ISSN: 09741011 (Open Access) NCAICN-2013, PRMITR,Badnera

[12] Analysis of Bayes, Neural Network and Tree Classifier of Classification Technique in Data Mining using WEKA by Yugal kumar1 and G. Sahoo2, 1Assistant Professor in CSE/IT Dept., Hindu College of Engineering, Industrial, Area, Sonepat, Haryana, India. yugalkumar.14@gmail.com, 2Professor in Dept. of Information Technology, Birla Institute of Technology, Mesra, Ranchi, Jhrakhand, India., gsahoo@bitmesra.ac.in, Computer Science & Information Technology (CS & IT), CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2236

[13]C4.5 algorithm and Multivariate Decision Trees by Thales Sehn Korting Image Processing Division, National Institute for Space Research – INPE, S˜ao Jos´e dos Campos – SP, Brazil.

[14] Human Talent Prediction in HRM using C4.5 Classification Algorithm by Hamidah Jantan, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Terengganu, 23000 Dungun, Terengganu, Malaysia Email: hamidahjtn@tganu.uitm.edu.my and Abdul Razak Hamdan and Zulaiha Ali Othman, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, Malaysia in Hamidah Jantan et al. / (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, 2526-2534, ISSN : 0975-3397

[15] Decision Tree Analysis using Weka, Machine Learning – Project II by Sam Drazin and Matt Montag, University of Miami

[16] Machine Learning Approach for Preoperative Anaesthetic Risk Prediction by Karpagavalli S1, Jamuna KS2, and Vijaya MS21,2 GR Govindarajulu School of Applied Computer Technology, PSGR Krishnammal College for Women, Coimbatore, India, Email: karpagam@grgsact.com in International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009

[17] Implementation of Decision Tree Algorithm to Analysis the Performance, Pooja Sharma¹, Asst. Prof. Rupali Bhartiya², Department Of Computer Science and Engineering,Shree Vaishnav Institute of Technology & Science, Indore, India¹² , International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 10, December 2012, ISSN (Print) : 2319-5940, ISSN (Online) : 2278-1021