TECHNOLOGICAL INSTITUTE OF THE PHILIPPINES
Quezon City

COLLEGE OF INFORMATION TECHNOLOGY EDUCATION
Computer Science Department

CS 307 – Thesis 1

Male Infertility Diagnosis & Recommendation System Using Machine Learning

Gabriel Joshua Miguel

Ria Marie Toledo

Michael Joe Refuerzo

CS 307 Thesis 1

Jasper S. Agustin
Adviser

October 2018

Chapter 1

INTRODUCTION

Infertility defines as a disease according to the World Health Organization (WHO). A couple who cannot conceive after 12 months of unprotected sexual intercourse is considered infertile (Sharon, Linda, 2006). Infertility causes several effects in different types of personal health: physical, mental, emotional, psychological, social and even religious, in the couples that suffer from it. A man's fertility generally relies on the quantity and quality of his sperm. If the number of sperm a man ejaculates is low or if the sperm are of poor quality, it will be difficult, and sometimes impossible, for him to cause pregnancy. It is estimated that one in 20 men suffer from being infertile with low numbers of sperm in his ejaculate. However, only about one in 100 men has no sperm in his ejaculate (Andrology Australia, 2018). Semen analysis is important for the evaluation of male fertility potential and can be used for the assessment of sperm donors; it is the cornerstone for evaluating men for fertility/infertility (Wang, Swerdloff, 2014). Male infertility has many causes, which include climatic conditions, lack of exercise, overweight, poor food choices, alcohol and cigarette use.

Most infertility cases 85 to 90 percent are treated with conventional therapies, such as drug treatment or surgical repair of reproductive organs. If fertility treatments are unsuccessful, it is possible to use sperm donated by a third party (American Society for Reproductive Medicine, 2012).

Machine Learning techniques have been applied as a decision support tool to assist domain expert in making accurate decisions. This runs mainly by supervised learning that allows

physicians to decide from limited sets of diagnoses or estimate patient risk based on symptoms and genetic information (Merkert, Mueler, Hubl, 2015).

Background of the Study

A recent study (Naina & Amit, 2015) has stated that increasing numbers of male infertility has been rising through the ages, which affects 8–12% of couples worldwide. This has been a major concern because many individuals are unable to afford medical checkups. The cost of a fertility test ranges from $50 to $300 ("Advanced Fertility Center of Chicago", n.d.). For male infertility checkups, the patient undergoes a semen analysis wherein two semen samples are collected on separate days by masturbation. The process is time consuming which is not simply reliable for patients seeking a faster solution (“Testing for Male Infertility,” n.d.). Other studies also state that one of the main reasons of divorce are due to infertility where in couples who have children are unlikely to file a divorce (Daniela, 2001; Ameneh et al., 2012). In spite of the fact that many people think of infertility as a problem for females, in about 40% of infertile couples, the man is the sole cause of Infertility. Male infertility has many causes: not only physical problems or hormonal imbalances but also behavioral and psychological problems. The aim of this study is to evaluate the importance of some environmental factors and lifestyle in order to predict if the patient is fertile or infertile (Koskas et al., 2014). The researchers want to develop a system that will detect possible fertility disorders using the current datasets provided.

Objectives

The study aims to effectively diagnose the fertility of a man from the data set. Specifically, this study aims to:

- To design and develop a system that will diagnose possible male infertility disorders using patient's parameters.
- To design and develop a diagnosis system for accurate results and medical recommendation for an effective treatment of male infertility.
- To predict the factor that leads to fertility using data mining
- Implement classification algorithms which will classify whether the patient is fertile or not
- Test the developed system using ISO 25010 with the software evaluation criteria Accuracy, Performance Efficiency, and Usability.

Significance of the Study

The study of the development of the proposed work will benefit the following:

Doctors. Predicting these kinds of fertility problems is a difficult task. Reproductive Endocrinologist, Urologist, and other specialist are well-trained and play an important role in understanding the complexities of the human reproductive system and fertility. The proposed work will benefit the doctors to be more efficient and convenient in their jobs with accurate results.

Future Researchers. The proposed study will be a useful reference for the researchers who would plan to make a study relative to this research in the future.

Health Agencies. The data of this study can be used as a reference for the health agencies in order to determine the number of people affected in a community and to give light and awareness to the communities about fertility/infertility.

Couples. The proposed study will benefit married couples who are having infertility cases. The system can develop a greater communication, cooperation and responsibility for partners that would like to treat infertility disorder especially if they would like to have children.

Scope and Limitations

Scope

The proposed work is focused on the development of a model that will be able to improve the accuracy in fertility prediction.

The study involves the development of Machine Learning that will be able to give efficient output in predicting normal and/or altered fertility.

The users will select eight symptoms in order for our system to predict if a male patient is fertile or infertile. Eight attributes are utilized, from them, all attributes will be considered as inputs which predict the future state of the attribute "Fertile/Infertile". The eight attributes needed for our prediction system are previous diseases, age, accident or serious trauma encounters, surgical intervention, high fevers in the previous year, the frequency of alcohol consumption, the number of hours spent sitting per day and the season which the analysis was performed.

If the system has predicted the user as Infertile, it will provide a recommendation in order to increase the sperm count of the patient There are many factors in increasing the sperm count such as limiting alcohol intake, maintaining a healthy weight and healthy diet, avoiding areas with high temperatures etc.

Limitations

The system is only limited in providing fertile or infertile results for male patients. An exact medical diagnosis will not be available for users who want to determine the disease linking to their symptoms and signs. An accurate sperm count won't be available due to semen analysis can only be measured and evaluated at a local hospital, clinic, etc.

Chapter 2

THEORETICAL FRAMEWORK

This chapter focuses on presenting and discussing other existing concepts and theories that are relevant to the study. These studies can further validate the study's background and provide supplementary information about existing studies and theories that can be related to the study designed. Data mining has been played an important role in the intelligent medical systems. The technical terms are also defined at the end of this chapter to help elaborate the study. These terms are defined conforming to the study's definition.

Review of Related Literature and Studies

A recent study (Narjes, Tina & Meimanat, 2017) discussed the related risk factors towards infertility. According to the researchers, the word "infertile" or "inability to give birth of procreate" presents so many clinical perspectives. A couple who is unable to achieve pregnancy after 12 months of unprotected sexual intercourse is considered infertile. This issue is affecting people globally whose cause and importance may vary according to the socio-economic condition. And according to the statistics, annually 60-80 million couples around the world suffer from this disease. Ten to 12% of couples around the world are suffering from infertility half of which, the man is the sole cause. Some major factors of this disease include the impact of the reproductive system disorders, hormonal disorders, reproductive system diseases, age, alcohol consumption, smoking, cell phone use, sexual violence, stress, obesity, nutrition, and any chronic disease reducing the change of a successful pregnancy.

