Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766235

META-ANALYSIS OF FEATURE SELECTION FOR PREDICTION OF GESTATIONAL DIABETES MELLITUS

Dr. Ranjit M. Gawande[1], Dr. Varsha H. Patil[2], Dr. Swati A. Bhavsar[3], Dr. Neeta A. Deshpande[4]

*[1] Asst. Prof. Department of Computer Engineering Matoshri College of Engineering & Research Center Nashik (M.S.) ranjitgawande@gmail.com, *[2] Vice-Principal & HOD Department of Computer Engineering Matoshri College of Engineering & Research Center Nashik (M.S.) varsha.patil@gmail.com, *[3] Asso. Prof. Department of Computer Engineering Matoshri College of Engineering & Research Center Nashik (M.S.) bhavsar.swati4@gmail.com, *[4] Asso. Prof. Department of Computer Engineering GES R.H. S. College of Engineering Nashik (M.S.)

Abstract

Gestational diabetes mellitus (GDM) is a major disease affecting pregnant women. Screening for GDM and applying adequate interventions may reduce the risk of adverse outcomes. The diagnosis of GDM depends on OGTT tests performed in the late second trimester. The goal of this study was to create a hybrid model for the prediction of GDM in early pregnancy in women using a machine learning algorithm using polling-based feature selection techniques.

Methods: Data on 1725 pregnant women in early gestation were used to fit the GDM risk-prediction model. Predictive maternal factors were selected through the poling method of the feature selection model. Predictive maternal factors were selected through the poling method of feature selection. Incorporated selected maternal factors into a modified Naive Bayes and decision tree. The area under the receiver operating characteristic curve (AUC) was used to assess discrimination.

Results: The risk of GDM could be predicted with OGTT zero min, C-peptide HOMA, maternal age, prepregnancy body mass index (BMI), parity, and offspring birth weight with a predictive accuracy of 87.92% and an AUC of 0.766 (95% CI 0.731, 0.801).

Conclusions: This GDM prediction model is potentially applicable to alternative decision support systems and women who plan to conceive a baby.

Introduction:

Gestational Diabetes Mellitus (GDM) is generally defined as "a condition in which a hormone made by the placenta prevents the body from using insulin effectively" [1]. The risk of GDM is increased with overweight and obesity, of which the global prevalence has increased substantially in the past decades [2]. GDM increases the risk of many maternal and neonatal complications such as gestational hypertension, polyhydramnios, Cesarean birth, premature delivery, large for gestational age, neonatal macrosomia, intensive care unit admission, hypoglycemia, and respiratory distress [3]. Moreover, GDM may predispose to long-term sequela for both mother and child including metabolic syndrome and type 2 diabetes mellitus, thus increasing later life chronic disease [4].

Research shows that interventions initiated early in pregnancy can reduce the rate of GDM in pregnant women with overweight or obesity [18]. However, applying interventions in every instance can be costly and time-consuming. A hybrid decision support system (HDSS) based

on machine learning and data mining can be helpful in providing a powerful computerized tool to assist clinicians in identifying women at risk of GDM. It would largely reduce the time and cost by allowing targeted intervention. The HDSS has great potential in clinical settings, especially under the circumstance that many clinicians have turned to telemedicine to maintain social distancing during the COVID-19 pandemic [19].

HDSSs have a great potential to improve healthcare delivery, though the literature on their successful adoption, especially that of machine learning and data mining-based HDSSs, is scarce. Researcher Sepulveda indicated that aside from system accuracy, efficiency, and usability are important for an HDSS to be accepted and integrated into clinical workflow. An HDSS should be time-saving, intuitive, and simple to use in order to obtain system outputs easily while juggling a heavy clinical workload. They also pointed out that black boxes are not acceptable for HDSSs. This is in line with researcher Antoniadi who indicated that explainability is a critical component for an HDSS to be adopted in practical use effectively. A famous example by researcher Caruana shows that a machine learning-based system can reflect the pattern in the training data but be inconsistent with medical knowledge and thus does not translate to clinical practice. Their system predicted that patients who had a history of asthma had a lower risk of dying from pneumonia than the general population. This is because patients who had asthma and present with pneumonia usually receive aggressive care, which lowers their risk. Even though the system truly captured the training data, it would be problematic if adopted in clinical practice without understanding why the model behaved this way. Such problems can be resolved using a hybrid model-building approach. Many benefits have been reported in the use of ensemble learning in HDSS, including enhancing decision confidence, generating the hypothesis about causality, and increasing the acceptability and trustworthiness of the system.

We aim to apply machine learning to develop an HDSS that predicts the risk of GDM in high-risk women with overweight and obesity to identify those who may benefit from prevention strategies early in pregnancy. We modeled baseline maternal characteristics, extracted from our unique approach to feature selection. The HOMA Calculator is used to calculate Insulin Resistance and pancreatic beta cell function. GDM is also associated with abnormalities of the placenta position and early pregnancy markers commonly used in pregnancy-associated plasma protein A (PAPP-A) and free β cell function have also been incorporated into predictive models.

