Scolarisation des jeunes filles et garçons au niveau de l'enseignement primaire au Bénin : approche par Apprentissage supervisé

Adjimon VITOFFODJI

Table des matières

INTRODUCTION GENERAL

Justification de l'étude

Au forum mondial sur l'éducation, qui s'est tenu à Dakar en avril 2000, la communauté internationale a réaffirmé son engagement à garantir l'accès pour tous à une éducation de base de qualité à l'horizon 2015. Les efforts accomplis ont entraîné une nette amélioration des taux de scolarisation. Malgré ces bons résultats, de très nombreux enfants n'ont toujours pas accès à l'éduction et d'autre abandonnent l'école. Il ne suffit donc pas d'augmenter les ressources et les capacités des systèmes éducatifs pour résoudre efficacement le problème.

L'évolution de la scolarisation et les progrès importants réalisés à l'égard de filles et garçons du Bénin, méritent une étude approfondie. C'est dans cette optique que nous jugé utile, dans le cadre de notre projet d'étude en Science de donnée de réfléchir sur le thème : << Scolarisation des jeunes filles et garçons au niveau de l'enseignement primaire au Bénin : approche par apprentissage supervisé >>.
La présente étude s'articule autour de deux (02) chapitres. Le premier met en relief le cadre théorique et méthodologique de l'étude, le second présente les résultats et interprétation.

CHAPITRE 1 : CADRE THEORIQUE ET METHODOLOGIE

Le présent chapitre expose le cadre théorique et méthodologique de l'étude. Il s'articule autour de deux sections à savoir : les Objectifs et hypothèses de recherche et la méthodologie.

Section 1 : Cadre de l'étude

A-Problématique

Depuis 1990 et la Conférence de Jomtien (Thaïlande), la scolarisation des filles mobilise les différents acteurs impliqués dans les systèmes éducatifs. Des programmes pour la promotion de la scolarisation des filles ont été mis en œuvre dans maints pays. Mais il reste des défis tout aussi importants à relever, particulièrement en Afrique subsaharienne, pour que tous les enfants notamment les filles, aient la possibilité d'accéder à un enseignement primaire obligatoire et gratuit de qualité et de suivre jusqu'à son terme, et pour éliminer les disparités entre les sexes dans l'enseignement primaire et secondaire et assurer aux filles l'accès équitable et sans restriction à une éducation de base de qualité avec les mêmes chances de réussite.

Malgré les différentes actions que mène le gouvernement, la scolarisation à l'éducation à la base des enfants surtout des filles reste des moindre.

B-Objectifs de l'étude

Cette étude vise à :

- Classer si 70% des individus ont achevé ou non leurs cursus
- Prédire si 70% des élèves arrivent à terminer leur cursus

Section 2 : Méthodologie de recherche

Dans cette partie, nous présenterons dans un premier temps, la description de notre de jeu donné, la nature et les sources de données puis, dans un second temps, les différentes méthodes de l'apprentissage non supervisé utilisées dans le cadre de notre étude.

Paragraphe 1 : Données utilisées et sources

A. Type et sources des données

Les données utilisées dans cette étude proviennent des indicateurs scolaires au niveau de l'enseignement primaire au Bénin de 2003 à 2017. La source Direction de la Programmation et de la Prospective (DPP) et du Ministère de l'Enseignement Maternel et Primaire (MEMP) du Bénin. Les données sont sur le site de l'Institut National de la Statistique et de l'Analyse économique (INSAE).

https://instad.bj/statistiques/statistiques-sociales

B. Description des variables utilisées

La variable cible d'intérêt de notre étude est la variable nominale Taux de réussite. Elle va nous permettre de prédire si soixante-dix (70%) des élèves arrivent à achever leurs cursus du primaire et si elle abandonne leur scolarité.

Les variables explicatives d'intérêt de notre étude sont : la variable catégorielle Sexe qui représente le sexe de l'élève, la variable quantitative Taux brut de scolarisation qui représente le pourcentage d'élève inscrit dans l'enseignement primaire durant chaque année de 2003 à 2017, la variable quantitative Taux d'achèvement qui représente le pourcentage des élèves qui arrivent à terminer leur scolarité et n'abandonne pas pour divers motifs.

La variables explicative Effectifs des élèves pour indiquer le nombre d'élève inscrit au cours de chaque année

Tableau : Variables et Modalités 1

Variables	Description	Modalité
Taux de réussite (Cible)	Est-ce que 70% des élèves inscrit ont pu achever leur année scolaire ?	Variable nominale 1- Oui 2- Non
Sexe	Sexe de élèves	Variable catégorielle 1- Masculin 2- Féminin

Taux bruite de scolarisation	Pourcentage d'élève inscrit par sexe et par année	Variable quantitative
Taux d'achèvement	Pourcentage d'élève par sexe qui arrive à terminer leur année	Variable quantitative
Effectifs d'élèves par sexe	Nombre d'élève inscrit par sexe durant chaque année	Variable quantitative

Paragraphe 2 : Méthode d'analyse

La méthodologie adoptée en vue d'atteindre notre objectif fait l'objet de ce paragraphe.

