Boston University College of Engineering

Static Code Analyzer for Python

Hong Xin

Introduction

Goal:
The goal of this project is to build a static codes analyzer that finds syntax errors in python without executing the source codes.

Product Goals

Users:
I want to find syntax errors in my python codes, specifically,

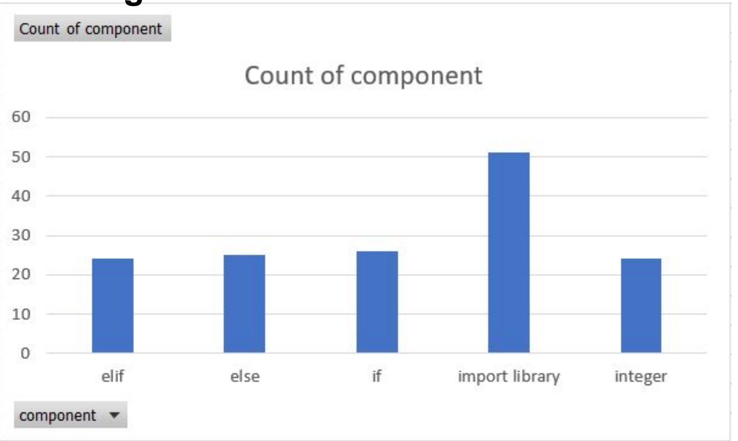
Instruction:
Copy and paste python codes to a CSV file. Each line of code will be placed in a cell. The program will read the CSV file. It will identify if a line of code contains syntax errors. The program will then calculate the accuracy of the predictive model and the accuracy of the codes.

System Components

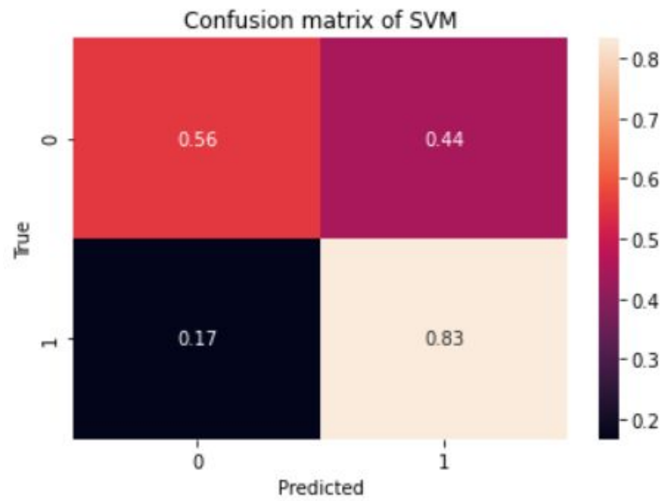
- Code
  - Model.ipynb
- Data
  - Dataset
    - dataset.csv
  - Test
    - test_if_else_elif.csv
    - test_import.csv
    - test_integer.csv
    -
- Visualization

Data

Training set:



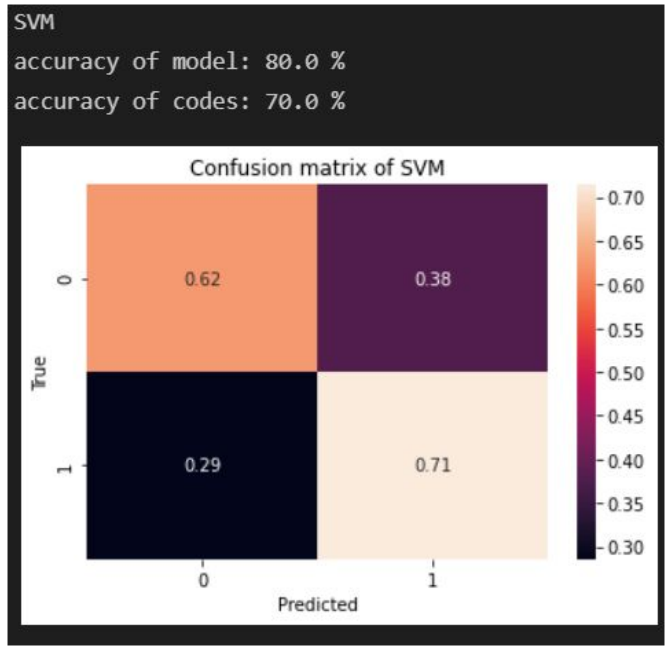
Confusion Matrix of import library:



Testing set (Test_import.csv)

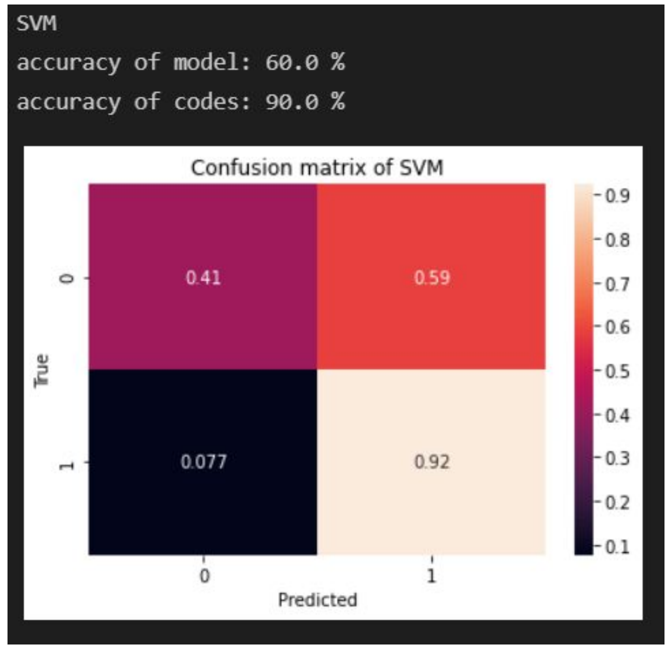
model_name	predicted_score	code	true_score	accuracy fo model(%)	accuracy of codes(%)
SVM	0	import	0	90	40
SVM	0	import	0	90	40
SVM	0	mport numpy	0	90	40
SVM	0	port pandas	0	90	40
SVM	0	port numpy	0	90	40
SVM	1	import numpy	1	90	40
SVM	1	import numpy as np	1	90	40
SVM	0	import csv	1	90	40
SVM	1	import pandas as pd	1	90	40
SVM	1	import pandas	1	90	40

Confusion Matrix of integer:



```
SVM
accuracy of model: 80.0 %
accuracy of codes: 70.0 %
```

Confusion Matrix of if_else_elif:



```
SVM
accuracy of model: 60.0 %
accuracy of codes: 90.0 %
```

Analysis

Classifiers:

1. Logistic Regression
2. Naive Bayes
3. KNN
4. SVM
5. Decision Tree

Analysis:

1. **Size & Accuracy:** Importing library has the most number of data entries and the highest prediction accuracy.

2. **Size VS Accuracy:** The size of if, else, and elif is about the same as the size of Integer. However, the accuracy of integer is the highest
   a. Comparing to less complicated syntax errors, more complicated syntax errors required a bigger dataset to achieve the same accuracy level

3. **Best Classifier:** The prediction that is made by SVM is highest

Next Steps

1. Create a Machine Learning to automatically generate syntax errors for various programming languages

2. If not, keep generating more datasets for complicated syntax errors and achieve 80% of accuracy