The feature selection process is the most crucial segment of model building. The probabilistic prediction models build with ranked features. Clinical usability was taken into account throughout the modeling process. Moreover, we applied Synthetic Minority Over-sampling Technique (SMOTE) to cope with class imbalance problems. Thus making them more acceptable and trustworthy for clinicians. The models were implemented using R and Weka. Our HDSS has the potential for clinicians to support decision-making. As well as it will help women who plan for conceiving a baby to identify the risk of GDM in early pregnancy.

The rest of the paper is organized as follows. Section “Related work” reviewed previous research done in the related field. The data, modeling process, and methodology used in this research are introduced in the Section “Methods”. Section “Results” describes our final models and their performance, majorly on white Caucasians and other populations such as the

Caribbean, African, and Indian. The discussion of our findings is presented in the Section "Discussion". Section "Conclusion" concludes this paper.

Related Work

We comprehensively reviewed research articles published between 2013 and 2020 on the use of machine learning to predict the risk of GDM. The search was performed on peer-reviewed journals. The search terms are: "GDM", "Data Mining", "Machine Learning", and "Early prediction".

Given the prevalence of GDM, there are usually many more non-GDM cases than GDM cases, leading to unbalanced datasets. Some studies have not successfully addressed this class imbalance problem, which may lead to the development of models that perform well for the majority class (non-GDM) only, that is high specificity but low -sensitivity [23]. In addition, most models have been designed to focus on GDM prediction in general pregnant women, whereas prediction in a woman who plans for conceiving a baby, has not been considered in this group. Only the researcher Artzi et al. considered the impact of the number of data features included in their research to build the model, but his research is restricted to the Israeli population, thus additional population is required to assess the real-world utility of his model.

To develop a prediction system for gestational diabetes mellitus using data mining and machine learning, we have carried out a detailed survey of existing methodology and approaches used in this domain. It is found that there are heterogeneous approaches adopted by the researchers. However, there is a need for a mixed-race data set to make a national-level prediction of the GDM field. There is a lack of studies that investigate the differences that might be associated with different cultural or ethnic backgrounds in GDM prediction. Additionally, there is a lack of implementation of the data repository or group of researchers who consistently work on machine learning and data mining-based prediction systems. The earlier research publications are based on EHR data sets or cohort study data sets. The output of these models is either predicted to class as GDM yes or no, or the probability of chances of suffering by GDM will be predicted. The evaluation of the machine learning model is very important because it influences how the performance of the ML classification algorithm is measured and compared. The techniques used for measuring the classification of performance metrics are mainly focused on Log-Loss, Accuracy, and ROC-AUC score (area under the curve). The log loss metrics are used to evaluate the performance of the binary classification algorithm, which predicts the probability of prediction class either 1 or 0 based on a threshold value. Furthermore, the confusion matrix is used by many researchers to find the correctness and accuracy of their prediction model. It is applicable when the generated output belongs to two or more types of classes.

Feature selection is a process of choosing relevant features from the input variables for building a machine-learning model. Further, feature selection is used to reduce the noise in the data, and help to solve complex problems. Researcher Yuhan Du removed unnecessary features to avoid "multidimensionality" and reduced the number of inputs required to use clinical models [32]. They identify redundant features with the help of the Pearson correlation coefficient technique and consider only those features whose correlation coefficient was found to be greater than 0.6. To reduce feature redundancy, they removed features with a high percentage of missing values before applying imputation. This ensured data reliability, saved time in computing, and also

reduced imputation effects. A cohort study in the Netherlands stated that less-educated women were at higher risk of GDM. But this feature is not applicable everywhere [28].