Spécification des clusters

Pour déterminer prédire le taux d'élèves qui arrivent à terminer leurs cursus primaires, nous allons dans un premier temps effectué un regroupent selon si deux classes connu d'avance que sont : si 70% des écoliers arrivent à achever leurs cursus du primaire.

Pour cela, avec Data Table, nous allons visualiser nos données dans Orange, ensuite nous allons ajouter les modèles classification suivante :

- L'algorithme des K plus proches voisins (KNN)
- Régression logistique
- Classifieur naïf de Bayes
- Machine à vecteurs de support
- Arbre de décision
- Forêt aléatoire
- Réseau de neurones

A la suite nous allons choisir le modèle le plus performant. Pour ce fait, nous allons utiliser Test and Score qui permet de générer une validation croisée des modèles. Ainsi pour choisir le modèle plus performant nous allons observer et comparer les accuracy CA, les f-mésure F1, la précision (Precision), et le rappel (Recall) de chaque modèle.

Dans notre étude, nous avons retenu le le Taux de prédiction correctes accuracy (CA) car ce dernier va nous permettre de prédire le taux d'écolier correcte suivant le sexe masculin et féminin qui arrive à terminer leurs cursus du primaire.

Validation du modèle

Nous allons utiliser la matrice de confusion à la suite de chaque modèle de classification pour retenir le modèle qui fait le moins d'erreurs et qui parait donc plus performant Ensuite étant donné que c'est le Taux de prédiction correcte qui sera retenu dans notre étude, nous allons comparer alors le Taux de prédiction de chaque modèle et retenu le modèle dont le Taux de prédiction correcte est le plus grand.

Ensuite, nous allons utiliser la Matrice de confusion pour visualiser les individus qui sont mal classé par les modèles. Nous allons à la suite utilisé un Scatter Plot qui prend en entré l'ensemble des données de files pour visualiser où se situe les individus mal classés. A la suite nous allons utiliser find informative Projections pour représenter les variables qui séparent au mieux les données selon le taux de réussite. Nous allons après refaire la même procédure pour en utilisant cette fois ci un positionnement multidimensionnel à la place du nuage de points pour vérifier si les résultats restent cohérents par rapport aux observation précédentes.

Enfin, une fois le meilleur modèle qui produit les meilleurs résultats identifiés, nous allons pourvoir faire les prédictions.

CHAPITRE 2 : PRÉSENTATION ET ANALYSE DES RÉSULTATS ET VALIDATIONS DU MODÈLE

Le présent chapitre expose les résultats obtenus après le traitement des données, leurs interprétations et la validation du modèle.

Nos différentes variables nous permettent d'avoir le nuage de points suivant :

Tableau 2 : File base 1

Columns (Double click to edit)

	Name	Type	Role	Values
1	Effectif des ...	N numeric	feature	
2	Taux brut de ...	N numeric	feature	
3	Taux ...	N numeric	feature	
4	Sexe	C categorical	feature	Féminin,Masculin
5	**Taux de ...**	**C categori...**	**target**	**Non,Oui**

Le tableau 1 montre le résumer de notre jeu de donnée.

Ainsi nous avons :

- 1 variable nominale (Taux réussite) qui indique si 70% des individus ont pu achever leur cursus (Oui) sinon (Non) ;
- 1 variable catégorielle sexe, qui indique le sexe de l'individu ;
- 3 variables quantitatives qui représentent l'effectif d'élève inscrit, le taux brut d'achèvement et le taux brut de scolarisation

