EcoSurvey Analysis Progress Report

Conor McNamara
University of Colorado Boulder
College of Engineering and Applied Science
Inquiry Hub
Conor.R.Mcnamara@colorado.edu

Shawn O'Grady
University of Colorado Boulder
College of Engineering and Applied Science
Shawn.Ogrady@colorado.edu

ABSTRACT

This paper will outline our groups proposed work for the remainder of the semester to analyze EcoSurvey models. In this outline we cover what knowledge can be mined from the EcoSurvey data, how data mining has been applied to this field previously, how we plan on working with our data set, and our expected steps required for completing this.

0.1 Questions to Answer

Our primary research question that we are investigating with this project was given the features of a subject organism and given the features of an object organism, can we, using classification algorithms, determine the relationship between the two organisms. A question that quickly became a part of this question as the project moved forward was what are the most predictive features, or in other words, which are the most important features for said classification. After drawing conclusions from our project, we can compare the results from the data from version two of EcoSurvey to the results of version one of EcoSurvey to answer the question what are the differences in your findings. These differences will help us answer the question which elements of the EcoSurvey application are effective and which elements are not as effective. After determining the elements of EcoSurvey that are most important, we can then attempt to answer the question, how do these important aspects of EcoSurvey relate to student learning. This result will help us answer the ultimate question: how can we use our findings to build more student centered curriculum for the future beyond EcoSurvey?

0.2 Summary of Results

Our results of this project have shown that we can predict the relationship between a subject and an object organism with seventy-eight percent accuracy with the data from version two of EcoSurvey. We found that the most predictive features of this data set were the organisms name, ID number, tag and role of both the object and the subject organism. When comparing these results to the previous finding of the analysis of the data from version one of EcoSurvey, we discovered that the most predictive features were identical and the percent accuracies between the two versions differed by merely two percent, version one being slightly more accurate than version two. We also determined that there was much more data in the version two of EcoSurvey, meaning that the students interacted with this version more than they did with the first version. The key changes that were made between the two versions were closing the relationships text field upon card creation, as well as implementing a visualization of the users ecosystem which they were creating. Given that there was positive feedback, namely more use, we found these changes to be effective in promoting student learning. In summation, having options instead of open text as well as having visualizations of their model were effective tools for promoting student learning.

CCS CONCEPTS

• **Computing methodologies → Feature selection**; **Model verification and validation**; *Supervised learning by classification*; *Classification and regression trees*; *Bayesian network models*; Ranking;

KEYWORDS

EcoSurvey, prediction algorithm, WEKA, Scientific Modeling, Collaborative Modeling, Teacher Differences, Classification

ACM Reference format:
Conor McNamara and Shawn O'Grady. 2017. EcoSurvey Analysis Progress Report. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 8 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Our research focuses on issues related to student scientific modeling in a student-centered, inquiry-based curriculum. Modeling practices are essential to student learning because they reveal new learning analytic methods to better characterize student scientific modeling practices and to examine classroom level differences. For this project, we have used computational techniques to study students' modeling practices at a detailed level, at unprecedented scale, and in new ways. Our computational techniques and interpretation of their results will be guided by theories and prior research on effective modeling practices arising from learning sciences research.

For this project, applied data mining techniques to an EcoSurvey model. EcoSurvey is an application that students use to build scientific models of their local ecosystems.

Scientific models are tools for explanation and prediction. A complete scientific explanation should explain observed relationships between variables and describe the mechanisms that support cause and effect inferences about them. Thus, to support student explanations, a scientific model of a phenomenon should include important

Conference'17, Washington, DC, USA
© 2017 ACM. 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

components (variables), their interactions (relationships), and define the mechanisms involved. When modeling an ecosystem, these correspond to the organisms in the ecosystem (animals, plants, insects, fungi, etc), how these organisms interact with each other and the environment (predator, prey, producer, de- composer, etc), and the involved processes (abiotic, biotic, etc). Professional biologists use this information to measure the biodiversity of an ecosystem in terms of species richness, evenness, and divergence [1].

