

**Institut Supérieur des Sciences Appliquées et Économiques**

**associé au Conservatoire National des Arts et Métiers**

**Entreposage et Fouille de données-STA 211**

**Devoir 4 :**

**Prétraitement des données**

**Prépare par : …Elio Bou Serhal………….**

**Date :…Le…Mardi 4 /4/2023…….**

# **Objective**

# **Résumé ou partie théorie**

# **Déscription de la base de données**

# **Représentation graphiques de la base de données**

# **Application la méthode sur la base de données**

# **Conclusion**

# **Références**

# **Annexes**

Table de matières

[**Objective** 1](#_Toc90386381)

[**Résumé ou partie théorie** 1](#_Toc90386382)

[**Déscription de la base de données** 1](#_Toc90386383)

[**Représentation graphiques pour les variables :** 1](#_Toc90386384)

[**Application la méthode sur la base de données** 1](#_Toc90386385)

[**Conclusion** 1](#_Toc90386386)

[**Réfèrences** 1](#_Toc90386387)

[**Liste des figures** 3](#_Toc90386388)

[**Liste des tableaux** 4](#_Toc90386389)

# **Objective**

# L'objectif de cette étude est d'appliquer le modèle SVM (Support Vector Machine) sur la base de données radar ionosphérique et d'évaluer sa performance pour la classification binaire des retours radar en "bons" et "mauvais".

# Importer la base des données "ionosphere data" suivante sur R.

# Faire les statistiques descriptives selon la variable cible.

# Appliquer la méthode SVM et faire plusieurs essaies pour choisir les meilleurs paramètres qui réduit l'erreur de classement.

# Valider les différents modèles trouvés par les indices de matrice de confusion: précision, taux d'erreur, rappel, kappa,

# Trouver l' AUC et représenter la courbe ROC des différents modèles.

# Conclure.

# Les résultats de cette étude pourraient avoir des implications importantes pour la surveillance et la détection des phénomènes ionosphériques à des fins de communication et de navigation.

# **Résumé ou partie théorie**

# Voici une brève récapitulation des SVM et des étapes nécessaires pour l’entraîner :

# Les SVM (Support Vector Machines) sont une méthode d'apprentissage supervisé utilisée pour la classification et la régression. L'objectif des SVM est de trouver un hyperplan qui sépare les données en deux classes aussi bien que possible. L'hyperplan est choisi de manière à maximiser la marge, qui est la distance entre l'hyperplan et les échantillons (appelés vecteurs de support) les plus proches de chaque classe.

# Voici les étapes pour entraîner un modèle SVM sur des données:

# Importez les données dans RStudio. Les données doivent être dans un format qui peut être utilisé par le modèle SVM, par exemple un fichier CSV.

# Faire les statistiques descriptives selon la variable cible.

# Divisez les données en un ensemble d'entraînement et un ensemble de test. L'ensemble d'entraînement est utilisé pour entraîner le modèle SVM, tandis que l'ensemble de test est utilisé pour évaluer les performances du modèle.

1. Vérifier si notre base de données est linéaire ou non-linéaire
2. Préparez les données pour l'entraînement en normalisant les variables et en effectuant d'autres transformations nécessaires.

# Entraînez le modèle SVM en utilisant la fonction svm() de la bibliothèque e1071. Cette fonction prend plusieurs paramètres, y compris les données d'entraînement, le type de noyau, les paramètres du noyau et les paramètres de coût.

# Évaluez les performances du modèle en utilisant l'ensemble de test. En calculant des mesures telles que l'exactitude, la précision, le rappel et la courbe ROC pour évaluer les performances du modèle.

# Optimisez les paramètres du modèle en utilisant une méthode telle que la validation croisée pour trouver les meilleurs paramètres pour le modèle SVM.

# **Déscription de la base de données**

**Un tableau récapitulatif sur la structure de notre base de données :**

1. La valeur de skewness (skew) est une mesure de l'asymétrie de la distribution d'une variable. Si sa valeur est proche de zéro (comme la plus part des valeurs de notre base de données), cela signifie que la distribution est approximativement symétrique. Si elle est négative (aussi comme la plus part de notre base de donnee), cela indique que la queue gauche de la distribution est plus longue et/ou plus étalée que la queue droite. En d'autres termes, cela signifie que la distribution est inclinée vers la gauche (elle a une queue gauche plus étalée).
2. Le kurtosis est une mesure de l'aplatissement de la distribution d'une variable par rapport à une distribution normale. Un kurtosis de zéro indique que la distribution a la même forme que la distribution normale, alors qu'un kurtosis supérieur à zéro indique que la distribution est plus concentrée autour de la moyenne et que les queues sont plus épaisses que celles d'une distribution normale. Un kurtosis inférieur à zéro indique que la distribution est moins concentrée autour de la moyenne et que les queues sont plus minces que celles d'une distribution normale.
3. Le Minimum et Maximum sont entre -1 et 1 ce qui nous montrent que les valeurs varient entre -1 et 1, de plus la médiane dans la plus part des variables est nulle pour lr groupe 0 alorsque celle du groupe 1 varie telle que la plus part sont plus grande que 0.5

