Présentations des résultats

ZAOUAM Sirageddine **FARHI Abdellah** **HADJ-SAID Mohand**

I. Intoduction :

Le but de cette partie est d'évaluer et de comparer les 3 algorithmes qu'on a décrit précédemment (KNN, SVM et l'arbre de décision) afin de choisir le modèle le plus performant : celui qui donne la plus forte précision (accuracy).

Avant que les données ne soient traitées, l'ensemble de données est divisé en deux parties selon un ratio de 75:25, dont 75% à l'entraînement et 25% aux tests. Les Données d'entraînement sont utilisées pour obtenir le modèle alors que les données de test y sont pour l'évaluer.

- un ensemble d'entraînement (x_train, y_train) : 584 tuples
- un ensemble de test (x-test,y_test) : 195 tuples

II. Résultats :

1) Méthode de la validation croisée avec GridSearch :

on a utilisé k-fold cross-validation avec k =5 , on a profité de toutes les données à disposition en les divisant en k parties égales (folds) sur lesquelles on entraîne et teste notre modèle pendant k itérations. A chaque itération, le modèle est entraîné sur k-1 folds et est testé sur le fold restant.

Le modèle qui a été construit est présenté dans le tableau 1.

Tableau 1: résultat du modèle

Algorithm	Result	Accuracy
KNN	K = 7	66,6 %
SVM	value C = 1000, kernel = « rbf »	74,49 % %
Decision tree	Max-depth = 8	68,15 %

2) Matrices de confusions :

Le modèle a ensuite été testé en utilisant des données de test pour connaître la précision de ce modèle. Le tableau 2 montre la matrice de confusion pour l'algorithme KNN, le tableau 3 montre le résultat de cette matrice pour l'algorithme SVM et le tableau 4 est le résultat de cette dernière pour l'algorithme de l'arbre de décision.

Tableau 2 : Matrice de confusion pour la méthode KNN

	Predicted label		
True label	Classe 1	Classe 2	Classe 3
Classe 1	96	0	19
Classe 2	10	4	2
Classe 3	32	0	32

Tableau 3 : Matrice de confusion pour la méthode SVM

	Predicted label		
True label	Classe 1	Classe 2	Classe 3
Classe 1	100	1	14
Classe 2	9	5	2
Classe 3	22	0	42

Tableau 4: Matrice de confusion pour l'arbre de décision

	Predicted label		
True label	Classe 1	Classe 2	Classe 3
Classe 1	100	6	9
Classe 2	8	6	2
Classe 3	34	2	28

Le tableau 5 montre la comparaison du résultat du test entre les algorithmes KNN, SVM et Decision Tree sur la matrice de confusion :

Tableau 5 : Résultat des matrices de confusion

	Indicateurs de performances				Precision			Accuracy
Algorithm	TP	TN	FP	FN	**Classe 1**	**Classe 2**	**Classe 3**	
					115	16	64	
KNN	96	38	42	19	0,70	1.00	0,6	0,68
SVM	100	49	31	15	**0,76**	**0,83**	**0,72**	**0,75**
Decision tree	100	38	42	15	0,70	0,43	0,72	0,69

III. Analyse et interprétations :

Le meilleur modèle pour l'algorithme KNN pour prédire les classes des produits chimiques d'après le tableau 1 est k (nombre de voisins) = 7 avec une précision de 66,6%, la valeur C = 1000 et kernel = « rbf » pour l'algorithme SVM avec une précision de 74,49% et max_depth = 8 pour l'algorithme de l'arbre de décision avec une précision de 68,15%. La comparaison des trois algorithmes montre que la meilleure précision est **l'algorithme SVM** . Ce modèle n'a pas encore été testé.

Après les tests, on constate d'après le tableau 5 que l'algorithme SVM peut toujours mieux prédire que l'algorithme KNN et l'arbre de décision avec une précision totale (accuracy) de 75 % par rapport à l'arbre de décision (69%) qui est légèrement plus précis, mais seulement 1% de différence par rapport à KNN (68%).

La comparaison avec la matrice de confusion montre aussi que l'algorithme SVM peut mieux prédire la classe 1 et la classe 2 avec des précisions correspondantes 76 % et 72 %. Tandis que le KNN de son coté il est très optimal pour prédire la classe 2 avec une précision de 100 %. L'arbre de décision est légèrment au dessus de KNN, car ce dernier peut prédire la classe 3 avec une précision 72 % contrairement à KNN avec 60 %.

IV. Avantages et inconvénients des méthodes (KNN, SVM et l'Arbre de décision) :

On peut conclure que le KNN est **très performant** sur un petit ensemble de donnée (16 tuples) , il a réussit à prédire avec succès tous les 16 produits chimique de classe 2 avec une précision de **100 %,** par contre son ralentissement est important à mesure que la taille des données utilisées augmente.

Le SVM en générale est très efficace en dimension élevé (10 variables explicatives dans notre cas) par contre choisir une bonne fonction de noyau ce n'est pas facile et il devient lent pour un ensemble très grand.

L'arbre de décision nécessite moins d'efforts pour la préparation de données par contre il nécessite plus d'informations sur les données pour donner une bonne précision (43 % de précision pour 16 tuples).

IV. Conclusion :

L'algorithme KNN peut mieux prédire la classe du produit avec k = 7. Le meilleur modèle d'algorithme SVM pour prédire les classes est d'utiliser la valeur de C = 100, kernel « rbf » et gamma =0,001 . Alors que si on utilise l'algorithme de l'arbre de décision, les meilleures prédictions si on utilise l'hyper-paramètre max_depth = 8 . La comparaison de trois algorithmes d'apprentissage automatique (KNN, SVM et Decision Tree) montre que SVM a la meilleure précision (75%) par rapport à Decision Tree (69%) et KNN (68%).

V. Références :

[1] INTERNATIONAL JOURNAL OF RESEARCH GRANTHAALAYAH : COMPARTIVE STUDY OF MACHINE LEARNING
[2] MACHINE LEARNING MATRICE DE CONFUSION EN PTTHON : ASKCODEZ.COM
[3] TOWARDS DATA SCIENCE : ACCURACY, RECALL, PRECISION ,F-SCORE, SPECIFICITY