L2 MIASHS
Science des données
2023

Classification supervisée pour une analyse sur le diabète et non supervisée pour l'analyse du recueil de poème”***Les Fleurs du Mal***”

Réalisé par : HAOUASNIA Mayssoune 22109716
LAASILI OMIA 22101501

Enseignant : Arnaud Sallaberry

Résumé du rapport:

Le rapport porte sur deux analyses, la première va se faire à l'aide d'une classification supervisée. Cette classification comporte sur les différents thèmes qui ressortent d'un recueil de poèmes. Le deuxième porte sur le diabète et va se faire par apprentissage supervisé. Dans les deux cas nous allons décrire la chaîne de traitement et tenter d'analyser les résultats et l'efficacité du modèle.

Table des matières

1-Apprentissage non supervisé

1.1 Jeu de données

Titre du Jeu de Données: Les Fleurs du mal

a) Objectif et Caractéristique du jeu de Données

Le jeu de données est un extrait du recueil de poème “Les Fleurs du mal” de C.Baudelaire. Il correspond à la partie Spleen et Idéale. Le but serait de voir quels sont les principaux thèmes abordés dans cette partie à l’aide d’un dendogramme.

Il y a au total 85 poèmes dont la longueur varie plus ou moins. Nous avons décidé de mettre dans les fichiers uniquement le contenu des poèmes, donc sans le titre la date ni l’auteur, pour les traiter uniquement en fonction de ça.

b) Source et Format des Données

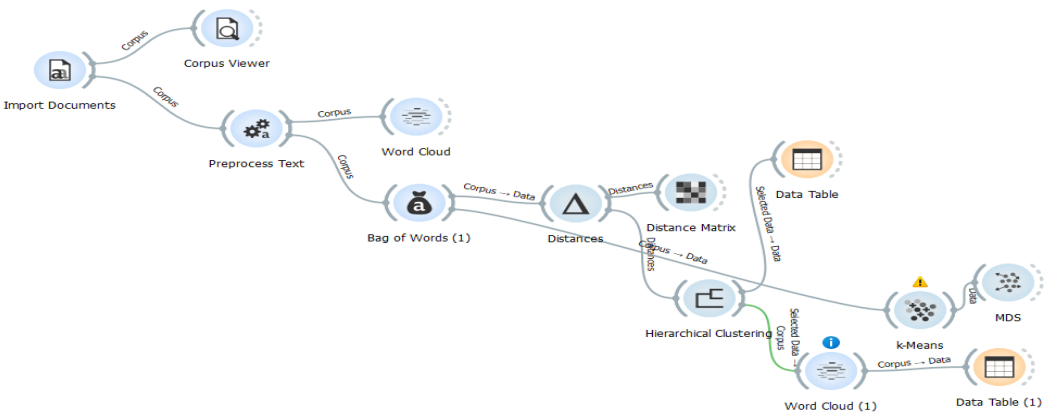
Les Données sont sous forme Word (.docx). Les poèmes ont été récupérés chacun sur internet et chaque poème est contenu dans un fichier, ils sont dans le document Les Fleur du mal.

1.2-Nettoyage et prétraitement des données

On a créer un fichier txt de mots d’arrêts avec tous les mots que nous n’avons pas jugé nécessaire à garder pour le traitement en se basant sur le nuage de mots. Nous avons aussi dû supprimer quelques poèmes pour analyser correctement les données.

1.3- Clustering

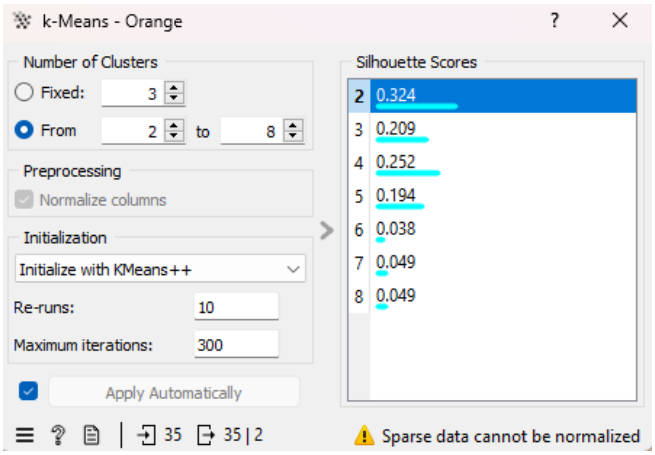
Figure [1]



