Projet Data mining

**Online Job Postings Mining**

**Remerciements :**

Au terme de ce travail, nous profitons de l’occasion pour remercier du fond du cœur notre professeur Mme.Benbrahim Houda pour ses précieux conseils et pour nous avoir donné l’occasion de mettre en application nos connaissances et de les approfondir à travers ce projet.

Nos remerciements vont aussi à tous ceux qui ont participé de près ou de loin à réalisation de ce projet et qui ont contribué à en faire une expérience enrichissante.

**Résumé**

Le sujet de notre projet datamining consistait à «Analyser les offres d’emploi ». Notre travail s’est fait en deux phases : Une phase d’analyse et une phase de réalisation.

à l’aide de notre professeur Mme BENBRAHIM Houda, nous avons appris à mener un des projets inhérents à notre branche d’ « Informatique Décisionnelle », le projet de Datamining, en s’appuyant sur la méthode CRISP-DM.

Nous vous présenterons donc tout au long de ce rapport, les étapes que nous avons suivies, les outils que nous avons utilisés pour réaliser notre projet et l’élaboration de différentes discussions sur les résultats obtenus.

Nous avons multiplié nos efforts afin d’atteindre notre objectif, ainsi d’améliorer nos connaissances d’une manière efficace et efficiente.

Mots clés : Datamining, CRISP-DM.

# **Abstract**

The subject of our datamining project was "Analyzing job offers". Our work was done in two phases: an analysis phase and an implementation phase.

With the help of our professor Mrs. BENBRAHIM Houda, we learned to lead one of the projects inherent to our "Business Intelligence" branch, the Datamining project, using the CRISP-DM method.

We will therefore present to you throughout this report, the steps we have taken, the tools we have used to carry out our project and the development of various discussions on the results obtained.

We have stepped up our efforts to achieve our goal, thereby improving our knowledge in an effective and efficient manner.

Key words: Data mining, CRISP-DM.

**CHAPITRE**

1

**Contexte**

**Général**

**Du projet**

# **Contexte général du projet**

Contexte

Les annonces d'emploi en ligne sont devenues le modèle dominant de recherche d'emploi et de jumelage employeur-employé dans la plupart des économies développées du monde et gagnent en popularité dans toutes les régions du monde. On estime qu'en 2014, plus de 70 % des offres d'emploi seront affichées en ligne aux États-Unis d'Amérique et qu'en consultant les données détaillées des offres d'emploi en ligne, les chercheurs pourront mieux comprendre le marché du travail.

Objectif du projet

Ce projet a analysé les offres d'emploi en ligne de 2004 à 2015 publiées sur le CareerCenter.am, un centre de ressources humaines arménien en ligne. Le principal objectif commercial est de comprendre la dynamique du marché du travail arménien et des questions commerciales pertinentes ont été définies. Un objectif secondaire est de mettre en œuvre une analyse de texte avancée comme preuve de concept pour créer une fonction de classification et de recherche d'information qui peut ajouter de la valeur ajoutée au portail de l'emploi.

Spécification du projet

Nous avons suivi la méthodologie CRISP-DM pour le projet. Après la compréhension de l'entreprise et des données avec visualisation, nous avons préparé les données. L'ensemble de données sur les emplois en Arménie a été nettoyé des données manquantes et des ensembles dupliqués.

**CHAPITRE**

2

**Les outils**

**De développement**

**Les outils de développement**

**IBM SPSS Modeler**

IBM SPSS Modeler (*Statistical Package for the Social Sciences*) est une plateforme d'analyse prédictive qui vous aide à créer rapidement des modèles d'analyse prédictive et intégrer les éléments prédictifs aux décisions des individus, des équipes, des systèmes et l'entreprise. Il fournit toute une gamme d'algorithmes et de techniques d'analyse avancés, dont l'analytique textuelle, l'analyse d'entité, la gestion et l'optimisation des décisions, et fournit des connaissances quasiment en temps réel. Il vous permet de prendre régulièrement de bonnes décisions, depuis votre poste de travail ou dans les systèmes opérationnels.



Figure 1:IBM SPSS MODELER

**Présentation de la méthode CRISP-DM**

CRISP-DM, qui signifie Cross-Industry Standard Process for Data Mining, est une méthode mise à l'épreuve sur le terrain permettant d'orienter vos travaux de Data mining.

* En tant que méthodologie, CRISP-DM comprend des descriptions des phases typiques d'un projet et des tâches comprises dans chaque phase, et une explication des relations entre ces tâches.
* En tant que modèle de processus, CRISP-DM offre un aperçu du cycle de vie du Data mining.

