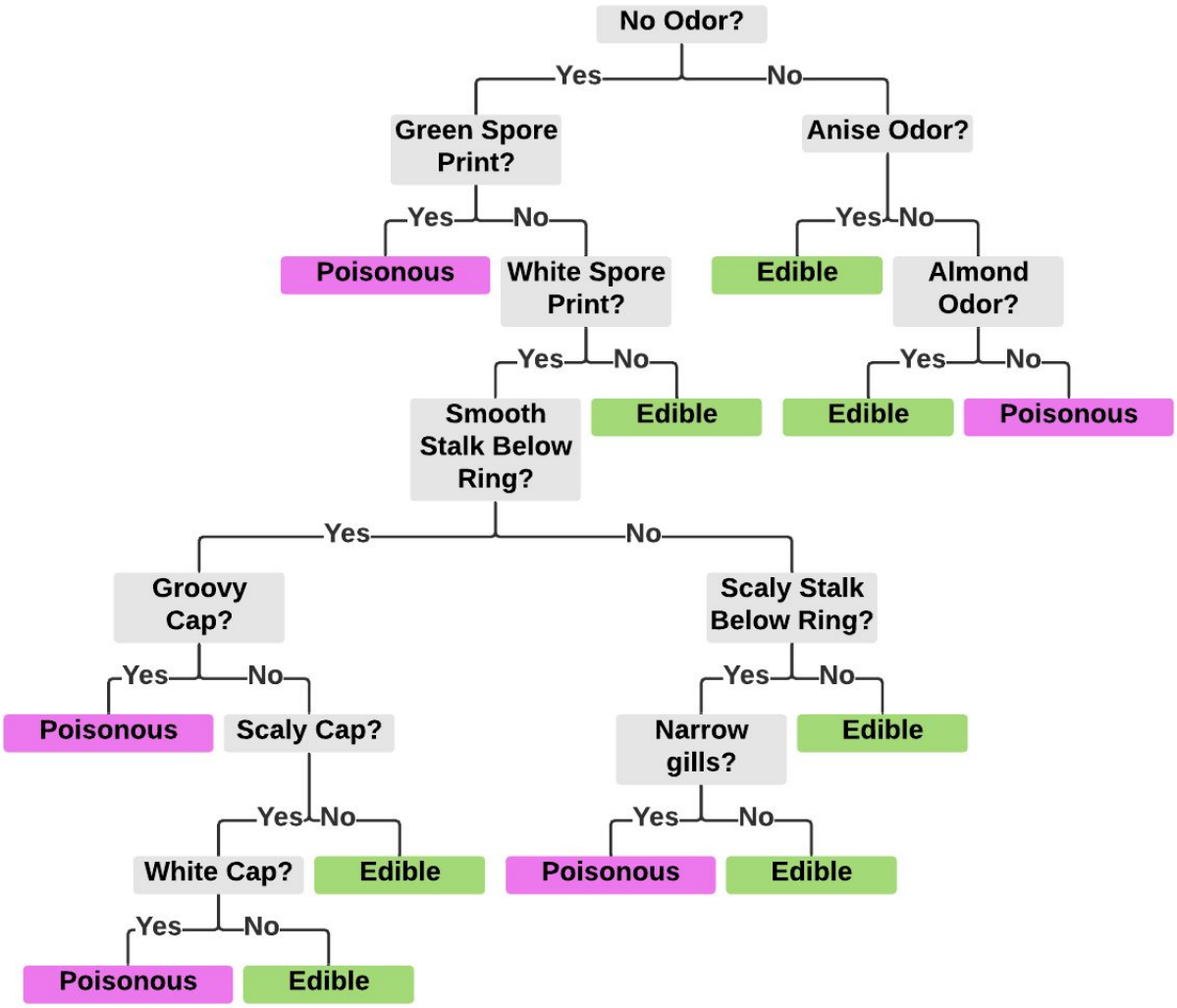
Final Report

Mushroom Classification



Sergio Robledo

December 4, 2019

Project Overview

Foraging for wild mushrooms, or mushroom picking, is the pastime of many people, and even though it does not seem like a dangerous hobby, consumption of a misidentified mushroom could cause death; it is therefore imperative for the person foraging for mushrooms to know the mushrooms' edibility. However, in order to become experienced enough to tell edible and poisonous mushrooms apart, a person would have to invest a good amount of time, because knowing the difference between an edible and fatal mushroom might be in the slightest detail. In order to make mushroom foraging more accessible to a wider range of people while preventing a possibly fatal intoxication, a reliable machine learning model capable of classifying mushroom edibility is needed. Below, the acquisition, cleaning and exploration of the mushroom dataset and the steps needed to create a machine learning model capable of correctly classifying the edibility of mushrooms 100% of the time is explained.