We also compared and contrasted the two different deployments of EcoSurvey to understand how the differences between the two versions influenced the students learning. In order to do this, we used data mining and machine learning techniques. By understanding and analyzing student models, we can develop theories in regards to the nature of these students learning and use this to help develop student-centered curriculum.

This analysis will have many potential uses, primarily to help develop future models for a more student-centered learning environment.

2 RELATED WORK

As our work is examines student centered learning, we have examined many similar topics. The most significant prior work that we looked at was self-determination theory, which examines peoples motivation behind their decisions without external influence. This is relevant to our research as we are determining how users interact with our application and exploring what motivates them to make their decisions when creating the model.

In addition to the self-determination theory, our work revolves around scientific modeling. As Homer states in his paper Why we iterate: scientific modeling in theory and practice, scientific modeling is distinguished from other practices due to it emphasis on iteration. By using iteration in scientific modeling, models can be revised constantly and, thus, develop into better representations. We implemented this iterative design process as a part of the EcoSurvey application, having users constantly iterate their models to help build the best possible scientific model as well as use the proper scientific modeling practices which Homer outlines [2].

Scientific models assist us in visualizing relationships between data as well as organizing the data in a way that makes the data easier to both mine and interpret. With the use of scientific models and the theories behind student centered learning, we hope to create an interpretation of our data that will help generated a more student centered curriculum.

We are also drawing a large part of our research and background from the work previously done by David Quigley, Jonathan Ostwald and Tamara Sumner called "Scientific Modeling: Using learning analytics to examine student practices and classroom variation." [1]

3 DATA SET

As previously stated, our data set comes from the EcoSurvey application, and contains data on both the organisms in the local ecosystem and the students interactions with the application. The data set is stored in a MySQL database, and has been organized in to forty different tables. All tables combined, we have 142354 rows and 433 columns, for a total of 61,639,282 data points. This should provide more than enough information to generate a prediction algorithm.

This is actually the second time analysis has been applied to EcoSurvey data, the first was performed by author Conor McNamara as a part of separate research. The data set that we are planning on analyzing for this project represents a second version, which was constructed from the results of the analysis on the first.

In addition, the data set requires authorization and training from the Collaborative Institutional Training Initiative (CITI) Program to access.

The important features that we selected from this data set were the features of data about the subject organism and the object organism as well as the features describing the relationship between the two organisms. We used queries to draw this data from our given relational database and found the following features to be the ones that we would conduct our test with:

- subject_time_stamp
- subject_name
- subject_description
- subject_url
- subject_lastmod
- subject_gid
- subject_owner_id
- subject_image
- subject_lat
- subject_longitude
- subject_cTmp
- subject_roleName
- subject_tagName
- obj_time_stamp
- obj_name
- obj_description
- obj_url
- obj_lastmod
- obj_gid
- obj_owner_id
- obj_image
- obj_lat
- obj_longitude
- obj_cTmp
- obj_roleName
- obj_tagName
- relationTableID
- obj_id
- subject_id
- tmp
- relationship_name
- relationship_description
- relationship_type
- relationship_inverse
- relationship_inverse_id

4 MAIN TECHNIQUES APPLIED

The phases of data mining and machine learning consist of exploring the data, data preprocessing, selecting features and analyzing methods. Once again, it is important to note that these phases are not linear. Thankfully, the entirety of our data set has been generated from the EcoSurvey application, so no additional data collection is required.

