



doi:10.5281/zenodo.18129642

C5.0 Decision Tree Classification Algorithm for Risk Prediction Among Pregnant Women in Ughelli North LGA of Delta State of Nigeria

Edafeajiroke Michael Favour¹ & Omoghenemuko Greg Imoniyovwe²

¹Department of Computer Science, University of Port Harcourt, Rivers State, Nigeria
michael.edafeajiroke@uniport.edu.ng

²Department of Computer Science, College of Education, Warri, Delta State, Nigeria

Corresponding Author: Omoghenemuko Greg Imoniyovwe
E- mail: omoghenemukog@mail.com

ABSTRACT

Complications during pregnancy are a significant concern for modern women, potentially leading to severe health issues and even the death of both mother and foetus. These complications have reached a staggering and unprecedented dimension. But if addressed through technological interventions in medical diagnostics, can greatly reduce maternal and fetal mortality in Nigeria. The Decision Tree Classification method, specifically the C5.0 Decision tree algorithm, is effective for classifying pregnancy-related data. This study focused on the importance of standardizing data parameters collected from a government general hospital in Ughelli North LGA of Delta State, Nigeria. By comparing the C5.0 classifier's performance on standardized and unstandardized datasets, it was found that the standardized dataset yielded higher classification accuracy (98.8701%) and faster training time (12.3237 seconds) compared to the unstandardized dataset (97.7401% accuracy and 14.0799 seconds training time). Therefore, standardization enhances the performance of the C5.0 classifier in predicting pregnancy complications.

Keywords: C5.0 Decision Tree Algorithm, risk prediction during pregnancy, pregnant women, classification algorithm, predicting pregnancy risk using C5.0

1 INTRODUCTION

Prediction of pregnancy is based on proper medical knowledge- a gynecologist who is an expert in this, can deploy the knowledge of pregnancy prediction to ameliorate complications among pregnant women and consequently reduce the astronomical rate of maternal and fetal mortality. Wu et al. (2024) and Lakshmi et al. (2016) state that complications in predicting pregnancy leads to increased rates of maternal and fetal mortality, and this has become a critical research focus, aiming to protect pregnant women in Ughelli North local government from potential deaths. Pregnancy, a delicate state for women, induces various health changes that can lead to acute complications, often resulting in maternal and fetal mortality in this region (Assaduzzaman et al., 2023). These complications are primarily caused by changes in physiological parameters (Macrohon et al., 2022). If unnoticed by medical practitioners or pregnant women, these changes can escalate emergencies (Islam et al., 2021). In emergencies, the health of the pregnant woman becomes difficult to normalize, often resulting in the loss of the mother, child, or both

(Aljameel et al., 2023). Each day, around 800 women die from pregnancy and childbirth-related causes, with approximately 99% of these deaths occurring in developing countries (Song, 2021). In 2013, 289,000 women died during and following pregnancy and childbirth, mostly in low-resource settings, and many of these deaths could have been prevented (James et al., 2022)

Preventing or controlling these complications is possible if pregnant women are warned before situations escalate, especially in remote areas like Ughelli North, which has limited maternity facilities and insufficient qualified staff (Inyang et al., 2020). Monitoring physiological parameters such as blood pressure, blood glucose level, and weight can help identify and prevent complications (Mecikalski et al., 2021).

These complications during pregnancy have become a significant problem for women, especially in the Ughelli metropolis of Delta State, Nigeria (Wu et al., 2024). In this region, it is challenging to find hospitals with fully qualified doctors to treat pregnant women, often necessitating transfers to central/general hospitals. Pregnant women frequently travel long distances to these hospitals, sometimes resulting in forced deliveries or severe complications, leading to many deaths. Aljameel et al. (2023) argue that pregnant women in this part of Nigeria need protection from stress and complications during pregnancy. The gestation period involves many physiological changes that can lead to severe health problems and the death of both mother and foetus. Assaduzzaman et al. (2023) note that technological interventions in medical diagnosis can significantly help address this problem, thereby reducing maternal and fetal mortalities. The Decision Tree Classification method is widely used in medical diagnosis due to its suitability Wu et al. (2024). Lakshmi et al. (2016) suggest that improvements on the C4.5 algorithm, to a C5.0 decision tree classifier, can be employed for better accuracy in predicting pregnancy risk levels. This study used a dataset collected from the Government Central/General Hospital in Ughelli, Delta State, Nigeria. The dataset was pre-processed and used to create a decision tree model for classification purposes. It aimed to classify and predict risk levels using both standardized and unstandardized databases, ensuring precise, valid, and reliable parameters for the study. Since the C5.0 algorithm, is a powerful Decision Tree Induction method, it was chosen for classification due to its superior performance in medical data classification (Zhang et al., 2021). This algorithm provided better results compared to other algorithms and has been found to outperform many others when applied to medical data. In other to achieve the study aim, acquired dataset was pre-processed using string to numeric conversion technique. Two-Way T-Test technique was employed as a features selection techniques on both the standardized and unstandardized. Optimal feature subset was used to build decision tree on both the standardized and unstandardized for classification process.

Purpose of the study

To investigate the efficacy of the C5.0 Decision Tree algorithm for classifying pregnancy-related data.

Scope of the study

The study covered C5.0 decision tree classification algorithm for risk prediction among pregnant women in Ugehelli North LGA of Delta State of Nigeria.