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **GROUP 0** | | | | | | | | | | | |
| **vars** | **mean** | **sd** | **median** | **trimmed** | **mad** | **min** | **max** | **range** | **skew** | **kurtosis** | **se** |
| cible | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | nan | nan | 0 |
| x1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | nan | nan | 0 |
| x2 | 0.21 | 0.81 | 0 | 0.25 | 1.48 | -1 | 1 | 2 | -0.38 | -1.42 | 0.13 |
| x3 | 0.05 | 0.84 | 0 | 0.06 | 1.48 | -1 | 1 | 2 | -0.09 | -1.6 | 0.14 |
| x4 | 0.37 | 0.71 | 0.5 | 0.44 | 0.74 | -1 | 1 | 2 | -0.63 | -0.89 | 0.12 |
| x5 | -0.05 | 0.8 | 0 | -0.06 | 1.48 | -1 | 1 | 2 | 0.09 | -1.49 | 0.13 |
| x6 | 0.24 | 0.71 | 0 | 0.28 | 1.48 | -1 | 1 | 2 | -0.35 | -1.06 | 0.12 |
| x7 | 0.08 | 0.75 | 0 | 0.09 | 1.48 | -1 | 1 | 2 | -0.12 | -1.26 | 0.12 |
| x8 | 0.24 | 0.75 | 0 | 0.28 | 1.48 | -1 | 1 | 2 | -0.39 | -1.19 | 0.12 |
| x9 | 0.25 | 0.73 | 0 | 0.3 | 1.48 | -1 | 1 | 2 | -0.38 | -1.14 | 0.12 |
| x10 | 0.42 | 0.64 | 0.5 | 0.5 | 0.74 | -1 | 1 | 2 | -0.61 | -0.7 | 0.1 |
| x11 | 0.05 | 0.77 | 0 | 0.06 | 1.48 | -1 | 1 | 2 | -0.08 | -1.35 | 0.12 |
| x12 | 0.22 | 0.8 | 0 | 0.26 | 1.48 | -1 | 1 | 2 | -0.37 | -1.41 | 0.13 |
| x13 | -0.19 | 0.8 | 0 | -0.22 | 1.48 | -1 | 1 | 2 | 0.34 | -1.4 | 0.13 |
| x14 | 0.13 | 0.7 | 0 | 0.16 | 0.74 | -1 | 1 | 2 | -0.18 | -1.04 | 0.11 |
| x15 | -0.06 | 0.7 | 0 | -0.07 | 0.23 | -1 | 1 | 2 | 0.1 | -1 | 0.11 |
| x16 | 0.28 | 0.77 | 0 | 0.33 | 1.48 | -1 | 1 | 2 | -0.48 | -1.25 | 0.13 |
| x17 | -0.11 | 0.82 | 0 | -0.13 | 1.48 | -1 | 1 | 2 | 0.18 | -1.56 | 0.13 |
| x18 | 0 | 0.84 | 0 | 0 | 1.48 | -1 | 1 | 2 | 0 | -1.61 | 0.14 |
| x19 | -0.05 | 0.84 | 0 | -0.06 | 1.48 | -1 | 1 | 2 | 0.09 | -1.6 | 0.14 |
| x20 | 0.03 | 0.81 | 0 | 0.04 | 1.48 | -1 | 1 | 2 | -0.04 | -1.55 | 0.13 |
| x21 | 0.24 | 0.79 | 0 | 0.28 | 1.48 | -1 | 1 | 2 | -0.42 | -1.31 | 0.13 |
| x22 | 0.34 | 0.75 | 0.5 | 0.41 | 0.74 | -1 | 1 | 2 | -0.61 | -1.02 | 0.12 |
| x23 | 0.07 | 0.81 | 0 | 0.08 | 1.48 | -1 | 1 | 2 | -0.14 | -1.51 | 0.13 |
| x24 | 0.37 | 0.79 | 1 | 0.44 | 0 | -1 | 1 | 2 | -0.71 | -1.06 | 0.13 |
| x25 | -0.29 | 0.8 | -0.5 | -0.34 | 0.74 | -1 | 1 | 2 | 0.54 | -1.28 | 0.13 |
| x26 | 0.84 | 0.55 | 1 | 1 | 0 | -1 | 1 | 2 | -3 | 7.19 | 0.09 |
| x27 | -0.05 | 0.98 | -0.5 | -0.06 | 0.74 | -1 | 1 | 2 | 0.1 | -1.99 | 0.16 |
| x28 | 0.16 | 0.82 | 0 | 0.19 | 1.48 | -1 | 1 | 2 | -0.28 | -1.51 | 0.13 |
| x29 | 0.15 | 0.82 | 0 | 0.18 | 1.48 | -1 | 1 | 2 | -0.28 | -1.49 | 0.13 |
| x30 | 0.08 | 0.85 | 0 | 0.09 | 1.48 | -1 | 1 | 2 | -0.14 | -1.64 | 0.14 |
| x31 | 0.14 | 0.81 | 0 | 0.17 | 1.48 | -1 | 1 | 2 | -0.27 | -1.48 | 0.13 |
| x32 | 0.11 | 0.69 | 0 | 0.12 | 0 | -1 | 1 | 2 | -0.13 | -0.96 | 0.11 |
| x33 | 0 | 0.7 | 0 | 0 | 0 | -1 | 1 | 2 | 0 | -1 | 0.11 |
| x34\* | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | NaN | NaN | 0 |
|  |  |  |  |  |  |  |  |  |  |  |  |
| **GROUP 1** | | | | | | | | | | | |
| **vars** | **mean** | **sd** | **median** | **trimmed** | **mad** | **min** | **max** | **range** | **skew** | **kurtosis** | **se** |
| cible | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | NaN | NaN | 0 |
| x1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | NaN | NaN | 0 |
| x2 | 0.69 | 0.42 | 0.88 | 0.78 | 0.18 | -1 | 1 | 2 | -2 | 4.33 | 0.02 |
| x3 | 0.04 | 0.37 | 0.02 | 0.06 | 0.17 | -1 | 1 | 2 | -0.34 | 2.28 | 0.02 |
| x4 | 0.63 | 0.49 | 0.81 | 0.72 | 0.28 | -1 | 1 | 2 | -1.77 | 2.81 | 0.03 |
| x5 | 0.14 | 0.4 | 0.03 | 0.13 | 0.2 | -1 | 1 | 2 | -0.11 | 1.28 | 0.02 |
| x6 | 0.59 | 0.45 | 0.76 | 0.65 | 0.36 | -1 | 1 | 2 | -1.36 | 1.86 | 0.03 |
| x7 | 0.13 | 0.49 | 0.02 | 0.14 | 0.25 | -1 | 1 | 2 | -0.21 | 0.21 | 0.03 |
| x8 | 0.54 | 0.46 | 0.69 | 0.6 | 0.44 | -1 | 1 | 2 | -1.04 | 0.56 | 0.03 |
| x9 | 0.17 | 0.45 | 0.02 | 0.17 | 0.2 | -1 | 1 | 2 | 0.06 | 0.28 | 0.03 |
| x10 | 0.48 | 0.55 | 0.67 | 0.57 | 0.46 | -1 | 1 | 2 | -1.09 | 0.24 | 0.03 |
| x11 | 0.17 | 0.45 | 0.04 | 0.16 | 0.29 | -1 | 1 | 2 | 0.08 | 0.05 | 0.03 |
| x12 | 0.42 | 0.6 | 0.65 | 0.51 | 0.51 | -1 | 1 | 2 | -0.91 | -0.32 | 0.03 |
| x13 | 0.13 | 0.43 | 0.05 | 0.13 | 0.26 | -1 | 1 | 2 | -0.05 | 0.41 | 0.02 |
| x14 | 0.37 | 0.64 | 0.63 | 0.45 | 0.54 | -1 | 1 | 2 | -0.85 | -0.59 | 0.04 |
| x15 | 0.09 | 0.42 | 0.01 | 0.09 | 0.24 | -1 | 1 | 2 | -0.01 | 0.64 | 0.02 |
| x16 | 0.39 | 0.6 | 0.59 | 0.47 | 0.57 | -1 | 1 | 2 | -0.84 | -0.51 | 0.03 |
| x17 | 0.01 | 0.44 | 0 | 0.01 | 0.29 | -1 | 1 | 2 | 0.07 | 0.21 | 0.02 |
| x18 | 0.4 | 0.58 | 0.6 | 0.48 | 0.53 | -1 | 1 | 2 | -0.89 | -0.35 | 0.03 |
| x19 | -0.02 | 0.47 | 0 | -0.02 | 0.26 | -1 | 1 | 2 | 0.07 | 0.13 | 0.03 |
| x20 | 0.37 | 0.57 | 0.53 | 0.43 | 0.67 | -1 | 1 | 2 | -0.68 | -0.67 | 0.03 |
| x21 | -0.02 | 0.47 | 0 | -0.02 | 0.29 | -1 | 1 | 2 | 0.01 | 0.07 | 0.03 |
| x22 | 0.36 | 0.59 | 0.53 | 0.43 | 0.66 | -1 | 1 | 2 | -0.75 | -0.54 | 0.03 |
| x23 | -0.07 | 0.48 | 0 | -0.07 | 0.37 | -1 | 1 | 2 | -0.03 | -0.11 | 0.03 |
| x24 | 0.4 | 0.55 | 0.54 | 0.47 | 0.57 | -1 | 1 | 2 | -0.87 | -0.13 | 0.03 |
| x25 | -0.04 | 0.46 | -0.01 | -0.05 | 0.32 | -1 | 1 | 2 | 0.08 | 0.09 | 0.03 |
| x26 | 0.51 | 0.5 | 0.63 | 0.58 | 0.47 | -1 | 1 | 2 | -1.15 | 0.7 | 0.03 |
| x27 | -0.07 | 0.47 | -0.02 | -0.07 | 0.29 | -1 | 1 | 2 | -0.04 | 0.05 | 0.03 |
| x28 | 0.41 | 0.53 | 0.52 | 0.48 | 0.58 | -1 | 1 | 2 | -0.91 | 0.15 | 0.03 |
| x29 | -0.05 | 0.45 | 0 | -0.04 | 0.29 | -1 | 1 | 2 | -0.07 | 0.27 | 0.03 |
| x30 | 0.39 | 0.52 | 0.46 | 0.45 | 0.61 | -1 | 1 | 2 | -0.79 | 0.06 | 0.03 |
| x31 | -0.02 | 0.46 | 0 | -0.02 | 0.3 | -1 | 1 | 2 | 0.04 | 0.21 | 0.03 |
| x32 | 0.38 | 0.49 | 0.45 | 0.42 | 0.64 | -1 | 1 | 2 | -0.59 | -0.27 | 0.03 |
| x33 | 0.02 | 0.43 | 0 | 0.01 | 0.25 | -1 | 1 | 2 | 0.17 | 0.4 | 0.02 |
| x34\* | 1.72 | 0.45 | 2 | 1.77 | 0 | 1 | 2 | 1 | -0.96 | -1.07 | 0.03 |