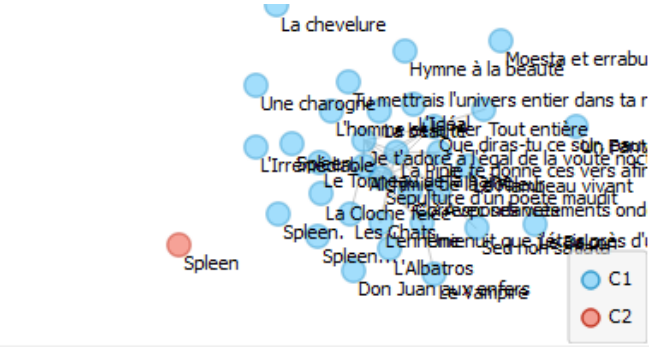
[1] Chaîne de traitement

Pour mettre en place cette chaîne, on a utilisé “Import Documents” afin d’importer le dossier avec tous les textes, puis on a effectué le prétraitement des données. On a placé “bag of Words” pour convertir ces textes en données numériques et pouvoir enchaîner avec la matrice de distances (pour calculer distance on a choisi d’utiliser cosinus). Enfin on a mis un clustering hiérarchique et un nuage de mots pour visualiser les différents clusters présents.

On a essayé comme on peut le voir de mettre un k-means pour essayer de déterminer le nombre de clusters intéressant à garder. Cependant il affiche un rendu pas très exploitable



Figure[2]



Figure[3]

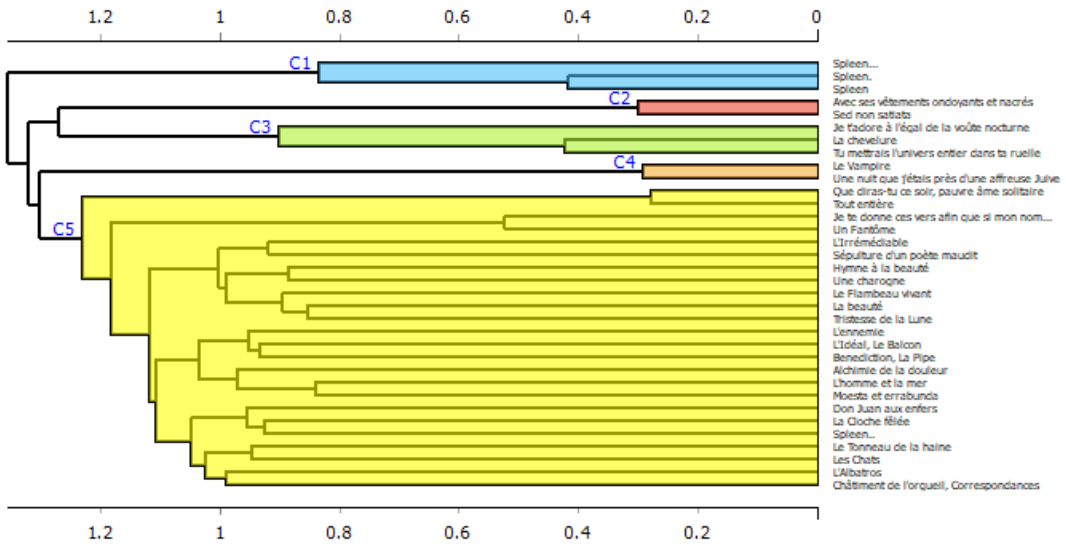


Figure [4]

[2] Résultat K-Means
[3] MDS lié au K-Means
[4] Dendogramme

Alors nous pensons que visuellement, séparer les clusters de cette façon est une bonne chose dans un premier temps. On se retrouve avec 5 clusters plus ou moins grands, qui se connectent entre eux plus ou moins tôt.

C1:

Figure[5]