2 Literature Review

Wu et al. (2024) aimed to develop a machine learning-based model to predict miscarriage risk for patients with immune-abnormal pregnancies. By analyzing data from 565 patients obtained from electronic medical records, the study sought to identify high-risk individuals, facilitating proactive interventions to mitigate adverse pregnancy outcomes. The predictive model was built using the XGBoost algorithm. Performance metrics included the Area Under the Curve (AUC), accuracy, precision, and F1 score. SHAP (SHapley Additive exPlanations) analysis was used to identify significant influencing factors. The model achieved accuracy, precision, and F1 scores of 0.3009, 0.1663, and 0.2852, respectively. Economic evaluation revealed substantial cost savings amounting to ¥7,485,865.7 attributable to the model's implementation. But sample size, potential biases in the electronic medical record data, and the model's generalizability to broader populations beyond the study sample remains the drawback of this study.

Assaduzzaman et al. (2023) focused on addressing maternal health issues, particularly in rural areas, where challenges such as lack of doctors, infrastructure, and transportation contribute to high rates of maternal and infant mortality. The study aimed to identify primary maternal risk factors using machine learning models and to propose improved data pre-processing methods. The study employed several machine-learning algorithms, including Cat Boost, Random Forest, XGBoost, Decision Tree, and Gradient Boost. The researchers focused on enhancing data pre-processing to improve model performance. Among the algorithms tested, the Random Forest algorithm emerged as the best performer in identifying primary maternal risk factors. Key factors included chronic conditions, age, nutrition, and the availability of medical assistance. data quality, the representativeness of the data sample, and the generalizability of the findings to different rural settings or broader populations remain a challenge.

Aljameel et al. (2023) addressed the increasing risk factors for pregnancy loss and emphasized the need for early prediction of miscarriage through an intelligent automated solution using machine learning (ML). The study aimed to assist obstetricians in accurately diagnosing and predicting miscarriage probability. The researchers developed a model utilizing four ML classifiers: decision tree, random forest, k-nearest neighbour, and gradient boosting. These classifiers were used to predict the probability of miscarriage. The model achieved an accuracy of 93.4% and an ROC-AUC of 97%, demonstrating high effectiveness in early identification of at-risk pregnant women. Dependency on the quality and comprehensiveness of the data used, potential biases in the data, and the need for validation in broader and more diverse populations beyond the initial study sample, remains a challenge.

Macrohon et al. (2022) emphasized the importance of early risk tagging for maternal health, particularly in regions with high fertility rates like the Philippines, where awareness of risks can greatly influence pregnancy outcomes and maternal mortality rates. The study aimed to compare various supervised machine learning algorithms to accurately predict high-risk pregnancies with limited data from the municipality of Daraga in Albay. Multiple supervised machine learning algorithms were compared to predict high-risk pregnancies. The Decision Tree algorithm emerged as the most accurate, achieving a test score of 93.70%. Additionally, a semi-supervised approach using a Self-Training model was applied to a modified version of the Decision Tree. The Decision Tree algorithm demonstrated a high level of accuracy, with a test score of 93.70%. After applying the semi-supervised approach using the Self-Training model to the modified Decision Tree, the accuracy rate improved to 97.01%. the generalizability of the findings beyond the specific municipality of Daraga in Albay, as well as the reliance on limited data for model development and evaluation. Further validation in diverse settings and with larger datasets may be necessary to confirm the robustness of the approach, remains challenging.

Islam et al. (2021) underscored the importance of early risk tagging for maternal health, particularly in regions with high fertility rates like the Philippines, where awareness of risks can greatly influence pregnancy outcomes and maternal mortality rates. The study aimed to compare various supervised machine learning algorithms to accurately predict high-risk pregnancies using limited data from the municipality of Daraga in Albay. Multiple supervised machine learning algorithms were compared to predict high-risk pregnancies. The Decision Tree algorithm emerged as the most accurate, achieving a test score of 93.70%. Additionally, a semi-supervised approach using a Self-Training model was applied to a modified version of the Decision Tree. The Decision Tree algorithm demonstrated a high level of accuracy, with a test score of 93.70%. After applying the semi-supervised approach using the Self-Training model to the modified Decision Tree, the accuracy rate improved to 97.01%. However, the representativeness of the data, the generalizability of the findings to other regions or populations, and the need for further validation in diverse healthcare settings remain a challenge.

Song et al. (2021) proposed an interpretable knowledge-based decision support system (IKBDSS) to help physicians predict disease risk levels during pregnancy. The system integrates historical cases and expert opinions, using the Multi-granularity Linguistic Term Sets (MLTS) model to address ambiguity. The system focuses on knowledge acquisition, similarity degree calculation, and consistency checking. It improves specificity, sensitivity, and F1 score compared to other methods. The decision-making process produces interpretability, increasing system reliability.

Inyang et al. (2020) investigated the performance of six classifiers and the effect of data balancing and formation approaches for predicting pregnancy outcomes. Synthetic minority oversampling technique (SMOTE) was used for data imbalance treatment. Random forest (RF) was evaluated on resampled datasets with four test modes. Results showed significant variation in mean performance across datasets, with RF being the best classifier-DB method pair. Train/test data modes insignificantly affected classification accuracy, but computational cost variations were noticeable.

Abbas et al. (2020) highlighted the utility of data mining and machine learning algorithms in healthcare, particularly for developing decision support systems (DSS), analyzing clinical factors, extracting valuable insights from historical data, and making predictions. The study aimed to utilize machine learning for classifying birth data using bagging and boosting algorithms. The researchers conducted a thorough comparison of these algorithms using birth data obtained from government hospitals in Muzaffarabad, Kashmir. The study focused on comparing bagging and boosting algorithms for classifying birth data. Bagging algorithms like Adabag and BagFda were examined in detail. The results indicated that bagging algorithms, particularly Adabag and BagFda, demonstrated slightly superior performance in terms of accuracy, precision, and recall compared to previous studies. potential constraints may include the representativeness of the data, the generalizability of the findings to other regions or populations, and the need for further validation in diverse healthcare settings.