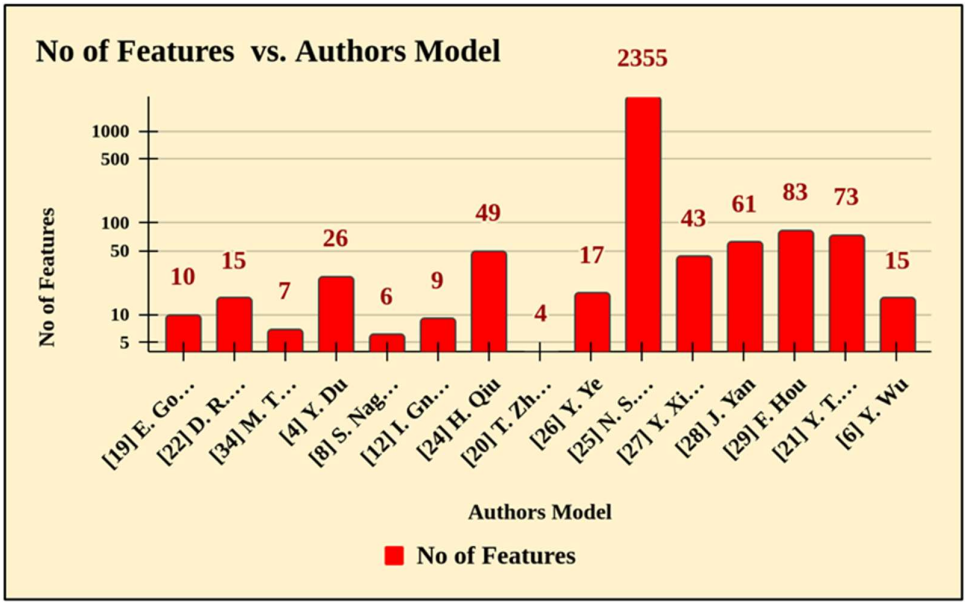


Figure 1: Number of features considered for model building

A study by Nitzan Shalom Artzi, suggests that the Shapley values feature attribute framework contributes well to gaining insight into the features that contribute most to model predictions [25]. Shapley's analysis can identify the most predictive features for GDM diagnosis. These included previous pregnancy GCT results, followed by maternal age and fasting blood glucose in the first trimester. He has considered all the 2355 features available in EHR data to build his predictive model. Tao Zheng has built a simple GDM model using only 4 features [33]. Figure 1 shows several features used by numerous researchers to build their predictive models.

Study design and data

This research is a study and analysis of the Cambridge Baby Growth data published under the title Acta Diabetologica with doi 10.1007/s00592-018-1162-7. The size of the standard Data Set is 204.2 Kb, and the file format is .csv type. In this data set, there are a total of 1724 records available with a total of 29 features considered for the collection of the data. The data creation year is between 2001 and 2009 at Rosie Maternity Hospital, Cambridge, United Kingdom. All study participants were over 16 years of age. The consent of the patient was taken to enroll in this cohort study. All methods were carried out by relevant guidelines and regulations.

In the CBGS dataset, there are 5 nominal attributes such as gestational diabetes, twin, ethnicity, sex of offspring, and smoking during pregnancy. The remaining 24 attributes are numeric. Our class or target variable is gestational diabetes Yes, No. The 42 mothers gave birth to twins and 1683 mothers with singleton pregnancies. The body mass index (BMI) is calculated as the self-reported pre-pregnancy body weight divided by the height squared. From the age of 18 to shortly before pregnancy, self-reported weight gain of 10 kg or more may be associated with an increased risk of GDM. On the other hand, if the weight loss continuously occurs even after the first trimester of pregnancy, then it is a cause for concern. As per the Institute of Medicine and National Research Council, the ideal BMI should be 25 Kg/m2.

During the CBGS, a variety of data were collected from these participants, including maternal characteristics, age at menarche, OGTT zero/sixty min glucose, twin, ethnicity, prepregnancy BMI, PAPPA concentration, age at delivery, sex of offspring, parity, smoking, Offspring birth weight, Gestational age birth. GDM was diagnosed according to the WHO 2013 criteria in approximately the second trimester. Characteristics of the Cambridge Baby Growth Study sample included in this analysis, and those who were excluded due to lack of prenatal questionnaire data. As a result, 1725 participants were included. As we aimed to predict GDM in early pregnancy, the pathological testing features used for data collection such as C peptide concentration, and C peptidogenic index. Along with these features few features like demographic characteristics, Gestational age, and Offspring birth weight.

Data preparation

Many features included in this research have missing values. We have created cases to treat missing values and tune the input data. In the first approach to handling missingness in the data, we imputed data by the mean-mode approach. The 10.32% GDM-positive cases were present in this first approach. In the second approach, we have dropped rows containing missing values: the resultant dataset contains 10.46% GDM-positive cases. While dealing with several imputation methods, it is observed that there is a need to handle the class imbalance problem.

Our goal is to take a poll for all attributes voted on by various feature selection techniques, with those attributes having the highest vote score, being tuned as inputs to the hybrid model.

Implementation of Feature Extraction Approaches

It is important to select the relevant features that may have an influence on the development of GDM. This can be done by using feature selection techniques in data mining and machine learning. We have used embedded feature selection methods, which adaptively found the optimal feature set from the CBGS data for each classification model, such as decision trees, logistic regression, Naive Bayes, and random forest. The methodology adopted for feature selection is based on voting, which means that the feature generated by each technique is compared with the raw data, imputed data, and SMOTE data. The feature variables that received the most votes in all datasets are selected for the final hybrid model. We used R-Studio and WEKA to perform feature selection processes. We obtained the optimal feature set for each dataset using the classification attribute evaluation filter, correlation attribute evaluation filter, correlation-based feature selection, and information gain attribute evaluation filter in WEKA, as well as performing experimentation on a decision tree, the Boruta algorithm, and a random forest using the integrated development environment (IDE) of R Studio. To get the rank of the features from the subset of data, we need to apply the variable ranking process and the "feature subset selection method". We have used numerous approaches for ranking the features, such as correlation ranking filter, information gain filter, wrapper subset evaluation, feature importance by a decision tree, Boruta algorithm, and random forest. These approaches help us identify optimal features. Further, the implementation details of each of these approaches are given.