Mushroom Dataset & Preprocessing

The mushroom dataset comes from the UCI Machine Learning repository and can be found here. The dataset records 22 characteristics of 8,124 mushrooms; the characteristics included range from mushroom odor to gill characteristics, and all mushrooms are labeled as either edible or poisonous. The only main preprocessing step the dataset required, was the renaming of the characteristics from single-letter representations to their full name; for example, for the possible types of surfaces seen on the mushrooms, an original value of 'f' was converted to 'fibrous', 'g' to 'grooves', 'y' to 'scaly', 's' to 'smooth' and 'k' to 'silky'.

Exploration & Visualizations

The exploration and visualization phase of the project began by looking at the distribution of the observed mushrooms' edibility; knowing the distribution of the dataset adds proves useful when deciding what machine learning algorithm to use, because some models do not handle imbalanced datasets unless the data is balanced

beforehand. The results can be seen below.

```
Total # of Observed Mushrooms: 8,124

                        Edible: 4,208 (52)%
          Poisonous Mushrooms: 3,916 (48)%
```

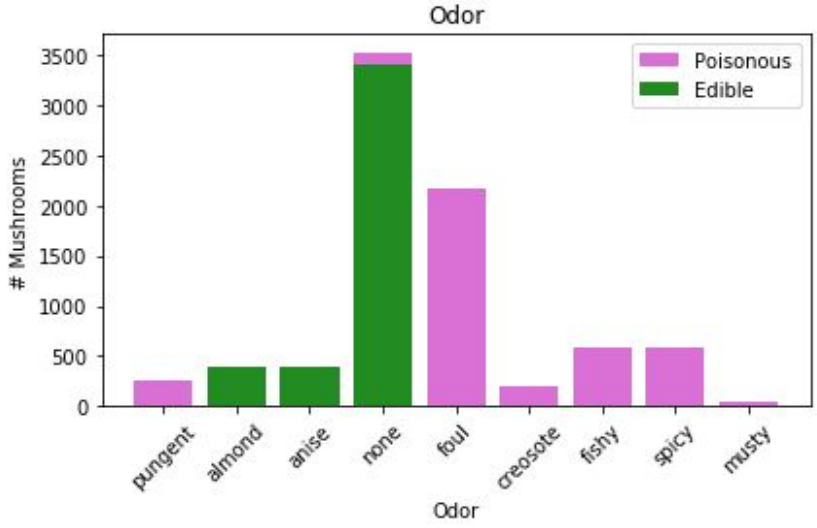
From above, we can see that there are a total of 8,124 mushrooms constituting the dataset, and about 52% are edible while the remaining 48% are poisonous; therefore, it can be concluded that the dataset is fairly balanced, so we do not need to do any further balancing.

Most Important Features:

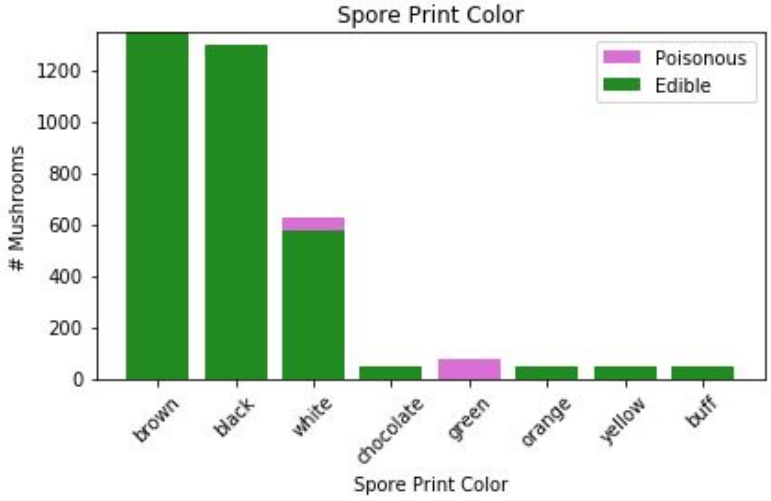
Next, using the Pearson correlation coefficient we are going to get the category with the highest correlation to the edibility classification of a mushroom and explore it further; below are the initial edibility correlations for all the mushroom properties.

```
Mushroom Edibility Correlations:
--------------------------------

odor_none                          0.785557
ring_type_pendant                  0.540469
gill_size_broad                    0.540024
bruises_True                       0.501530
stalk_surface_above_ring_smooth    0.491314
                                      ...
gill_size_narrow                  -0.540024
stalk_surface_below_ring_silky    -0.573524
stalk_surface_above_ring_silky    -0.587658
odor_foul                         -0.623842
veil_type_partial                       NaN
```