Other studies (Miina, & Anitta, 2014) provide an insight with infertility as an experience and its effects on a relationship. The study discusses how the disease has affected an infertile couple with their relationship. Infertility affects infertile couples causing feelings of jealousy, shame, disbelief and anger. This can lead to withdrawal from social contacts. And undeniably this has become a difficult problem to talk about. Questions to infertile couples about having children are often painful as well with visits in an infertility clinic. Problems experienced by the couples feel like they are never ending and not being resolved.

Medical examination of infertility is started if pregnancy has not occurred after a year of unprotected regular intercourse. Sometimes, couples are not required to wait for a year, if coupe has e.g. history of irregular periods or amenorrhea. The first steps in investigating the disease is for both partners to consult a general health care center at the same time. A nurse will chart the couple's sexual and gynecological history. (Tiitinen, 2009).

Before a couple gets access with infertility treatments, couples undergo medical examinations which show the cause of infertility. After the examinations, treatments are then suggested to the couple which they have agreed on. The doctor will take into account what are the reasons behind the diagnosis before starting the treatments. Typical treatments are surgical treatments (removing myomas, endometriosis), hormonal treatments and assisted reproductive treatments (in vitro fertilization – IVF). (Tulppala, 2007).

Data Mining is a non-trivial extraction of implicit, previously unknown and potential useful information about data (Frawley, & Piatestsky, 1996). In short, it is a process of data being analyzed from different perspectives, and gathering the knowledge from it. A study has

been improved to improve heart disease prediction using data mining classification techniques. The researchers used a number of factors in their system to identify if a patient is diagnosed with heart disease which are family history, smoking, poor diet, high blood pressure, high blood cholesterol, obesity, physical inactivity and hypertension. Factors mentioned are used to analyze the heart disease. In many situations, diagnosis is generally based on the patient's current test result. Thus the diagnosis is a complex task that requires much experience and high skill. The datasets used by the researchers consist of total 573 records in the Heart disease database. The records were divided into two data sets were one data set was used for training which consisted of 303 records and the other 270 records were used for testing. The overall objective or the researchers were to predict the presence of heart disease more accurately. During their study, obesity and smoking were used to get more accurate results. Three data mining classification techniques were used namely Decision trees, Naïve Bayes and Neural Networks. From the results, it has been seen that Neural Networks provided accurate outputs compared to Decision trees and Naïve Bayes. (Chaitrali, & Sulabha. 2012).

Lung cancer has been one of the causes of leading cancer deaths worldwide. Treatment depends on the histological type of cancer, stage, and the patient's performance status. The treatments towards this disease include chemotherapy, surgery and radiotherapy. Survival rate depends on the overall health and stage of the patient. The researchers focused on using data mining classification in order to predict lung cancer. The classification algorithm used in the microarray analysis belonged to four categories: IF-THEN rule, Decision tree, Bayesian classifiers and Neural network. The researchers concluded that the most effective algorithm used

in their system is the Naïve Bayes followed by the IF-THEN rule, decision trees and neural network in terms of accuracy. (Krishnaiah, Narsimha, & Chandra. 2013).

A recent study discusses breast cancer as the second main cause of death among women. It is cancer that forms in the cells of the breast in men or women. There is a mighty tool called WEKA it holds supervised learning as well as unsupervised learning methods include classification, clustering, association mining, future selection, and data visualization. Using the powerful tool is to help the researchers to implement and analogize data mining strategies very easily on real or synthetic data. According to Dr. Chinyere Akpanika a Gynaecologist of University of Calabar Teaching Hospital that eighty-five percent of women do not have a family history of having a breast cancer. For the analysis of diabetic disease, a breast cancer dataset has been produced. The dataset consist of 768 instances and eight attributes are used for the comparative analysis such as menstrual and reproductive history, certain genome, dense breast tissue, population, tissue preparation, immunohistochemistry, image acquisition, quantification with stereology test grid, and visual evaluation.

Another related studies T. Marikani worked on data mining techniques to predict heart disease using supervised learning. The result of this study gives accuracy of Decision tree 95 % , Naive Bayes 81.7 %, K-NN 75.7 %. (T. Marikani, K.Shyamala 2017).

(Jarad, A., Katkar, R., Shaikh, A., & Salvem A. 2015) used database software called MongoDB with Naive Bayes, Decision Tree, and KNN algorithm to predicting patient's heart disease. They also used sample dataset from UCI with 13 out of 76 parameters available. The

result is the accuracy of algorithms used: Naive Bayes 52.33 %, Decision Tree 52 %, and KNN gives 45.67 %.

There are several major kinds of classification of algorithms such as Naïve Bayes and J48 which were used for the breast cancer prediction system. The comparison of the two algorithms was done based on performance factors in terms of classification accuracy and execution time. From the results, the Naive Bayes was accurate thus it is considered the best classifier when compared with the J48 algorithm. (Meera, 2018)

Algorithms Used

Decision Tree

Decision Trees are supervised learning method used for classification and regression. The Goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

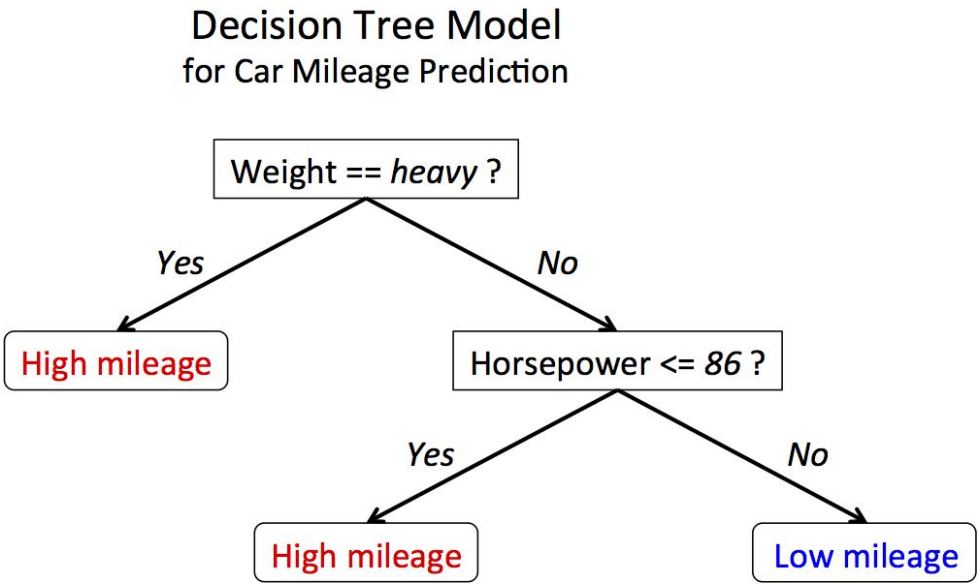

Figure 1. Decision Tree

In the figure we can use the decision tree to predict any unlabeled data. If we are given a sample data like Weight = Not Heavy and Horsepower <= 86, we can simply say by looking at the decision tree that the car has high mileage