3 METHODOLOGY

Design of the study

The design approach adopted in the study was the machine learning design approach- which used experimentation solution framework to the problem specified by the requirements of the proposed system. The experiments were designed to evaluate the classification of the pregnancy dataset using a 2-way T-Test filter selection technique extraction and C5.0 decision tree classifier to classify standardized and unstandardized datasets. The system followed a modeling and developing phase to validate the integrity and efficiency of the developed model/system. The implementation of the development was segmented into five stages. In the first stage, a dataset of pregnancy on different pregnant patients in the clinic was collected from Government Central/General Hospital Ughelli, Delta State, Nigeria. The second stage was to pre-process the dataset and also filter it. The third stage involved the feature selection with the 2-way T-Test statistical algorithm. Dataset was partitioned into training, testing, and classification. The final stage involved results evaluation of the developed system as depicted in see Figure 3.1.

Data collection

The data set was collected at the Government Hospital in Ughelli Delta State. The dataset consisted of twelve attributes and one class label and 237 instances. The class label are grouped into low risk, medium risk and high risk. The description of the dataset is shown in Table 3.1.

Table 3.1

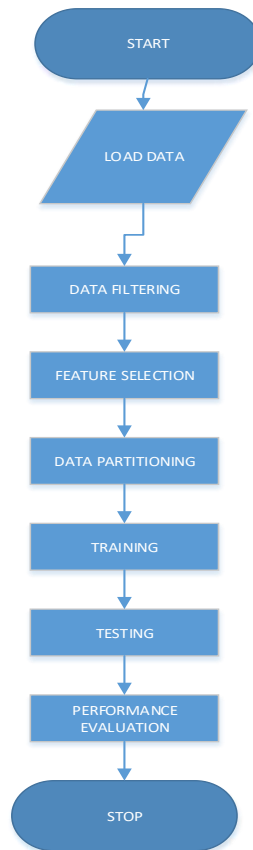
Dataset description

S/N	Variable	Description	Domain
1.	Age	Age of the pregnant women in years	Numeric
2.	Pregnancy parity	Present pregnancy number	Textual
3.	History of Eclampsia	Did the pregnant woman experience Pre-eclampsia state in previous pregnancies? (Yes/No)	Textual
4.	Mother/Sister Preclampsia	Did the pregnant woman's mother/sister experience Pre-preclampsia state in any of their pregnancies? (Yes/No)	Textual
5.	Mother/Sister Gestational Diabetes	Did the pregnant woman's mother/sister experience gestational? (Yes/No)	Textual
6.	Health State	The present health status of the pregnant woman (Hypertensive/Normal/Overweight/Underweight)	Textual
7.	Trimester()	Present trimester of pregnancy (2/3)	Numeric
8.	Present Month	Present month of pregnancy (4/5/6/7/8/9)	Numeric
9.	Blood Pressure	Blood Pressure recorded presently in mm/Hg	Numeric
10.	Presence of Gestational diabetes	Blood Sugar Levels recorded presently in mg/dl	Numeric
11.	Weight	Present weight gain from the previous month in Kgs	Numeric
12.	Class Group	Low Risk, Medium Risk, High Risk	Textual

The model for the developed system is as shown in figure 3.1. Inconsistency in the acquired dataset was removed and converted into numeric variables for proper labelling. A 2-way T-Test filter technique was then used to select features optimal features from the dataset. The 2-way t-test works by finding the significant difference between the predictors and the class label. The significance level is set at a 95% confidence level. After the feature selection, the data was partitioned into the training and testing set to create a robust knowledge discovery for the C5.0 decision tree classifier and validate the classifier. The data was grouped into training and testing samples at a ratio of 75% to 25%. After data portioning into the training and validating set, the training set was introduced into the C5.0 decision tree classifier for pattern

discovery from the selected features, this helps to create an experimental self-learning of the dataset. The standardized data and unstandardized data were introduced to the classifier for system experimentation; to determine the classification accuracy and predictive capability of the classifier. The model was validated with 25% of the dataset that was held out during data partitioning. This helped to determine the performance of the model and the classification accuracy. The developed model was evaluated using machine learning statistical parameters and was implemented on MATLAB R2016a (9.0.0.341360).

Figure 3.1
Flowchart of risk prediction model for the developed system



4 RESULTS AND DISCUSSIONS

System experimental setup

The developed system examined and tested the application of classification algorithms for risk prediction during pregnancy. The algorithms used were the Statistical T-Test and C5.0, applied to both standardized and unstandardized datasets. Data was collected from a domain expert at the Government Central/General Hospital in Ughelli, Delta State, Nigeria, and then pre-processed and filtered for experimentation. The experimental approach included data collection, processing, filtering, feature selection with a two-way t-test, data partitioning, and classification using the C5.0 decision tree classifier. The experiments were conducted using Matlab (MATLAB 2016A), where various functions were developed and linked to a graphical user interface for user interaction. The system utilized multiple components within Matlab to develop and present the results of the data mining tasks.

4.3 Dataset Settings

The dataset was obtained from an expert at Government General Hospital Ughelli, Delta State, Nigeria, and contains 237 instances, 12 attributes, and one class label. Table 4.1 details the names of the 12 attributes, which are the predicting variables, and the class label, which is the response variable.