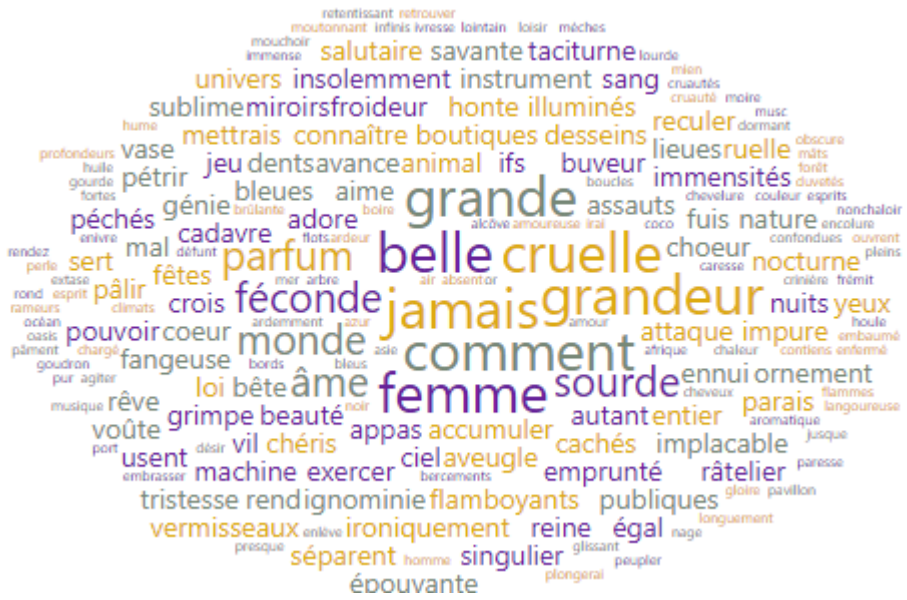
Le premier cluster a un nuage de mots avec des termes assez **funestes**, ils ne se connectent pas immédiatement entre eux tout de suite mais les différences entre les textes n'en restent pas moins petites.C'est les textes "spleen" qui sont réunis, cependant il y en a un qui a été placé dans le dernier cluster, on essayera de voir pourquoi en analysant ce dernier.

Le deuxième cluster je dirais que c'est plus la nature qui est mise en avant :

Figure[6]

[5] Nuage de mots C1
[6] Nuage de mots C2

ce qui est intéressant c'est que dans ce cluster, il y a seulement 2 poèmes. Sa prochaine connexion se fait avec le **cluster 3** bien plus tard, celui-ci ressort un thème qui se rapproche plus de la beauté de la femme.



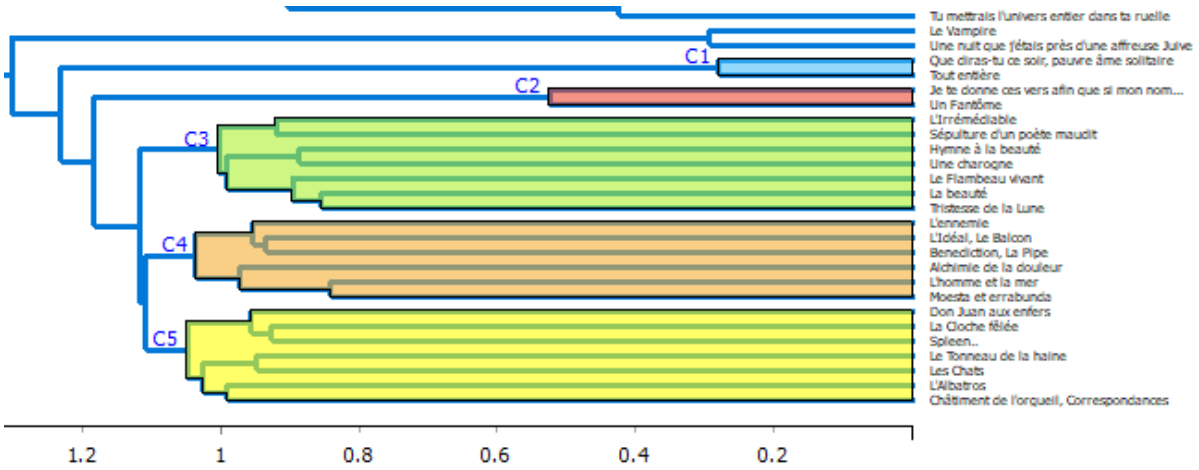
Figure[7]

Et quand on regarde le dernier **cluster (le 5eme**), on se rend compte que le thème est approximativement le même.

Figure[8]

Les principaux mots se ressemble à la seule différence que ce nuage de mot ne contient pas le mot femme. Si on analyse un peu plus ce cluster :

