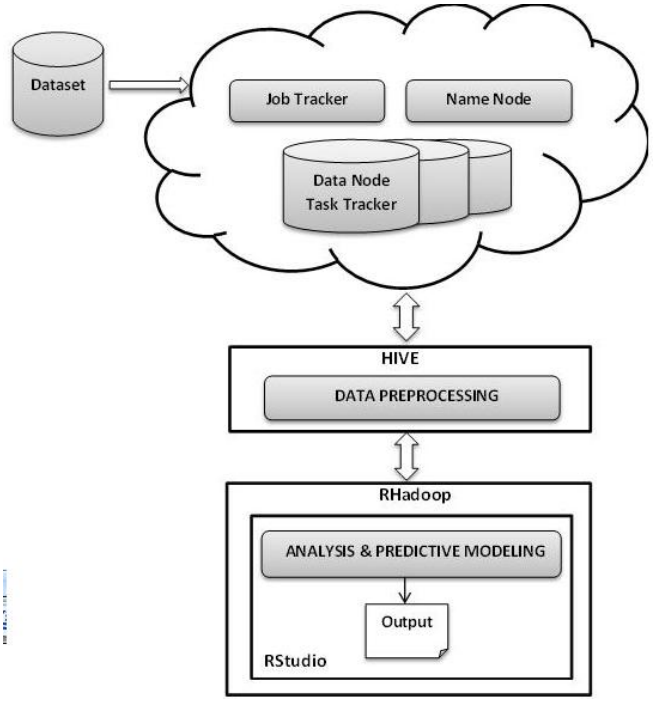

Figure 1: Architecture

Some classification rules for diabetes datasets from the decision tree are as follows,

1. Plasma glucose <127.5 AND Age <28.5 AND Body mass index <26.35 then 'tested_negative' for readmission (39/2).
2. Plasma glucose <127.5 AND Age <28.5 AND Body mass index <26.35 AND plasma glucose <99.5 AND pedigree <0.561 then 'tested_positive' for readmission (9/25).
3. Plasma glucose <127.5 AND Body mass index < 29.25 AND Plasma glucose <145.5 then 'tested_negative' for readmission (35/6).
4. Plasma glucose <127.5 AND Body mass index < 29.25 AND Plasma glucose <145.5 AND AGE <25.5 then 'tested_negative' for readmission (4/0).
5. Plasma glucose <127.5 AND Body mass index < 29.25 AND Plasma glucose<157.5 AND Age <30.5 then _tested_positive‘ for readmission (18/47).

To improve our model performance, next step is to find out the importance of each variable in diabetes dataset. All measures of importance are scaled to have a maximum value of 100. As show in Figure.3 Plasma glucose have highest variable importance i.e. 0.7881 followed by age (0.687), body mass index (0.686) while the insulin attribute is the least important. Now we detect the risk predictors in our diabetes dataset by using recursive feature elimination function. We rank each predictor based on its importance in the model. Based on recursive feature elimination function, we discover top 5 variables out of 8 variables i.e. plasma glucose, body mass index, age, pregnant, pedigree function are top predictors in our model. These are the top 5 readmission predictors in diabetic dataset.

3. Result

The result of correlation matrix is shown in Table.2. In the process of calculating correlation, we find that feature plasma glucose is highly correlated with readmission feature. From the Table.2 we can also find that Blood Pressure (0.065) is least correlated with readmission. But for building the decision tree classification model all the 8 features are included. Based on these features a Decision Tree has been generated. The decision tree is shown in Figure 2.

Table 2: Correlation between 8 features and readmission

Feature	Value of correlation
Pregnant	0.2218982
Plasma glucose	0.4665814
Blood pressure	0.06506836
Skin	0.07475223
Insulin	0.130548
Body Mass Index	0.2926947
Pedigree Function	0.1738441
Age	0.238356