Table: 4.1
Dataset attributes

Attributes
Age
Pregnancy parity
History of eclampsia
History of gestational diabetes
Mother/sister preeclampsia
Gestational diabetes
Health state
Trimester ()
Present
Blood pressure
Presence of gestational diabetes
Weight
Class(low Risk, medium Risk, High Risk)

4.4 Interactive Developmental Stage

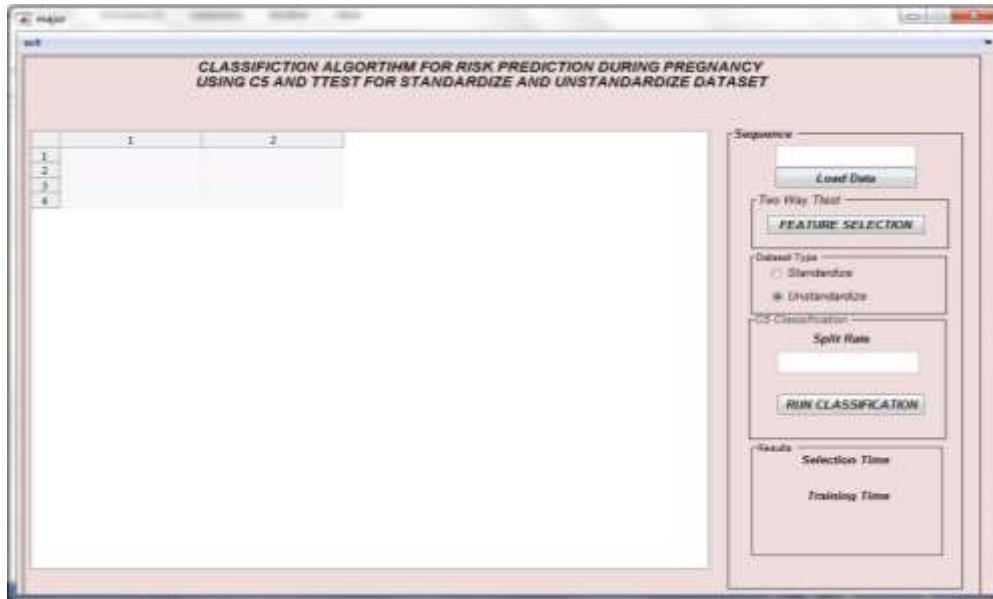
The developed system at run time is shown in the figure 4.3. The combined model for risk classification for pregnant women follows the experimental procedure below.

1. Dataset Filtering
2. Dataset Feature Selection
3. Dataset Partitioning
4. Classification

Performance evaluation

Figure 4.3 describe the outlook of the initial start up the application at run time gives a platform to automate all the working process of developed platform.

Figure 4.3
Initial start-up



4.5 PRESENTATION OF RESULTS

This section presents the stepwise results for the classification of pregnancy risk in women.

i. Data Filtering

The filtered data helps to present well-formatted data into the system the data was filtered by converting string variables to numeric variables. A sample case study is shown in Figure 4.4.

ii. Loading of the Dataset

Figure 4.4 shows the loading of the predicting variables which are also the independent variables that gives precise information about the target which is also the response variable. A total number of 237 instances, 12 attributes, and one class label was loaded into the system for evaluation.

