Universidad del Valle de Guatemala
Facultad de Ingeniería
Departamento de Ciencias de la Computación
CC3094 - Security Data Science

Laboratorio 03

Evelyn Andrea Amaya Malin - 19357
Brandon Josue Hernández Marroquín - 19376

Guatemala, Ciudad de Guatemala 16 de marzo de 2023

¿Se lograron obtener mejores métricas que las obtenidas en el artículo para la clasificación de malware?

Se implementaron los modelos: Random Forest Classifier y Support Vector Machine. En el artículo se menciona que el RF les dio una precisión de 0.82 y SVM les dio una precisión de 0.91 en un set con datos balanceados y, una precisión de 0.84 y 0.94 respectivamente con datos no balanceados..

	RF	**SVM**
Precision Balanced Data	0.82	0.91
Precision Imbalanced Data	**0.84**	**0.94**

Tabla 1: Resultados Teóricos

Al igual que como se menciona en el paper, realizamos un dataset balanceado con todos los virus con más de 80 datos y guardando como máximo 300 por cada uno de ellos.

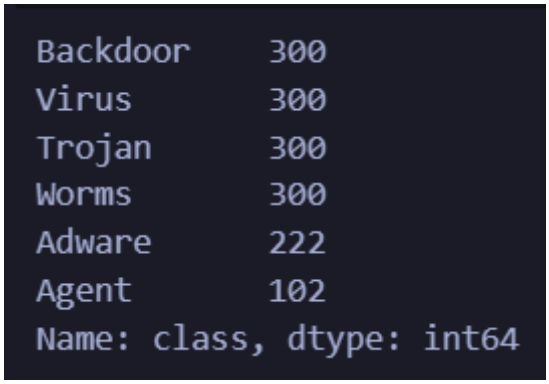


Figura 1: Dataset balanceado

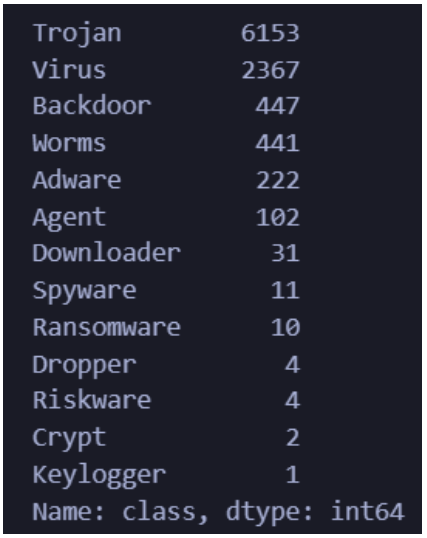
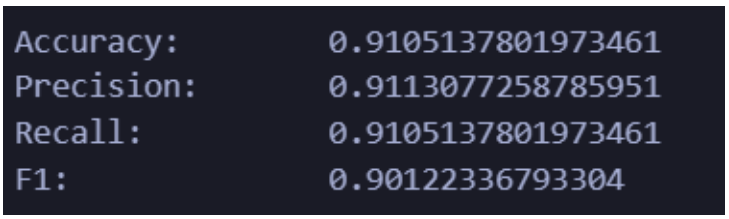


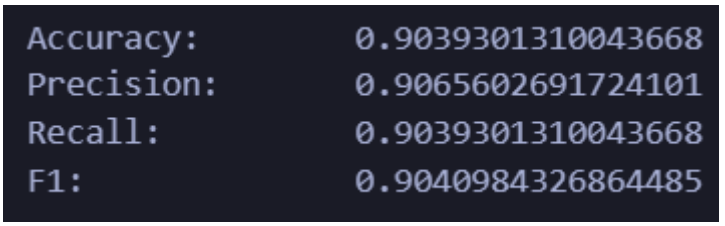
Figura 2: Dataset no balanceado

Resultados Random Forest Classifier

Se creó una bag of words para tener un identificador por cada palabra y así tener datos únicamente cuantitativos. Se hizo uso de CountVectorizer el cual es un método para convertir de texto a datos numéricos. Luego se separó cada dataset en 70% entrenamiento y 30% pruebas.