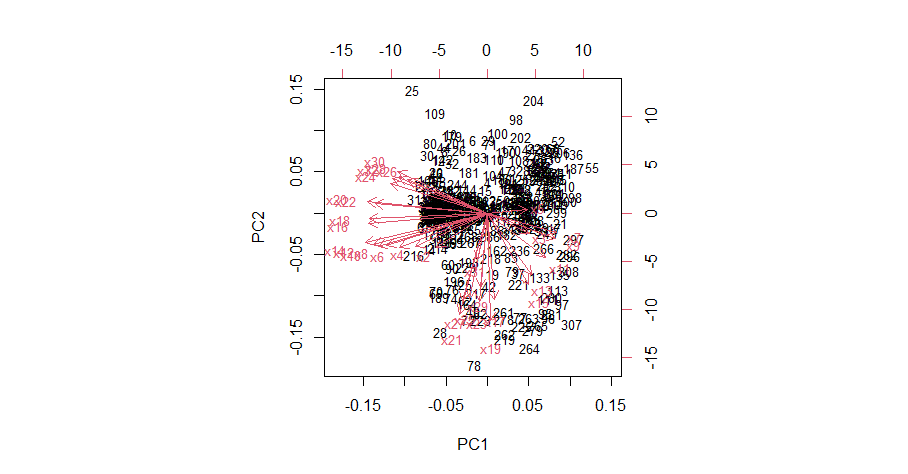
**Etudier si notre base de données est linéaire ou non-linéaire :**