Formula for making a decision tree

Information gain:

$$I\ (p,n) = \frac{-p}{p+n}\left(\frac{p}{p+n}\right) - \frac{-n}{p+n}\left(\frac{n}{p+n}\right)$$

Entropy:

$$E\ (A) = \sum_{i=1}^{n} \frac{pi+ni}{p+n} * I\ (pi, ni)$$

Gain:

$$Gain\ (A) = I\ (p,n) - E\ (A)$$

The Decision Tree algorithm uses the information gain and entropy. The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain while the entropy use to calculate the homogeneity of sample if the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one

Naive Bayes Classifier

Naive Bayes classifiers are a unit extremely ascendible, requiring a variety of parameters linear within the number of variables (features/predictors) in a very learning downside. Maximum-likelihood coaching is done by evaluating a closed-form expression that takes linear

time, instead of by high-ticket repetitive approximation as used for several different forms of classifiers.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

Conditional Probability: It help us to find the probability that something will happen given that something else has happened.

$$P\ (A\ and\ B) = P\ (A) * P\ (B \mid A)$$

Bayes rule: The rule helps us to know how often A happens given that B has already happened P (A | B), when we know how often B happens given that A has already happened P (B | A)

$$P\ (A \mid B) = \frac{P\ (B \mid A) * P\ (A)}{P\ (B)}$$

Naïve Bayes Classifier finds the probability of every feature then it selects the outcome with highest probability.

K Nearest Neighbors

K – nearest neighbors algorithm (k-NN is a non – parametric method used for classification and regression. The input consists of the k closest training example in the feature space. The output depends on whether K- NN is used for classification or regression:

- In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
- In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

Calculating distance: We can find the distance between two points using the Euclidean distance formula.

$$\sqrt{\sum_{i=1}^{k} (Xi - Y\,i)\ \hat{}\ 2}$$

Euclidean distance or Euclidean metric is the "ordinary" straight-line distance between two points in Euclidean space. With this distance, Euclidean space becomes a metric space. The associated norm is called the Euclidean norm. Older literature refers to the metric as the Pythagorean metric

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms

Design Trade Offs

Based on the constraints of the three algorithm designs, the tradeoffs can be the basis on which design is the most reliable. Each of the constraints were collected from related literature each implementing the three algorithms. The three constraints were measured in terms of Accuracy, speed, and size.

Algorithm	Design Constraints		
	Accuracy(%)	Training time (s)	Testing time (s)
Decision Tree	95.0000	0.0006	0.0001
Naïve Bayes	81.0000	0.0003	0.0008
K-NN	75.0000	0.0000	0.0012

Table 1. Design Constraints

Table 1 shows the comparison of three algorithms based on their accuracy, speed and size. The researchers chose the criteria's suitable for the prediction algorithms based on the related studies; the criteria presented are accuracy and speed.

The chosen algorithms will satisfy the chosen criterion, each algorithm will be ranked depending on the compatibility of the algorithm in the system. The subordinate rank would be computer based on the value of the percentage difference.

Percentage Difference Formula:

$$\%\text{difference} = \frac{(\text{Higher Value} - \text{Lower Value})}{\text{Higher Value}}$$

Subordinate Rank Formula:

Subordinate Rank = Governing Rank - [(% difference) x 10]

Designs	Accuracy(%)	%Difference	Subordinate Rank	Subordinate Rank Round Off	Percentage
Decision Tree	95.0000	-0.1728	11.7284	10.0000	0.0000
Naïve Bayes	81.0000	0.0000	10.0000	10.0000	0.0000
K-NN	75.0000	0.0741	9.2593	10.0000	0.0000

Table 2. Accuracy

Table 2 shows the comparison of three algorithms based on their accuracy, it will determine in terms of what algorithm has the highest percentage of providing the accurate output.

Computation of ranking for trade-off the Decision Tree and K-NN:

$$\%\text{difference} = \frac{(95.00 - 75.00)}{95.00}$$

%difference = 0.0741

Subordinate rank = 10 - [(0.741 x 10)]

Subordinate rank = 10.00

Designs	Training time (s)	%Difference	Subordinate Rank	Subordinate Rank Round Off	Percentage
Decision Tree	0.0006	0.9833	0.1667	9.0000	10.0000
Naïve Bayes	0.0003	0.9667	0.3333	9.0000	10.0000
K-NN	0.0000	0.0000	10.0000	10.0000	0.0000

Table 3. Training time

Table 3 shows the comparison of three algorithms based on their speed, it will determine in terms of what algorithm has the highest percentage of providing the fastest output.

Designs	Testing time (s)	%Difference	Subordinate Rank	Subordinate Rank Round Off	Percentage
Decision Tree	0.0001	0.0000	10.0000	10.0000	0.0000
Naïve Bayes	0.0008	0.8750	1.2500	10.0000	0.0000
K-NN	0.0012	0.9167	0.8333	9.0000	10.0000

Table 4. Testing time

Table 4 shows the comparison of three algorithms based on their testing time, it will determine in terms of what algorithm has the highest percentage of providing the fastest testing time.

Designs	Criterion's Importance	Ability to satisfy the criterion (On Scale of 1 to 10)					
	(On Scale of 1 to 10)	Decision Tree		Naïve Bayes		K-NN	
Accuracy(%)	10.0000	95.0000	950.0000	81.0000	810.0000	75.0000	750.0000
Training time (s)	9.0000	0.0006	0.0054	0.0003	0.0027	0.0000	0.0001
Testing time (s)	8.0000	0.0001	0.0008	0.0008	0.0064	0.0012	0.0096
			950.0062		810.0091		750.0097

Designs	Criterion's Importance	Ability to satisfy the criterion (On Scale of 1 to 10)					
	(On Scale of 1 to 10)	Decision Tree		Naïve Bayes		K-NN	
Accuracy(%)	8.0000	95.0000	760.0000	81.0000	648.0000	75.0000	600.0000
Training time (s)	10.0000	0.0006	0.0060	0.0003	0.0030	0.0000	0.0001
Testing time (s)	9.0000	0.0001	0.0009	0.0008	0.0072	0.0012	0.0108
			760.0069		648.0102		600.0109

Designs	Criterion's Importance	Ability to satisfy the criterion (On Scale of 1 to 10)					
	(On Scale of 1 to 10)	Decision Tree		Naïve Bayes		K-NN	
Accuracy(%)	9.0000	95.0000	855.0000	81.0000	729.0000	75.0000	675.0000
Training time (s)	8.0000	0.0006	0.0048	0.0003	0.0024	0.0000	0.0001
Testing time (s)	10.0000	0.0001	0.0010	0.0008	0.0080	0.0012	0.0120
			855.0058		729.0104		675.0121

Designs	Criterion's Importance	Ability to satisfy the criterion (On Scale of 1 to 10)					
	(On Scale of 1 to 10)	Decision Tree		Naïve Bayes		K-NN	
Accuracy(%)	10.0000	95.0000	950.0000	81.0000	810.0000	75.0000	750.0000
Training time (s)	8.0000	0.0006	0.0048	0.0003	0.0024	0.0000	0.0001
Testing time (s)	9.0000	0.0001	0.0009	0.0008	0.0072	0.0012	0.0108
			950.0057		810.0096		750.0109