4.1 Tools

In order to analyze the data from the EcoSurvey model, we will utilize the following programs:

(1) GitHub: We will be using GitHub to manage our code for the project, as well as any required documentation and write ups.
(2) MySQL: All of our data is currently located in a MySQL database, which we will need to continually access throughout the course of this project
(3) Microsoft Excel: In order to better visualize our data, we will use Microsoft Excel's various plotting features
(4) Python: We will be using Python scripts to take the data from the MySQL database, convert it to a .csv file, then convert it to .arff file
(5) Sublime: Clearly, to create Python scripts, we will need to use a text editor such as Sublime
(6) WEKA: The majority of the actual data analysis portion of our project will be performed using WEKA. WEKA allows us to use various attribute selection and classification schemes in order to develop the optimal prediction algorithm for our data set. WEKA will be an extremely valuable tool for this project, particularly for comparing several prediction algorithms.
(7) Microsoft Word and LaTeX : To formally present our findings and describe our methods, we will need to make use of a word processor or typesetting system.

By exploring the data provided in this way, we were able to determine particular parts of the data which we needed, an other parts which we did not. We were also able to make a plan as to how we would join particular queries to get the data that we needed for our analysis. By visualizing the data this way, join keys became evident, which helped a great deal in our query plan. We took this knowledge into our 'Data Preprocessing' process.

4.2 Exploring the Data

We wrote queries in mysql to create new tables with the particular features which were relevant to our analysis. After writing the queries and altering the tables into more clear and concise data tables, we translated these mysql tables to a comma separated value file, better known as CSV.

Next, we wrote a python script which took this CSV file as an input and altered it into an attribute relation file format, better known as ARFF. ARFF files are the only type of file that WEKA accepts as an input into its software. This ARFF creation took a large amount of effort and processing of the data, but eventually came to fruition.

After loading the data in WEKA, our preprocessing truly began. In order to run the classification algorithms we wanted to on the data set, we needed to convert all of the attributes from type 'String' to type 'Nominal.' After accomplishing that, we discussed the relevance and importance of each attribute and deleted the ones that we deemed unnecessary in predicting the relationship between the object and the subject organism. We used WEKA's preprocess function to do this.

4.3 Data Preprocessing

The next step is data fitting, where we used Python to write code that selects the correct and relevant data.

4.4 Feature Selection

After the data had been properly preprocessed, we continued using WEKA to discover the most predictive features of our data set, which were derived from a training data set. The AttributeSelection feature of WEKA allowed us to use our training data to determine which attributes we will use to perform our classification, and how to rank them. As one of our project goals was to determine which features of the data set are the most predictive of the relationship between two organisms, this WEKA feature was very helpful in our analysis.

4.5 Analyzing Methods

Finally, we used Weka to analyze that data and determine the accuracy of the fit. In order to do this, we used supervised methods for data analysis, namely classification and clustering in Weka. We then recorded our findings and proceeded using the exact same process as used for the data of the first version of EcoSurvey. We then compared our findings for the first version's data to the findings from the second version's data, and determined the significant differences. We then recorded these differences, and used them to determine our ultimate analysis. This analysis will help improve both the EcoSurvey tool and assist in the development of a foundation for a more student centered curriculum.

5 KEY RESULTS

5.1 Increase in the Number of Interactions

Our analysis of version two of EcoSurvey revealed some extremely interesting key findings. The first key finding of note is the number of instances that occurred during the use of second version of EcoSurvey. The first version of EcoSurvey only had three hundred and three instances. However, the second version of EcoSurvey had five thousand, seven hundred, and thirty-five instances. This shows that the second version of EcoSurvey had many more organisms charted than version one of EcoSurvey. This shows that the students using version two were creating more organism cards and, therefore, interacting with version two a great deal more than version one, particularly 18.92 times more. This shows that version two of EcoSurvey was much more effective at drawing the students attention and, therefore causing them to interact with it more than version one.