**Méthode 1 :** Effectuer une Analyse Factorielle « ACP »

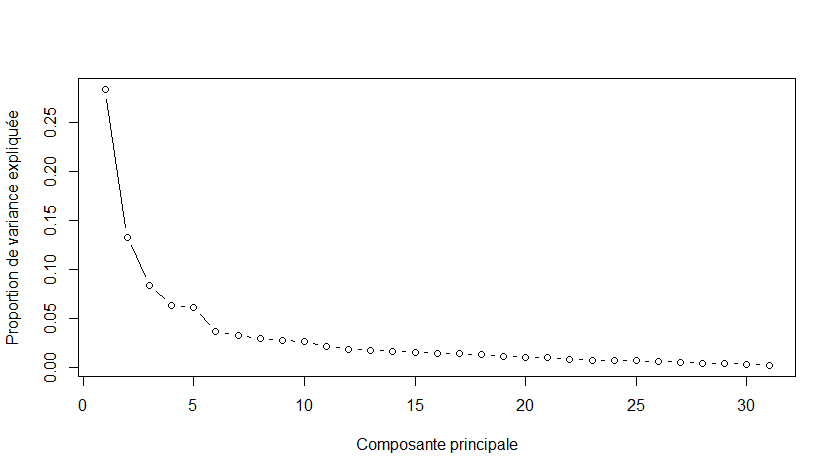
Le biplot généré par la fonction biplot() affiche les données dans un espace de deux dimensions représenté par les deux premières composantes principales. Les flèches indiquent la direction de chaque variable dans cet espace, tandis que les cercles représentent les points de données. Si les flèches sont approximativement parallèles ou pointent dans la même direction, cela indique une forte linéarité des données par rapport à ces variables. Si les flèches pointent dans des directions différentes, cela indique une faible linéarité des données par rapport à ces variables.

Pour analyser un biplot et déterminer si les variables sont linéaires ou non, on peut suivre ces étapes :

1. Examiner la direction et la longueur des flèches : Si les flèches pointent dans des directions différentes et ont des longueurs similaires, cela indique que les variables ne sont pas linéairement liées ce qui est présent dans notre cas. En revanche, si les flèches pointent dans la même direction et ont des longueurs similaires, cela indique que les variables sont linéairement liées.
2. Examiner la distance entre les points de données : Si les points de données sont éloignés les uns des autres, cela indique une forte variation et une forte corrélation entre les variables. En revanche, si les points de données sont proches les uns des autres, cela indique une faible variation et une faible corrélation entre les variables ce qui est présent dans notre cas.
3. Examiner les angles entre les flèches : Si les angles entre les flèches sont proches de 90 degrés, cela indique que les variables sont indépendantes ce qui est present dans notre cas. En revanche, si les angles entre les flèches sont proches de zéro degré, cela indique que les variables sont fortement corrélées.



1. Examiner la proportion de variance expliquée : La proportion de variance expliquée par les premières composantes principales peut indiquer si une ou plusieurs composantes principales expliquent la majorité de la variance des données. Si une seule composante principale explique la majorité de la variance (par exemple, plus de 70 %), cela peut indiquer que les variables sont fortement corrélées. Mais dans notre cas il n’existe pas une composante principale qui explique la majorite de la variance (la plus grande proportion de variance est de 28%).