Designs	Criterion's Importance	Ability to satisfy the criterion (On Scale of 1 to 10)					
	(On Scale of 1 to 10)	Decision Tree		Naïve Bayes		K-NN	
Accuracy(%)	9.0000	95.0000	855.0000	81.0000	729.0000	75.0000	675.0000
Training time (s)	10.0000	0.0006	0.0060	0.0003	0.0030	0.0000	0.0001
Testing time (s)	8.0000	0.0001	0.0008	0.0008	0.0064	0.0012	0.0096
			855.0068		729.0094		675.0097

Designs	Criterion's Importance	Ability to satisfy the criterion (On Scale of 1 to 10)					
	(On Scale of 1 to 10)	Decision Tree		Naïve Bayes		K-NN	
Accuracy(%)	8.0000	95.0000	760.0000	81.0000	648.0000	75.0000	600.0000
Training time (s)	9.0000	0.0006	0.0054	0.0003	0.0027	0.0000	0.0001
Testing time (s)	10.0000	0.0001	0.0010	0.0008	0.0080	0.0012	0.0120
			760.0064		648.0107		600.0121

Table 5. Sensitivity Analysis

Table 5 shows the sensitivity analysis combinations. By ranking each algorithm on a scale of 1 to 10 based on design criterions accuracy and hardware requirement. This is to show that there is no bias in choosing the algorithm, the researchers ranked each algorithm randomly from 8 to 10. 10 being the highest priority of criterion and 8 being the lowest.

Sensitivity Analysis Combination	Decision Tree	Naïve Bayes	K-NN
Trial 1	950.0062	810.0091	750.0097
Trial 2	760.0069	648.0102	600.0109
Trial 3	855.0058	729.0104	675.0121
Trial 4	950.0057	810.0096	750.0109
Trial 5	855.0068	729.0094	675.0097
Trial 6	760.0064	648.0107	600.0121

Table 6. Summary of Sensitivity Analysis

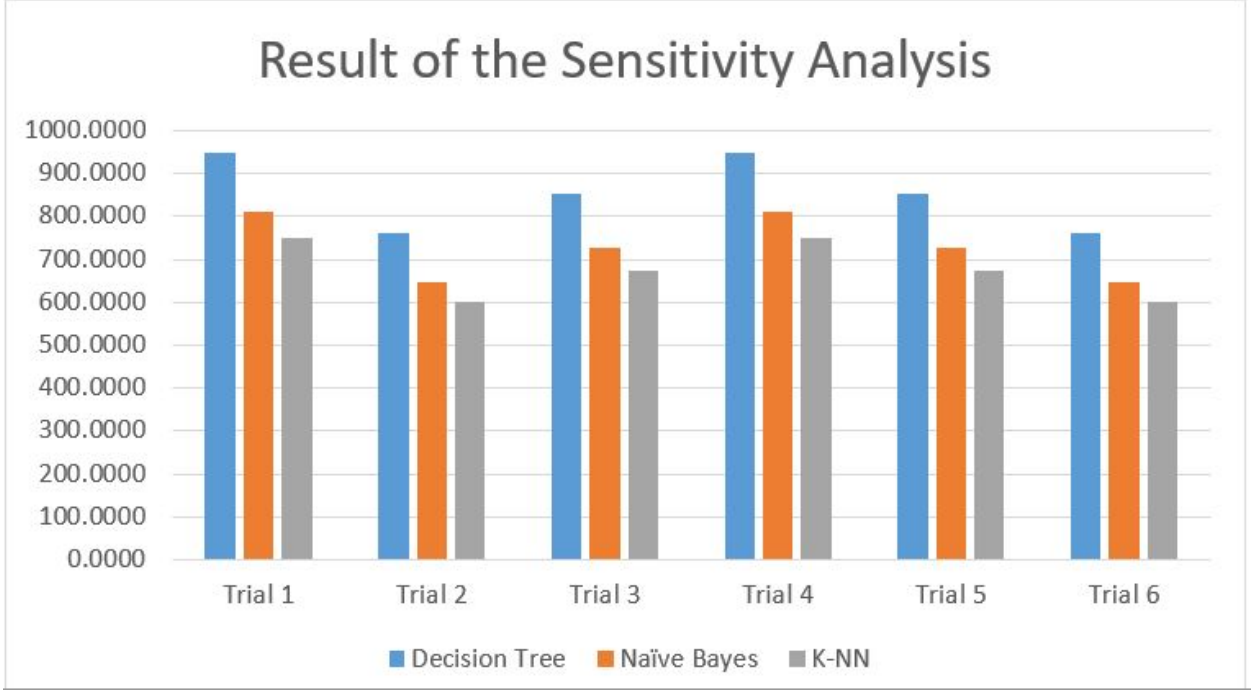

Figure 2. Sensitivity Analysis Result

The table shows the summary of the combinations of sensitivity analysis. Decision Tree has the highest value against K-NN and Naïve Bayes.

Conceptual Framework

Below is an Input Process Output (IPO) diagram that lists the inputs and outputs of each individual process.