5.2 New 'Mutually Benefits' Reltionship Type

Our second key finding was that the addition of the 'Mutually Benefits' relationship did not greatly change the percent accuracy of the classification of our model. This was not at all used in the first version of EcoSurvey. The users would input the relationships and that data was then normalized to the relationships that you see in the version one confusion matrix alone. However, after adding closed text boxes to the relationship field in the second version of EcoSurvey and having the users select one of these options, 'Mutually Benefits' was used a total of four hundred and eighty-nine times. This finding was interesting as it shows that the relationship that was added to the system was indeed used. However, when observing the percent accuracies of the classifications on version one and version two, one would notice that they are fairly comparable, namely version one being about 80% accurate and version two being seventy-eight percent accurate. This can be shown using the confusion matrices for versions one and two, shown in Figure 1 and Figure 2 respectively.

5.3 Accuracy of Classification of Relationship Types

Another key finding is the distribution of percent accuracy among the various relationships. For example, in version one of EcoSurvey, 'Unknown' was only sixty-one percent accurate(Figure 3). However in version two of EcoSurvey, 'Unknown' was twenty four percent accurate(Figure 4). In the first version of EcoSurvey, 'Preys Upon' was classified with ninety-one percent accuracy(Figure 3). However, in version two of EcoSurvey, 'Preys Upon' was classified with eighty-seven percent accuracy(Figure 4). Both 'Unknown' and 'Preys Upon' were classified less accurately in version two than in version one.

However, 'Competes' and 'Supports' were classified better in version two than version one. 'Competes' was classified with forty-one percent accuracy in version one(Figure 3), whereas in version two, it was classified with sixty-eight percent accuracy(Figure 4). 'Supports' was classified in version one at fifty-seven percent accuracy(Figure 3), where in version two of EcoSurvey, it was classified with eighty-seven percent accuracy(Figure 4).

5.4 Percent Accuracy Analysis

One of the hallmark findings that we discovered with our project is that the total percent accuracy of the classification of the first version of the data set was 80% and the percent accuracy of the classification of the second data set was 78%. It is interesting how comparable the predictability of the organism's relationships are, in both version one and version two of EcoSurvey.

5.5 Most Predictive Features

Additionally, our research found that the most predictive features of determining relationships were both the subject and object organisms ID, name, role, and tag. We determined this using Weka?s feature selection function. This was interesting because they were the same as the most predictive features found in the analysis of version one. This shows that these features are extremely important to keep in the application in the future.

5.6 Distribution of Relationship Types

Our analysis on version two of EcoSurvey also revealed to us interesting information regarding the distribution of relationship types. Figure 6 visually depicts this distribution of how many times users categorized organism relationships into the various relationship types. What is important to note here is that the bulk of the relationships between organisms were categorized as 'Supports.' The second most categorization of relationship types was 'Preys Upon.' In addition, the least categorized relationship type was 'Unknown' with merely one percent.

Figure 5 visually depicts this exact same distribution for version one of the EcoSurvey application. There are some key differences here between the distribution of relationship types between version one and version two. In version one, 'Preys Upon' is the most categorized relationship as opposed to 'Supports' in version two. However, 'Supports' is the second most prominent relationship in version one. It is also interesting to note the 'Unknown' is still the least categorized relationship, making it the least used relationship type in both version one and version two.

5.7 Analysis of Overfitting

An area that we wanted to explore was if our classification algorithms were overfitting or not when we were determining our results. To visualize this better than the confusion matrices seen above, we created Figure 8. The bar graph clearly shows where overfitting is occurring and where it is not. As it depicts, there is not great degree of overfitting with our analysis of the version two data set. It may be over classifying 'Preys Up,' but it not by enough to determine that our classification algorithms are overfitting.

We also decided to analyze the overfitting of the data from the version one data set. As seen in Figure 7, the classification algorithms ran on version one also over fit 'Preys Up.' However, it over-fit it a bit more than we did in our analysis of version two. This shows that our choice in using the Naïve Bayes classification algorithm in version two was a good one as we got strong results and did not over-fit.