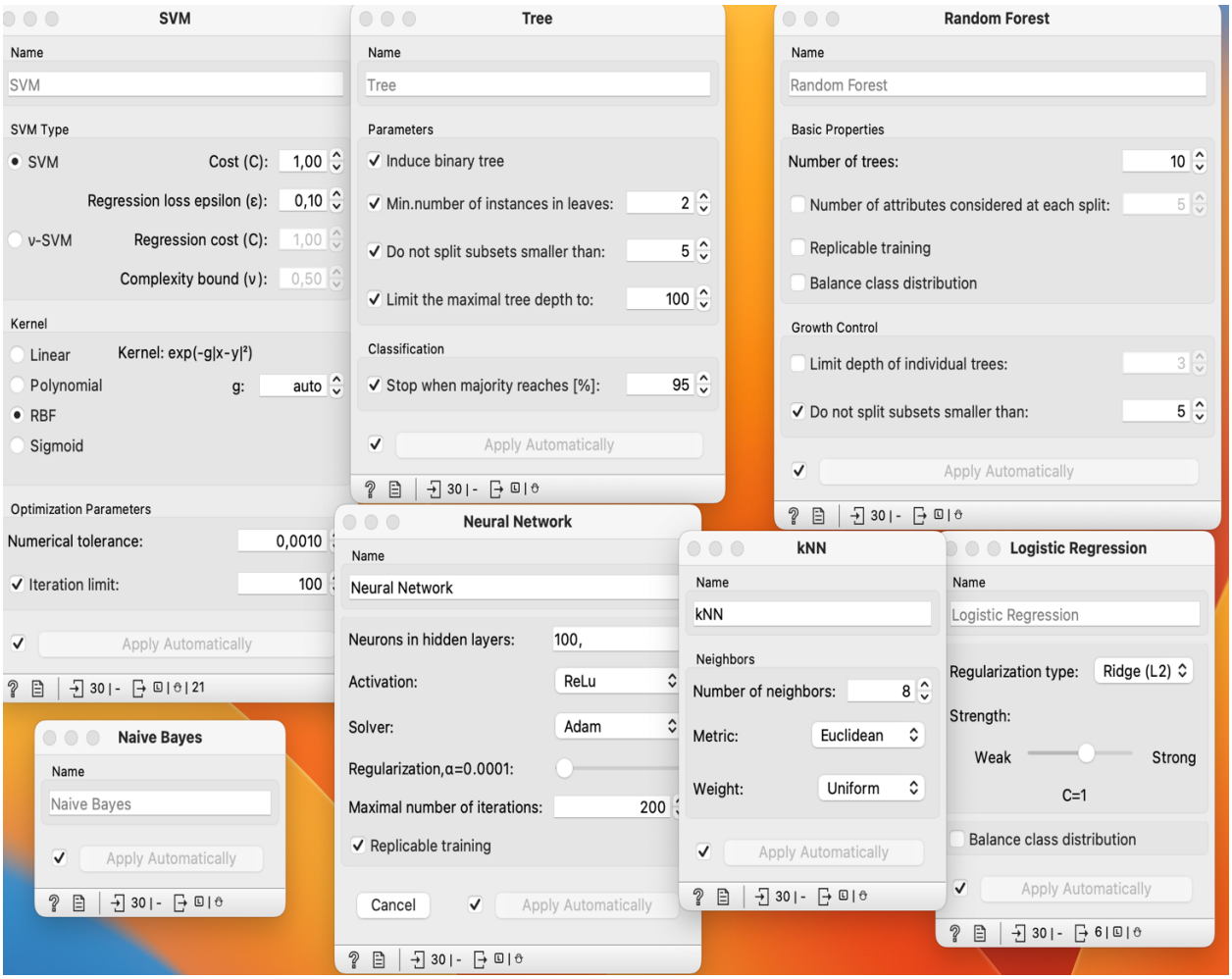
Tableau 3 : Data table 1

	Taux de réussite	ffectif des élèves	brut de scolarisa	aux d'achèvemen	Sexe
1	Non	1259537	107.10	61.10	Masculin
2	Non	1319648	107.80	61.60	Masculin
3	Non	1318140	105.22	77.50	Masculin
4	Oui	1356818	99.59	75.60	Masculin
5	Oui	1474206	104.33	75.85	Masculin
6	Oui	1601146	109.42	68.37	Masculin
7	Non	1719390	113.46	56.83	Masculin
8	Non	1787940	114.02	70.35	Masculin
9	Oui	1869379	115.93	73.15	Masculin
10	Oui	1987182	122.45	76.80	Masculin
11	Oui	2064031	123.37	81.60	Masculin
12	Oui	2133330	123.68	81.28	Masculin
13	Oui	2238185	126.16	82.03	Masculin
14	Oui	2267835	119.12	74.81	Masculin
15	Oui	2246949	117.12	63.51	Masculin
16	Non	528424	80.50	36.10	Féminin
17	Non	564661	83.60	36.60	Féminin
18	Non	574103	84.09	54.00	Féminin
19	Non	602703	85.88	54.00	Féminin
20	Non	663563	92.17	56.05	Féminin
21	Non	728987	98.71	52.24	Féminin
22	Non	791429	104.46	65.00	Féminin
23	Non	830785	106.86	57.30	Féminin
24	Non	870213	109.05	73.15	Féminin
25	Oui	931875	116.78	65.77	Féminin
26	Non	972325	118.71	71.72	Féminin
27	Oui	1010029	120.18	73.66	Féminin
28	Oui	1063070	123.37	76.18	Féminin
29	Oui	1080674	111.58	67.67	Féminin
30	Non	1069769	109.31	56.85	Féminin

Le tableau2 montre une visualisation numérique des variables de notre jeu de données.