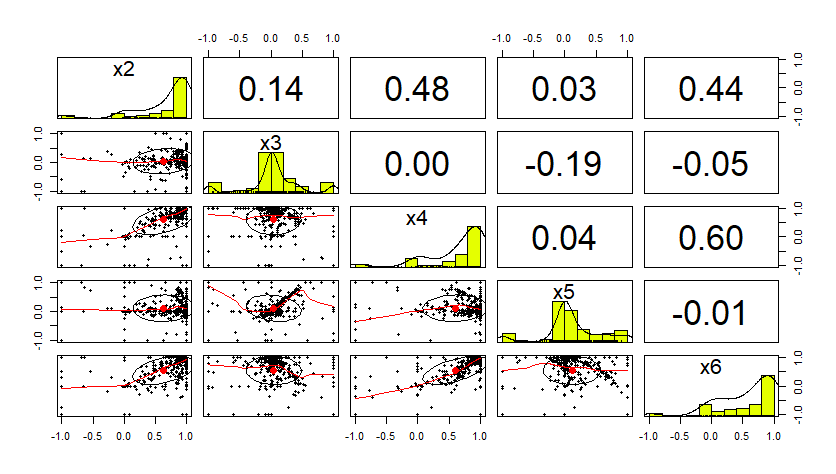
Bref : Dans notre cas la figure montre qu’il n’existe pas une linéarité entre ces variables alors on va appliquer le modèle SVM dans le cas non-linéaire.

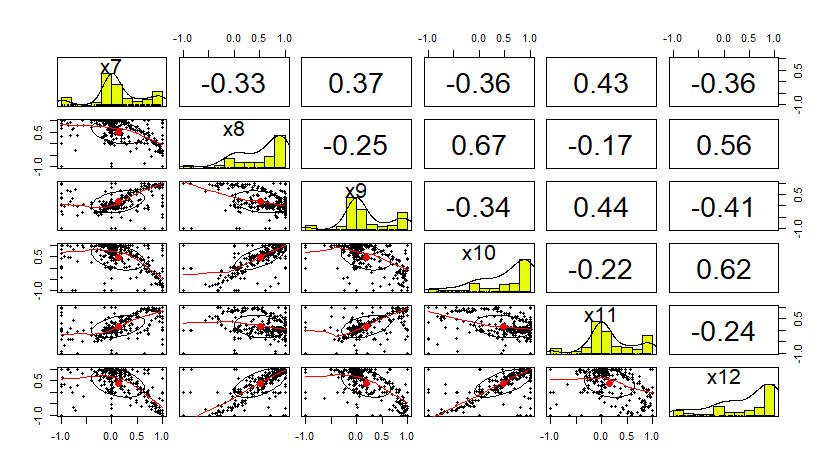
**Méthode 2 :** Trouver la corrélation entre les variables

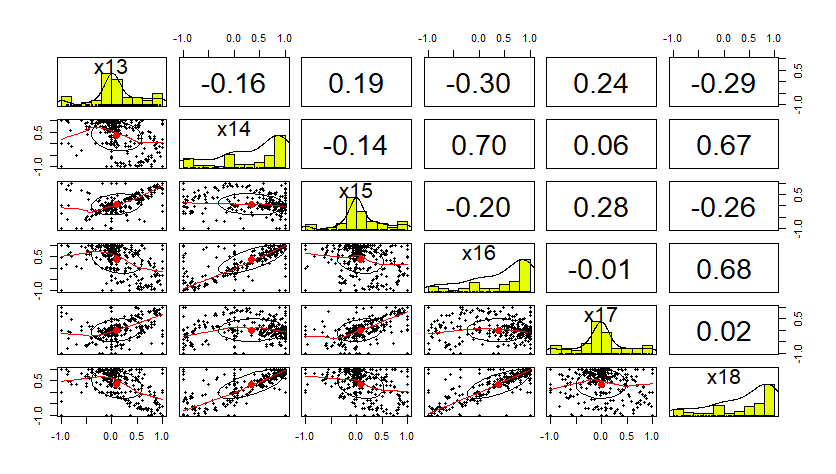
Dans les figures suivantes vous pouvez voir :

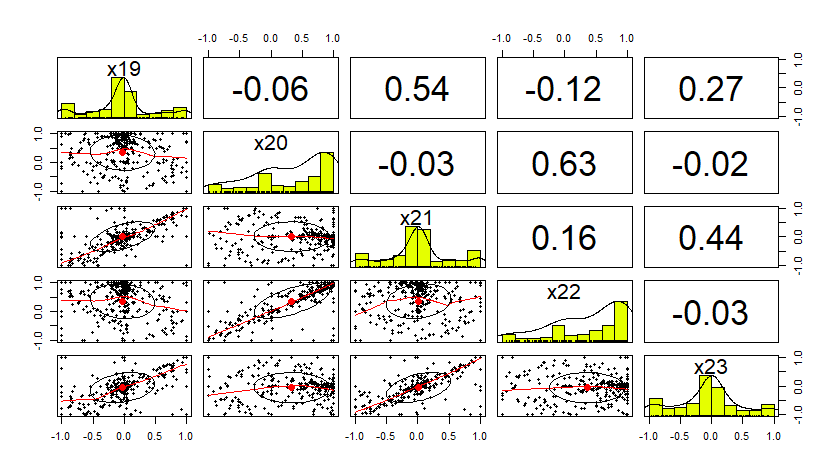
1. Coefficients de corrélation : Les panneaux triangulaires supérieurs du graphique affichent les coefficients de corrélation de Pearson entre les paires de variables, ce qui suggère que les variables sont faiblement liées entre elles. (valeurs proches de 0)
2. Nuages de points : Les panneaux triangulaires inférieurs du graphique affichent les nuages de points de paires de variables,qui nous montrent des relations non linéaires.
3. Diagrammes de densité : Les panneaux diagonaux du graphique affichent les diagrammes de densité des variables individuelles,qui nous montrent la forme des distributions : symetrie ou asymetrie, de plus s’il contient des valeurs aberrntes.
4. Ellipses : Les panneaux triangulaires inférieurs du graphique affichent également des ellipses de corrélation pour cherchez la forme et l'orientation des ellipses, qui peuvent nous donner des informations sur la force et la direction de la corrélation entre les variables.