6 APPLICATIONS

The ultimate goal for our project is to help create a more student-centered curriculum. To show how our findings relate to this, we will be being with our results. Our results show that in the the second version of EcoSurvey, there was a great increase in interactions with the application. There were many changes that were made from the first version to the second version, and these changes in the format of the application clearly resulted in more student interactions with the application. The key changes that were made from version one to version two was closing some of the fields from open text to having the user select certain options. Another major change was that the relationships graph was made a much more prominent part of the application so that users would be able to visualize the ecosystem model that they created. As a result of these findings, it is reasonable to conclude that by adding the relationship graph as a more prominent part of the application and by closing the text fields, users were able to interact better with the application in a more effective and efficient way. Therefore, when building applications for students in the future, using these two methods would be effective in having the students interact more with the application.

A more direct application of our project is to help develop the next version of the EcoSurvey application. By observing our conclusions, we are able to identify what changes to EcoSurvey were positive and what changes were negative. Knowing what worked well and what did not will be very helpful in building the next version of EcoSurvey.

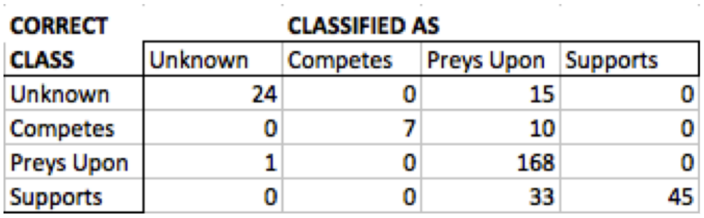
Our research also helped determine which features of organisms were most important in predicting the relationships between organisms. As a result, our findings can be used to determine which features should remain in EcoSurvey, and which features are less crucial to the application. This can help build a better application with less unneeded data and allow users to focus more on the data that is the most important to the project.

7 REFERENCES

REFERENCES

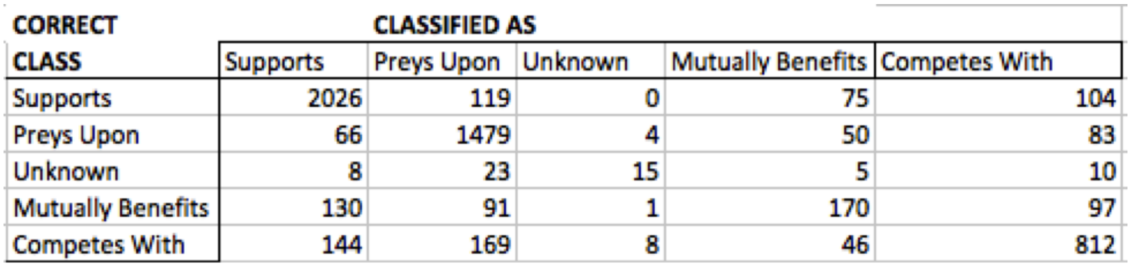
[1] David Quigley, Jonathan Ostwald, Tamara Summer (2016) "Scientific Modeling: Using learning analytics to examine student practices and classroom variation", University of Colorado Boulder

[2] J. B. Homer. Why We Iterate: Scientific Modeling in Theory and Practice. System Dynamics Review, 12(1):1?19, 1996



CORRECT	CLASSIFIED AS			
CLASS	Unknown	Competes	Preys Upon	Supports
Unknown	24	0	15	0
Competes	0	7	10	0
Preys Upon	1	0	168	0
Supports	0	0	33	45

Figure 1: Confusion Matrix of the normalized relationship types from version 1



CORRECT	CLASSIFIED AS				
CLASS	Supports	Preys Upon	Unknown	Mutually Benefits	Competes With
Supports	2026	119	0	75	104
Preys Upon	66	1479	4	50	83
Unknown	8	23	15	5	10
Mutually Benefits	130	91	1	170	97
Competes With	144	169	8	46	812