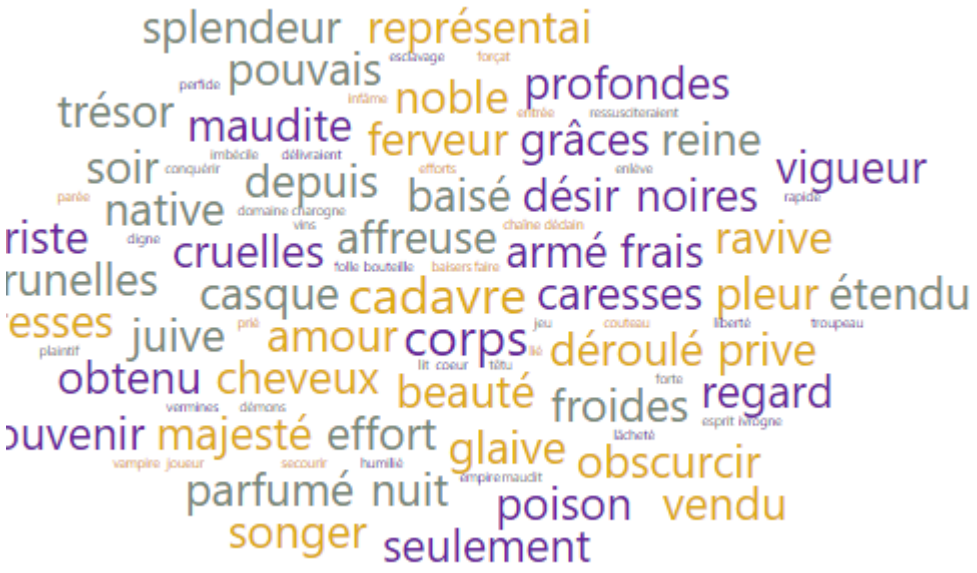
[7] Nuage de mots C3
[8] Nuage de mots C5



Figure[9]

On remarque que C5/C5 contient des textes plutôt sombre avec des mots qui se démarquent comme : haine, sombre, noir.
Le C5/C4 lui contient des mots principaux imageant la profondeur comme : coeur, mer, paradis, douleur.
On comprend déjà mieux pourquoi ces poèmes sont reliés ensembles.
La connexion avec le C5/C3 arrive juste après, cependant dans ce cluster on a plutôt des mots spirituelles et qui représentent aussi la beauté(ce qui fait que ce cluster se rapproche de celui C5/C4(paradis, douleur)) comme : yeux, beauté, ange, ciel, soleil.
De même pour le C5/C2, les mots les plus retrouvés n'ont pas beaucoup de rapport avec les précédents: grand, lecteur,mémoire, profond.
A nouveau dans le C5/C1 on retrouve des termes se rapprochant de la beauté ce qui explique la connexion mais pourquoi est-elle si tardive ?

On vient maintenant sur le **cluster 4 :**



[9] Séparation du C5 en petits clusters

Figure[10]

On peut qualifier le thème de ce cluster d'amour décadent, on retrouve en même temps des mots comme désir, amour ou même beauté, mais aussi des termes comme cadavre, affreuse et cruelles qui eux représentent peut être un peu plus la tristesse de l'amour par la violence des mots.

Résultat :
Les thèmes qui réunissent tous ces textes sont :

- La femme et la beauté
- La tristesse
- La nature
- L'amour décadent

Cependant je pense tout de même que le traitement n'est pas satisfaisant à 100%, peut être effectuer un prétraitement plus intensif était nécessaire pour trouver les thèmes des poèmes et surtout les classer correctement.

[10] Nuage de mots C4

2 -Apprentissage supervisé

2.1 Jeu de données

a) Description du jeu de données étiquetées

Pour l'apprentissage supervisé, nous avons choisi un jeu de données qui contient les informations de 520 individus. Il présente les données des signes et symptômes des patients nouvellement diabétiques ou potentiellement diabétiques (" Sign and Symptom Data of Newly Diabetic or would be Diabetic Patient".
Le jeu de données contient 17 variables dont une variable quantitative : l'âge de chaque individu ainsi que 16 variables qualitatives, toutes dichotomiques : le genre avec les modalités "Male"/"Female" , 14 variables qui correspondent aux symptômes que l'on a tendance à retrouver chez les personnes diabétiques (liste des symptômes : Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, Obesity) qui ont les modalités "Yes"/"No" ,ainsi que la variable class qui possède les modalités "Positive"/"Négative", cette variable prédit si la personne est malade ou non en fonction des symptômes.

Les données sont dans un fichier CSV qui a été récupéré sur Internet à partir de ce lien*. Il se trouve dans le document "diabetes_data_upload.csv" dans le dossier supervisé.

*(
https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset/
)

b) Description du jeu de données à prédire

On a mis les informations étiquetées de 98 individus sur un fichier CSV. Il présente exactement les mêmes variables que le fichier téléchargé sur Internet. Les données sont dans un fichier CSV nommé "diabetes_data_1-99.csv" dans le dossier Haouasnia_Mayssoune.Laasili_Omia_data21.