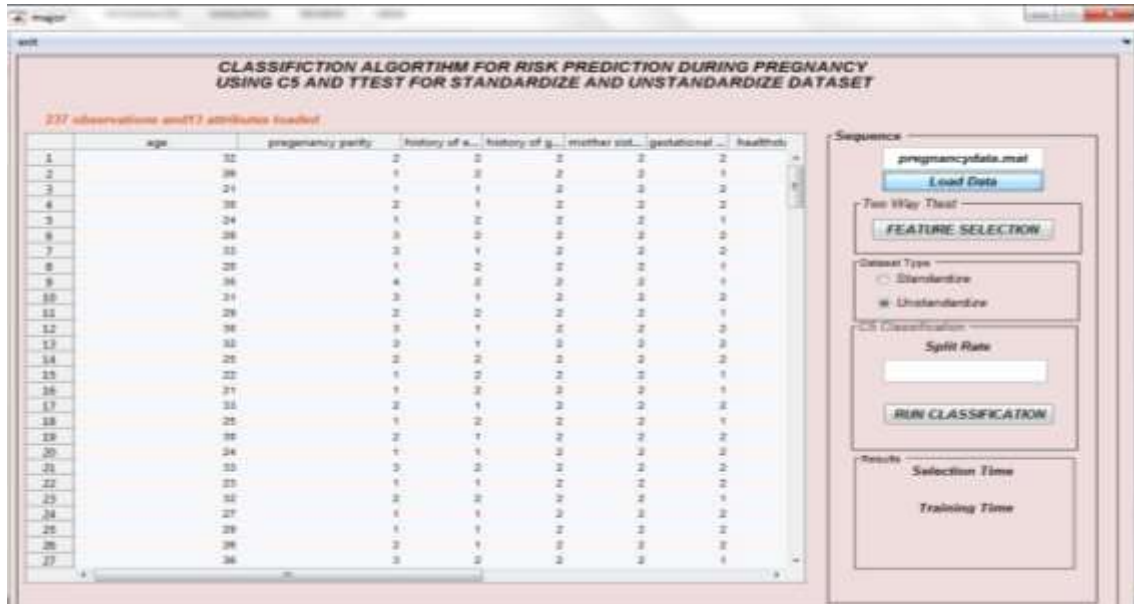
Figure 4.4
Dataset normalization

	A	B	C	D	E	F	G	H	I	J	K	L
1	age	pregnancy parity	history of eclampsia	history of gestational diabetes	mother/sister preclampsia	gestational diabetes	health sta	trimester	present	blood pre	presence	weight
2	32	2	2		2	2	1	8	2	1	9.3	11.2
3	26	1	2		2	2	1	9	2	1	8.5	11.07
4	21	1	1		3	2	1	6	1	1	10.1	10.25
5	30	2	1		2	2	1	11	2	1	9.4	11.1
6	24	1	2		2	1	1	9	2	1	8.4	10.7
7	28	3	2		2	2	1	12	1	1	9.3	9.25
8	33	3	1		2	2	1	11	2	1	11.1	11.1
9	25	1	2		2	1	1	8	1	1	9.3	11.05
10	35	4	2		2	1	1	12	1	1	8.5	11.2
11	31	3	1		2	2	1	11	2	1	10.2	9.8
12	29	2	2		2	1	1	10	2	1	8.3	11.2
13	30	3	1		3	2	1	8	2	1	10.5	9.8
14	32	3	1		2	2	1	11	2	1	9.3	11.01
15	25	2	2		2	2	1	10	2	1	8.1	9.2
16	22	1	2		2	1	1	8	2	1	9.3	11.2
17	21	1	2		2	1	1	12	1	1	8.5	11.2
18	33	2	1		2	2	1	9	2	1	7.5	10.3
19	25	1	2		2	1	1	8	2	1	10.2	10.5
20	30	2	1		2	2	1	7	2	1	9.2	11.2
21	24	1	1		2	2	1	10	1	1	11.1	8.9
22	33	3	2		2	2	1	11	1	1	9.4	11.2
23	23	1	1		2	2	1	4	1	1	9.3	11.2
24	26	2	2		2	1	1	8	2	1	8.4	10.7

iii. Loading of the Dataset

Figure 4.5 shows the loading of the predicting variables which are also the independent variables that give precise information about the target which is also the response variable. A total number of 237 instances, 12 attributes, and one class label were loaded into the system for evaluation.

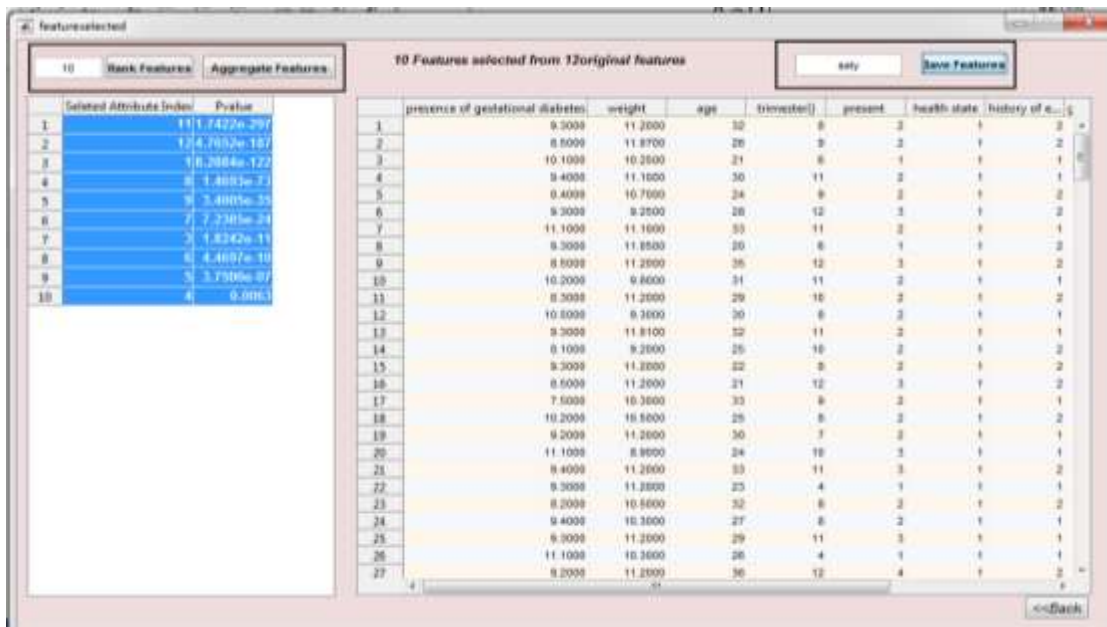
Figure 4.5
Loading of dataset into the application



iv. Feature selection

The two-sample test was carried out to find the relationship between the predicting variables and the response variables, each of the predicting variables was compared to the response variable to check their level of significance, and the level of significance was set at 0.05 confidence interval level. A total of ten attributes were selected from the main 12 attributes see Figure 4.6.

Figure 4.6
Feature selected



The selected features with their corresponding ranking ability of the 2-way test value are shown in Table 4.2.

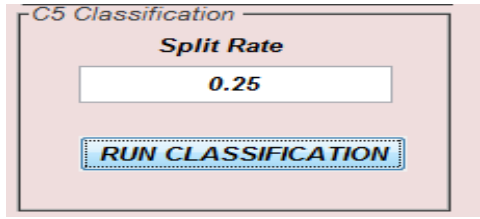
Table: 4.2 Selected Features

Attributes Selected	P-value
11	1.74E-297
12	4.77E-187
1	8.29E-122
8	1.47E-73
9	3.40E-35
7	7.24E-24
3	1.82E-11
6	4.47E-10
5	3.75E-07
4	0.0063

v. Classification Results

The C5.0 decision tree was used to classify both standardized and unstandardized data. Data standardization involved rescaling attributes to have a mean of 0 and a standard deviation of 1, assuming a Gaussian distribution. Before classification, the data was partitioned into training and testing sets in a 75% to 25% ratio. See Figure 4.7.

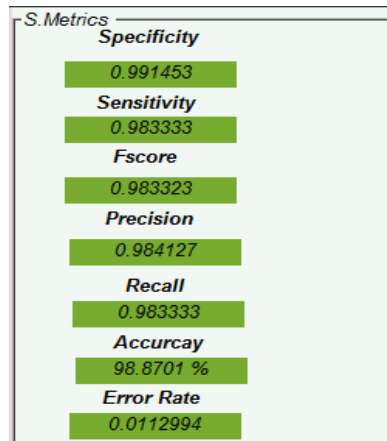
Figure 4.7
C5.0 standardize and unstandardized classification



vi. System evaluation
Standardized dataset

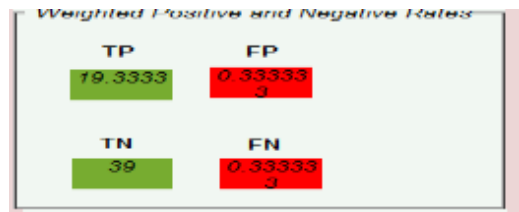
The classification results for the standardized dataset are Sensitivity (true positive rate or recall) measures the proportion of correctly identified positives. Specificity (true negative rate) measures the proportion of correctly identified negatives. F Score (or F Measure) is calculated as $2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$, indicating the balance between precision and recall. The standardized results are illustrated in Figure 4.8

Figure 4.8 System evaluation



The true positive rate shows the Positive and Negative class correctly identified as Positive and Negative class and the Positive and Negative class incorrectly identified Positive and Negative class. See Figure 4.9.

