

Data Driven Prediction and Classification of Student Success

Kirk, R.A.; Wu, T.-F.; Fales-Williams, A.J.; Danielson, J.A.

Office of Curricular and Student Assessment, ISUCVM; Human-Computer Interaction Program; Veterinary Pathology; Psychology

Introduction

At the Iowa State University College of Veterinary Medicine (ISUCVM), student, faculty, alumni and employer performance and satisfaction data has been collected in order to assess the curriculum. Categories can be created a theory driven top-down approach to identify constructs. The internal consistency of these categories can be found using statistical analysis. This allows for the creation of data-based statistical clusters. Finally, once categories are established, tools from artificial intelligence and machine learning can be used to discover which categories are the most important for determining critical student success outcomes.

Methods

Every year, data is collected at the ISUCVM from the employers of ISU students who graduated the previous year. The survey is sent out to all known good addresses and the data used for this survey represents all responses across all three years where data has been collected (n=106). Establishing the initial, hierarchical structure of categories was based upon:

1. External, standardized guidelines
2. Establishing internal best practices
3. Experts in the field

Categories were validated using techniques such as regression analysis, Cronbach's alpha and factor analysis.. The algorithms used included: SVMs, Decision Trees, ANNs and NBNs (see Fig1). Linear Regression and chance were used as baselines for success. Special attention is give to the decision tree because of its current use in practice within Veterinary Medicine as a tool for economic, trade-off analysis in selecting treatment recommendations. The DTA algorithm used was similar to the C4.5 optimal, depth limited to 6 utilizing gain index to train.

Results

| Name | Abrv. | Accuracy +/- Std. Error | SVM | DTA | ANN | NBN | LinReg |
|------------------------|--------|-------------------------|-----|-------|-------|-------|--------|
| Support Vector Machine | SVM | 0.642 +/- 0.480 | | 0.674 | 0.333 | 0.405 | 0.059 |
| Decision Tree | DTA | 0.613 +/- 0.487 | | | 0.581 | 0.678 | 0.17 |
| Neural Network | ANN | 0.575 +/- 0.494 | | | | 0.891 | 0.376 |
| Naïve Bayes | NBN | 0.585 +/- 0.493 | | | | | 0.433 |
| Linear Regression | LinReg | 0.518 +/- 0.431 | | | | | |
| Chance | | 54.00 +/- 0.000 | | | | | |

*Significance asserted at alpha = 0.05

Figure 1: algorithm accuracy and their significance

Accuracy was calculated by using hold-one-out analysis within a cross-validation technique for each algorithm. A t-test was then used to calculate the significance of difference between various algorithms. Linear regression and chance are the two baselines of comparison. Here, chance is the prior odds of the most prevalent outcome (Very Satisfied) which represented roughly 54% of the data.

Conclusions

No algorithms were statistically significantly more accurate than any other algorithm or from the baseline, linear regression (see Fig 1). Nonetheless, decision trees can be recommended because :

1. They are relatively accurate (in comparison to other algorithms)
2. They are computationally less expensive than other algorithms such as SVMs and ANNs.
3. They visualize nicely and can be interpreted.

Here, the decision tree makes it clear that people skills and problem solving skills are very important for categorizing overall employer satisfaction with students. (see Fig 2).

Decision trees can be created to determine the concepts important to particular outcomes. Thus, outcomes can be clearly mapped to their important, constituent constructs and categories.

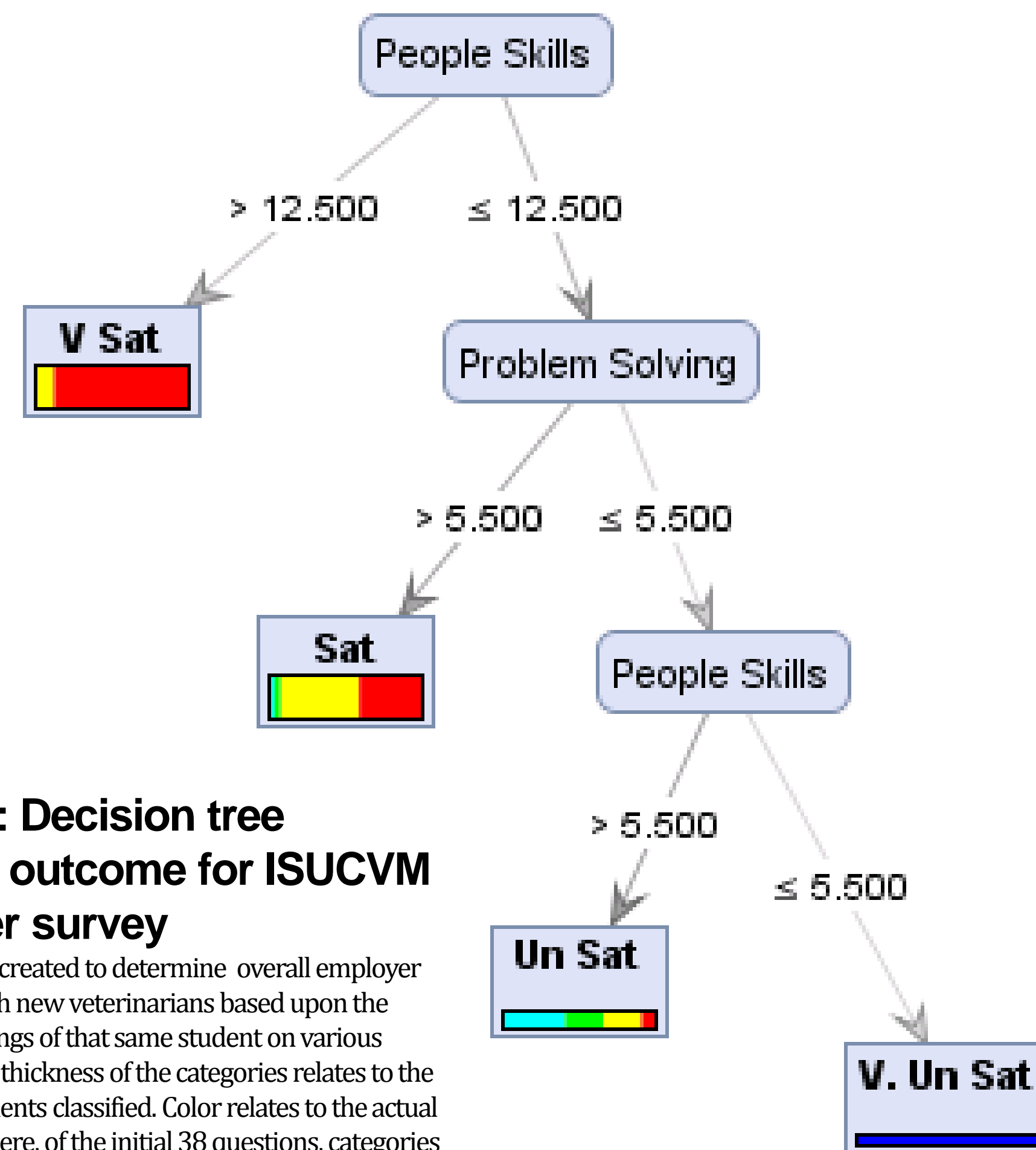


Figure 2: Decision tree analysis outcome for ISUCVM employer survey

A decision tree created to determine overall employer satisfaction with new veterinarians based upon the employer's ratings of that same student on various categories. The thickness of the categories relates to the number of students classified. Color relates to the actual classification. Here, of the initial 38 questions, categories have been statistically clustered as a form of pre-processing.