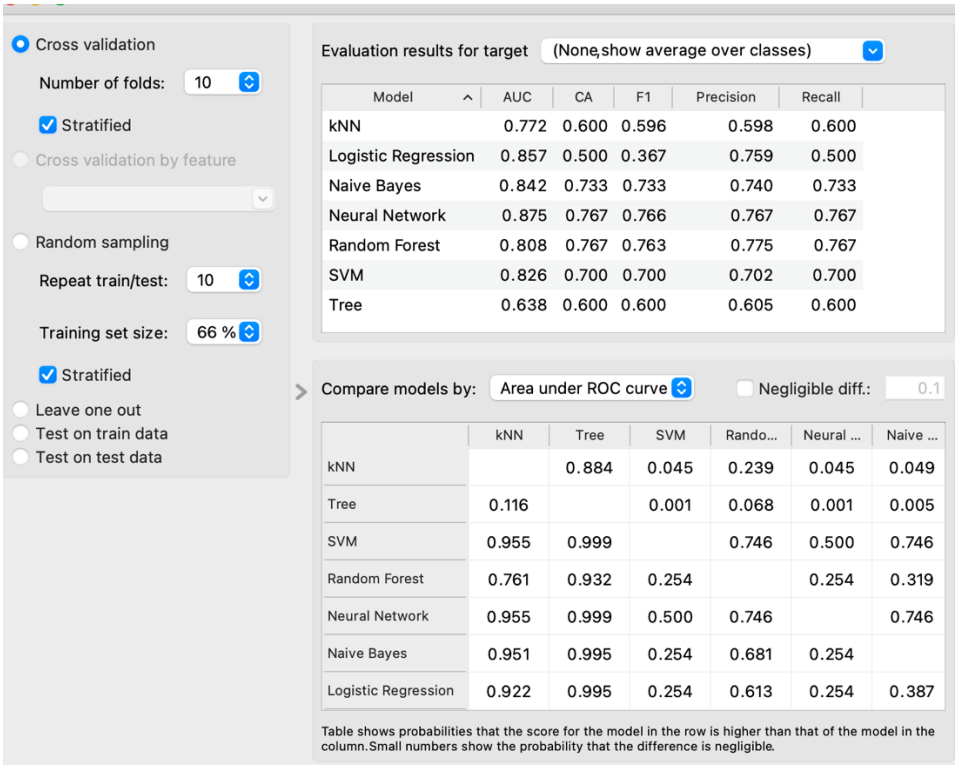
On observe 30 observation, 04 varaiables.

Tableau 4 : Modèles de classification 1



Le tableau 3 montre les modèles de classifications. On note que pour le modèle des KNN, les k plus proches voisins sont 8, la méthode de distance calculé est la distance euclidienne, la distribution uniforme est utilisée.

Tableau 5: Test and Score



Cross validation
Number of folds: 10
Stratified
Cross validation by feature
Random sampling
Repeat train/test: 10
Training set size: 66 %
Stratified
Leave one out
Test on train data
Test on test data

Evaluation results for target (None,show average over classes)

Model	AUC	CA	F1	Precision	Recall
kNN	0.772	0.600	0.596	0.598	0.600
Logistic Regression	0.857	0.500	0.367	0.759	0.500
Naive Bayes	0.842	0.733	0.733	0.740	0.733
Neural Network	0.875	0.767	0.766	0.767	0.767
Random Forest	0.808	0.767	0.763	0.775	0.767
SVM	0.826	0.700	0.700	0.702	0.700
Tree	0.638	0.600	0.600	0.605	0.600

Compare models by: Area under ROC curve Negligible diff.: 0.1

	kNN	Tree	SVM	Rando...	Neural ...	Naive ...
kNN		0.884	0.045	0.239	0.045	0.049
Tree	0.116		0.001	0.068	0.001	0.005
SVM	0.955	0.999		0.746	0.500	0.746
Random Forest	0.761	0.932	0.254		0.254	0.319
Neural Network	0.955	0.999	0.500	0.746		0.746
Naive Bayes	0.951	0.995	0.254	0.681	0.254	
Logistic Regression	0.922	0.995	0.254	0.613	0.254	0.387

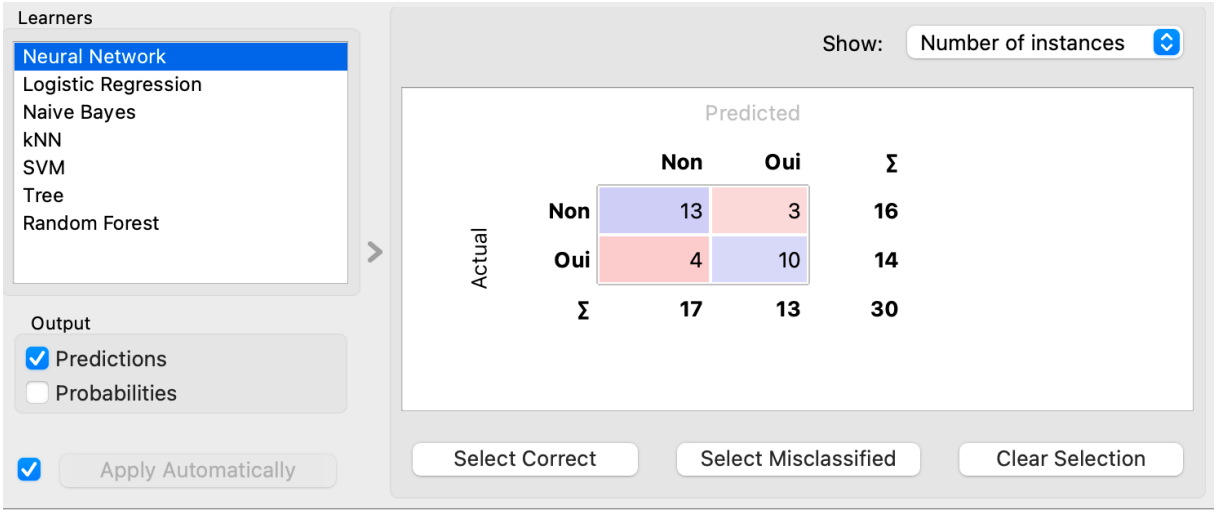
Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