Figure 4.9 System evaluation

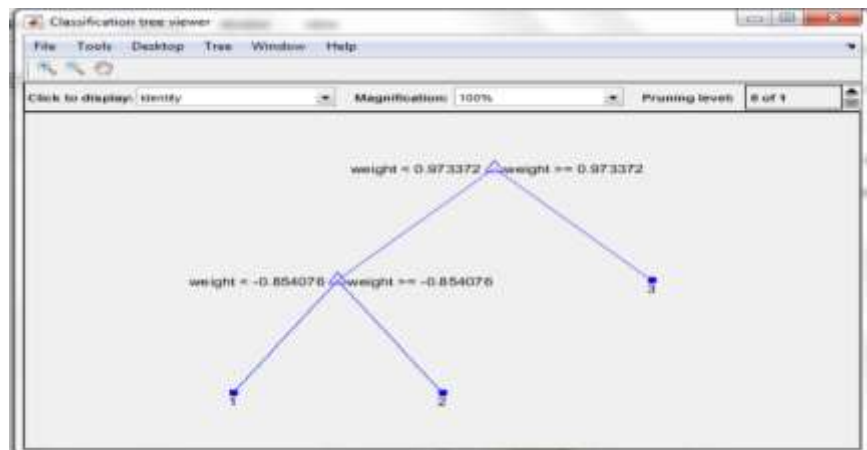


The confusion matrix is an indication of the correctly and incorrectly classified class. The class 1 which represents the low risk shows a total of 20 observations for the validation set. A total of 19 were classified as low risk and 1 was classified as medium, while the class 2 represents the medium risk, which shows a total of 20 observations for the validation set. 20 were classified correctly and none as incorrect. While the class 3 represents high risk with a total of 20 observations, a total of 20 were classified as high risk with none as misclassified. The confusion matrix obtained is shown in Figure 4.10. The pruned tree is described in Figure 4.11.

Figure 4.10 Confusion matrix

	1	2	3
1	19	0	0
2	0	20	0
3	0	1	19

Figure 4.11 Pruned Tree



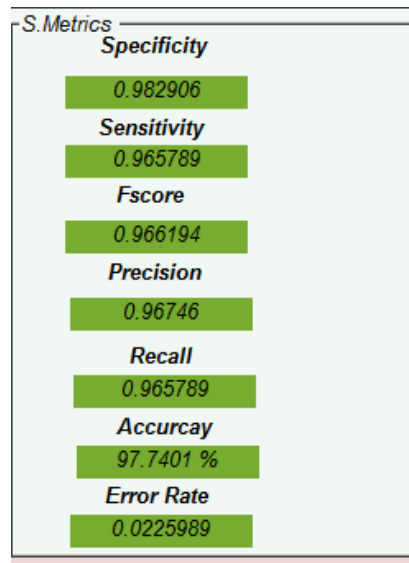
Unstandardized dataset

The classification results for the unstandardized dataset are Sensitivity (true positive rate or recall) measures the proportion of correctly identified positives. Specificity (true negative rate) measures the proportion of correctly identified negatives.

F Score (or F Measure) is calculated as:

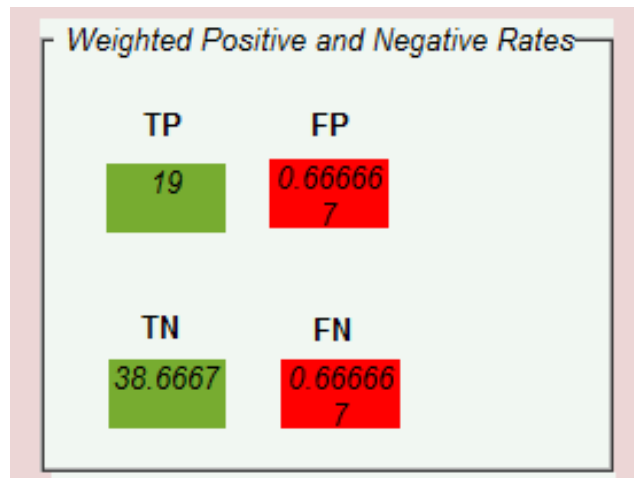
$2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$, indicating the balance between precision and recall. The unstandardized results are illustrated in Figure 4.12.

Figure 4.12
System evaluation



The true positive rate shows the Positive and Negative class correctly identified as the Positive and Negative class and the Positive and Negative class incorrectly identified as the Positive and Negative class. See Figure 4.13.

