# Predicting behavioral data using ID3

[1]Dr. R. Rajani, [2]B. Phani Krishna, [3]Sk.Fareeda, [4]A.E. Kokila
Narayana Engineering College, Nellore.

***Abstract*—Predicting human behavioral data is challenging due to its characteristics like huge in size of data, different behaviors and interest outcomes of every individual is imbalance in state. Due to this, predicting an accurate model for identifying the behavior of human beings is a biggest challenge. To address this challenge we can depend upon various statistical models to describe about the behavioral data of individuals. Here, we consider an algorithm ID3 (Iterative Dichotomiser 3) Decision Tree which is a variant of decision tree algorithms. It is the most suited algorithm for identifying the categorical data values.***

***Index terms*— ID3, categorical data values, Human behavioral data, Feature selection**

## I. INTRODUCTION

In recent years, prediction has become a biggest challenge for evaluating social data using supervisory models. This issue is resolved by integrating various algorithms such as Machine Learning, Decision Trees and Support Vector Machines (SVM). These algorithms have the ability to accurately predict untrained data over trained interpretable models.

Interpreting behavioral data of each and every individual is successfully done by using Artificial Intelligence techniques. We know that behavioral data is massive in nature which consists of many individual records, each with a large number of potentially highly correlated features. However, the data is also different and imbalanced. Also the data is composed of subclasses that vary widely in their behavioral characteristics. For example, online users have very less number of followers and indeed they post a fewer number of messages. However, few users have millions of potential followers. Avoiding sparse behavioral characteristics of many individuals may lead data analysts to predict inaccurate conclusions due to various statistical procedures.

Online communities, Machine learning and data science have proposed a number of approaches to understand the data using supervised models. The popular models are regression methods, decision trees and their ensemble variants, such as random forests and boosting methods. However, these approaches are dominant one upon another. For example, Regression models (e.g., Lasso, Logistic Regression, Linear Regression, Elastic Net) others limited interpretability due to their failure to capture relationships in data that do not adhere to this form, and thus can be ineffective at adequately representing the data. Tree-based methods are very effective at capturing non-linear and imbalanced data, yet have limited interpretability. However, they provide a metric of feature importance, the relationship between the response and features is more ambiguous as it requires moving towards the depths of many trees, potentially with the same features appearing at different levels.

## II. RELATED WORK

Explanations of human and social phenomena that provide interpretable causal mechanisms often ignore their predictive accuracy. However, we argue that the increasing computational nature of social science is beginning to reverse this conventional bias against prediction; and, it has also highlighted three important issues that require resolution. Firstly, current practices for evaluating predictions must be better standardized. Theoretical limits to predictive accuracy in online systems must be optimized, thereby setting forecasting for what can be predicted or explained as a second issue. As a third challenge, predictive accuracy and interpretability must be recognized as complements. Resolving these three issues will lead to better, more replicable, and more useful social science.

Jon Kleinberg, Himabindu Lakkaraju et. al, in their paper came to a conclusion that judges may fail in offering a good decision about the defendants. The authors analyzed a case study about prediction of judgement over many criminals. The judges generally take decisions based on prior judge decisions. This makes it hard to evaluate counterfactual decision based on various parameters such as judges may have a broader set of preferences; for instance, judges may care specifically about violent crimes or about racial inequities.

## III. METHOD

Motivated by the need for algorithms that perform strongly the goal of prediction, we propose an ***Iterative Dichotomiser 3 (***ID3), a mathematically principled method for learning interpretable statistical models of behavioral data. The algorithm, which is a variant of decision trees, transforms raw data to rule based decision making trees. It is both highly interpretable and can be used for out-of-sample prediction. In addition, the learned models can be used to visualize data.

ID3 works as follows: First of all, Dichotomization means dividing data into two completely opposite things. The algorithm iteratively divides attributes into two groups which are the most dominant attribute and others to construct a tree. After that it calculates the entropy and information gain of each attribute. In this way, the most dominant attribute ca be determined. Then the most dominant one is put on the tree as a decision node. Thereafter, entropy and gain scores would be calculated again among other attributes. Thus the next most dominant attribute is found. Finally, this procedure continues until reaching a decision for that branch.