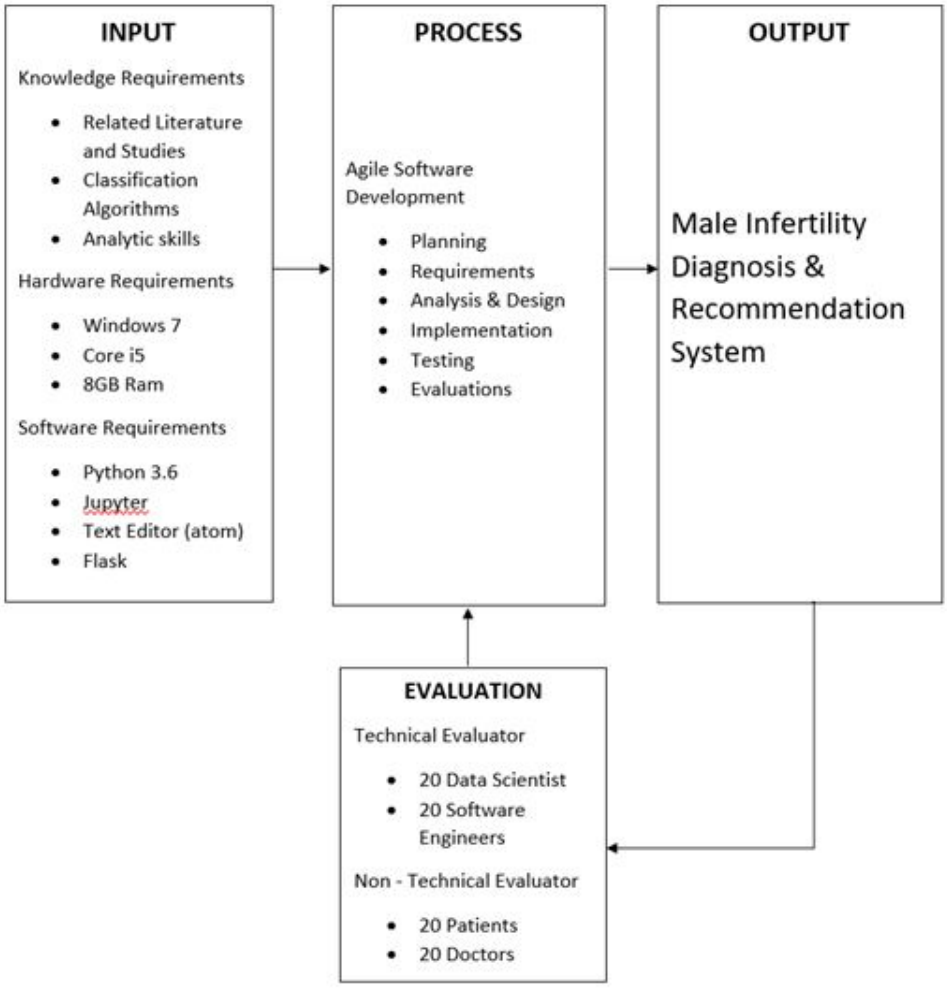

Figure 3. Conceptual Framework

The diagram shows the input, process then eventually the output. In the input part we has three requirements. The researches will plan and gather all the requirements necessary for them to analyze and predict the data. After gathering the researchers will start to train the data.

Dataset

Data Set Characteristics:	Multivariate	**Number of Instances:**	100
Attribute Characteristics:	Real	**Number of Attributes:**	10
Associated Tasks:	Classification	**Missing Values?**	N/A

Table 7. Datasets

Attribute	**Values**
Season in which the analysis was performed	• Winter • Spring • Summer • Fall (-1, -0.33, 0.33, 1)
Age at the time of analysis	18-36 (0, 1)
Childish diseases	• Yes = 1 • No = 0
Accident or serious trauma	• Yes = 1 • No = 0
Surgical intervention	• Yes = 1 • No = 0
High fevers in the last year	• Less than three months ago = -1 • More than three months ago = 0 • No = 1
Frequency of alcohol consumption	• Several times a day = -1 • Every day = 0 • Several times a week = 1 • Once a week = 2 • Hardly ever or never = 3
Number of hours spent sitting per day	• 0 • 1
Diagnosis	• Normal • Altered

Table 8. Datasets

100 health individual data obtained from University of California Irvin (UCI) machine learning repository that were collected and shared in 2013 by the department of Biotechnology of University of Alicante. The data sets consist of 100 instances with 10 attributes (seasons in

which analysis was performed, age, childish disease, accidents or serious trauma, surgical intervention, high fever in the last one year, frequency of alcohol consumption, smoking habits, and number of hours spent sitting per day. Provides a semen sample that was analyzed according to World Health Organization (WHO) 2010 criteria sperm concentration is related to socio-demographic data, environmental factors, health status, and lifestyle habits.

System Architecture. It shows the structure and behavior of the system being developed. The three-tier application is used.

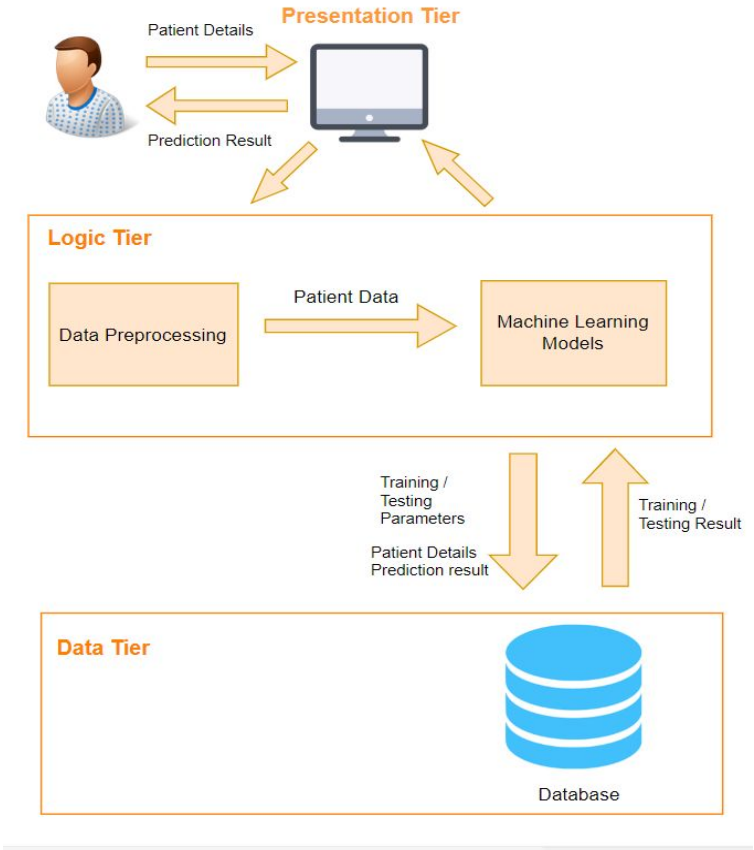

Figure 4. System Architecture

Presentation Tier. This tier displays information of patient and the prediction result

Logic Tier. This tier controls the functionality of the system by performing data preprocessing and data mining.

Data Tier. The data tier is represents the database where the patient details prediction result and training/testing parameters are stored and retrieved.

Definition of Terms

Couples. Two individuals who are considered together romantically or sexually.

Algorithm. It is a step by step instructions and is need to be done in order to accomplish task.

Doctors. Is a qualified practitioner of medicine; specifically a Reproductive Endocrinologist.

Website. It is a collection of publicly accessible, interlinked Web pages that share a single domain name.

Chapter 3

METHODOLOGY

The Chapter focuses on discussing the research design to be used in developing the application. This chapter also includes the different diagrams to present the workflow of the system and the software process model.

Project Design

The project design includes the three (3) diagrams to describe the system which includes Use Case Diagram, Context Diagram and Algorithm Design.

Use Case Diagram

The use case diagram presents the roles of what different users to the system.