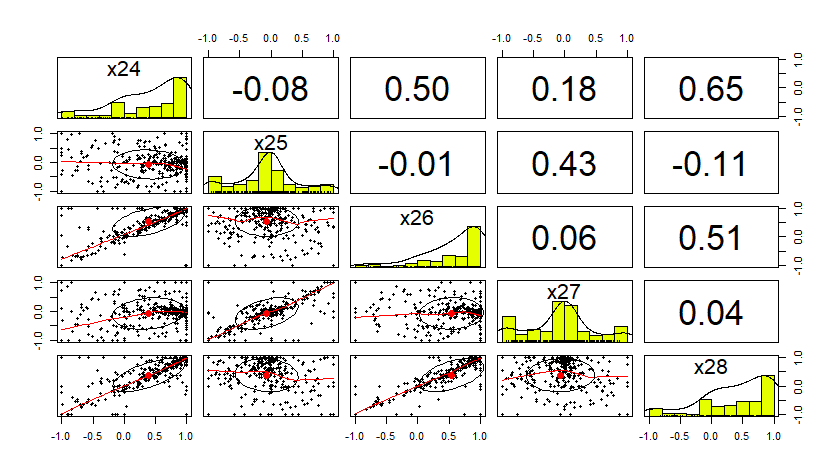
Bref : Auncun de ces 4 facteurs de la fonction « pairs.panels » montrent que notre base de donnee est lineaire.

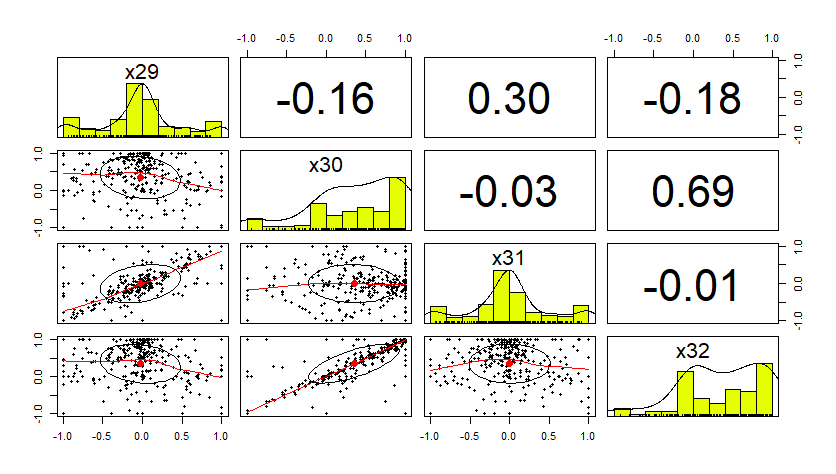










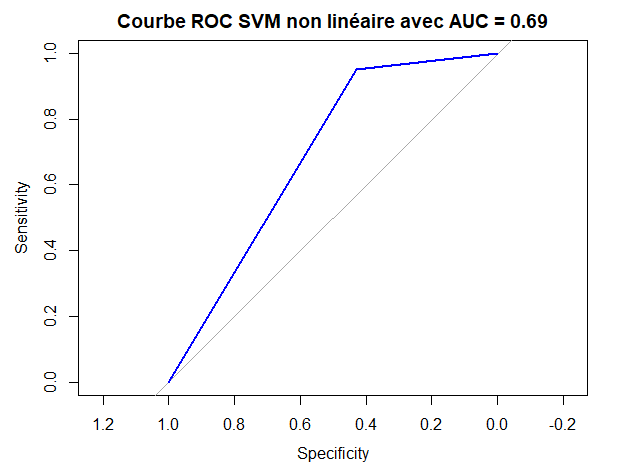


# **Représentation graphiques de la base de données**

# **Application la méthode sur la base de données**

**Cas 1 du modèle SVM : En utilisant la librairie « Tune »**

1. Le modèle SVM s'agit d'un modèle de classification qui divise les données en deux classes en utilisant un hyperplan.
2. La fonction "tune" est utilisée pour trouver les meilleurs paramètres pour le modèle SVM en utilisant l'ensemble d'apprentissage.
3. Les paramètres optimaux trouvés sont utilisés pour créer le modèle final.
4. Les résultats de la fonction "summary" pour montrer combien de vecteurs de support ont été utilisés dans le modèle final.
5. Les prévisions pour l'ensemble de données de test en utilisant la fonction "predict".
6. Les trois méthodes d'évaluation de la performance : la matrice de confusion qui nous a donnee un accuracy de 89,86%, le taux d'erreur de 10,14% et la courbe ROC d’AUC 69%.