To execute the proposed method, we apply it to model a variety of datasets, from most significant to large-scale heterogeneous behavioral data that we collect from social platforms, including Twitter, Facebook etc. To better understand the behavior of human beings, ID3 uses Entropy. Entropy is a measure for calculating the certainty of an element i.e., for example, the element can provide a response of type "yes" or "no". The formula for Entropy is:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Where P= $p_1, p_2, \ldots . p_n$

The next step of ID3 is Information Gain. Gain is determined by partitioning a set T into subsets $T_1$, $T_2$, .., $T_n$ .

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

To implement ID3 algorithm, attributes play a big role. Every attribute needs to help the algorithm with distinguishing between different query contexts, or no information gain will result from a split on it. To help with distinguishing between query contexts, the time of the query and the query string were used. Given these attributes, it is possible to calculate or to predict the human behavior.

## IV. CONCLUSION

Examination of current research practice suggests that, by and large, researches opt for a different framework like usage of Deep Learning, Recurrent Neural Networks, Data Science etc. to predict "What is going on inside people's head's, as a basis for predicting future behavior. In particular, we are planning to make use of ID3 algorithm for human behavioral prediction.

## REFERENCES


1. Lipton ZC (2018) The mythos of model interpretability. ACM Queue 16(3):30. https://doi.org/10.1145/3236386.3241340
2. Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4(1):5580
3. Alipourfard N, Fennell PG, Lerman K (2018) Can you trust the trend: discovering Simpson's paradoxes in social data. In: Proceedings of the eleventh ACM international conference on web search and data mining—WSDM'18. ACM Press, New York, pp 19–27. https://doi.org/10.1145/3159652.3159684. 1801.04385
4. Fennell PG (2018) GitHub. https://github.com/peterfennell/S3D
5. Hofman JM, Sharma A, Watts DJ (2017) Prediction and explanation in social systems. Science 355(6324):486–488. https://doi.org/10.1126/science.aal3856
6. Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2017) Human decisions and machine predictions. Q J Econ 133(1):237–293. https://doi.org/10.1093/qje/qjx032
7. Dheeru D, Karra Taniskidou E (2017) {UCI} Machine Learning Repository. http://archive.ics.uci.edu/ml
8. Candanedo LM, Feldheim V, Deramaix D (2017) Data driven prediction models of energy use of appliances in a low-energy house. Energy Build 140:81–97. https://doi.org/10.1016/j.enbuild.2017.01.083
9. Settles B, Meeder B (2016) A trainable spaced repetition model for language learning. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Stroudsburg, pp 1848–1858. https://doi.org/10.18653/v1/P16-1174
10. Dorie V, Harada M, Carnegie NB, Hill J (2016) A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. Stat Med 35(20):3453–3470. https://doi.org/10.1002/sim.6973

## V. BIOGRAPHY

**Dr. R. RAJANI** is a Professor and heading the department of MCA, Narayana Engineering College, Nellore, AP, India. She pursued her Ph.D from Sri Padmavathi Mahila University, Tirupati, India. She guided many projects for B.Tech and PG students. Her research interests include Data mining, Query Optimization, Computer Networks and Software Engineering etc.,

**B. Phani Krishna** had completed B.Tech and M.Tech from Jawaharlal Nehru Technological University. He is currently working as an Associate Professor in Narayana Engineering College, Nellore, Andhra Pradesh. His areas of interest are Wireless Networks and IoT. Member of IAENG.

**Sk.Fareeda** had completed B.Tech in Computer science and M.Tech in Computer science from Jawaharlal Nehru Technological University. She is currently working as an Assistant Professor in Narayana Engineering College, Nellore. Her areas of interest are cloud computing and computer networks.

**A.E.Kokila** had completed B.Tech and M.Tech from Jawaharlal Nehru Technological University. She is currently working as an Assistant Professor in Narayana Engineering College, Nellore. Her areas of interest are Wireless Networks and Online Coding Competitions.