

MACHINE LEARNING MODEL

FOR GESTATIONAL AGE AND GROWTH CURVES BASED ON FETAL BIOMETRY FROM 14 TO 40 WEEKS OF GESTATION IN NORTHEAST MEXICO

OBJECTIVE

Determine gestational age with regression algorithms and describe the reference curves of fetal biometric parameters appropriate to our population of fetal biometry parameters.

AUTHORS:

Sara María Villarreal Alfaro
José Anastacio Hernández Saldaña
Roberto Abraham Vargas Loza
Andrés Sánchez Domínguez

MATERIAL AND METHODS

A cross-sectional, descriptive, retrospective and observational study was conducted, including all pregnant women who attended the Institute of Maternal-Fetal Medicine of the Hospital of Gynecology and Obstetrics of Monterrey “Ginequito” during the period from January 2019 to January 2024, healthy, of indistinct age, with a normo- evolutionary singleton pregnancy, (who did not present maternal or fetal complications) and who were between 14 and 40 weeks of gestation established by date of last reliable and safe menstruation, or gestational age adequately corrected by ultrasound.

To estimate the gestational week value of a fetus based on ultrasound measurements, the available variables are: **biparietal diameter (BPD)**, **head circumference (HC)**, **abdominal circumference (AC)**, **femur length (FL)**, and **estimated fetal weight (EFW)**.

To model the gestational weeks as a numerical value using these variables, we will employ **regression models by testing linear regression, K-nearest neighbors, and decision tree models.**

These models will be implemented using the scikit-learn library, and pandas will be used to represent and manage the data. The code and models can be found at the following link:
<https://gitlab.com/ppGodel/ginecology/>.



The values obtained from the ultrasound will be our control variables and the gestation weeks will be our dependent variable.

The models will be trained first, using the cross validation method, the cross validation takes a percentage of the data, in our case the 80 percent of the data, and that 80 percent set will divide it into 5 parts, where 4 will be used to train a models and the fifth part will be used to test it. This will be done by each part, training 5 models with 5 parts that will be used as test.

This to prevent the model from having biases to raise the sets, in the end the average of the adjustment of the models obtained in the 5 cases will be used, this being our training value.

The models adjustment measure is the determination coefficient or R^2 , this coefficient determines the quality of the model to replicate results and that the data vary as a function depending on the model, r^2 takes values of 0 when the model explains the data as well as the average and a value of 1 shows us that the model explains exactly the data.

Each model needs parameters to adjust to work, and we want to select the parameters that give us the best adjustment, in the case of linear regression we use the minimum square method to find the line that explains the behavior of our dependent variable which does not need parameters.

For closest K-nearest neighbors, we adjust the value of K for values 3, 5, 7, 10, 15 neighbors, this parameter adjusts how many neighbors we want to group to generate an average of that data.

Finally, for the decision trees, 2 parameters, maximum and minimum samples per leaf were adjusted, these parameters control the depth of the tree or the number of comparisons that are they will be made to the data and the number of values that will be grouped to generate the average of the data, the values used were 5,15,20,25 of depth levels and 3,5,10,15 deep depth of the sample.

Then each model with its parameters will be tested with cross validation to obtain its adjustment measure, and from all of them the model that has had the best adjustment measure will have been chosen.

Finally, these models are compared to 20% of the remaining data, known as the validation set, to see how the model with new data behaves. Already with the train and validation measures, we choose the best model.

RESULTS

We found **4179** studies of pregnant patients, which met the selection criteria, in which the model was trained in **80%** and model validation testing was performed in **20%** of the cases.

LINEAR REGRESSION

The linear regression model gave us a value of R^2 during the 0.93 test and a value of 0.95 in the validation set, this shows us that the variables can explain the behavior of the dependent variable weeks of gestation, suggesting linear behavior between the variables given the values of r^2 greater than 0.9.

CLOSEST K-CLOSE ONES

The closest K- Close ones model gave us a value of R^2 during the 0.95 test and a validation value of 0.94 for the 5 neighbors model. This model gives us better results than the linear regression model explaining the data with a minor variance.

DECISION TREE MODEL

The decision tree model gave us an R^2 value of 0.95 during testing and a value of 0.96 for the validation set, this for the parameters of 10 depth levels and 10 samples per leaf, this being the best of the models obtained and the one suggested to use for estimating gestational age.

CONCLUSIONS

In this project, we present the study with the largest population reported at the national level, suggesting the continuation of this line of research as it would benefit the population from which the sample was drawn and later could be expanded to other populations, as it will help us have greater confidence in the predictions of gestational age using this model.

This study will also be useful for accurately determining gestational ages in patients who, due to socioeconomic factors, cannot maintain adequate prenatal care, allowing for the timely detection of pathology associated with fetal growth. All the models showed high good values for both the test and validation, so we do not expect to have overfitted or underfitted models.