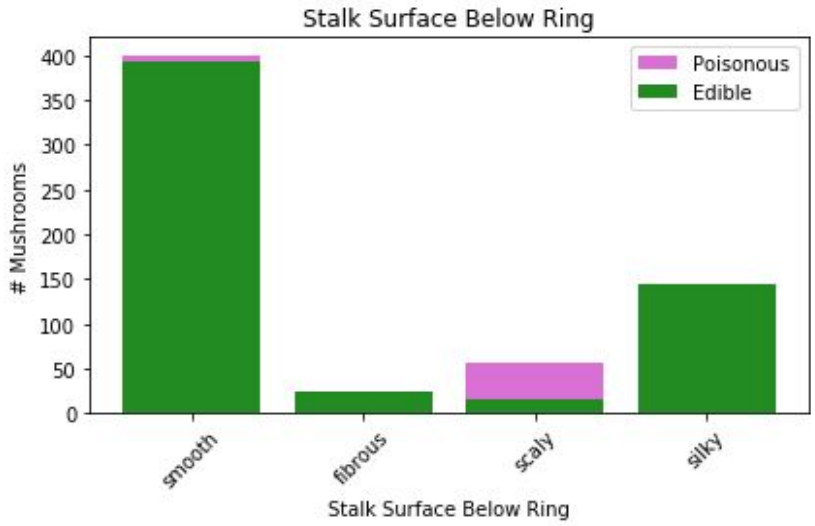
Characteristics with positive correlation values indicate that the probability of a mushroom being edible increases when it has the characteristic; on the other hand, negative correlations indicate the mushroom is more likely to be poisonous if it has such characteristic. Also, “veil_type_partial” has a value of NaN, because all mushrooms observed had partial veil types; this means that there is no correlation between the veil type of mushrooms observed and their edibility. We will now dig deeper into the dataset in an attempt to understand what rules govern the edibility of mushrooms; we will start by comparing the odors of edible and poisonous mushrooms.



The odor characteristic of the mushrooms turns out to be a very important starting feature of the mushrooms, since it can be used reliably to distinguish between edible and poisonous mushrooms; all edible mushrooms are either almond or anise smelling, and poisonous mushrooms smell pungent, foul, creosote, fishy, spicy or musty. However, mushrooms with no odor, while mostly edible, have a possibility of being poisonous, and because even one misclassification of edibility is unacceptable due to the possibly lethal consequences, we have to look at mushrooms with no odor more closely to see what sets them apart. After filtering the mushrooms to include only mushrooms with no odor, the new strongest correlation with edibility was the spore print color of the mushroom.