Figure 2: Confusion Matrix of the normalized relationship types from version 2

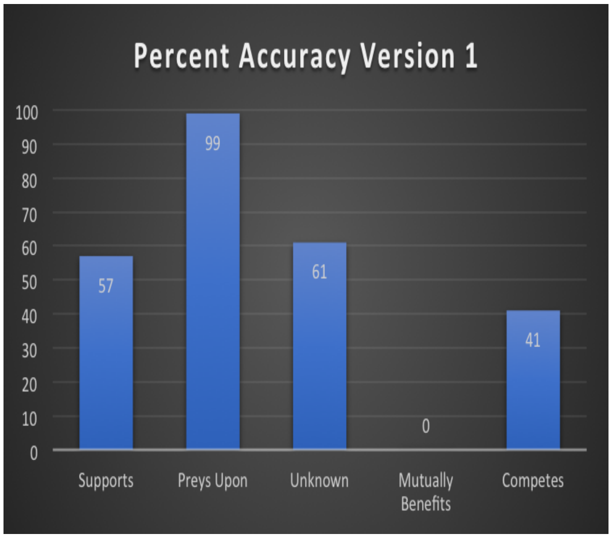


Figure 3: Percent Accuracy Bar Chart of the normalized relationship types from version 1

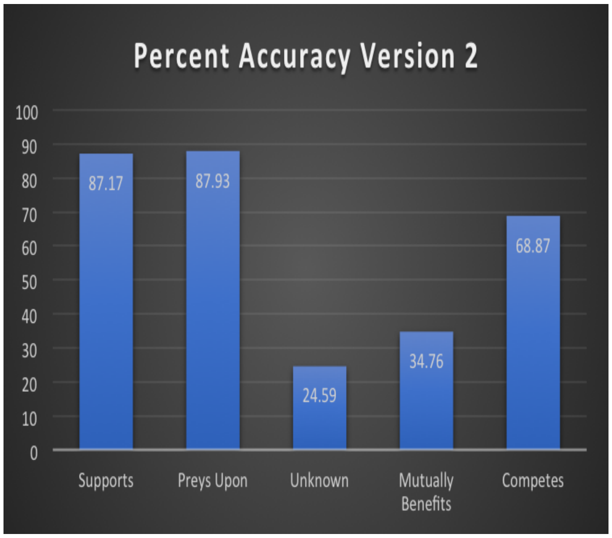


Figure 4: Percent Accuracy Bar Chart of the normalized relationship types from version 2