Le tableau 4 montre le résultat des modèles de classification.

De n'analyse de ce graphique il ressort que tous les modèles sont performants à l'exception de modèle Logistic Regression qui n'est pas bon avec un accuracy (0.500) et F1 = 0.367 malgré une précision élevée. On remarque que les modèles les mieux performant sont les modèles Neural Network et Randon Forest qui ont même accuracy et rappel (0.767) et même rappel (Recall, 0.767). Le modèle Random forest est meilleur en précision par contre le modèle Neural Network est meilleur en AUC et en F1.

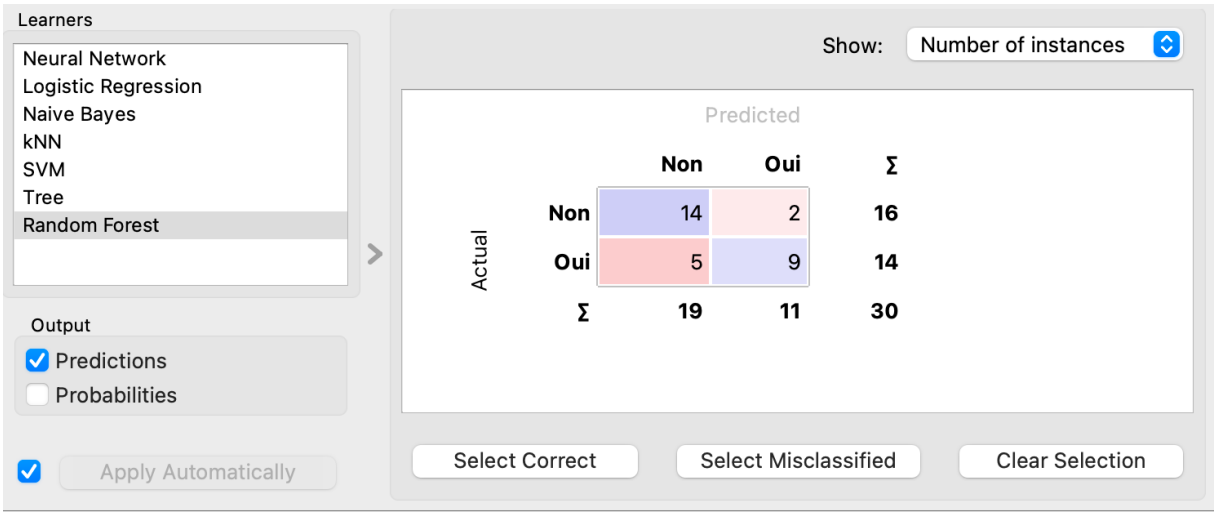
Ainsi le modèle Neural Network sera retenu dans notre étude.

Graphique 1 : Confusion Matrix modèle NNK



Le graphique 1 montre les résultats de la matrix de confusions. De ce graphique on note que le modèle Neural Network parait être le plus performant. Ainsi on remarque que 13 individus dont le taux de réussite inférieur à 70% sont bien classé et 10 individus dont le taux de réussite supérieur à 70% sont bien classé. Par contre 3 individus dont le taux de réussite est inférieur à 70% est classé comme individu dont le taux de réussite est supérieur à 70%, et 4 individus dont le taux de réussite est supérieur à 70% est classé comme individus dont le taux de réussite est inférieur à 70%.

Graphique 2 : Confusion Matrix Random Forest



Le graphique 2 montre les résultats de la matrix de confusions du modèle Random forest. De ce graphique on note que le modèle Random Forest est plutôt bon mais loin d'être infaillible. On remarque que 14 individus dont le taux de réussite inférieur à 70% sont bien classé et 9 individus dont le taux de réussite supérieur à 70% sont bien classé. Par contre 2 individus dont le taux de réussite est inférieur à 70% est classé comme individu dont le taux de réussite est supérieur à 70%, et 5 individus dont le taux de réussite est supérieur à 70% est classé comme individus dont le taux de réussite est inférieur à 70%.

Prédiction

Maintenant que nous avons comparé les modèles de classification et identifier celui qui produit les meilleurs résultats, nous allons pouvoir faire la prédiction.