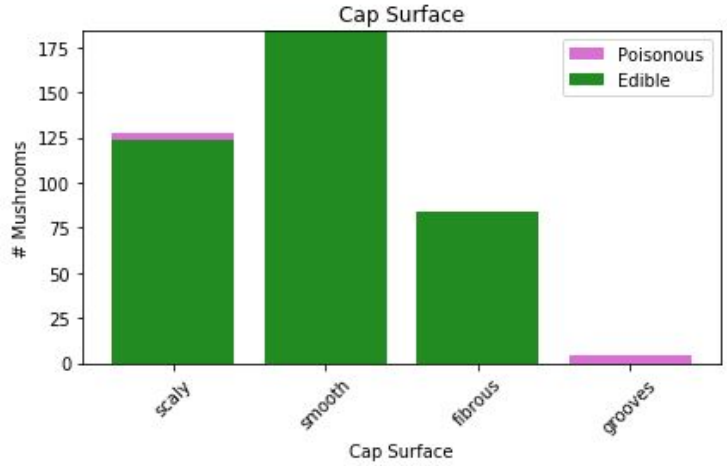


Of the mushrooms with no odor, the ones with a green spore print were all poisonous; using spore print color in conjunction with odor, we were able to reduce the number of misunderstood poisonous mushrooms from 120 to 48. However, we still have a mix of edible and poisonous mushrooms when an odorless mushrooms has a white spore print. Now, after filtering for odorless mushrooms with a white spore print, we got that the next strongest correlation to the edibility of a mushroom was the stalk surface below the ring.

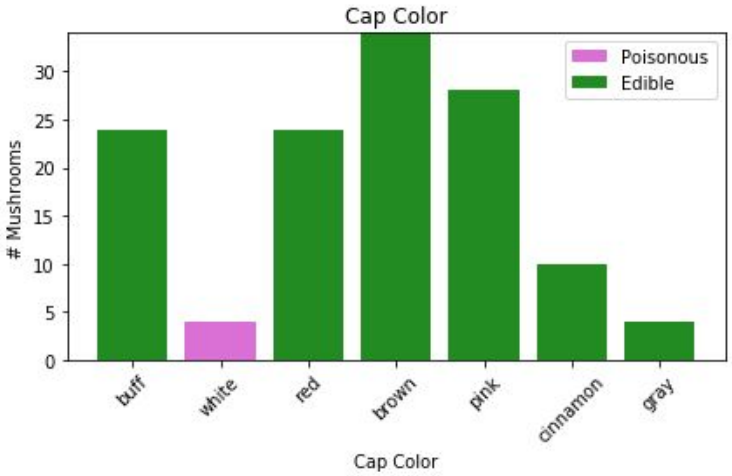


Even though this lead did not reduce the number of misunderstood poisonous mushrooms, it did reduce the pool of mushrooms that we have to further analyze by telling us that all odorless mushrooms with a white spore print and either fibrous or silky stalk surfaces below their rings are edible. Since the poisonous mushrooms mixed in with the edible mushrooms now fall under two categories, we will have to explore both separately; we will first filter the remaining mushrooms further to include ones with a smooth stalk surface below their rings. Once filtered, the strongest edibility correlation was the cap surface of the mushroom.

```
--------------------------------------------------

--- Stalk Surface Below Ring (Part 1) ---

--------------------------------------------------

Stalk Surface Below Ring (Smooth) Mushrooms: 400
Edible: 392
Poison: 8
```

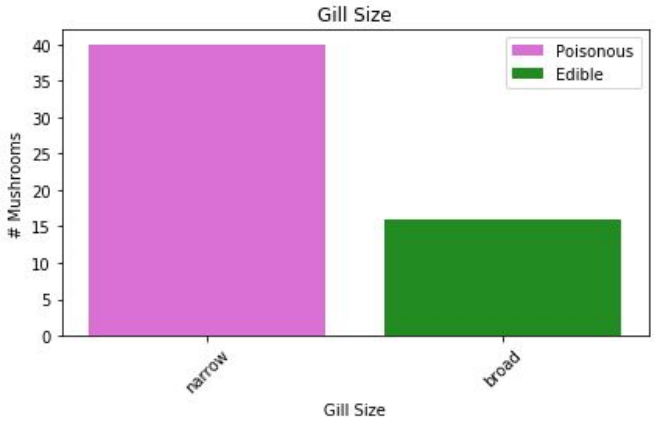


By filtering odorless mushrooms with white spore prints to smooth stalk surfaces before their rings, we can see that if their cap surface is groovy, they will all be poisonous; we are getting closer to learning the main differences between edible and poisonous mushrooms, but we still have mushrooms with a scaly cap surface to look at. Filtering by the cap color will give us the last characteristic needed to separate edible and poisonous mushrooms with smooth stalk surfaces and scaly cap surfaces.