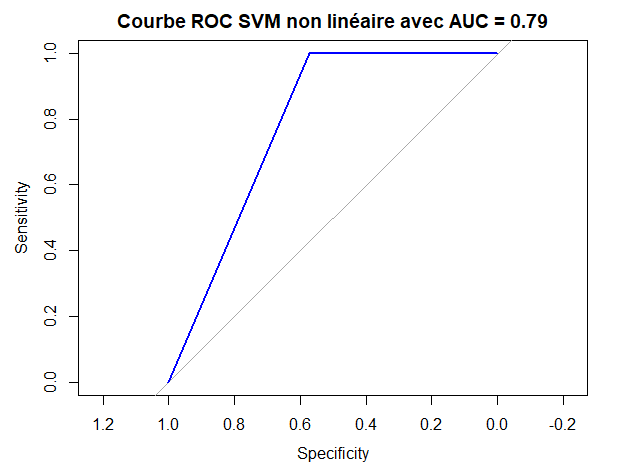


1. Représenter la courbe ROC en utilisant la fonction "roc" de la bibliothèque "pROC".
2. Expliquez comment la fonction "ci" de la bibliothèque "pROC" est utilisée pour calculer l'intervalle de confiance (95% CI: 0.4903-0.8899) de l'AUC.
3. Expliquez comment la fonction "coords" de la bibliothèque "pROC" est utilisée pour trouver le seuil optimal de la courbe ROC en fonction de la spécificité qui a pour valeur 42.85% et de la sensibilité qui a pour valeur 95.16%.

**Cas 2 du modèle SVM : En considérant que la base de donnée est linéaire pour kernel= « linéaire »**

Dans ce cas on a considère que notre base de donnée est linéaire avec le paramètre kernel= « linéaire » alors que les autre paramètres restent telle qu’elles sont.

1. Le modèle est construit à l'aide de la bibliothèque R "e1071".
2. Le code commence par diviser les données en deux ensembles: un ensemble d'entraînement ("train") et un ensemble de test ("test").
3. Ensuite, un SVM linéaire est construit à l'aide de la fonction "svm". Les paramètres de coût et de gamma sont spécifiés comme 0,1 et 0,5, respectivement.
4. La fonction "summary" est utilisée pour afficher les résultats de l'apprentissage. Ensuite, les prédictions sont faites sur l'ensemble de test et stockées dans une variable appelée "pred".
5. Trois méthodes d'évaluation de la performance du modèle sont utilisées: la matrice de confusion qui nous a donnée un accuracy de 95,65%, le taux d'erreur de 4,34% et la courbe ROC d’AUC 79%.



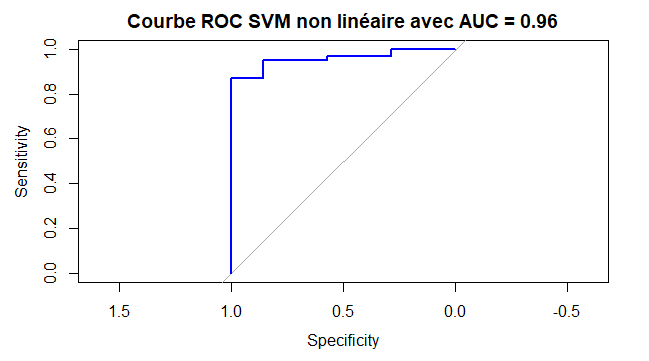
1. Expliquez comment la fonction "ci" de la bibliothèque "pROC" est utilisée pour calculer l'intervalle de confiance (95% CI: 0.5877-0.9837) de l'AUC.
2. Expliquez comment la fonction "coords" de la bibliothèque "pROC" est utilisée pour trouver le seuil optimal de la courbe ROC en fonction de la spécificité qui a pour valeur 57.14% et de la sensibilité qui a pour valeur 100%.

**Conclusion :**

On a obtenu un résultat inattendu puisqu’on a obtenu un meilleur résultat avec le kernel linéaire qu'avec le kernel polynomial pour une base de données qui n'est pas linéaire, cela peut être surprenant car le kernel linéaire ne peut pas modéliser des relations non linéaires dans les données, tandis que le kernel polynomial peut. Pour cela on va essayer une autre méthode de SVM pour faire la comparaison et choisir la meilleure méthode.

**Cas 3 du modèle SVM : En Utilisant la librairie « Rminer »**

1. Tout d’abord créer une partition aléatoire des données en un ensemble d'entraînement et un ensemble de test.
2. Supprimer la colonne "x1" de l'ensemble d'entraînement et de l'ensemble de test puisqu’elle contient la même valeur (une constante).
3. Supprimer les lignes contenant des valeurs manquantes dans l'ensemble d'entraînement et l'ensemble de test, puis entraîner un modèle SVM non-linéaire avec un noyau RBF. Le modèle est entraîné sur l'ensemble d'entraînement en utilisant la fonction fit().
4. Effectuer des prédictions sur l'ensemble de test à l'aide du modèle SVM non-linéaire entraîné, calculer le taux d'erreur de 2% en comparant les prédictions aux valeurs réelles de l'ensemble de test.
5. charger la librairie pROC, et créer un objet de courbe ROC à l'aide des prédictions et des vraies valeurs de l'ensemble de test, pour calculer l'aire sous la courbe ROC (AUC de 96%) à l'aide de la fonction auc()
6. tracent la courbe ROC à l'aide de la fonction plot().