Tableau 5 : Résumer de la base de prédiction

Tableau 6 : Base de prédiction 1

	Name	Type	Role	Values
1	Effectif des ...	N numeric	feature	
2	Taux brut de ...	N numeric	feature	
3	Taux ...	N numeric	feature	
4	Sexe	C categorical	feature	Féminin,Masculin

Le tableau 5 : présente le résumer de la base de prédiction contenant 03 variables quantitative explicative et une variable catégorielle explicative.

Tableau 7 : Résumer _ base de prédiction 1

	ffectif des élèves	brut de scolarisa	aux d'achèvemen	Sexe
1	1259537	107.10	61.10	Masculin
2	1319648	107.80	61.60	Masculin
3	1318140	105.22	77.50	Masculin
4	1356818	99.59	75.60	Masculin
5	1474206	104.33	75.85	Masculin
6	1601146	109.42	68.37	Masculin
7	1719390	113.46	56.83	Masculin
8	1787940	114.02	70.35	Masculin
9	1869379	115.93	73.15	Masculin
10	1987182	122.45	76.80	Masculin
11	2064031	123.37	81.60	Masculin
12	2133330	123.68	81.28	Masculin
13	2238185	126.16	82.03	Masculin
14	2267835	119.12	74.81	Masculin
15	2246949	117.12	63.51	Masculin
16	528424	80.50	36.10	Féminin
17	564661	83.60	36.60	Féminin
18	574103	84.09	54.00	Féminin
19	602703	85.88	54.00	Féminin
20	663563	92.17	56.05	Féminin
21	728987	98.71	52.24	Féminin
22	791429	104.46	65.00	Féminin
23	830785	106.86	57.30	Féminin
24	870213	109.05	73.15	Féminin
25	931875	116.78	65.77	Féminin
26	972325	118.71	71.72	Féminin
27	1010029	120.18	73.66	Féminin
28	1063070	123.37	76.18	Féminin
29	1080674	111.58	67.67	Féminin
30	1069769	109.31	56.85	Féminin

Le tableau 6 montre que le fichier contient 30 individus pour lesquels les 04 variables caractéristiques ont été spécifiées. Par contre, il ne contient par la variable Taux de succès qui représente si 70% des individus ont achever leurs cursus du primaire ou non.

Tableau 8 : Tableau de prédiction

Predictions

Show probabilities for (None) Restore Original Order

	Tree	Random Forest	SVM	kNN	Naive Bayes	Logistic Regression	Neural Network	Effectif des élèves	Taux brut de scolarisation	Taux d'achèvement	Sexe
2	Non	Non	Non	Non	Non	Oui	Non	1319648	107.80	61.60	Masculin
3	Oui	Non	Oui	Non	Oui	Oui	Oui	1318140	105.22	77.50	Masculin
4	Oui	Oui	Oui	Non	Oui	Oui	Oui	1356818	99.59	75.60	Masculin
5	Oui	Oui	Oui	Non	Oui	Oui	Oui	1474206	104.33	75.85	Masculin
6	Oui	Oui	Oui	Non	Oui	Oui	Oui	1601146	109.42	68.37	Masculin
7	Non	Non	Non	Oui	Non	Oui	Non	1719390	113.46	56.83	Masculin
8	Non	Non	Oui	Oui	Oui	Oui	Oui	1787940	114.02	70.35	Masculin
9	Oui	Oui	Oui	Oui	Oui	Oui	Oui	1869379	115.93	73.15	Masculin
10	Oui	Oui	Oui	Oui	Oui	Oui	Oui	1987182	122.45	76.80	Masculin
11	Oui	Oui	Oui	Oui	Oui	Oui	Oui	2064031	123.37	81.60	Masculin
12	Oui	Oui	Oui	Oui	Oui	Oui	Oui	2133330	123.68	81.28	Masculin
13	Oui	Oui	Oui	Oui	Oui	Oui	Oui	2238185	126.16	82.03	Masculin
14	Oui	Oui	Oui	Oui	Oui	Oui	Oui	2267835	119.12	74.81	Masculin
15	Oui	Oui	Oui	Oui	Oui	Oui	Oui	2246949	117.12	63.51	Masculin
16	Non	Non	Non	Non	Non	Non	Non	528424	80.50	36.10	Féminin
17	Non	Non	Non	Non	Non	Non	Non	564661	83.60	36.60	Féminin
18	Non	Non	Non	Non	Non	Non	Non	574103	84.09	54.00	Féminin
19	Non	Non	Non	Non	Non	Non	Non	602703	85.88	54.00	Féminin
20	Non	Non	Non	Non	Non	Non	Non	663563	92.17	56.05	Féminin
21	Non	Non	Non	Non	Non	Non	Non	728987	98.71	52.24	Féminin
22	Non	Non	Non	Non	Non	Non	Non	791429	104.46	65.00	Féminin
23	Non	Non	Non	Non	Non	Non	Non	830785	106.86	57.30	Féminin
24	Non	Non	Non	Non	Non	Non	Non	870213	109.05	73.15	Féminin
25	Oui	Oui	Non	Non	Non	Non	Oui	931875	116.78	65.77	Féminin
26	Non	Oui	Oui	Non	Oui	Non	Oui	972325	118.71	71.72	Féminin
27	Oui	Oui	Oui	Non	Oui	Non	Oui	1010029	120.18	73.66	Féminin
28	Oui	Oui	Oui	Non	Oui	Non	Oui	1063070	123.37	76.18	Féminin
29	Oui	Non	Non	Non	Non	Non	Oui	1080674	111.58	67.67	Féminin
30	Non	Non	Non	Non	Non	Non	Non	1069769	109.31	56.85	Féminin

Le tableau 7 présente la prédiction du taux de réussite de 70% des individus suivant les modèles de classification.