1) Correlation Ranking Filter:- The correlation ranking filter evaluates the worthiness of an attribute by estimating the correlation between attributes and the target class [13]. While applying this filter, we loaded the CBGS training dataset and then selected the "Correlation Ranking Filter" from the Filter tab. We have selected the "rank by" option as Ranker-T's 1.79

coefficient and the number of attributes to be selected as 28. After applying the filter to the training dataset, the output (Figure 2) shows the attributes ranked by correlation coefficient. We have shortlisted the attributes with the highest correlation coefficients for the polling system.

Figure 2: Attribute Ranking

2) Information Gain Filter:- Another popular feature selection technique is information gain computation. We calculated the information gain (also called entropy) for each attribute for the output variable. Access values range from 0 (no information) to 1 (maximum information). Attributes that contribute more information have a higher information value and are selected, while those that do not add more information have a lower score and may therefore be removed. Weka supported feature selection by gaining information using the Weka-supervised feature "InfoGainAttributeEval" attribute evaluator. Like the correlation technique above, the ranker search method is used. Figure 3 depicts the results of the information gain filter.

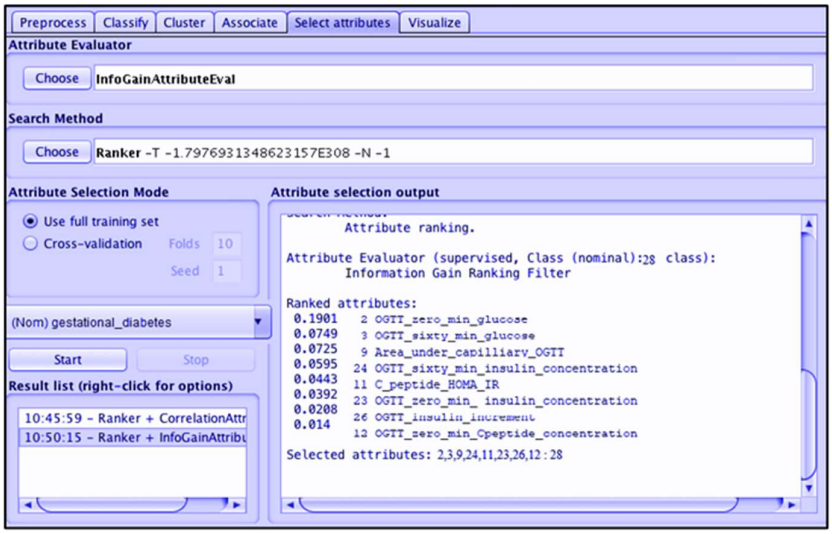


Figure 3: Attribute selection using Information Gain

3) Wrapper Subset Evaluation: Previously used feature selection techniques were voted to be OGTT 0 min glucose (mmol/L), OGTT 60 min glucose (mmol/L), Area under the capillary glucose OGTT, OGTT 60 min insulin concentration (pmol/L), C-peptide-derived HOMA-IR, OGTT insulin increment (pmol/L), and OGTT 0 min insulin concentration (pmol/L). So far, these parameters have received a total of 2 votes. We now proceed to apply the wrapper subset evaluation method to support our earlier task of feature extraction. The goal is to take a poll for all attributes voted on by various feature selection techniques, with those attributes having the highest vote score being tuned as inputs to the hybrid model. We configure "WrapperSubsetEval" with a classifier of "J48" which is available under the tree, and select "Search Method" as "Best First" The feature selection has been carried out on training sets, which contains 70% of the total data. The OGTT 0 min glucose (mmol/L) and OGTT 60 min glucose (mmol/L) heavily contributed to the target variable.

4) Feature Importance by Decision Tree: The way we chose to assess the importance of each feature in the decision tree is by looking at the weight, or "Importance" of each feature in the training set. This importance can be measured using various measures, such as impurity reduction or the total number of times a feature is used to split the data in the tree. Generally, features with higher importance scores are more likely to be used in decision trees. We have constructed a decision tree using the CRAN (Comprehensive R Archive Network) Party Library [17].