After looking at the cap color of the remaining scaly cap mushrooms, we can finally confidently tell apart the remaining poisonous mushrooms from the other 124 edible mushrooms with the same previously looked at characteristics. The last thing we need to do is find the differences between odorless mushrooms with white spore prints and scaly stalk surfaces below their rings, since there is still a mixture of 56 edible and poisonous mushrooms with those characteristics.

```
------------------------------------------------

--- Stalk Surface Below Ring (Part 2) ---

------------------------------------------------

Stalk Surface Below Ring (Scaly) Mushrooms: 56
Edible: 16
Poison: 40
```



Finally, after filtering odorless mushrooms with white spore prints and scaly stalks below their rings by their gill size, we were able to separate the last 56 mushrooms with mixed edibility classifications into separate classifications; mushrooms with narrow gill are all poisonous, and mushrooms with broad gills are all edible. By learning what the most important characteristics are when separating edible and poisonous mushrooms, we will be able to reduce the amount of data the machine learning model will have to understand and in turn decrease the complexity of the model.

Final Dataset

The dataset we will pass to the machine learning model will only use the 6 most

important features found during exploration of the mushroom dataset; these include the mushrooms' odor, spore print color, stalk surface below their rings, cap surface, cap color and gill size. Since machine learning models do not accept name categories as input, we will have to one-hot encode the columns; one-hot encoding involves creating a new column for each category of a characteristic and setting the column values to 1 if the characteristic is present and 0 otherwise. Below we can see the top five observations before and after one-hot encoding.

	odor
0	pungent
1	almond
2	anise
3	pungent
4	none

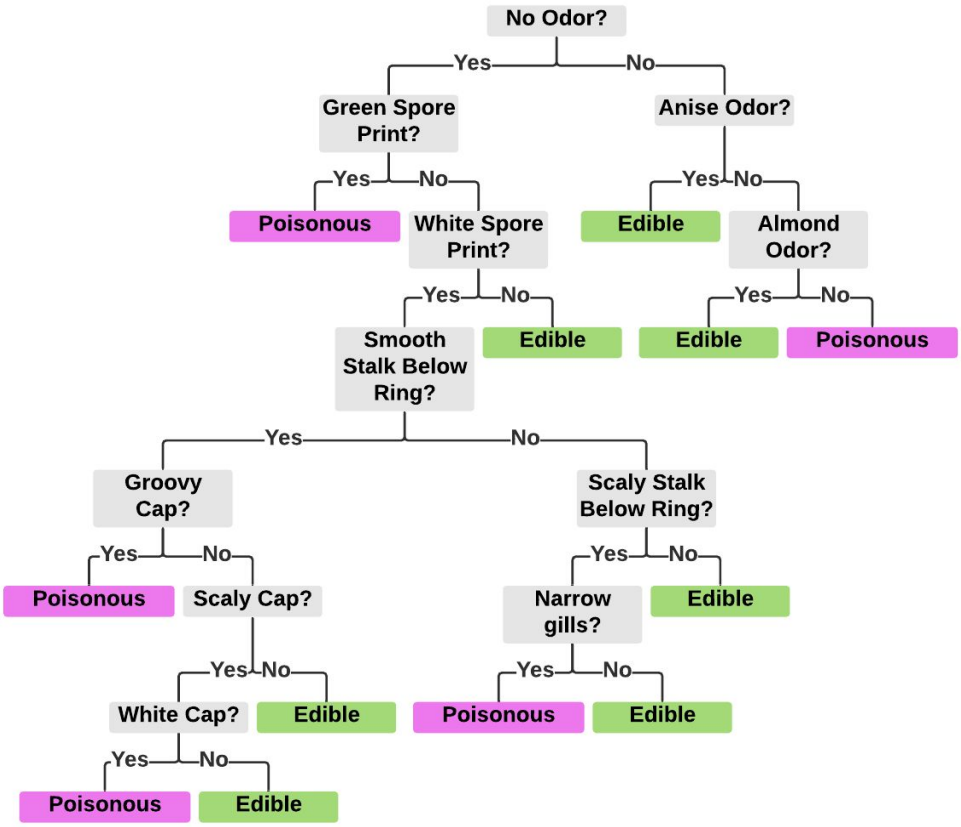
	odor_anise	odor_creosote	odor_fishy	odor_foul	odor_musty	odor_none	odor_pungent	odor_spicy
0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0
4	0	0	0	0	0	1	0	0