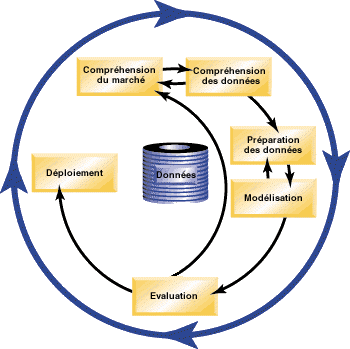


Figure 2: étapes du CRISP-DM

Le modèle de cycle de vie comporte six phases dotées de flèches indiquant les dépendances les plus importantes et les plus fréquentes entre les phases. La séquence des phases n'est pas strictement établie. De fait, les projets, pour la plupart, passent d'une phase à l'autre en fonction des besoins.

**Les Phases de CRISP-DM :**

|  |  |
| --- | --- |
| **Phases** | **Tâches** |
| **- Compréhension du problème** | * Déterminer les objectifs commerciaux. * Evaluer la situation. * Déterminer les objectives du Data Mining. * Produire d'un plan du projet. |
| **- Compréhension des données** | * Collecte des données initiales * Description des données * exploration des données * Vérification de la qualité des données |
| **- Préparation des données** | * Sélection des données * Nettoyage des données * Construction de nouvelles données * Intégration des données * Formatage des données |
| **- Modélisation** | * Sélection des techniques de modélisation. * Génération d'une conception de test. * Création des modèles. * Evaluation de modèles. |
| **- Evaluation** | * Evaluation de résultats * Processus de révision * Détermination des étapes suivantes |
| **- Déploiement** | * Planification du déploiement * Planification de surveillance et maintenance * Production de rapport final * Exécution d'une révision de projet final |

**CHAPITRE**

3

**Mise en**

**Œuvre**

**Mise en œuvre**

En ce qui concerne la mise en œuvre du projet, on procèdera moyennant la méthode CRISP-DM Cross-Industry Standard Process for Data Mining, qui nous a permis d’orienter notre travail de Data Mining, où nous allons décrire les différentes étapes constituant cette dernière :

Les étapes de CRISP-DM sont les suivantes :

Compréhension du problème :

Nos principaux objectifs commerciaux sont de comprendre la dynamique du marché du travail arménien en utilisant le portail de l'emploi en ligne comme proxy. L'objectif secondaire est de mettre en œuvre une analyse de texte avancée comme preuve de concept afin de créer des fonctions supplémentaires telles qu'une fonction de recherche améliorée qui peut ajouter de la valeur pour les utilisateurs du portail de l'emploi.

Les

Nature des emplois, profils des compagnies

Skills requis par catégorie d’emploi

Classification des jobs IT

Compréhension des données :

Cette base de données sur laquelle nous allons travailler regroupe 19.000 offres d'emploi postées sur le site CareerCenter. Elle comporte des informations telles que le salaire offert, le nom du poste, les capacités requises, etc ...

jobpost – The original job post   
date – Date it was posted in the group   
Title – Job title   
Company - employer   
AnnouncementCode – Announcement code (some internal code, is usually missing)   
Term – Full-Time, Part-time, etc   
Eligibility -- Eligibility of the candidates   
Audience --- Who can apply?   
StartDate – Start date of work   
Duration - Duration of the employment   
Location – Employment location   
JobDescription – Job Description   
JobRequirment - Job requirements   
RequiredQual -Required Qualification   
Salary - Salary   
ApplicationP – Application Procedure   
OpeningDate – Opening date of the job announcement   
Deadline – Deadline for the job announcement   
Notes - Additional Notes   
AboutC - About the company   
Attach - Attachments   
Year - Year of the announcement (derived from the field date)   
Month - Month of the announcement (derived from the field date)   
IT – TRUE if the job is an IT job. This variable is created by a simple search of IT job titles within column “Title”

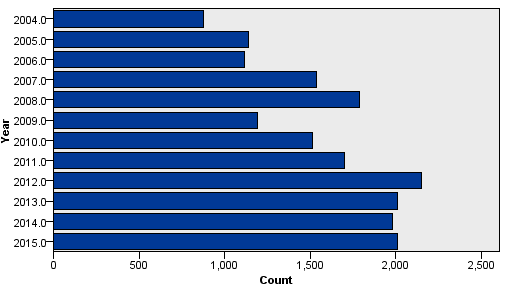


Figure : Offres d'emploi par année

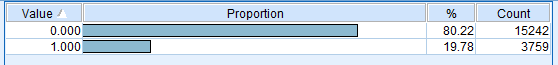


Figure : Nombre d'entreprises informatiques

Préparation des données :

1. Sélection et construction des données

On a choisit de supprimer les champs pour lequel le pourcentage de completude inferieur a 75%. Et aussi supprimer toutes les lignes contenant des valeurs null pour une des variables auxquelles nous sommes intéressés.