Resultados Dataset No Balanceado



Resultados Dataset Balanceado

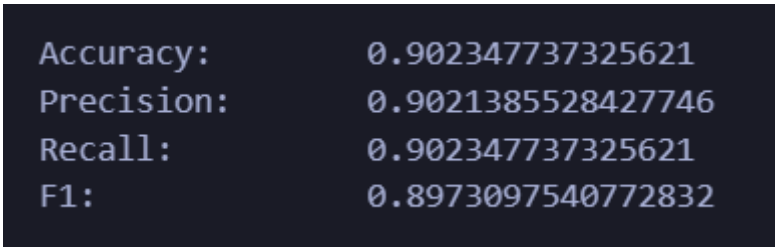
La precisión se refiere a la habilidad que tiene el modelo para no marcar un valor negativo como positivo. El recall se refiere a la habilidad del modelo para encontrar todos los valores positivos. Y, el accuracy hace referencia a predecir un valor igual a su valor correspondiente.

Es decir, nuestro modelo clasificará 91% de los valores de forma correcta (TP o TN), un recall 0.91 nos asegura que tenemos pocos FN y una precisión de 0.911 nos asegura que tenemos pocos FP. En este caso, al ser malware, es mejor tener más falsos negativos, pues se estarán clasificando de forma incorrecta goodware, mientras en el caso de falsos positivos tendremos malware clasificado como goodware.

Al igual que en el paper, nuestro modelo de Random Forest tiene un mejor resultado con la data no balanceada. Y, tiene mejores métricas, teniendo un accuracy de 0.91 comparado contra un 0.84 del paper.

Resultados SVM

El segundo modelo generado fue un Support Vector Machine, el cual fue el modelo con mejores resultados en el paper. Al comparar nuestros resultados, el Random Forest Classifier no balanceado tuvo mejores métricas que los dos resultados del SVM.



Resultados Dataset No Balanceado

```
Accuracy:       0.9017467248908297
Precision:      0.9017521889701989
Recall:         0.9017467248908297
F1:             0.9009649858099997
```

Resultados Dataset Balanceado

Es decir, nuestro modelo clasificará 90% de los valores de forma correcta (TP o TN), un recall 0.90 nos asegura que tenemos pocos FN y una precisión de 0.902 nos asegura que tenemos pocos FP. En este caso, al ser malware, es mejor tener más falsos negativos, pues se estarán clasificando de forma incorrecta goodware, mientras en el caso de falsos positivos tendremos malware clasificado como goodware.

	RF	**SVM**
Precision Balanced Data	0.90393	0.90174
Precision Imbalanced Data	**0.9105**	**0.90234**

Tabla 2: Resultados Prácticos

Resultados Cross-Validation con K folds = 10

- Random Forest Classifier
    - Imbalanced

```
Cross-Validation [0.93197279 0.91836735 0.92517007 0.92517007 0.92517007 0.92517007
 0.92517007 0.92517007 0.93537415 0.92832765]
Accuracy: 0.93 (+/- 0.01)
```

    - Balanced

```
Cross-Validation [0.89130435 0.91304348 0.86956522 0.91304348 0.86956522 0.91304348
 0.93478261 0.89130435 0.93333333 0.91111111]
Accuracy: 0.90 (+/- 0.04)
```

- Support Vector Machine
    - Imbalanced

```
Cross-Validation [0.86394558 0.84013605 0.86054422 0.86394558 0.86054422 0.85714286
 0.83673469 0.85034014 0.86054422 0.88054608]
Accuracy: 0.86 (+/- 0.02)
```

    - Balanced

```
Cross-Validation [0.86956522 0.91304348 0.89130435 0.84782609 0.89130435 0.93478261
 0.95652174 0.89130435 0.91111111 0.91111111]
Accuracy: 0.90 (+/- 0.06)
```

Al comparar todas las cross-validations se puede observar que el mejor resultado lo obtiene el Random Forest Classifier no balanceado teniendo una desviación de solo +-0.01. Esto significa que tendrá resultados correctos entre un 0.92 y 0.94. Mientras que el modelo con peor desempeño es el SVM no balanceado con un rango entre 0.84 y 0.88 de resultados correctos.

Sin duda, el RFC devolverá resultados más confiables al poder devolver hasta 0.94 TP o TN.