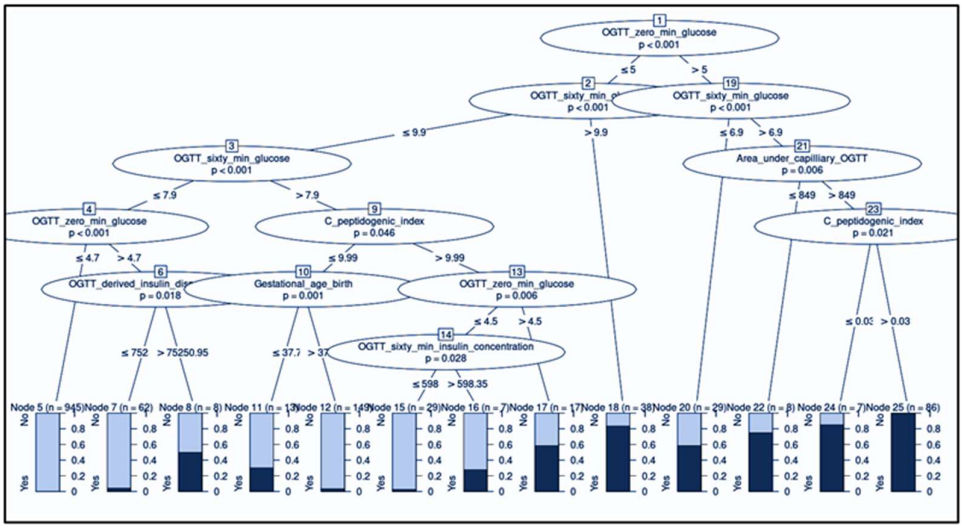


Figure 4: Weight and score of variables by a decision tree

It is an algorithm for recursive partitioning based on parametric models. It calculates feature importance by keeping the best-performing features as close as possible to the root of the tree. The variable importance in the decision tree is shown in Figure 4. At the root node, the OGTT 0 min glucose (mmol/L) variable is found, thus supporting the earlier feature selection techniques.

5) Boruta Algorithm:- The Boruta algorithm is a feature selection method that can be used to identify the most important features in a dataset. It is a wrapper approach that uses the random forest algorithm to evaluate the importance of each feature and then identifies the features that are most likely to be relevant to the target variable. We have used the libraries Boruta, mlbench,

caret, and randomForest to implement this algorithm in R. The feature selection has been done using the set.seed() function, and then by applying the plot function, the resultant features are extracted from the training set. The output is shown in Figure 5.

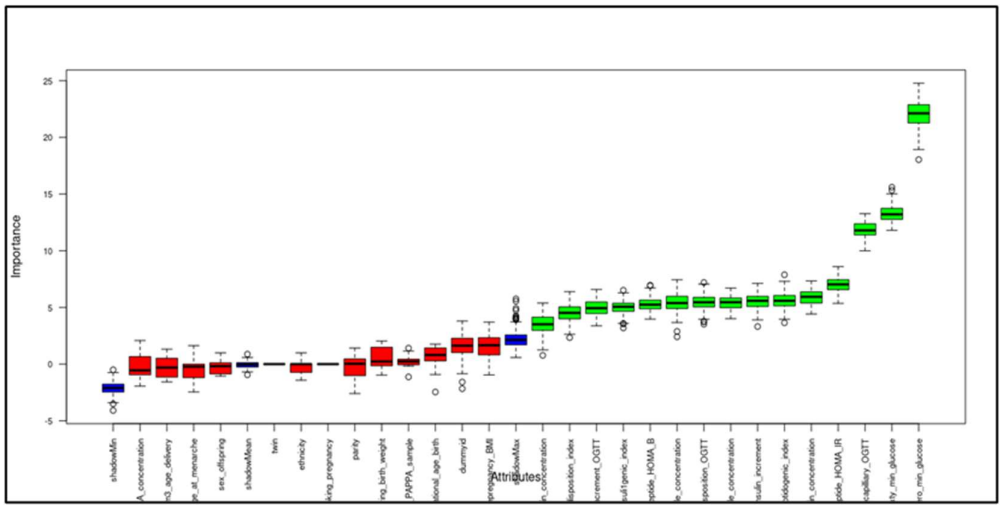


Figure 5: Feature selection using Boruta algorithm

6) Random Forest:- A supervised learning technique that may be applied to both classification and regression problems is the random forest algorithm. In terms of feature selection for data related to gestational diabetes, we employed a random forest technique to pinpoint the most significant features or variables in the data.

To apply the random forest algorithm for feature selection, first we split the data into training and testing sets and then trained a random forest model on the training data using the most relevant features or variables as input. This trained model delivers importance scores for the individual features; further, it helps identify the most relevant features for predicting gestational diabetes. These are the top 10 features by RF (Figure 6), It has to be read from top to bottom, at the top most contributing parameter at the bottom is a comparatively less contributing parameter. All 10 features secure the most votes among the remaining variables using all the feature selection techniques discussed earlier.

