
# SUPERVISEDLEARNING

## SVM- RANDOM FOREST

AGHILAS SMAIL

MATSER 2 SID

# Content

1 Data Overview: ..... 2
2 Data preprocessing ..... 2
2 Choice of Evaluation Metrics ..... 2
3 Comparison Between SVM and Random Forest ..... 3
3.1 Before Hyperparameter Tuning ..... 3
3.1.1 SVM ..... 3
3.1.2 Random forest ..... 3
3.2 After Hyperparameter Tuning ..... 4
3.2.1 SVM ..... 4
3.2.2 Random forest ..... 4
4 Summary of results ..... 4

# 1 Data Overview:

Le jeu de données en question contient des informations relatives à des individus adultes et à divers attributs socio-économiques. Le schéma du jeu de données est le suivant :

**Age, Workclass, Fnlwgt, Education, Education_num, Marital_status, Occupation, Relationship, Race, Sex, Capital_gain, Capital_loss, Hours_per_week, Native_country, Income**

# 2 Data preprocessing

Pour préparer les données à l'entraînement des modèles, les variables catégorielles telles que workclass, education, marital_status, occupation, relationship, race and sex. sont converties en représentations numériques à l'aide de StringIndexer, qui attribue un index unique à chaque catégorie.

Ces indices sont ensuite transformés en un vecteur binaire clairsemé avec OneHotEncoder, créant un format plus adapté pour les modèles d'apprentissage automatique. Toutes les colonnes de caractéristiques, y compris les nouvelles encodées, sont combinées en un seul vecteur à l'aide de VectorAssembler.

Ce vecteur consolidé, nommé 'features', est prêt pour l'entraînement du modèle. Enfin, le jeu de données est divisé en ensembles d'entraînement et de test pour évaluer la performance du modèle et garantir qu'il se généralise bien aux nouvelles données, permettant une analyse approfondie de ses capacités prédictives.

# 2 Choice of Evaluation Metrics

Lors de l'évaluation de la performance des modèles de random forest et de SVM dans le contexte de données déséquilibrées, le choix des métriques est crucial pour assurer une compréhension complète de leurs capacités. Voici une explication détaillée de la raison pour laquelle chaque métrique a été sélectionnée :

**Area under ROC curve:**

est une mesure de performance pour les modèles de classification à deux classes. Elle représente la probabilité qu'un modèle classe correctement un exemple positif par rapport à un exemple négatif, avec une valeur allant de 0 à 1, où 1 indique une performance parfaite.

**Accuracy et Recall:**

La précision est la proportion de prédictions correctes par rapport au nombre total de prédictions, reflétant la fréquence à laquelle le modèle est correct. Le rappel, ou sensibilité, mesure la proportion de vrais positifs correctement identifiés, indiquant la capacité du modèle à trouver tous les cas pertinents.

**F1-Score:**

Le F1-Score est une mesure équilibrée qui combine la précision et le rappel. Dans les jeux de données déséquilibrés, obtenir un F1-Score élevé indique l'efficacité d'un modèle à

gérer à la fois les faux positifs et les faux négatifs. Cela est particulièrement pertinent lorsque le coût d'une mauvaise classification est élevé pour les deux types d'erreurs.

# 3 Comparison Between SVM and Random Forest

## 3.1 Before Hyperparameter Tuning

### 3.1.1 SVM

```
▸ (6) Spark Jobs

Area under ROC curve: 0.884278072268483
Accuracy: 0.7702390131071704
recall : 0.9995925020374898
F1 Score: 0.6829980614324778
```

### 3.1.2 Random forest

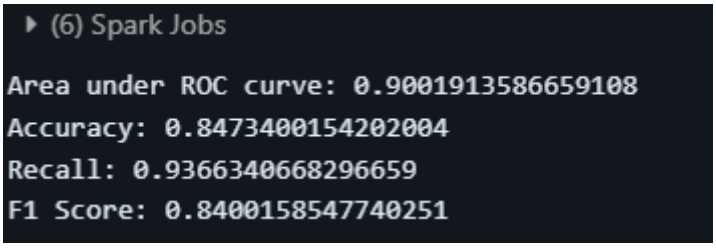
```
Area under ROC curve: 0.8860698875801749
Accuracy: 0.825905936777178
Recall: 0.825905936777178
F1 Score: 0.801125231731012
```

**Overall Observations:**

- The Random Forest model consistently outperforms the SVM model in terms of predictive accuracy on the imbalanced dataset.
- The higher F1 Score of the Random Forest model suggests a better balance between precision and recall, making it more suitable for handling imbalanced data.
- Both models exhibit relatively high Weighted Precision, indicating accuracy in positive predictions.
- The Random Forest model demonstrates superior performance in capturing instances of the minority class (1), as evident from the lower count of False Negatives.

## 3.2 After Hyperparameter Tuning

### 3.2.1 SVM

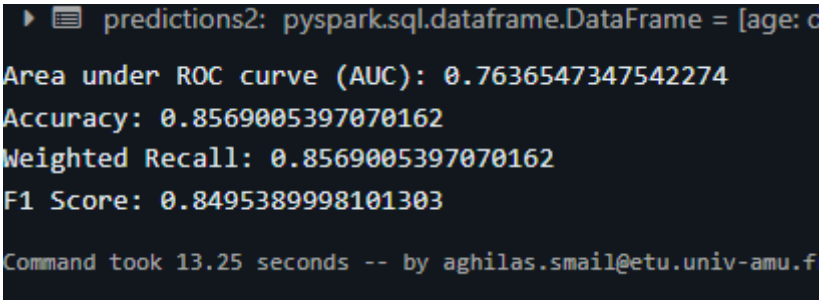



Le modèle SVM, après ajustement des hyperparamètres, présente des améliorations notables de la performance par rapport à l'évaluation initiale :

Interprétation :

- Le processus d'ajustement des hyperparamètres a considérablement amélioré la performance globale du modèle SVM sur diverses métriques.

### 3.2.2 Random forest




Le modèle de random forest montre également une amélioration considérable après l'ajustement des hyperparamètres.

Interpretation:

- De manière similaire au modèle SVM, le processus d'ajustement des hyperparamètres a conduit à des améliorations substantielles de la performance du modèle Random Forest.

- Un F1 Score plus élevé indique un meilleur compromis entre précision et rappel, démontrant l'efficacité du modèle Random Forest dans la gestion des données déséquilibrées.

# 4 Summary of results

Dans le domaine de la classification binaire pour des données déséquilibrées, les processus d'évaluation et de réglage des hyperparamètres ont fourni des aperçus précieux sur la performance des modèles de SVM et de random forest.

| SVM | Random forest |
|---|---|
| **Before Tuning** | |
| • Area under ROC curve: 0.8842<br>• Accuracy: 0.7702<br>• recall : 0.9995<br>• F1 Score: 0.6829 | • Area under ROC curve: 0.8860<br>• Accuracy: 0.8259<br>• Recall: 0.8259<br>• F1 Score: 0.8011 |
| **After Tuning** | |
| • Area under ROC curve: 0.9001<br>• Accuracy: 0.8473<br>• Recall: 0.9366<br>• F1 Score: 0.8400 | • Area under ROC curve : 0.7636<br>• Accuracy: 0.8569<br>• Recall: 0.8569<br>• F1 Score : 0.8495 |