Visualizing Diagnosis: Decision Tree vs. Naive Bayes for Diabetes Prediction

Aaryaman Jaising & Rustom M. Dubash

Dataset: We will use the Pima Indians Diabetes Dataset, originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset includes diagnostic measurements such as number of pregnancies, BMI, insulin levels, age, and more. All patients are women aged 21 or older of Pima Indian heritage. The target variable is Outcome, indicating the presence (1) or absence (0) of diabetes.

Dataset link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Methods: We will use scikit-learn to implement a Decision Tree Classifier and compare its performance and interpretability to a Naive Bayes Classifier. Decision Trees will be emphasised for their ability to provide visual explanations that are similar to a doctor's diagnostic reasoning. Naive Bayes will serve as a baseline true classifier.

Scientific Question: Would a Decision Tree Classifier be as informative and accurate to a patient about their diagnosis as a Naive Bayes Classifier?

We hypothesize that while Naive Bayes may be more accurate, Decision Trees will offer easier interpretability and patient understanding.

Evaluation and Visualization:

For Naive Bayes Classifier: Confusion matrices, accuracy, precision, recall, and false negative rate.

For Decision Trees: visualization of the full tree using sklearn built in tree function to simulate a flowchart like diagnostic tool.

References: Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus.” *Proceedings of the Symposium on Computer Applications and Medical Care* (1988).