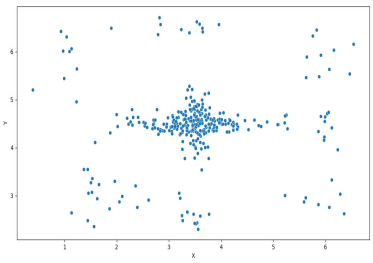
Rapport - Projet Intelligence Artificielle

Hélène SIVRIC - Stefan UNGUREANU

1 Données

1. Dans ce jeu de données, il y a 250 données dans la classe inlier et 80 données dans la classe outlier.
2. Ci joint, une capture pour visualiser les données:



3. On remarque un grand cluster central d'inlier et les outlier, eux, sont éparpillés de part et d'autre autour de ce cluster.

2 Évaluation

1. En lisant les coefficients, le modèle associé semble bon, car on peut lire 1000 et 5 données sur la diagonale des True Negative et des True Positive, contre 30 et 2 données de False Positive et False Negative, ce qui à vue d'œil nous indique qu'on aura un bon modèle.
2. Calculons l'exactitude:

$$\frac{1000+5}{1000+2+30+5} = 0,96$$

Puis calculons l'exactitude pondérée:

$$\frac{\frac{1000}{1000+2}+\frac{5}{30+5}}{2} = 0,57$$

En effectuant ces calculs on remarque que l'exactitude pondérée est bien plus basse que l'exactitude.

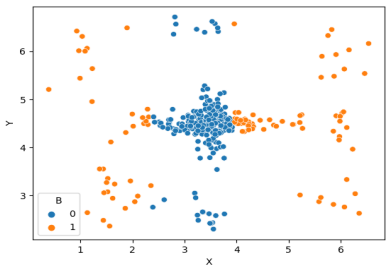
3. L'exactitude donne un bon score car il y a un déséquilibre entre les deux classes inlier et outlier. En effet, il y a beaucoup plus de données présentes dans la classe inlier.

4. L'exactitude n'est pas pertinente dans notre cas car nous voulons plutôt trouver les choses pertinentes dans notre modèle et relever les vrais points pertinents (donc les True Negative et les True Positive). Il faudrait donc plus mettre l'accent sur les métriques de précision et de rappel.

3 Algorithme

3.1 Arbre réduit à 1 feuille

Voici une capture de nos données:

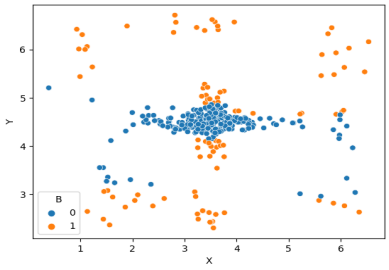


La matrice associée à ce modèle est la suivante: $\begin{pmatrix} 204 & 46 \\ 24 & 56 \end{pmatrix}$

Aux niveau des métriques, nous avons une exactitude de 0,78, une exactitude pondérée de 0,75, une précision de 0,54 et un rappel de 0,7. Nous pouvons donc observer que nous avons une exactitude et exactitude pondérée plus proche, et notre rappel est plus élevé, ce qui veut dire que nous avons réussi à attraper plus de outlier.

3.2 Arbre Superficiel

Nous avons utilisé 3 structures de données: **DecisionLeaf** qui est une feuille de notre arbre, elle permet de comparer les valeurs pour savoir si c'est un outlier grâce à un attribut choisi; **Node** qui représente un nœud de l'arbre, c'est un centre qui nous renvoie soit vers un ature noœud de l'arbre, soit dans la classe **DecisionLeaf** ou dans une classe **DirectDecision**, qui elle représente notre cas d'arrêt. Voici une capture de nos données:



La matrice associée à ce modèle est la suivante: $\begin{pmatrix} 214 & 36 \\ 21 & 59 \end{pmatrix}$

Aux niveau des métriques, nous avons une exactitude de 0,82, une exactitude pondérée de 0,79, une précision de 0,62 et un rappel de 0,73. Notre rappel est plus élevé que précédemment ainsi que la précision, ce qui veut dire que nous avons réussi à attraper plus de outlier et moins de faux positif.

3.3 Arbre généralisé

Hauteur	1	2	3	4
Exactitude pondérée	0,75	0,79	0,62	0,5
Précision	0,54	0,62	0,31	0,24
Rappel	0,7	0,73	0,82	0,8

La hauteur 2 semble donner les meilleurs résultats. En effet, il y a un bon compromis entre la précision et le rappel et l'exactitude pondérée est la plus haute. Pour les hauteurs 3 et 4 le rappel augmente mais la précision baisse significativement ce qui veut dire qu'on attrapera beaucoup des faux positifs.