Data Mining

Propriétaire : Bosco KUATE •••
Dernière mise à jour : le oct. 31, 2023

1-Collecte des Donnée

Question: Où pourriez-vous trouver les données à collecter ?

Les sources que nous envisageons d'utiliser pour la collecte des informations sur les professionnels du bâtiment sont :

- l'Ordre National des Ingénieurs en Génie Civil du Cameroun,
- le ministère des marchés publics et le ministère des travaux public
- enquête
- …

2-Préparation des Données

Question: Lorsqu'on nettoie les données, c'est pour "nettoyer", corriger ou éliminer quoi par exemple ?

Le nettoyage consiste à identifier et corriger les données altérées, inexactes ou non pertinentes. Par exemple un professionnel ayant 150 ans d'âge, de sexe féminin et s'appelant Tanfotro Jean de Dieu. Un professionnel qui est à la fois dans le fichier de l'ONIGC et les ministères et dont le matricule dans les deux fichiers se différentie par un espace supplémentaire dans le fichier du ministère, on corrige le matricule et on supprime les doublons.

3-Exploration des Données

Question:

1-Quelles techniques sont souvent utilisées pour identifier ces tendances et schémas, lorsqu'on a plein de données. Concrètement, ce sont des colonnes (critères) avec des milliers de lignes

2-Donnez un exemple de la structure que pourrait avoir les colonnes et les lignes et comment on s'y prendrait pour trouver des tendances et schémas.

L'exploration des données regroupe les trois principales techniques suivantes:

- Descriptive: cette approche nous permet d'avoir des similitudes entre les professionnels. (regroupement, détection de modèle). Par exemple, les professionnels spécialisé dans le gros œuvre ou bien les finitions.
- Diagnostic: cette approche nous permet de savoir quels sont les caractéristiques des professionnels de bâtiment spécialisé dans le gros œuvre par exemple. (Classification)
- Prédictive: cette approche nous permet d'estimer la probabilité qu'un professionnel ayant un certain nombre de caractéristique soit spécialiste de gros œuvre. (prédiction)

Dans le cadre de ce projet nous utiliserons la technique prédictive, car le but sera de déterminer si un professionnel ayant certaines caractéristiques sera à même de respecter ses engagements.

Le tableau ci-dessous montre un échantillon des professionnels du bâtiments avec leur évaluation. Les colonnes sont les caractéristiques, sauf la dernière qui est le résultat de l'évaluation. Chaque ligne correspond à un professionnel.

Matricule PB	Nom	Expériences (ans)	Nombre de projet réalisé	Inscrit à l'ONIGC ?	…	Est certifié ISO ….	Niveau d'étude le plus élevé	Evaluation(0-100)
1PB1787	Tatuenne	10	25	oui		non	Bac+5	80
5PB1917	Tanfotro	6	15	oui		non	Bac+5	75
…								

8PB2119	KAMA	5	7	oui		oui	Bac+5	87

Pour découvrir les similitudes qu'il peut avoir entre les professionnels de l'échantillon, la technique descriptive est utilisé. Pour ce faire, en dehors de la colonne sur l'évaluation, toutes les colonnes seront utilisés dans le cadre d'un apprentissage non supervisé.

4-Modélisation

Question: Comment on s'y prendrait ? Dites le juste ou montrez un exemple

Dans le but de mettre en œuvre notre système d'aide à l'évaluation des professionnels du bâtiment, les étapes à suivre sont les suivantes:

- collecter et nettoyer les informations sur les professionnels du bâtiment. Constituer trois échantillon (training set, test set, validation set)
- analyser les données afin d'identifier les caractéristiques les plus pertinentes
- construire un modèle sur la base des techniques de régression logistique ou ANN (Artificial Neural Network)

le training set est utilisé pour entrainer le modèle

le test set est utiliser pour tester le modèle obtenu

le validation set est utiliser pour valider le modèle.

Pour définir si un professionnel est bon ou mauvais, on définit un seuil. Par exemple 55%.

si P(professionnel) >55% alors c'est un bon professionnel car la probabilité qui assure ses engagements est élevé

si P(professionnel) <=55% alors c'est un bon professionnel car la probabilité qui assure ses engagements est élevé

- produire la matrice de confusion (Faux positif et faux négatif) pour évaluer la performance du modèle obtenu.

Matrice de confusion		Prédiction	
		Négatif	Positif
Réel	Négatif (mauvais)	TN	FP (Faux Positif)
	Positif (bon)	FN (Faux négatif)	TP

En comparant les taux de FP et FN, on décide si le modèle sera utiliser ou pas.

Par exemple si le taux de FP est considérable, le modèle ne sera pas utiliser. L'utiliser reviendrai à qualifier des professionnel de bon alors qu'en réalité ils sont mauvais.

5-Interprétation et Validation

Voir matrice de confusion. Notons également que certains modèle sont plus explicite que d'autre.

6-Implémentation

Question: Donnez 3 autres exemples de ce type ou plus

- Outils d'aide à la décision, sur la base du résultat de l'évaluation d'un professionnel on peut décider s'il faut l'engager ou pas
- Système de notation des professionnels
-

7-Evaluation et Amélioration Continue

Refaire le training du modèle et s'assurer que la matrice de confusion obtenu est toujours acceptable