On obtient ainsi 6 champs:

1- Title.

2- JobRequirement.

3- JobDescription.

4- RequiredQual.

5- AboutC.

6- IT.

7-Year.

* Nettoyage des données:

On a supprimé les postes d'emploi en double, car ils sont probablement attribuables à l'affichage de la même affiche à plusieurs reprises lorsque le poste n'est pas comblé. La condition de détection des doublons était fondée sur la prise en compte à la fois de l'affichage de poste et de la colonne de titre. Nous n'avons tout simplement pas tenu compte des colonnes étiquetées NA dans le cadre de l'exploration de texte.

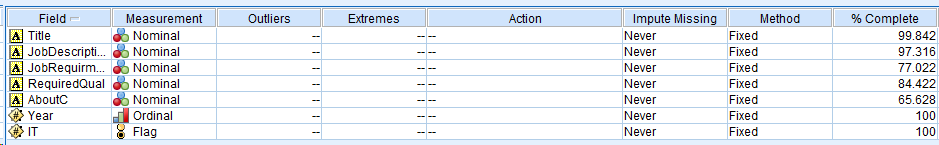


Figure : audit initial



Figure : filtrer figure : sélectionner

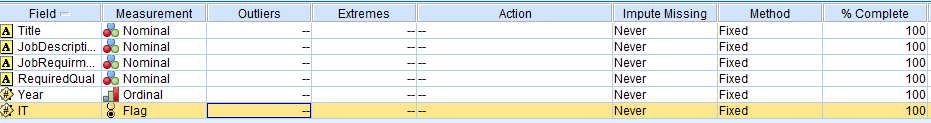
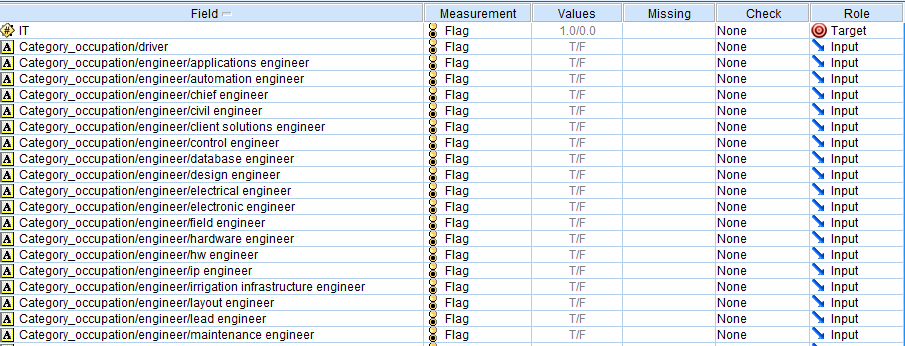


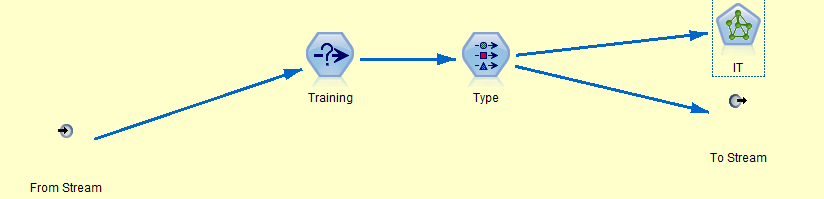
Figure : audit final

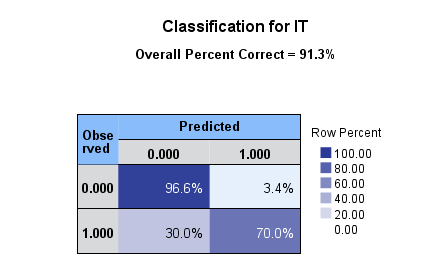
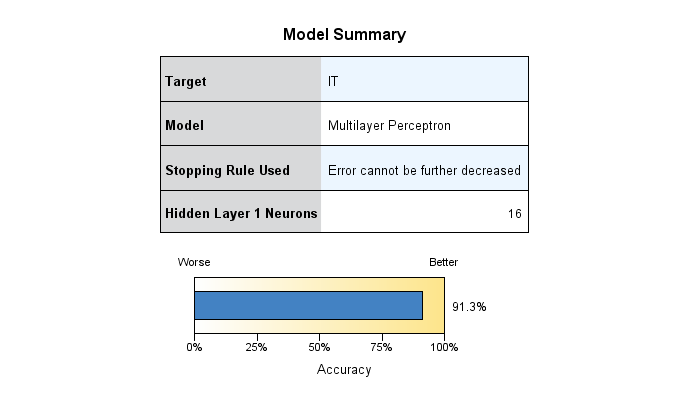
Modélisation :

* Pour la première problématique, concernant l’aspect prédictif, nous avons choisi comme attributs (Variables indépendantes) tous les concepts dans la catégorie Occupation. Puisque, cette catégorie est celle contenant le plus des concepts. Donc, cette catégorie est assez représentative du champ ‘Type’.