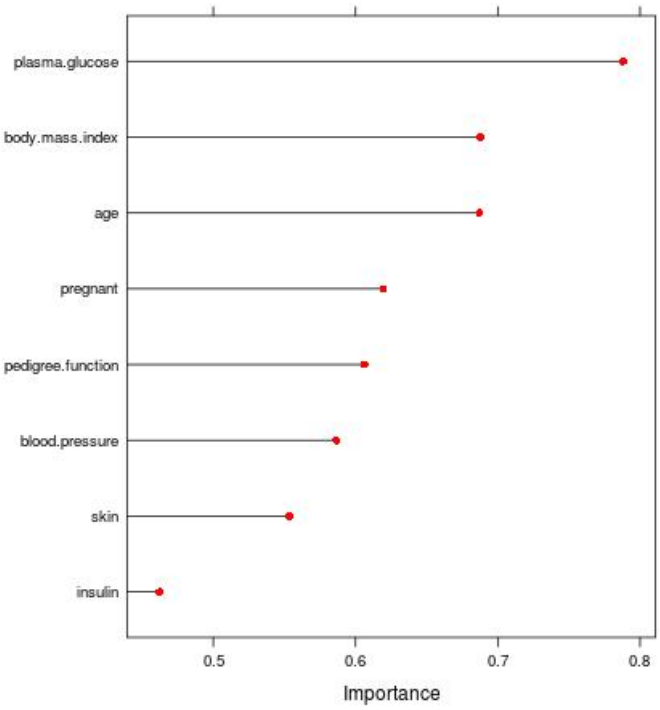

Figure 3: Variable Importance

Licensed Under Creative Commons Attribution CC BY

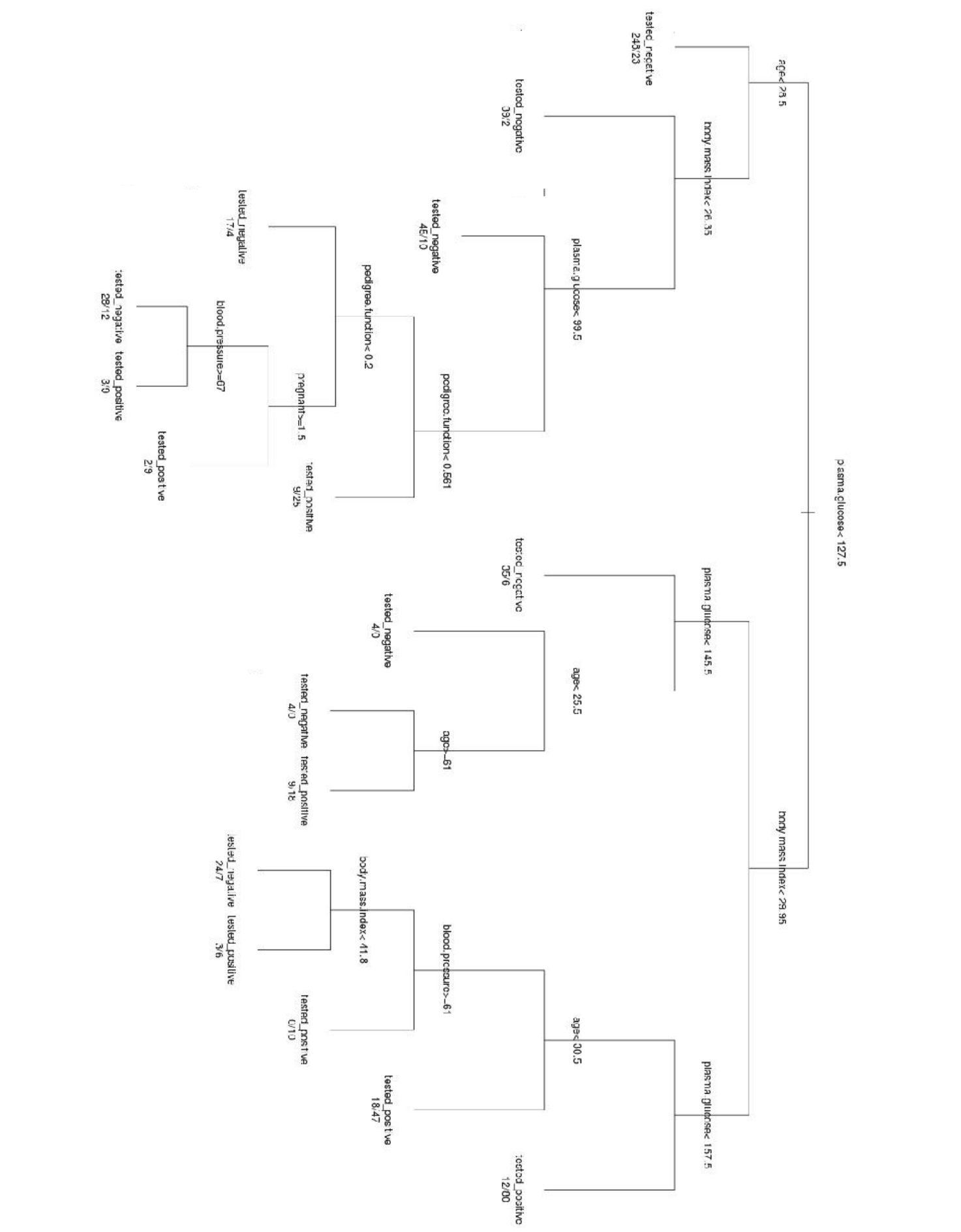

Figure 2: Decision Tree

Licensed Under Creative Commons Attribution CC BY

4. Conclusion

Big Data analytics have been applied to evaluate the risk of readmission for diabetes patients. Predictive modeling has been employed by applying decision tree classification method. It has been observed that chance of readmission in diabetic patient is successfully predicted using the above analysis. Many analysis methods can be explored to improve the accuracy of the existing system.

References

[1] Donzé J. Aujesky D., Williams D., Schnipper J.L, MD. ―Potentially avoidable 30-day hospital readmissions in medical patients: Derivation and validation of a prediction model. JAMA Internal Medicine,‖173(8):632-638, Apr. 2013.
[2] ―Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications,‖ Report of a WHO Consultation Part 1: Diagnosis and Classification of Diabetes Mellitus World Health Organization Department of Non communicable Disease Surveillance, Geneva, 1999.
[3] UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html.
[4] Manyika J., Chui M., Brown B., and Bughin J. and Dobbs R. ―Big data: The next frontier for innovation competition and productivity,‖ McKinsey Global Institute, 2012.
[5] The Apache Software Foundation, http://hadoop.apache.org/common/credits.html.
[6] Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., And Murthy, R Hive—a warehousing solution over a Map-Reduce framework. In VLDB, 2009.
[7] H.W. Ian, E.F., ―Data mining: Practical machine learning tools and techniques,‖ 2005: Morgan Kaufmann.
[8] Sung-Hyuk Cha, and Charles Tappert, "A genetic algorithm for constructing compact binary decision trees,‖ Journal of Pattern Recognition Research, vol. 4, no.1, pp. 1-13, 2009.
[9] Ottenbacher K., Smith P., Illig S., Linn R., Fiedler R., and Granger C. ―Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke,‖ Journal of clinical epidemiology, 54(11):1159-1165, 2001.
[10] Murdoch T., Detsky A., ―The Inevitable Application of Big Data to Health Care,‖ JAMA. 2013; 1351- 1352. doi:10.1001/jama.2013.393.

Licensed Under Creative Commons Attribution CC BY