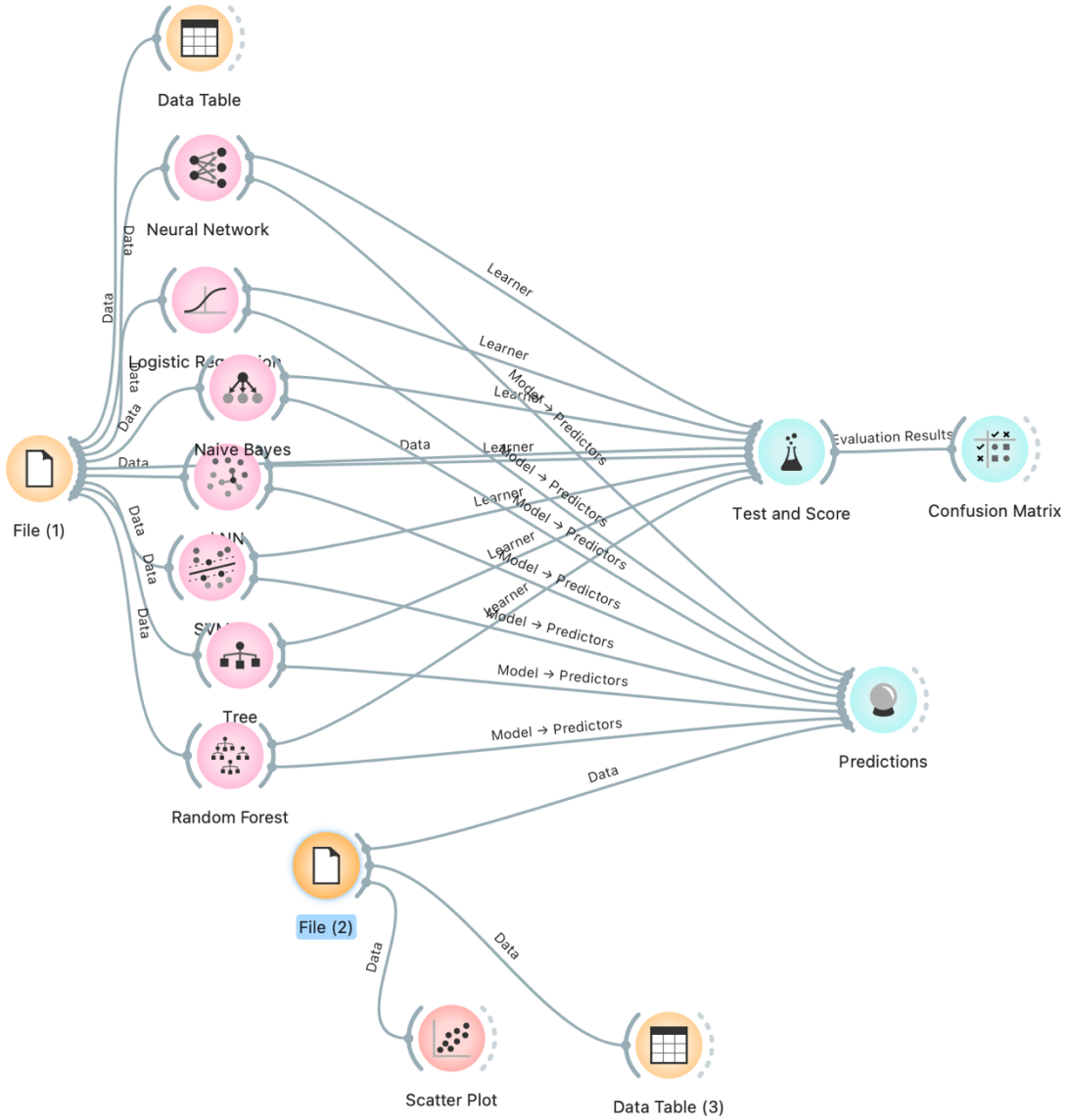
Le modèle Neural Network étant celui qui est retenu dans notre étude, il ressort de ce modèle que la première ligne 1 a été classé parmi les individus dont le taux de réussite est inférieur à 70%(Non) ce qui parait correcte, par contre les lignes 2,3 devrait être classé parmi les individus dont le taux de réussite est inférieur à 70% (Non). Ces erreurs montrent qu'un modèle de classification n'est sûr à 100% et qu'une vérification manuelle reste toujours nécessaire.