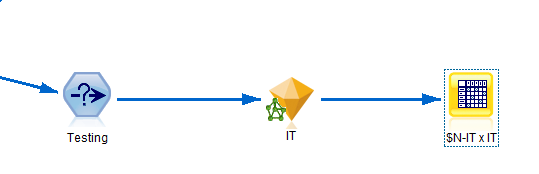


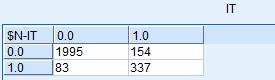
Et pour le modèle nous avons choisi un réseau de neurones puisque tous les valeurs champs sont dans l’intervalle [0.1], et parce qu’il donne des bons résultats généralement:



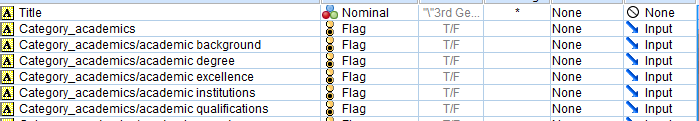
On obtient les resultats suivants :

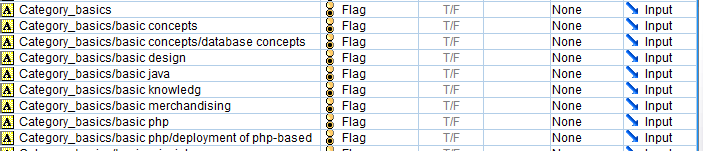
Pour l’ensemble de test on a la matrice de contingence qui donne :

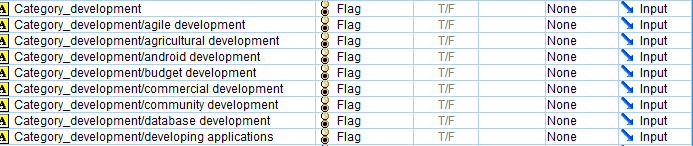




* Pour la deuxième problématique, qui concerne l’aspect descriptif nous avons choisi des attributs (variables indépendantes) à partir de plusieurs (515 attributs en total) :