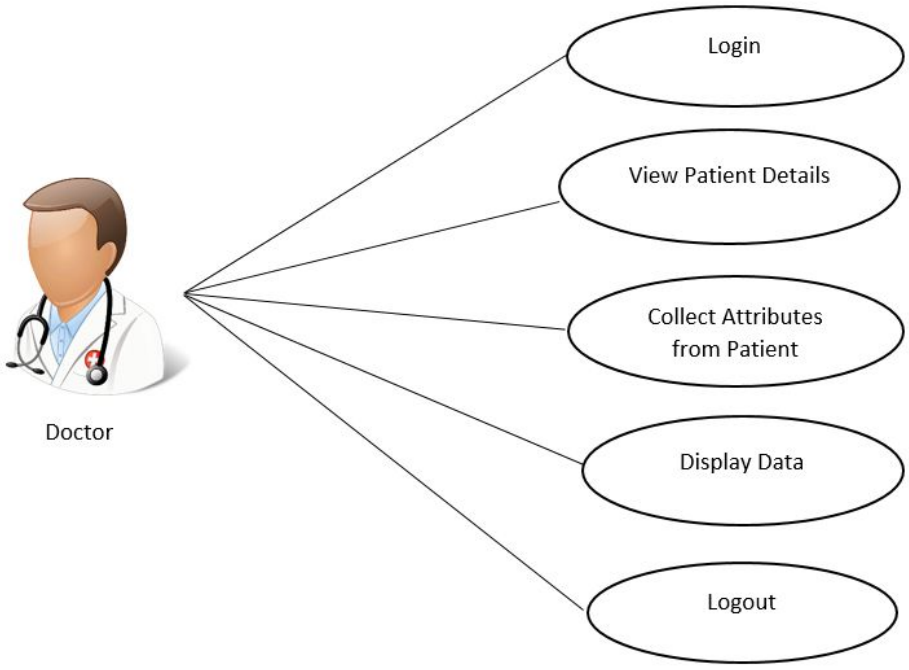

Figure 5. Use Case Diagram – Doctor

Figure 1 shows the capabilities of a doctor using the system. The Doctor should be able to register, login, view patient details, collect attributes, display data, and logout. Once a doctor

has logged in he/she can view the patient details and results. Once the result of a patient is altered, a doctor will give health recommendations that will help the patient with his treatment.

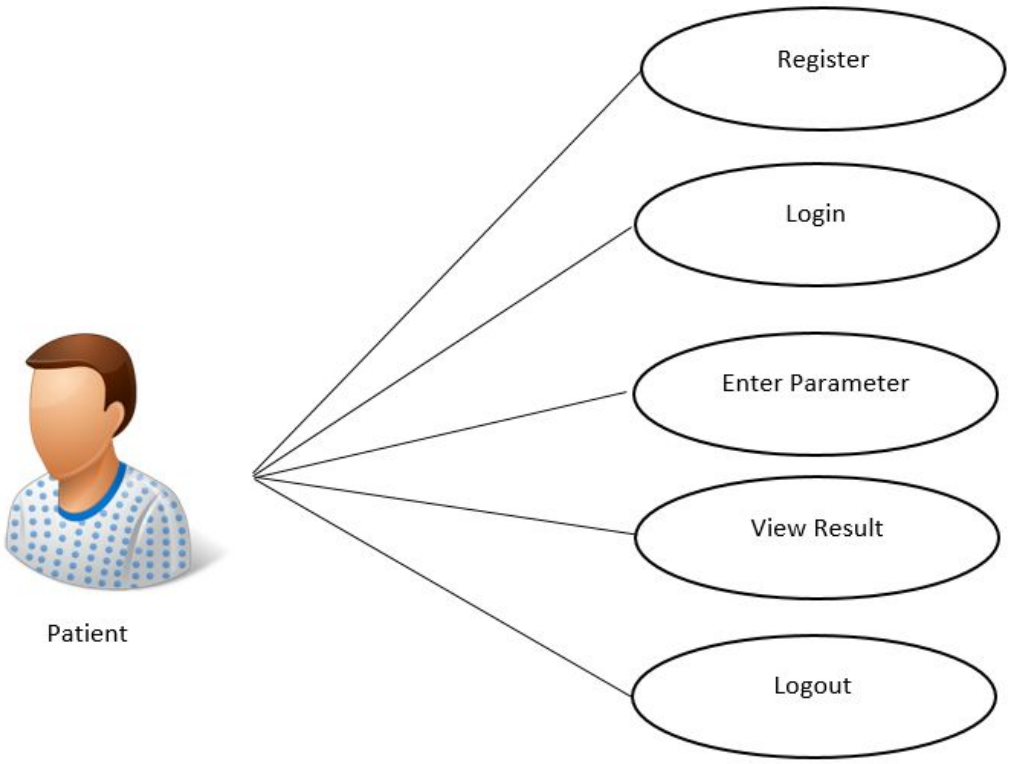

Figure 6. Use Case Diagram – Patient

Figure 2 shows what the patient can perform using the system. The patient should be able to Log in, Enter health details, view results and log out. After he has logged in, a patient should input or answer the following questions regarding his health background, lifestyle activities, and habits. This information will help the prediction of whether a patient has a normal or altered fertility.

Context Diagram. The following figure displays the relationship between the user, health centers, and the application.

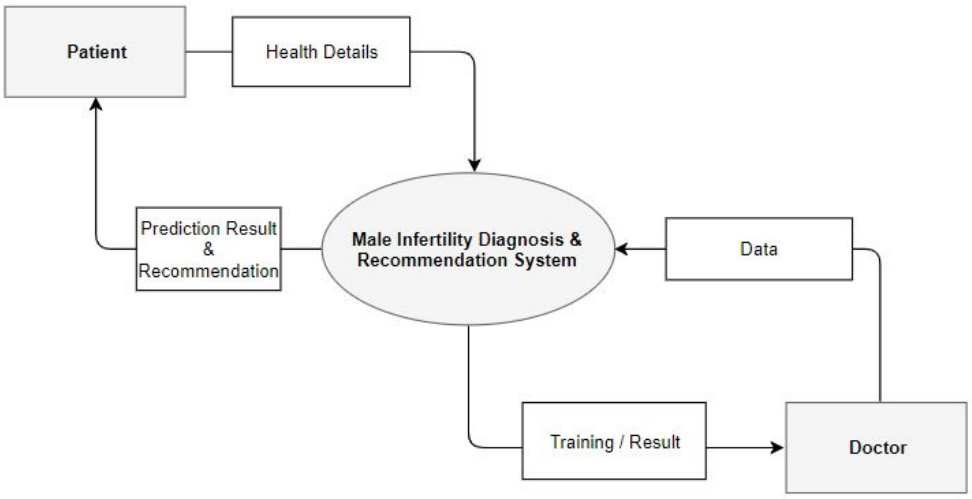

Figure 7. Context Diagram

Figure 3 shows how the patient of Male Infertility Diagnosis & Recommendation System can interact with the System. First, the patient inputs the symptoms required by the Male Infertility Diagnosis & Recommendation System . The data gathered will then be used to predict if the Patient is fertile or altered. The system will then provide the prediction result to the Patient. The data provided by the patient will be received by the doctor for evaluation. Once an evaluation is done, the doctor will send the recommendation back to our system.