1. Expliquez comment la fonction "ci" de la bibliothèque "pROC" est utilisée pour calculer l'intervalle de confiance (95% CI: 0.911-1) de l'AUC.
2. Expliquez comment la fonction "coords" de la bibliothèque "pROC" est utilisée pour trouver le seuil optimal de la courbe ROC en fonction de la spécificité qui a pour valeur 100% et de la sensibilité qui a pour valeur 87.09%.

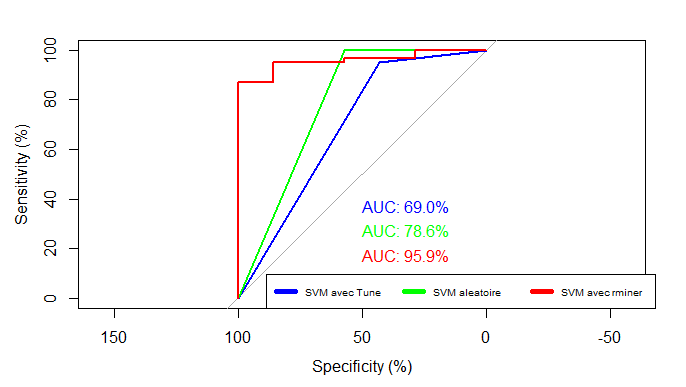
# **Conclusion**

Nous avons implémenté trois modèles SVM pour la classification de données.

Le premier modèle a utilisé le noyau linéaire, le deuxième modèle a utilisé le noyau gaussien et le troisième modèle a utilisé la librairie Rminer pour déterminer automatiquement les paramètres du modèle.

Nous avons évalué la performance des trois modèles en utilisant le taux d'erreur et la courbe ROC.

Et alors voici une figure qui montre les trois courbes de Roc ensemble, ce qui nous permette de comparer l’AUC et choisir le meilleur qui correspond au 3eme cas qui a pour AUC 95.9%

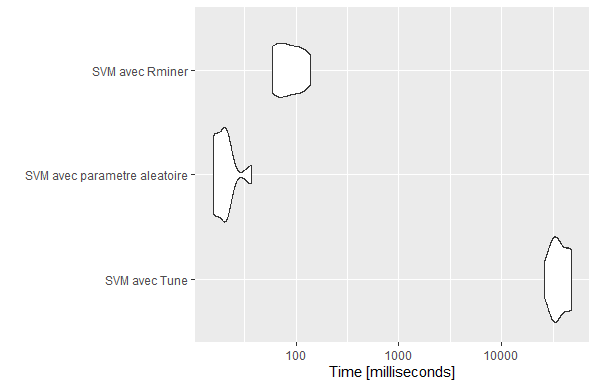


|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **min** | **lq** | **mean** | **median** | **uq** | **max** | **neval** |
| SVM avec Tune | 26042.411 | 31727.948 | 37379.505 | 36029.058 | 47223.304 | 48034.51 | 10 |
| SVM avec parametre aleatoire | 15.7084 | 15.8441 | 20.4252 | 20.1975 | 21.2908 | 36.4355 | 10 |
| SVM avec Rminer | 58.3472 | 61.8002 | 88.7624 | 81.0443 | 104.9699 | 138.1629 | 10 |

En regardant les résultats de ce tableau, nous pouvons voir que l'algorithme SVM avec réglage de paramètres à l'aide de Tune a une performance significativement meilleure que les deux autres algorithmes. En effet, il a la plus petite valeur de temps d'exécution moyen, ce qui indique une rapidité d'exécution, ainsi qu'un écart interquartile (Q3-Q1) plus petit que les autres algorithmes, ce qui indique une plus grande stabilité de la performance.

D'un autre côté, nous pouvons voir que l'algorithme SVM avec rminer a le temps d'exécution moyen le plus élevé, avec une variance plus élevée, ce qui indique une certaine instabilité dans la performance.

Enfin, l'algorithme SVM avec paramètres aléatoires a le temps d'exécution moyen le plus court, mais sa performance est considérablement inférieure à celle des deux autres algorithmes, avec une grande variance dans le temps d'exécution.



**Bref :**

En appliquant les 3 cas sur les données, nous avons obtenu des modèles SVM avec des performances différentes.

Le premier modèle SVM utilisant la méthode "tune" pour optimiser les paramètres a donné un taux d'erreur de 10,14% et une AUC de 69%.

Le deuxième modèle SVM utilisant des paramètres aléatoires a donné un taux d'erreur de 4,34% et une AUC de 79%.

Le troisième modèle SVM utilisant la bibliothèque "rminer" a donné un taux d'erreur de 2% et une AUC de 96%.

Nous pouvons conclure que le troisième modèle SVM est légèrement meilleur que les deux autres modèles car il a donné le taux d'erreur le plus faible et l'AUC la plus élevée. Cependant, les différences entre les performances des trois modèles ne sont pas très grandes, ce qui suggère que le choix de la méthode de réglage des paramètres pourrait ne pas avoir un impact majeur sur la performance du modèle.