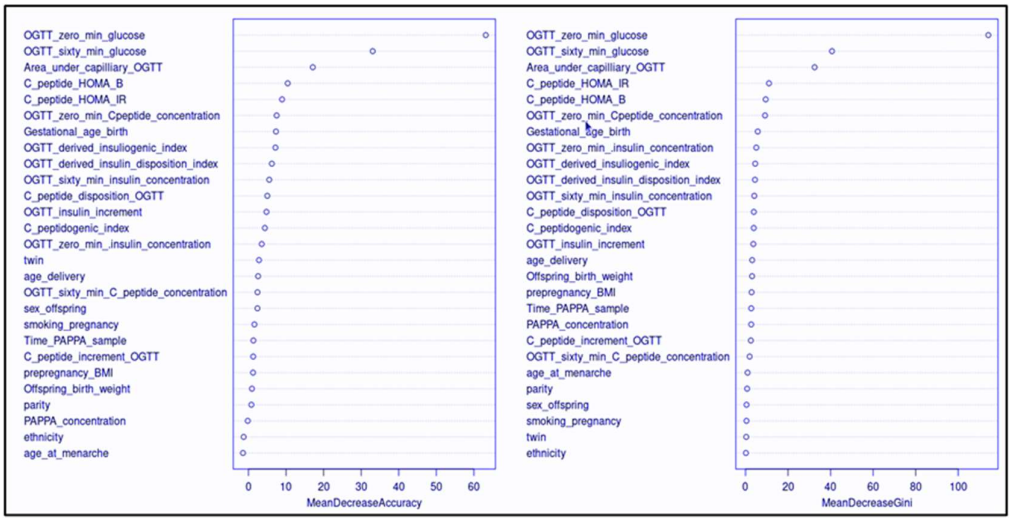


Figure 6: Mean decrease accuracy and Gini

The top 10 features are shown in Table 1. These top 10 features secured maximum votes in almost all the feature selection techniques, and they contributed a lot to the hybrid prediction model of gestational diabetes. As an assessment of whether the feature selection method used was correct or not, we have verified it using the correlation matrix between independent variables and the target variable.

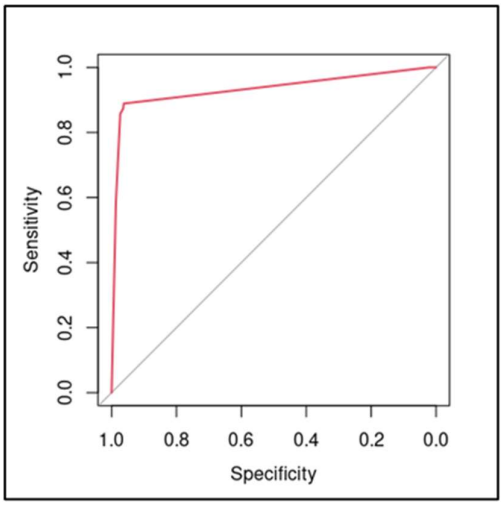
Table 1: Top 10 features of CBGS Dataset

S. No.	Attribute Name	Type of Attribute	Rank
4	OGTT 0 min. glucose ($mmol/L$)	Numeric Real	1
5	OGTT 60 min. glucose ($mmol/L$)	Numeric Real	2
11	Area under the capillary glucose OGTT	Numeric Real	3
12	C-peptide derived HOMA B	Numeric Real	4
13	C-peptide derived HOMA IR	Numeric Real	5
14	OGTT 0 min C-peptide concentration	Numeric Real	6
24	Gestational age at birth (weeks)	Numeric Real	7
26	OGTT 60 min insulin concentration ($pmol/L$)	Numeric Real	8
27	OGTT-derived insulin disposition index	Numeric Real	9
29	OGTT-derived insulinogenic index	Numeric Real	10

This polling approach has helped to improve model performance and reduce the risk of overfitting.

Result Analysis

The beauty of the polling-based feature selection approach is that it allows for personalized predictions and can adapt to changes in a patient's glucose patterns over time. It also provides valuable insights into the factors that contribute to glucose variability, which can be used to inform treatment decisions and improve overall diabetes management. The embedded feature selection techniques, such as decision trees, logistic regression, Naive Bayes, and random forests, were utilized to adaptively identify the most appropriate feature set from the CBGS data for each classification model. The voting-based methodology chosen for feature selection entails comparing the features produced by each technique to the raw_data, imputed_data, and SMOTE_data. The final hybrid model is built using the feature variables that obtained the most votes across all datasets.



This HDSS approach was further validated using 10-fold cross-validation techniques to verify its accuracy for future data. The ROC curve plotting the true positive rate (sensitivity) on the y-axis and the false positive rate (1-specificity) on the x-axis for the test set is shown in Figure 6. Sensitivity refers to the proportion of actual positive cases correctly identified as positive (true positive rate). Specificity refers to the proportion of actual negative cases correctly identified as negative (true negative rate). The ROC curve visually represents the trade-off between sensitivity and specificity. A classifier with effective diagnostic ability will have a ROC curve close to the top left corner of the plot. This indicates high sensitivity and specificity. Thus, the HDSS approach combines the strengths of multiple classifiers. ROC curves are useful for evaluating binary classifier system performance.

Conclusion

The Cambridge baby growth study (CBGS) data source is used to build a predictive model. This model can be used to identify women at high risk of GDM and require further screening by healthcare professionals. In addition, the system can help inform health promotion and prevention initiatives as well as provide more accurate GDM diagnoses. To investigate the performance of HDSS techniques for gestational diabetes prediction, we conducted experiments with numerous feature selection techniques. We selected the most robust model only after deliberately testing it on test sets with a different set of feature vectors. The hybrid decision support system (HDSS) performs well with an accuracy of 87.92% on the Smote_balanced_data. Likewise, the random forest and decision tree independently perform well with scores of 84.41 and 83.89% respectively.