Algorithm Design

The researchers have decided to use Decision Tree as the algorithm for the proposed system. Based on the tradeoffs Decision Tree has shown better accuracy and slow training time among the three. Decision trees can be visualized and are simple to understand and interpret it's require very little data preparation whereas other techniques often require data normalization, the creation of dummy variables and removal of blank values. The cost of using the tree (for predicting is logarithmic in the number of data points used to train the tree,

Performance Measure

Confusion matrix is a table with 4 different combinations of predicted and actual values that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

		Predicted	
		Negative	**Positive**
Actual	**Negative**	TN	FP
	Positive	FN	TP

Table 9. Confusion Matrix

True Negative (TN) : Correct predicted negative values which means that the value of actual class is no and value of predicted class is also no.

True Positive (TP) : Correct predicted values which means that the value of actual class is yes and the value of predicted value class is also yes.

False Positive (FP) (Type 1 Error): When actual class is no and predicted class is yes.

False Negative (FN) (Type 2 Error): When actual class is yes and predicted class is no.

Accuracy - It Refers to the total number of records that are correctly classified by the classifier.

Precision - is the fraction of retrieved instances that are relevant.

Recall - is the fraction of relevant instances that are retrieved.

F1 score - is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.

Project Development

In this section is the discussion about the project development of the system which includes the data gathering procedures to be conducted, testing and operation procedures, and the software model to be used.

Software Process Model

The software process model used in the study is Agile Development Model. Agile Development model is an approach to software development under which requirements and solutions evolve through the collaborative of self-organizing and cross-functional.

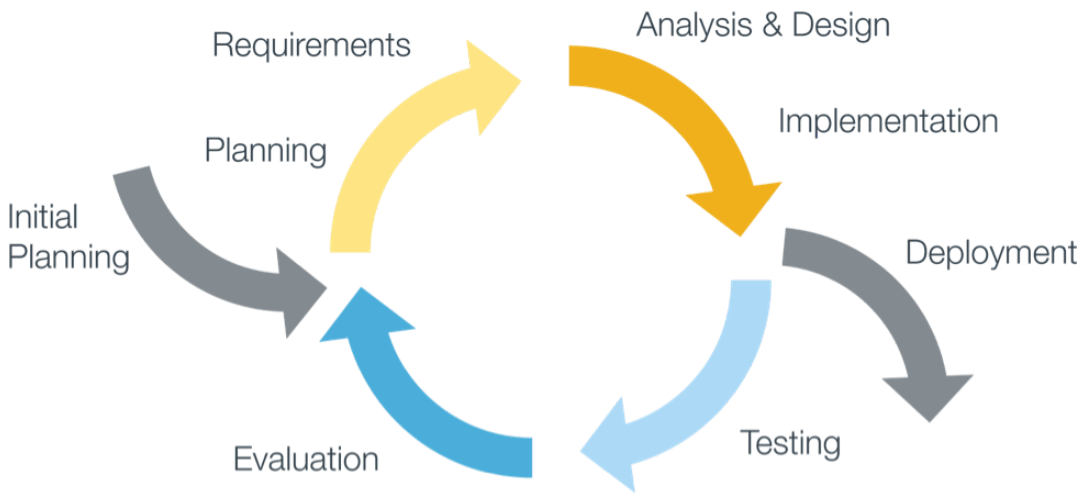

Figure 8. Agile Development Model

1st Iteration:

Planning. The researchers will analyze the gathered information, data and understand the flow of the system.

Implementation. The researchers will implement all the features, functionalities needs for diagnosing infertility.

Testing. The researcher will conduct testing to see if the system meet the all requirements needed for diagnosing infertility.

2nd Iteration:

Planning. The researchers will improve the features, functionalities and design of the system.

Implementation. The researchers will implement the improve features, functionalities and design

Testing. The researcher will conducting testing again to see the improvement of the system

3rd Iteration:

Planning. The different algorithms were compared to identify the best accuracy level in terms of prediction

Implementation. The researchers will further implement the best algorithm for the system.

Testing. The developers will perform evaluation and testing to have a better accuracy in doing the prediction.

Evaluation. The researchers conducted surveys for the technical users and non-technical users.

Testing and Operating Procedure

To determine the reliability of the software, the developers and quality assurance technicians will perform the testing of the software to find the errors and fix it before presenting it to the target users. There will be testing method that will be conducted.

Unit Testing. In unit testing, each part of the system must be tested, to know if every functions or features of the system works properly. Also the developer will be able to encounter some errors and bugs that must be fixed to improve the System

Integration Testing. In integration testing, This test checks whether the relationship between modules are well-related. Also it evaluates the communication between each modules and how they are related to each other.

System Testing. In system testing, the System will be tested especially the behavior, if the system crashes in a short period of time or having bugs and errors with the functions during the runtime process.

Table 10. Test Plan

Test No.	Module	Action	Expected Result	Pass?
1	User Registration	User creating an account and filling up the required information	User can successfully create or register account	
2	Login (admin)	Logging In	Should get to the Admin Panel	
3	Login (admin)	Filling up the required fields	System should validate if username and password matches, and if it is existing or non-existing	
4	Login(user/patient)	Logging In	Should get to the Patient Profile	
5	Login(user/patient)	Filling up the required fields	System should validate if username and password matches, and if it is existing or non-existing	
6	Admin Panel	View Patients information	Should get to the View Patients for viewing the correct data of patients info and prediction result	
7	Patient Profile	Send Health Details	System must allow the user to send a details that can help in predicting fertility	
8	Patient Profile	View result	Should get to the View Result module and can be able to view the prediction result	
9	Patient Profile	Edit Info	Should get to the Edit Info module and the system should allow the user to edit the information on their profile	

Project Evaluation

Stratified Sampling. The researchers divide the evaluators into separate groups, the technical and non-technical evaluator. From the formed group, the researcher will randomly select the sample proportionally and will draw the conclusion based on the gathered data from the survey.

ISO/IEC 25010:2011. This standard defines a product quality model composed of three characteristics that relate to static properties of software and dynamic properties of the computer system. The model is applicable to both computer systems and software products.

- **Accuracy.** This characteristic represents the degree to which a system provides accurate results, measurements and calculations
- **Performance Efficiency.** This characteristic represents the performance of a system which the processing and response time of a system meets the requirements based on its functions.
- **Usability.** The system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

For the decision criteria, the Likert's Scale was used in interpreting the survey results where five (6) is the highest and one (1) is the lowest rate.

The following statistical tools were used for the interpretation of data:

Mean. The mean formula is one of the most useful and widely used methods to find out the average in statistics. The mean can be computed by adding up all the numbers and dividing that sum by the total number of numbers.

$$X = Sfx\ N$$

Where: X = Arithmetic Mean

S f x = Sum of products of frequency by midpoint

N = Number of Respondents

Percentage. The percentage represents the portion of the item in a given category and is expressed in hundreds. The percentage of relative frequency will be computed as follows:

$$P = F\ N * 100$$

Where: P = Percentage

F = Frequency of the Responses

N = Total Number of Respondents

Respondents

Following are the persons who will be invited to evaluate the system that is being developed by the researchers.