Figure 4.13
System evaluation



Confusion Matrix

The confusion matrix is an indication of the correctly and incorrectly classified class. Class 1 which represents the low risk shows a total observation of 20 observations for the validation set. A total of 20

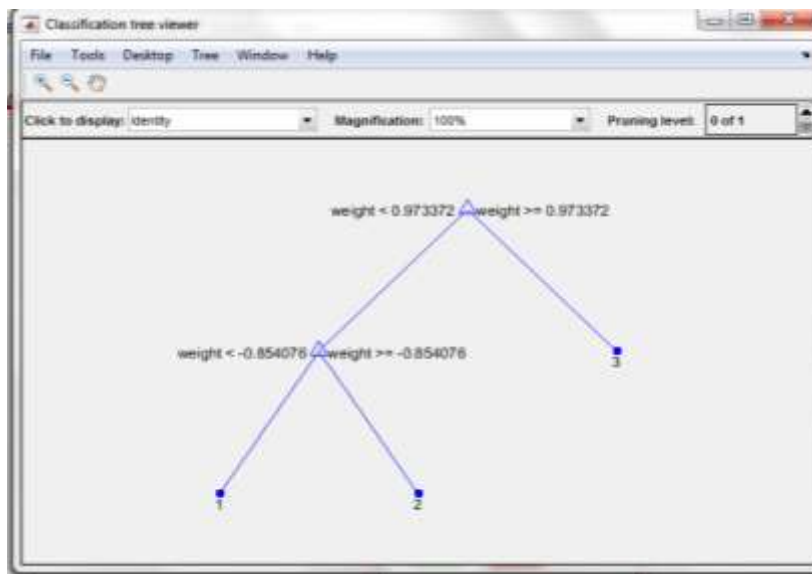
were classified as low risk and 0 was classified as medium, while class 2 represents the medium risk, which shows a total of 20 observations for the validation set, 19 were classified correctly and none as incorrect. While class 3 represents high risk with a total of 19 observations, a total of 18 were classified as high risk and one was classified as misclassified. This is depicted in Figure 4.14. The pruned tree is explained in Figure 4.15.

Figure 4.14
Confusion matrix

	1	2	3
1	20	0	0
2	1	19	0
3	0	1	18

Figure 4.15
Confusion matrix

Pruned tree



The result obtained for both the standardized and Unstandardized Statistical Metrics is as explained in the Table 4.3.

Table 4.3
Standardized and unstandardized statistical metrics

Classification Metrics	Standardized Results	Unstandardized Results
True Positive Rate	19.3333	19
False Positive Rate	0.3333	0.6667
True Negative Rate	39	38.667
False Negative Rate	0.3333	0.6667
SENSITIVITY = True positive rate = TP / (TP + FN)	0.9833	0.9658
SPECIFICITY = True negative rate = TN / (TN + FP)	0.9915	0.9829
ACCURACY = TP + TN / (FP + FN + TP +TN)	98.8701 %	97.7401 %
RECALL	0.9833	0.9658
FSCORE	0.9833	0.965789
TRAINING TIME	12.3237secs	14.0799secs

Figure 4.16 describes the graphical illustration for the classification accuracy. It shows the correct classification and misclassification rate attained by the C5.0 algorithm for both the standardized and unstandardized datasets, which is an indication of the system performance.

Figure 4.16
Classification accuracy

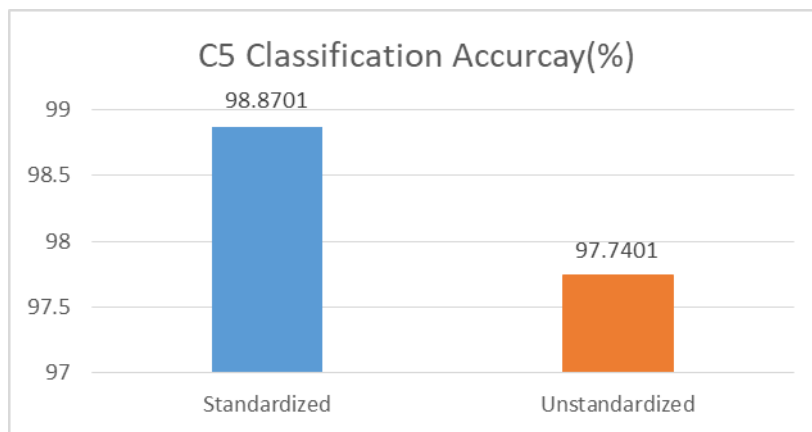


Figure 4.17 describes the graphical illustration for the sensitivity and specificity of both the standardized and unstandardized datasets

Figure 4.17
Classification accuracy

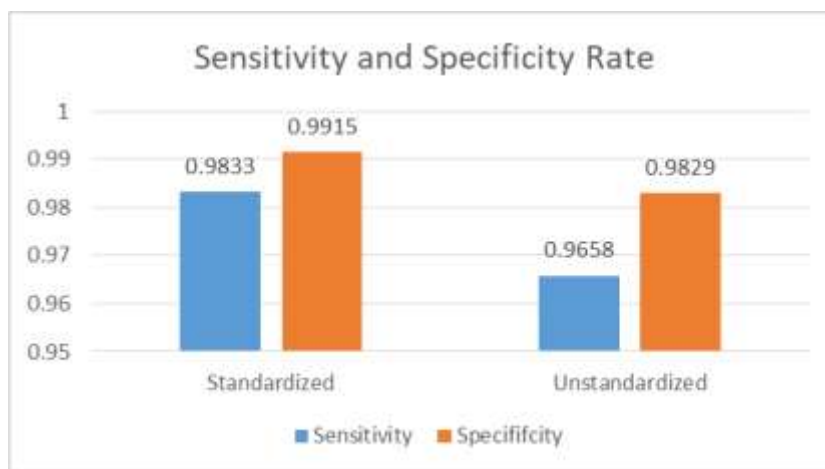
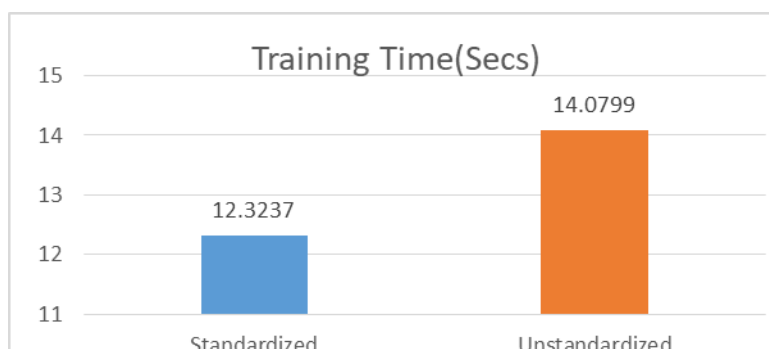


Figure 4.18 shows the training time that is the time taken by the model to create a knowledge retention of the data supplied to the classifier. The training time is taken for both the standardized and unstandardized datasets.