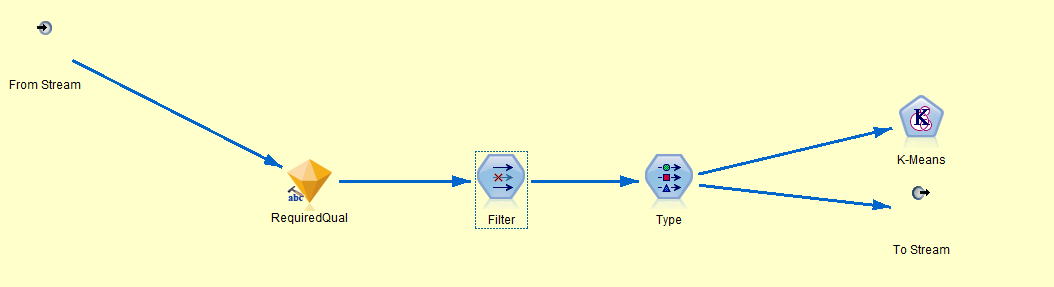


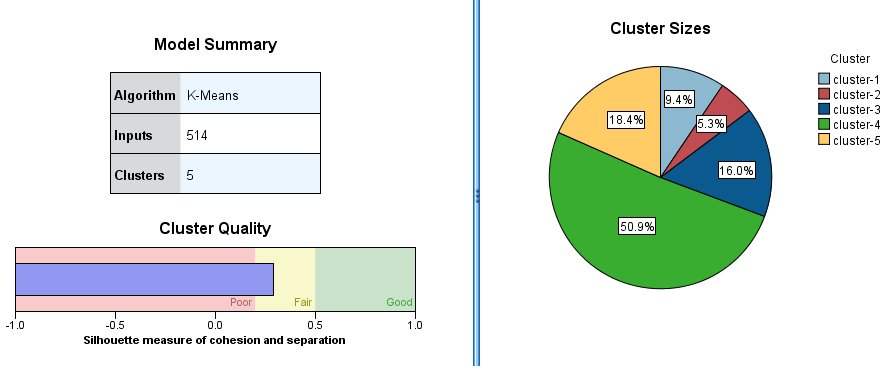




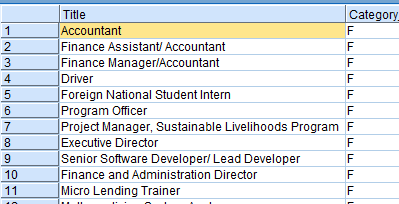
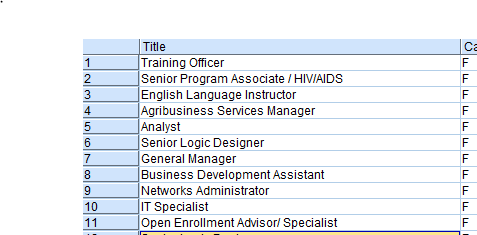
Et pour le modèle nous avons choisi l’algorithme de K-Means avec k = 5 :

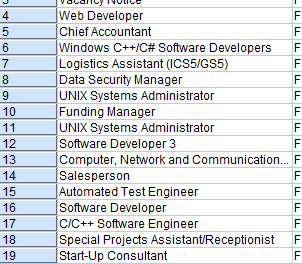
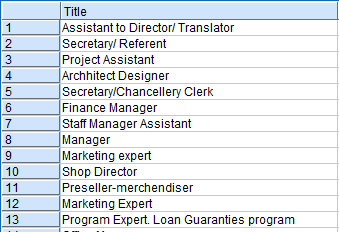
Pour d’autres choix de K nous avons trouvé des modèle de qualité (Silhouette measure) mauvaise, et parfois des clusters qui sont très réduit.

On obtient les résultats suivants :

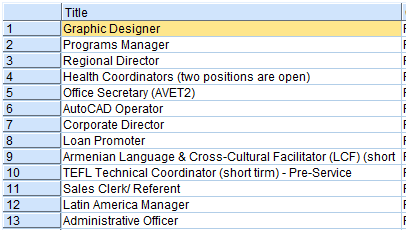
* 5 clusters :

On obtient une qualité de regroupement assez bonne et 5 clusters de taille assez large.

 Accounting & Finance cluster Business & Management cluster

Marketing & sales cluster  IT cluster

Autres



Dans chaque cluster on remarque une tendance générale (non parfaite) a un certain domaine de métier. Ainsi on peut libeller les différents cluster, comme ci-dessus.

Evaluation :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Job field | IT(1) | Marketing Sales(5) | Business & Management(2) | Accounting & Finance(3) | Others (4) |
| Requirements | Academic degree (50%)  Finance Knowledge (10%)  Development (25%)  Economics knowledge (10%) | Academic degree (5%)  Finance Knowledge (15%)  Development (5%)  Economics knowledge (20%) | Academic degree (25%)  Finance Knowledge (10%)  Development (5%)  Economics knowledge (10%) | Academic degree (5%)  Finance Knowledge (65%)  Development (5%)  Economics knowledge (30%) | Academic degree (15%)  Finance Knowledge (0%)  Development (60%)  Economics knowledge (30%) |

On obtient la table suivante par analyse des différents clusters. Pour chaque cellule on retrouve le taux de sollicitation d’une certaine capabilité dans un domaine de travail donné.

Ceci est fait sur le premier plan de la catégorie nous pouvons explorer des sous catégories pour obtenir des ‘skills’ plus spécifiques pour des métiers plus spécifiques.

Donc nous avons pu répondre à la problématique initiale de pouvoir recommander des ‘skills’ pour des domaines de travail donnes.

On peut effectuer un déploiement sur des sites web ou applications bureau sous forme de questionnaires qui fournissent des réponses ou même des ChatBots.

# **CONCLUSION**

A travers ces différents chapitres que nous avons dévoilé, en première partie, le contexte du projet et sa finalité. Dans la deuxième partie, nous avons défini les outils de développement utilisés pour réaliser le projet.

Enfin, nous avons entamé la mise en œuvre de l’application et les solutions proposées pour les deux problématiques que nous avons déjà mentionner. Nous nous sommes basées également sur des « captures-écran » afin de concrétiser la démarche adoptée.

Ce projet nous a aidé à améliorer les connaissances nécessaires à l’élaboration des projets datamining suivant la démarche CRISP-DM et nous a donné l’opportunité d’apprécier le travail en groupe qui s’avère très enrichissant pour ce genre de projets.