Puis,on a mis les informations des 422 individus restants sur un fichier CSV. Il présente les mêmes variables que le fichier téléchargé à l'exception de la variable class que l'on a supprimé. Les données sont dans un fichier CSV nommé "diabetes_data_100-500.csv" dans le dossier Haouasnia_Mayssoune.Laasili_Omia_data22.
Notre objectif est de nous appuyer sur des méthodes d'apprentissage supervisées comme les modèles (SVM,logistic regression etc.) pour prédire avec précision si les individus de ce dernier fichier sont potentiellement diabétiques ou non en se basant les informations données.

2.2 Modèles d'apprentissage

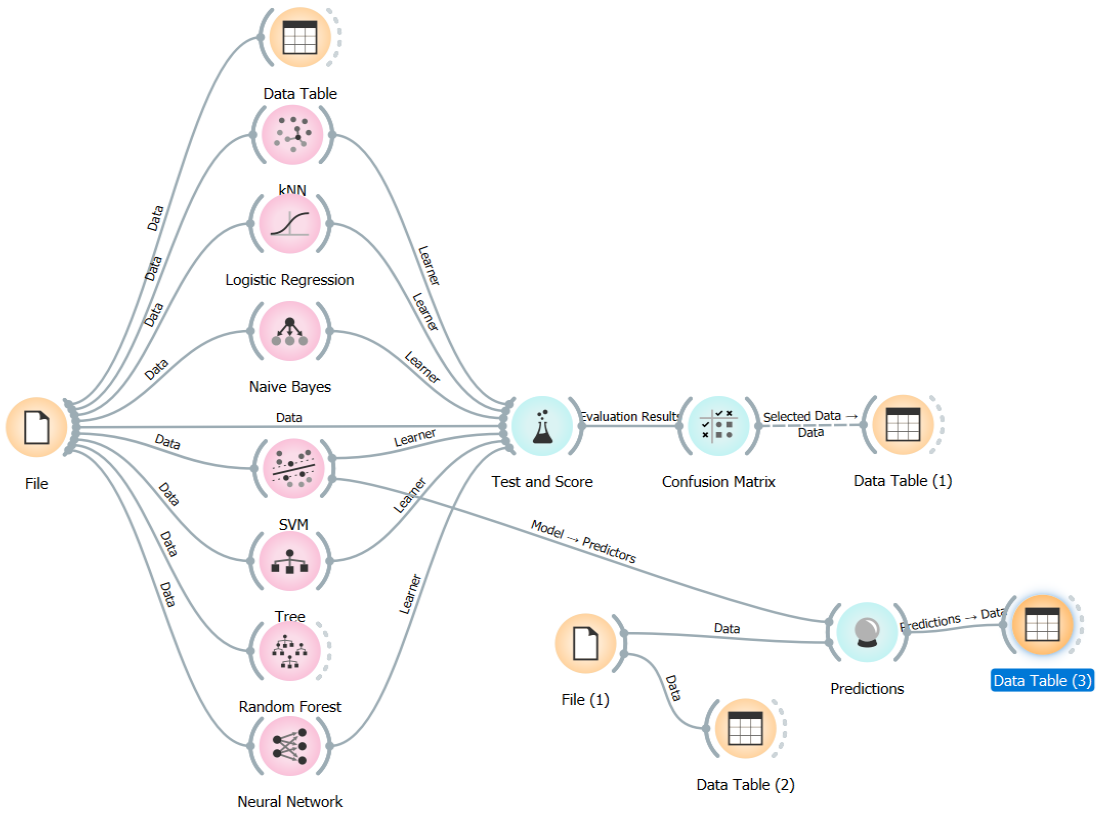


Figure 1 - Vue générale de la chaîne mise en place

Pour mettre en place la chaîne de la Figure 1, on a utilisé l'outil "File" afin d'importer le fichier "diabetes_data_1-99.csv" avec les données étiquetées des 98 individus. En sortie de "File", nous avons ajouté l' outil "Data table" qui nous permet de visualiser les données importées ainsi que les 7 modèles d' apprentissage suivants : ????.
Afin d'évaluer la performance de ces modèles, nous avons intégré l'outil "Test and Score" qui prend en entrée ces modèles.