The left image shows the odor column of the original dataset, and the right image shows the one-hot encoded odor columns of the processed dataset; as it can be seen by comparing both datasets, the one-hot encoded dataset creates a column for every type of odor and marks a 1 if that type of odor was observed for a given mushroom, and since there can only be one odor recorded per mushroom, only one column will contain a value of 1 and the rest will contain 0s for that row. Once we have one-hot encoded the original dataset containing the 6 important mushroom characteristics, we just need to split the one-hot encoded dataset into training and testing datasets; the training set will contain 75% of the mushrooms observed, and the test dataset will contain the remaining 25% observations which will not be used at all during the machine learning model's training phase. This ensures that the final performance achieved by the learning model is valid, because it confirms that the model is not cheating its way to a stellar test score by having seen the answers to the test dataset before the testing phase. The dataset is now ready, and the machine learning model can be created.

Machine Learning Model

When hypothesizing what a good starting classifier would be, the Random Forest

Classifier came to mind, because it is made up of decision trees; decision trees begin by looking at one of the mushroom's characteristics and depending on what value is in that column, it moves on to what it considers the mushroom's next most important feature that will help with the classification of the mushroom. For example, from the exploration of the mushroom dataset we know that mushrooms with no odor are mostly edible with the exception of a few mushrooms that are poisonous, so just looking at the odor of an odorless mushroom is not enough to separate the edible and poisonous mushrooms; therefore, the task of the random forest classifier would be to find what features will help it move forward in the edibility classification of such mushrooms. Random forests are made up of multiple decision trees that each look at various combinations of mushroom characteristics to arrive at a classification, and by choosing the most popular classification among the decision trees, the random forest classifier is able to be a fairly accurate classifier; below is an example of what an optimal decision tree would look like if it used the most important characteristics found during the dataset's exploration.



If a random forest classifier contained this single decision tree, it would be able to

correctly classify the edibility of mushrooms 100% of the time given the current dataset. Due to the Random Forest classifier's ability to handle one-hot encoded data very well, which is what the mushroom dataset exclusively contains, it was selected for the first draft of what was to become the final machine learning model. Not surprisingly, after training the scikit-learn's default Random Forest classifier, it was able to correctly classify the edibility of the testing dataset's mushrooms; in other words, the out of the box classifier was able to classify the edibility of 2,031 never-before seen mushrooms without a single misclassification. When trying to arrive at an optimal machine learning model, the best performing model with the lowest level of complexity should be chosen, and since scikit-learn's base Random Forest classifier was able to achieve the accuracy desired, 100% accuracy, there was no need to further fine-tune the model. If there had been a need to fine-tune the random forest classifier, the two most important parameters of the random forest are the number of estimators, a.k.a. number of decision trees, and the maximum number of features a single decision tree could choose from as its next splitting node; as the number of estimators increase, the model's performance increases at the expense of efficiency and simplicity, and as the number of max features the model can consider for its next internal node decreases, its ability to generalize to new observations improves, because it is forced to learn from multiple features instead of the same ones over and over again. Scikit-learn's default Random Forest classifier sets the number of estimators to 10 the maximum number of features considered to the square root of the number of features in the dataset; even though only 6 mushroom features are really being used, because of one-hot encoding which uses a column for every category within a feature, the dataset passed to the model contains 32 features, and therefore, the maximum number of features the default random forest classifier can consider for each split is 5.

Conclusion

The most important lesson to be learned from the completion of this project is how far thorough exploration of the dataset can take you, because by taking the time to understand what mushroom characteristics separated edible and poisonous mushrooms the best, we were able to use a very simple machine learning model that required very little preparation; after exploring the mushroom dataset, we were able to reduce the

number of features from 22 to 6, and the only step required by the random forest classifier was the one-hot encoding of the categorical dataset. After training the final random forest classifier on the training dataset, it achieved a 100% accuracy score on the remaining 2,031 mushrooms found in the testing dataset; no fine-tuning was needed since the default random forest configuration was able to achieve the desired accuracy score set at the beginning of the project. In the future, as more mushrooms are observed, an occasion may arise where the current features and model are no longer able to fully capture the growing complexity of the mushroom dataset; in the case such occasion occurs, a greater number of mushroom features or a more complex machine learning model will have to be used, or both.