Figure 4.18 Classification training time



5 CONCLUSION

The results indicate that the C5.0 decision tree classifier exhibits higher accuracy in predicting risk levels during pregnancy when applied to standardized data compared to unstandardized data. The graphical representations establishes that the C5.0 classifier performs better with standardized datasets than with unstandardized datasets. Consequently, the C5.0 classifier demonstrates superior performance when standardization is implemented before analysis, as evidenced by results from both datasets in this study. This study evaluated the C5.0 classifier's accuracy and highlighted the impact of standardization on accuracy, aiming to improve prediction accuracy for pregnancy risk. While other classification techniques could be explored for analyzing pregnancy data, the C5.0 classifier was chosen for its potency, popularity, efficiency, and relevance to the intricacies of the pregnancy debacles.

5.1 Future Research Direction

Further research could enhance accuracy even more post-standardization, providing pregnant women with precise risk levels for a safer and healthier pregnancy period.

5.2 RECOMMENDATIONS

This research is highly recommended to medical doctors in public and private hospitals for their effective medical practice because the findings of the research highlight the potential for improving pregnancy risk prediction accuracy through a combination of standardized data pre-processing and thoughtful selection of classification techniques. Future research endeavours should be embarked on; to advance the field of predictive analytics in maternal healthcare, ultimately leading to better healthcare outcomes for pregnant women and their infants globally.

REFERENCES

- Abbas, S. A., Rehman, A. U., Majeed, F., Majid, A., Malik, M. S. A., Kazmi, Z. H., & Zafar, S. (2020). Performance Analysis of Classification Algorithms on Birth Dataset. *IEEE Access*, 8, 102146–102154. <https://doi.org/10.1109/access.2020.2999899>
- Aljameel, S. S., Aljabri, M., Aslam, N., Alomari, D. M., Alyahya, A., Alfaris, S., Balharith, M., Abahussain, H., Boujlea, D., & Alsulmi, E. S. (2023). An automated system for early prediction of miscarriage in the first trimester using machine learning. *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, 75(1), 1291–1304. <https://doi.org/10.32604/cmc.2023.03571>
- Assaduzzaman; M. D. Abdullah A. M.; Zahid H. M. (2023). Early Prediction of Maternal Health Risk Factors Using Machine Learning Techniques. *International Conference for Advancement in Technology (ICONAT)*. DOI: 10.1109/ICONAT57137.2023.100807 00
- Breiman, L., Friedman, J., Olsen, R., and Stone, C. (1984). A Classification and Regression Trees, *Wadsworth International Group*.
- Inyang, U. G., Francis, B., Imo, J., Adenrele, A., & Chukwudi, O. (2020). Comparative analytics of classifiers on resampled datasets for pregnancy outcome prediction. *International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications*, 11(6). <https://doi.org/10.14569/ijacsa.2020.0110662>.
- Islam, M. N., Mahmud, T., Khan, N. I., Mustafina, S. N., & Islam, A. K. M. N. (2021). Exploring machine learning algorithms to find the best features for predicting modes of childbirth. *IEEE Access*, 9, 1680–1692. <https://doi.org/10.1109/access.2020.3045469>
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2022). An Introduction to Statistical Learning with Applications in R, Springer New York, ISBN:978-1-4614-7137-0.
- Lakshmi, B. S., Indumathi, T. S., & Ravi, N. (2016). A study on C5.0 Decision Tree Classification algorithm for risk predictions during pregnancy. *Procedia Technology*, 24, 1542–1549. <https://doi.org/10.1016/j.protcy.2016.05.128>
- Macrohon, J. J. E., Villavicencio, C. N., Inbaraj, X. A., & Jeng, J. (2022). A Semi-Supervised Machine Learning approach in predicting High-Risk pregnancies in the Philippines. *Diagnostics*, 12(11), 2782. <https://doi.org/10.3390/diagnostics121127>
- Mecikalski, J. R., Sandmal, T. N., Murillo, E. M., Homeyer, C. R., Bedka, K. M., Apke, J. M., & Jewett, C. P. (2021). A random-forest model to assess predictor importance and nowcast severe storms using high-resolution radar goes satellite lightning observations. *Monthly Weather Review*, 149(6), 1725–1746. <https://doi.org/10.1175/MWR-D-19-0274.1>
- Song, K., Zeng, X., Zhang, Y., De Jonckheere, J., Yuan, X., & Koehl, L. (2021). An interpretable knowledge-based decision support system and its applications in pregnancy diagnosis. *Knowledge-based Systems*, 221, 106835. <https://doi.org/10.1016/j.knosys.2021.106835>.
- Wu, Y., Yu, X., Li, M., Zhu, J., Yue, J., Wang, Y., Man, Y., Zhou, C., Tong, R., & Wu, X. (2024). Risk prediction model based on machine learning for predicting miscarriage among pregnant patients with immune abnormalities. *Frontiers in Pharmacology*, 15. <https://doi.org/10.3389/fphar.2024.1366529>
- Zhang, X., Yang, Z., & Cordes, D. (2021). A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping*, 41(13), 3807–3833. <https://doi.org/10.1002/HBM.25090>