Evaluation results for target: Positive

Model	AUC	CA	F1	Prec	Recall	MCC
Naive Bayes	0.922	0.837	0.881	0.952	0.819	0.645
Logistic Regression	0.929	0.867	0.909	0.915	0.903	0.664
kNN	0.929	0.878	0.912	0.969	0.861	0.727
SVM	0.973	0.898	0.932	0.919	0.944	0.733
Tree	0.911	0.888	0.921	0.955	0.889	0.734
Neural Network	0.954	0.898	0.931	0.931	0.931	0.738

Figure 2 - Résultat d'évaluation des modèles

Comme on peut le voir sur la Figure 2, on a choisi à colonne rappel “Recall” , en effet, il est plus judicieux de prendre des précautions pour ne pas obtenir des faux négatifs, il est préférable d’avoir autant de vrais positifs que possible même si cela implique qu’il y ait des chances de se retrouver avec davantage de faux positifs.

On observe alors que sur la colonne rappel, le modèle avec le meilleur score est le SVM (Support Vector Machine) avec la cible “Positive”, le score dans ce cas vaut 0.944, tandis que si on sélectionne la cible “Négative” ou “None, show average over classes”, les scores les plus élevés diminuent respectivement à 0.923 et 0.898. On choisit donc la cible “Positive” et le modèle SVM.

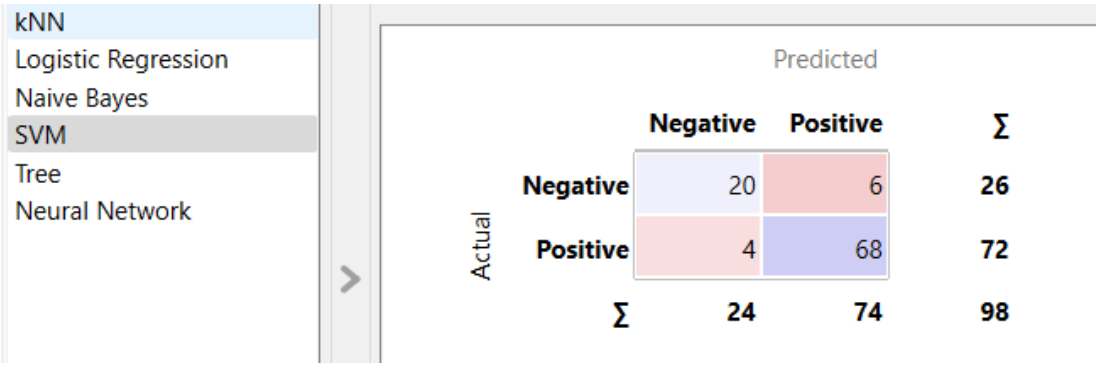


Figure 3 - Matrice de confusion SVM

Ensuite, nous avons ajouté l’outil “Confusion Matrix” pour visualiser les résultats de prédictions obtenus.

On remarque qu'il y a seulement 6 individus qui ont été prédits positifs alors qu'ils étaient en réalité négatifs, d'autre part, 4 individus ont été prédits négatifs alors qu'ils sont positifs.
Nous avons ensuite ajouté "Data Table" pour nous permettre de visualiser les données prédites correctement ou non, séparément ou simultanément.

	class	class(SVM)	Age	Gender	Polyuria	Polydipsia
1	Positive	Negative	40	Male	No	Yes
2	Positive	Negative	45	Male	No	No
3	Negative	Positive	49	Male	No	Yes
4	Positive	Negative	37	Male	No	No
5	Positive	Negative	41	Male	Yes	No
6	Negative	Positive	55	Male	No	No
7	Negative	Positive	72	Male	Yes	No
8	Negative	Positive	50	Male	No	No
9	Negative	Positive	56	Male	No	Yes
10	Negative	Positive	40	Male	No	No

Figure 4 - Liste des individus mal classés

Bien qu'il y ait des conclusions incorrectes, 10 individus sur 98 étant mal classés, cela équivaut à seulement environ 10,20% des individus, le résultat est donc plutôt satisfaisant.

2.3 Prédictions

Pour cette partie, nous avons ajouté l'outil "Prédictions" en sortie de "SVM", "Prédictions" prend en entrée l'outil "File" qui contient cette fois le fichier "diabetes_data_100-500.csv" avec les informations des 422 individus restants pour lesquelles il faut prédire la variable class. L' outil "Data table" en sortie de "File" nous permet de visualiser les données importées.

Pareillement, nous avons ajouté "Data Table" en sortie de "Prédictions" afin de visualiser les résultats prédits pour les individus par le modèle d' apprentissage SVM, le modèle idéal pour ce jeu de données.