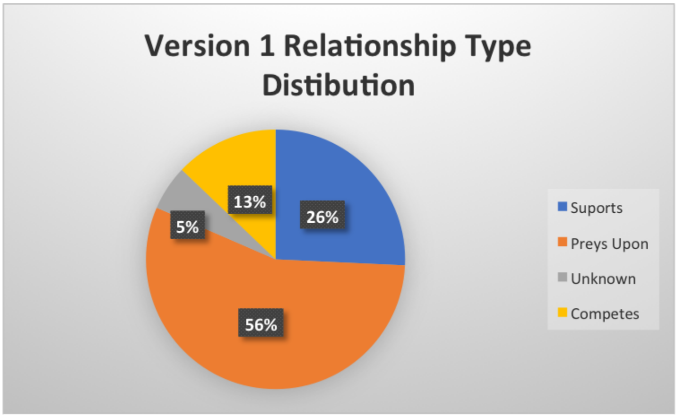


Figure 5: Relationship Type Distribution for version 1

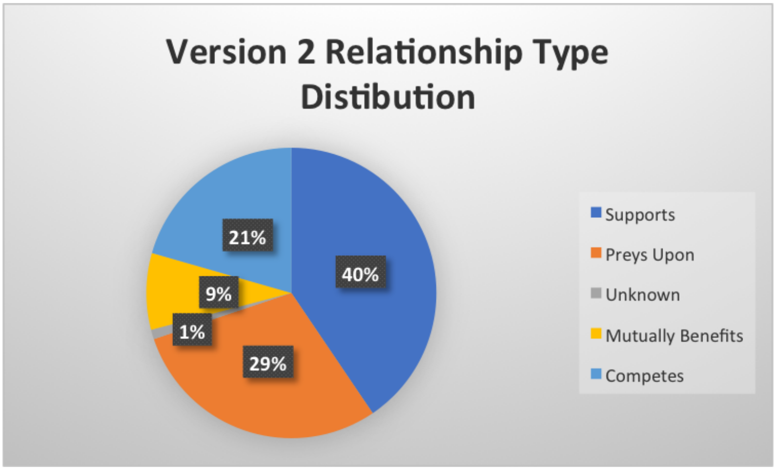


Figure 6: Relationship Type Distribution for version 2

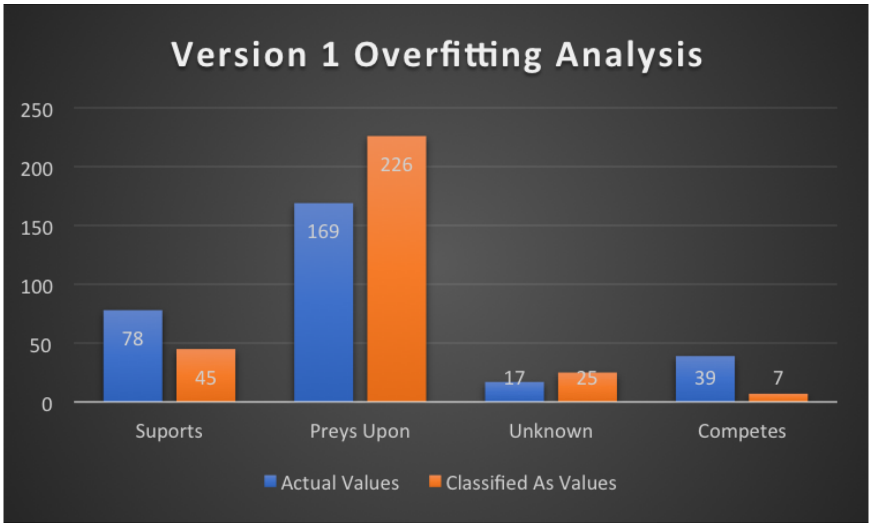


Figure 7: Overfitting Analysis for version 1

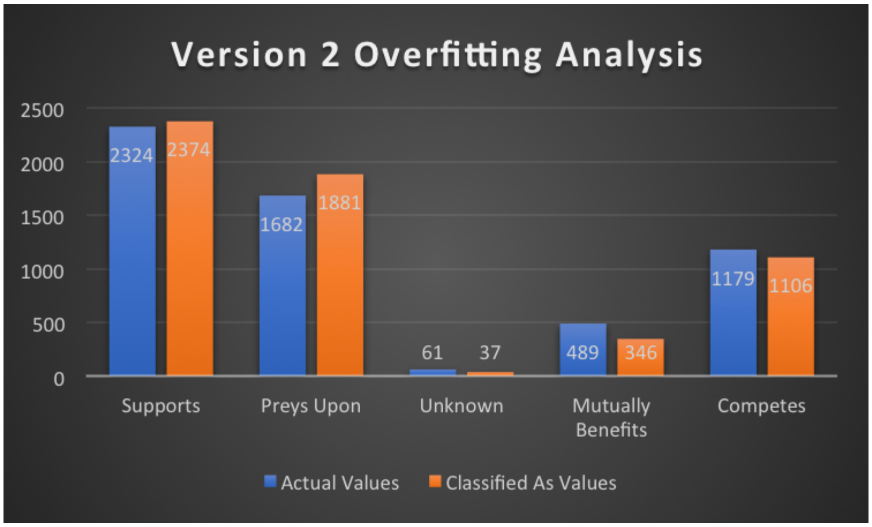


Figure 8: Overfitting Analysis for version 2