Diabetes analysis
Weka & Machine Learning

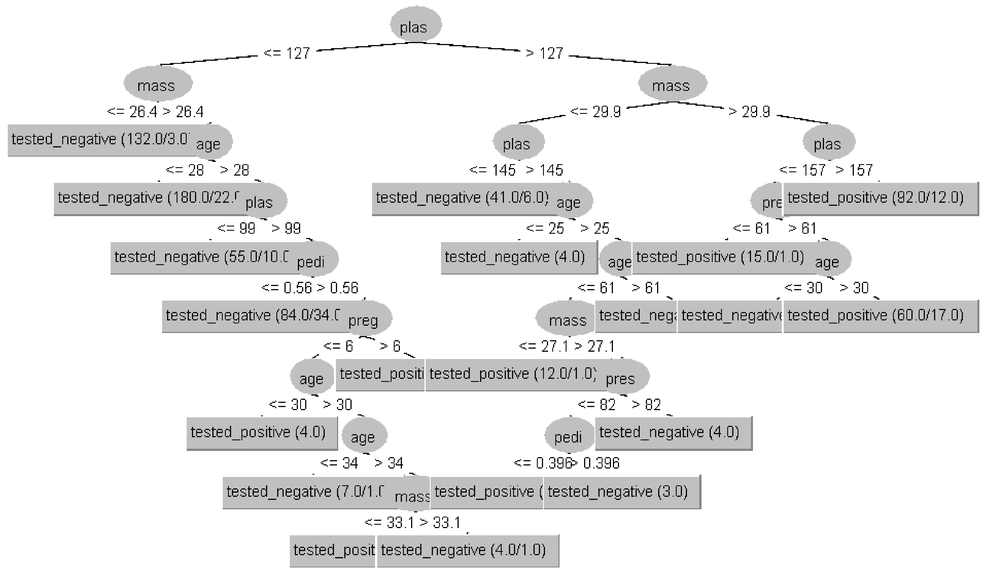
Joel Dos Santos Iraha

s545242@nwmissouri.edu / (203)-300-9996

NORTHWEST MISSOURI STATE UNIVERSITY

The Pima Indian population have been heavily studied since 1965 on account of high rate of diabetes. This dataset contains measurements for 768 female subjects, all aged 21 and above. Can we figure out if someone will likely have diabetes just by taking a few of these measurements?

- Data: Labeled dataset with nominal features. Use of training set as testing option.
- Algorithm: Supervised learning algorithm using J48 decision tree to visualize classification accuracy.



The model performs with an accuracy of 84.11%, correctly classifying instances. Precision and recall metrics are well-balanced, with precision at 84.2% and recall at 84.1%. Kappa statistic of 0.6319 indicates substantial agreement beyond prediction. False positives and false negatives are minimized, showcasing the model's reliability. The model offers strong predictive power, showing accurate and balanced outcomes for tested_negative and tested_positive classes.

The variables involved in this dataset are structured in a nominal classified way which allows us to work with using supervised algorithm training to figure out if someone will likely have diabetes based on the selected attributes.

However, validation and point metrics suggests that the model may be overfit. a relative absolute error of 52.43% means that the predictions of our model are off by approximately 52.43% compared to the actual values, thus indicating high level of model's prediction.

This is a public training dataset offered by Weka.

https://storm.cis.fordham.edu/~gweiss/data-mining/datasets.html

Project made for Intro to data science and mining at Northwest Missouri State University

https://joeldossantospersonal.github.io/DosSantosJ.github.io/