References

[1]
Mithal A, Bansal B, Kalra S. Gestational diabetes in India: Science and society. Indian J Endocr Metab 2015;19:701-4.

[2] A. M. Al-Khasawneh, “Decision Support System for Diabetes Classification Using Data Mining Techniques,” vol. 11, no. 3, pp. 281–303, 2018, doi: 10.4018/978-1-5225-5460-8.ch012.

[3] McIntyre H. David, Kapur Anil, Divakar Hema, Hod Moshe TITLE=Gestational Diabetes Mellitus—Innovative Approach to Prediction, Diagnosis, Management, and

Prevention of Future NCD—Mother and Offspring JOURNAL=Frontiers in Endocrinology VOLUME=11 YEAR=2020 DOI=10.3389/fendo.2020.614533 ISSN=1664-2392

[4] K. V. Lakshmi, "Modeling an Expert System for Diagnosis of Gestational Diabetes Mellitus Based On Risk Factors," IOSR J. Comput. Eng., vol. 8, no. 3, pp. 29–32, 2013, doi: 10.9790/0661-0832932.

[5] Loriaux D;Lynn MD. Diabetes and The Ebers Papyrus: 1552 B.C. Endocrinologist. 2006; 16;(2): 55-56

[6] ZajacJ, Shrestha A, Patel P, Poretsky L. The Main Events in the History of Diabetes Mellitus. In: Poretsky L editor. Principles of diabetes mellitus.2nded. New York, NY, USA:Springer Verlag. 2010; 3-16.

[7] Burhan Ahmed. A Detailed History of Diabetes.Medicalopedia. 2012; (April)

[8] Carpenter S, Rigaud M, Barile M, Priest T J, Perez L, Ferguson JB. An Interlinear transliteration and English translation of portions of the Ebers Papyrus, possibly having to do with diabetes mellitus. Annandale-on-Hudson, NY, United States, Bard College, 1998.

[9] G. P. 2017 Arora, "Gestational Diabetes Mellitus in North India," 2017.

[10] Leonid Poretsky, editor. Principles of diabetes mellitus, 2nd ed. New York: Springer, 2009; 3. ISBN 978-0-387-09840-1.

[11] "IDF Diabetes Atlas, 10th edition." [Online]. Available: www.diabetesatlas.org 2021

[12] Hoet JP, Lukens FD. Carbohydrate metabolism during pregnancy. Diabetes. 1954; 3:1-12.

[13] O'Sullivan JB. Gestational diabetes. Unsuspected, asymptomatic diabetes in pregnancy. N Engl J Med. 1961; 264:182-1085.

[14] "About diabetes". World Health Organization. Published on 31 March 2014

[15] Diabetes Fact sheet N "312". WHO. October 2013

[16] World Health Organization. Diagnostic criteria and classification of hyperglycemia first detected in pregnancy: a World Health Organization Guideline. Diabetes Research and Clinical Practice. 2014;103 (3):341-63.

[17] WHO Consultation: definition, Diagnosis and Classification of Diabetes Mellitus and Its Complications: Report of a WHO Consultation. Part 1: Diagnosis and Classification of Diabetes Mellitus. Geneva, WHO/NCD/NCS/99.2; World Health Org., 1999.

[18] American Diabetes Association. Classification and Diagnosis of Diabetes. Diabetes Care. 2017;40 (Suppl 1):S11-S24.

[19] E. Gomes Filho, P. R. Pinheiro, M. C. D. Pinheiro, L. C. Nunes, and L. B. G. Gomes, "Heterogeneous Methodology to Support the Early Diagnosis of Gestational Diabetes," IEEE Access, vol. 7, no. March, pp. 67190–67199, 2019, doi: 10.1109/ACCESS.2019.2903691.

[20] P. Prentice, C. L. Acerini, A. Eleftheriou, I. A. Hughes, K. K. Ong, and D. B. Dunger, "Cohort Profile: the Cambridge Baby Growth Study (CBGS)", doi: 10.1093/ije/dyv318.

[21] E. G. Filho, "Support to Early Diagnosis of Gestational Diabetes Aided by Bayesian Networks," vol. 985, pp. 277–286, 2019, doi: 10.1007/978-3-030-19810-7.

[22] D. R. Krishnan et al., "Evaluation of predisposing factors of Diabetes Mellitus post Gestational Diabetes Mellitus using Machine Learning Techniques," in IEEE Student Conference on Research and Development, Oct. 2019, pp. 81–85. doi: 10.1109/SCORED.2019.8896323.