Technical Evaluation

The system will be evaluated by technical experts such as Data Scientist who are capable in analytical and technical abilities to extract meanings or insights from massive data sets, Software Engineers who are knowledgeable of applying principles and techniques of engineering, mathematics, and computer science to the design, development, and testing of

software applications for computers. We need 20 Data Scientist and 20 Software Engineers to test and analyze the system.

Evaluation form based from ISO 25010:2011

Table 11. Technical User Evaluation form

Accuracy	6	5	4	3	2	1
The system is able to provide accurate result in classifying fertility						
The system shows proper choices from the questions/symptoms provided						
The system provides genuine recommendations for the user						
The questions provided are relevant in diagnosing fertility						
The system only provides recommendations when the patient is diagnosed as infertile						
Performance Efficiency	6	5	4	3	2	1
The system does not consume an abundant amount of power						
The system does not get errors						
The system performs based on its function						
The system can be learned in a short period of time						
The system reacts timely with no delays						
The system can operate with other						

applications running in the background						
Usability	6	5	4	3	2	1
The system provides user-friendly interfaces.						
The system is easy to operate.						
The system keeps the user against making errors.						
The system provides useful information that is relative to the need of the user						
The system provides texts that are easy to read and understand						

Table 12. Likert's Scale

Rating	Scale	Interpretation
6	**5.45 - 6.00**	**Strongly Agree**
5	**4.45 – 5.44**	**Agree**
4	**3.45 – 4.44**	**Slightly Agree**
3	**2.45 - 3.44**	**Slightly Disagree**
2	**1.45 – 2.44**	**Disagree**
1	**1.00 – 1.44**	**Strongly Disagree**

Table 12 shows the rating used in the questionnaire. It represents the scale and ratings towards the corresponding interpretations.

Work Plan

T#	TASKS	July-18				August-18				September-18				October-18				Assigned to
		Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	
	System Planning																	
1	Task Assignment and Scheduling																	All
2	Data Gathering																	All
3	Chapter 1																	Toledo, Ruferzo
4	Software Requirements Specification																	Miguel
	System Analysis and Design																	
5	Creation of Diagrams																	Miguel
6	Chapter2 and 3																	All
7	Oral Defense																	All
	System Development																	
8	Planning																	All
9	Requirements																	All
10	Implementation																	Miguel
11	Testing																	All
T#		November-18				December-18				January-19				February-19				Assigned to
		Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	
	System Development																	
8	Planning																	Miguel
9	Requirements																	Miguel
10	Implementation																	Miguel
11	Testing																	All
12	Chapter 4 and 5																	Toledo, Ruferzo
13	Final Presentation																	All

Table 12. Work Plan

Figure 6 shows how the tasks is distributed among the members and the scheduled frame used to reach the project within its deadline. The blue highlights cover System Planning which was finished at the month of July. The light blue green highlights covered System Analysis and Design which covered the months August and September. The other colors highlighted focuses in System development.

References

Priya, N. (2017). *Improve Machine Learning Results for semen analysis using ensemble meta classification*. SDNB Vaishnav College for Women.

Fabio, E., Mendoza-Palechor, & Marlon, Piñeres Melo. (2016). *Fertility Analysis Method Based on Supervised and Unsupervised Data Mining Techniques.* Universidad de la Costa, Colombia.

Roseline, O., Osaseri, & Agharese, R., Usiobaifo. (2016). *Predicting Male Fertility using Soft Computing Approach.* University of Benin, P.M.B. 1154, Benin City, Nigeria

Florence, Koskas, & Yoann, Buratti. (2014*). Semen fertility prediction based on lifestyle factors*. Stanford University.

Zarinara, Ali-Reza, & Zeraati, Hojjat. (2016*). Models Predicting Success of Infertility Treatment: A Systematic Review.* Tehran University of Medical Sciences.

Meera, D. & Nalini, Dr. C. (2018*). Breast cancer prediction system using Data mining methods.* Bharath University.

Dangare, s., Chaitrali, & Apte, S., Sulabha, PhD. (2012). *Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques.* Walchand Institute of Technology. Solapur, Maharashtra, India

Deyhoul, Narjes, & Mohamaddoost, Tina. (2017*). Infertility-Related Risk Factors: A Systematic Review.* International Journal of Women's Health and Reproduction Sciences.

Osaseri, Roseline. (2016). *Predicting Male Fertility Using Soft Computing Approach.* University of Benin, Nigeria, Benin City.

Sudha G, Reddy KS. (2013) *Causes of male infertility: a crosssectional study.* International Journal of Latest Research in Science and Technology.

Auger, J., Kunstmann, J. M., Czyglik, F., & Jouannet, P. (1995*). Decline in semen quality among fertile men in Paris during the past 20 years.* New England Journal of Medicine.

Carlsen, E., Giwercman, A., Keiding, N., & Skakkebæk, N. E. (1992). *Evidence for decreasing quality of semen during past 50 years.*

Gil, D., Girela, J. L., De Juan, J., Gomez-Torres, M. J.,& Johnsson, M. (2012). *Predicting seminal quality with artificial intelligence methods.* Expert Systems with Applications.

Sami N, Saeed AT, Wasim S, Saleem S. (2012). *Risk factors for secondary infertility among women in Karachi, Pakistan.*

Jaiswal D, Trivedi S, Agrawal NK, Singh K. (2015). *Association of polymorphism in cell death pathway gene FASLG with human male infertility.* Asian Pacific Journal of Reproduction.

Li C, Meng CX, Zhao WH, Lu HQ, Shi W, Zhang J. *Risk factors for ectopic pregnancy in women with planned pregnancy: a case-control study.* Eur J Obstet Gynecol Reprod Biol.

Johanes M, Marcus M, Marvin H. (2015). *A survey of the Application of Machine Learning in Decision Support Systems.* Association for Information Systems

Gil, D., Girela. (2016). *Machine Learning Repository: Fertility Dataset.* http://archive.ics.uci.edu/ml/datasets/Fertility

Jarad, A., Katkar, R., Shaikh, A. R., & Salve, A. (2015) *Intelligent Heart Disease Prediction System With MongoDB..* A International Journal of Emerging Trends & Technology in Computer Science, Volume 4, Issue 1

Jason R. Kovac, MD, PhD, FRCSC, Abhinav Khanna, MD, and Larry I. Lipshultz, MD (2015) *The Effects of Cigarette Smoking on Male Fertility*

Saaranen M, Suonio S, Kauhanen O, Saarikoski S. *Cigarette smoking and semen quality in men of reproductive age.* Andrologia. 1987;19(6):670–670

Jessie PN (2013), Psychosocial Aspects of Infertility, Indian Journal of Applied Research, 3 (8), 634-636.

Geneva, Swtizerland: World Health Organization (2010). *World Health Organization Laboratory Manual for the Examination and Processing of Human Semen.* 5.

MacLeod J, Wang Y. (1979) *Male fertility potential in terms of semen quality: a review of the past, a study of the present.* Fertil Steril.

Feixiang Huang, Shengyong Wang, Chien Chung Chan (2012) *Predicting disease by using data mining based on healthcare information system,* IEE.