Conclusion

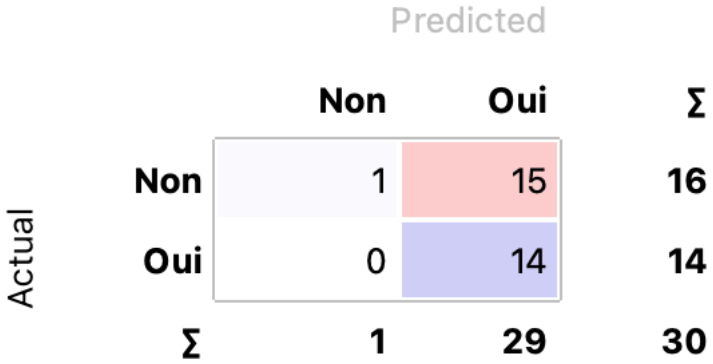
Tout au long de ce document, l'analyse a été centré, sur la classification supervisée de la scolarisation des jeunes filles et garçons au bénin. La présente étude a eu pour objectif d'analyser le taux de réussite d'achèvement des écoliers est supérieur ou non à 70%. Elle s'est prioritairement intéressée à partir des modèles d'apprentissage au validation croisé et aux matrices de confusion à une prédiction du taux de réussite des élèves. Pour cela nous avons utilisé après validation du modèle Neural Network à la prédiction à l'aide du logiciel Orage. Des différents résultats, on retient que 13 individus dont le taux de réussite est inférieur à 70% ont été bien prédit et 10 individus dont le taux de réussite est supérieur à 70% ont été bien prédite aussi. Par contre 3 individus ont été classé parmi ceux dont le taux de réussite est supérieur à 70% or leur taux de réussite est inférieur à 70%, de même pour 04 individus dont le taux de réussite supérieur à 70% ont été classé par ceux dont le taux de réussite est inférieur à 70%.

Annexe :

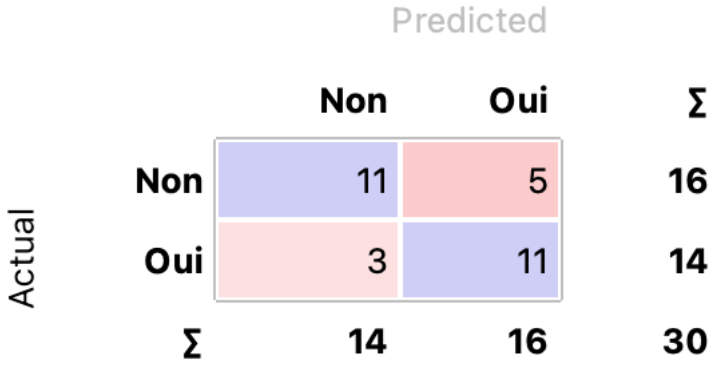
1-Résumé de la chaine mise en place



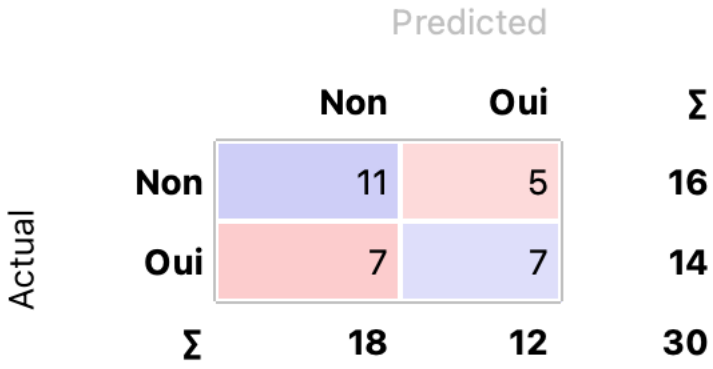
2-Matrix de confusion Logistic Regression



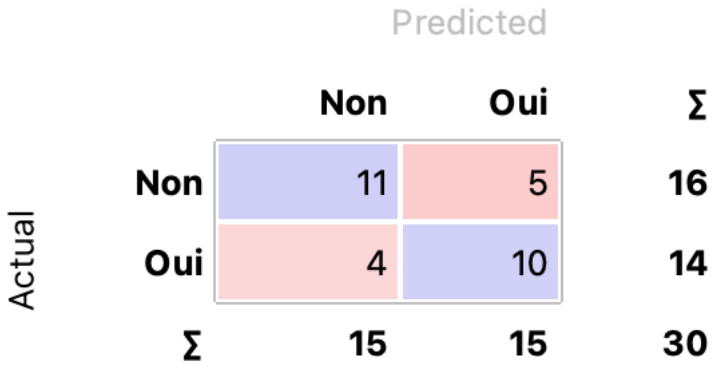
3- Matrix de confusion Naive Bayes



4- Matrix de confusion KNN



5- Matrix de confusion SVN



6- Matrice de confusion Tree