[23] P. Prentice, C. L. Acerini, A. Eleftheriou, I. A. Hughes, K. K. Ong, and D. B. Dunger, "Cohort Profile: the Cambridge Baby Growth Study (CBGS)", doi: 10.1093/ije/dyv318.
[24] H. Qiu et al., "Electronic Health Record Driven Prediction for Gestational Diabetes Mellitus in Early Pregnancy," Sci. Rep., vol. 7, no. 1, pp. 1–13, 2017, doi: 10.1038/ s41598-017-16665-y.
[25] N. S. Artzi et al., "Prediction of gestational diabetes based on nationwide electronic health records," Nat. Med., vol. 26, no. 1, pp. 71–76, 2020, doi: 10.1038 /S41591- 019-0724-8.
[26] Y. Ye et al., "Comparison of Machine Learning Methods and Conventional Logistic Regressions for Predicting Gestational Diabetes Using Routine Clinical Data: A Retrospective Cohort Study," J. Diabetes Res., vol. 2020, 2020, doi: 10.1155/2020/4168340.
[27] Y. Xiong et al., "Prediction of gestational diabetes mellitus in the first 19 weeks of pregnancy using machine learning techniques," J. Matern. Neonatal Med., pp. 1–7, 2020, doi: 10.1080/14767058.2020.1786517.
[28] J. Yan et al., "A Prediction Model of Gestational Diabetes Mellitus Based on First Pregnancy Test Index," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12435 LNCS, pp. 121–132, 2020, doi: 10.1007/978-3-030-61951-0_12.
[29] F. Hou, Z. X. Cheng, L. Y. Kang, and W. Zheng, "Prediction of Gestational Diabetes Based on LightGBM," ACM Int. Conf. Proceeding Ser., pp. 161–165, 2020, doi: 10.1145/3433996.3434025.
[30] Y. T. Wu et al., "Early Prediction of Gestational Diabetes Mellitus in the Chinese Population via Advanced Machine Learning," J. Clin. Endocrinol. Metab., vol. 106, no. 3, pp. e1191–e1205, 2021, doi: 10.1210/CLINEM/DGAA899.
[31] Y. Wu et al., "A risk prediction model of gestational diabetes mellitus before 16 gestational weeks in Chinese pregnant women," Diabetes Res. Clin. Pract., vol. 179, p. 109001, 2021, doi: 10.1016/j.diabres.2021.109001.
[32] W. Volanski et al., "D-GDM: A mobile diagnostic decision support system for gestational diabetes," Arch. Endocrinol. Metab., vol. 63, no. 5, pp. 524–530, 2019, doi: 10.20945/2359-3997000000171.
[33] Y. Srivastava, P. Khanna, and S. Kumar, "Estimation of Gestational Diabetes Mellitus using Azure AI Services," Proc. - 2019 Amity Int. Conf. Artif. Intell. AICAI 2019, pp. 321–326, 2019, doi: 10.1109/AICAI.2019.8701307.
[34] M. Thiyagarajan, C. Raveendra, P. Thulasi, and S. K. Priya, "Role of Association Rules in Medical Examination Records of Gestational Diabetes Mellitus," iccca, vol. 2017-Janua, pp. 78–81, 2017, doi: 10.1109/CCAA.2017.8229775.
[35] C. J. Petry, K. K. Ong, I. A. Hughes, C. L. Acerini, J. Frystyk, and D. B. Dunger, "Early pregnancy-associated plasma protein a concentrations are associated with third trimester insulin sensitivity," J.Clin Endocrinol. Metab., vol. 102, no. 6, pp. 2000–2008, 2017, doi: 10.1210/jc.2017-00272.
[36] H. D. Mcintyre, P. Catalano, C. Zhang, and G. Desoye, "Gestational diabetes mellitus," Springer US, 2019. doi: 10.1038/s41572-019-0098-8.
[37] V. Sigillito, "Pima Indians Diabetes Data Set," UCI Mach. Learn. Repos., pp. 1–2, 1990, [Online]. Available: https://archive.ics.uci.edu /ml/datasets/Pima +Indians +Diabetes

[38] R. A. Rahman, N. S. A. Aziz, M. Kassim and M. I. Yusof, "IoT-based personal health care monitoring device for diabetic patients," 2017 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), 2017, pp. 168-173, doi: 10.1109/ISCAIE.2017.8074971.

A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the p-value, the greater the statistical significance of the observed difference. A p-value of 0.05 or lower is generally considered statistically significant
What does the P statistic tell you?
weka how to add SignificanceAttributeEval
https://www.analyticsvidhya.com/blog/author/sharoon-saxena/
https://www.analyticsvidhya.com/blog/2019/09/everything-know-about-p-value-from-scratch-data-science/
the second trimester of pregnancy means
Each trimester is roughly 14 weeks long. When you enter your second trimester, you are around 14 weeks pregnant. This middle trimester will last from week 14 to the end of week 27. During your second trimester of pregnancy, you'll start looking and feeling more pregnant.
When the OGTT test carried out during pregnancy
The American College of Obstetricians and Gynecologists recommends performing a one-hour blood glucose challenge test to screen for gestational diabetes in low-risk pregnant women between 24 and 28 weeks of pregnancy