Université de Lille

Projet de datascience

2019-2020

Régression Logistique

Realisé par :
Nasradine IBRAHIM
Pierre KREPA

Encadré par :
Pr. Jocelyne LEGRAIN

21 janvier 2020

TABLE DES MATIÈRES

Remerciement

Tout d'abord, ce travail est le fruit d'une collaboration avec *Nasradine et Pierre* tous les deux étudiants parcours MIASHS de l'UFR MIME à l'Université de Lille 3 .Je remercie énormément Madame ***LEGRAIN*** son soutien qu'il nous a donné et également pour sa gentillesse .

CHAPITRE 1

INTRODUCTION

Les systèmes d'informations se sont beaucoup développés pour assurer la liaison entre les décisions et l'apprentissage machine .Cette dernière est fait pour la prise des décisions de manière automatique et a permis à des entreprises de prendre des choix stratégiquement forts .

Cependant,des points importants se sont poses suite à des erreurs dues au biais des machines dans la société et pour savoir notamment si les décisions que prendraient les machines sont équitables ,impartiales vis à vis des individus ,des groupes lorsque la variable dépendante s'agit d'une variable binaire et qu'on essaye de classer des individus selon leurs comportements : d'où l'utilité de l'un des algorithmes de la classification supervisée ***La régression logistique*** .

Il est important de se demander les questions suivantes :

-C'est quoi une classification supervisée ?
-Que s'agit-il la régression logistique ?
-En quoi la courbe ROC est-elle si important ?
Nous allons répondre étape par étape à ces questions importantes .

CHAPITRE 2

APPRENTISSAGE STATISTIQUE

2.1 Apprentissage Supervisée

La classification supervisée est une méthode d'apprentissage statistique ou l'on cherche à produire des règles automatiques à partir d'une bases de données d'apprentissage contenant des classes .

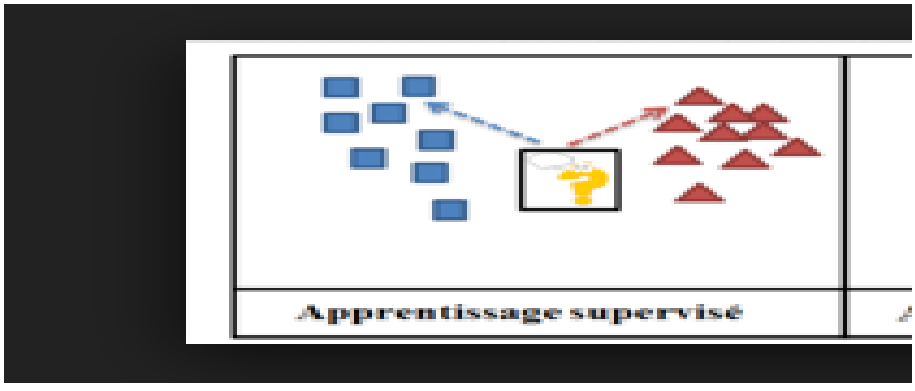

FIGURE 2.1 – Classification supervisée

L'objectif principal de cette technique consiste à trouver des règles permettant de classer des objets dans des ensembles à partir des variables continues ou catégoriques caractérisant ces objets.

Par exemple , dans la figure ci-dessus ,on cherche à classer si cet objet appartient au classe bleue ou rouge . Dans notre cas , nous nous trouvons dans la même situation c'est à dire on essaye de savoir si par exemple l'individu 'a accorder un prêt bancaire est de sexe féminin ou masculin . On considère la variable $\hat{Y}$ la décision d'accorder un prêt bancaire et le sexe va ˆetre considère comme la variable qualitative X = {"H" , "F"} contenant deux classes (homme et femme) . Plusieurs méthodes sont utilisées pour le problème de classification supervisée ,les algorithmes de K-NN(k-plus proches voisins) ,SVM(support vector machine),arbre de décision ,foret aléatoire etc ....

2.2 Définition de la Matrice de Confusion

La matrice de confusion est un bon outil utilisé dans l'apprentissage supervisée servant à donner des mesures sur la qualité du classfication .

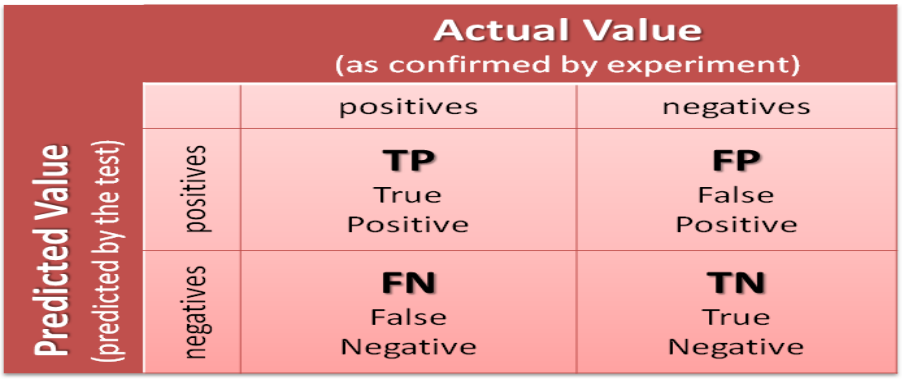

FIGURE 2.2 – Matrice de confusion

Le but essentiel de l'utilisation de cette matrice est qu'elle montre de manière efficace la précision de notre système de classification .

Comme illustré sur cette figure ci-dessus c'est croisant les valeurs prédites et les valeurs réelles par les modèles d'apprentissage statistique utilisé (SVM ,ou arbre de décision , etc...) qu'on met en place la matrice de confusion .
-TP(vrai positif) indique la vraie valeur attendue et bien prédite par le système .
-TN (vrai négatif) est la fausse valeur attendue et bien prédite par le système .
-FN(faux négatif) explique la vraie valeur attendue alors que le système n'a pas prédit .
-FP(faux positif) nous donne la valeur prédite par le système alors que l'on s'attendait une fausse valeur .

Pour expliquer , la table de confusion ,un exemple fera l'affaire. Imaginons si un individu fait une de prêt bancaire auprès d'une banque .La matrice de confusion affichée par le système sera composée des éléments suivants
-TP indique la décision d'accorder le prêt et bien prédite par le système .
-TN est la décision de ne pas accorder le prˆet et bien prédite par le système .
-FN explique la décision d'accorder le prêt alors que le système n'a pas prédit .
-FPnous donne la valeur prédite par le système alors qu'en réalité on n'aurait pas accordé le prêt .

2.3 Regression logistique

La régression logistique est un type d'algorithme de classification supervisée puisqu'on connaît la variable cible .Elle a été utilisée dans les sciences biologiques au début du XXe siècle. Il a ensuite été utilisé dans de nombreuses applications des sciences sociales. Dans un contexte ou elle a eté inventée ,la régression logistique est utilisée lorsque la variable dépendante (cible) est catégorielle.

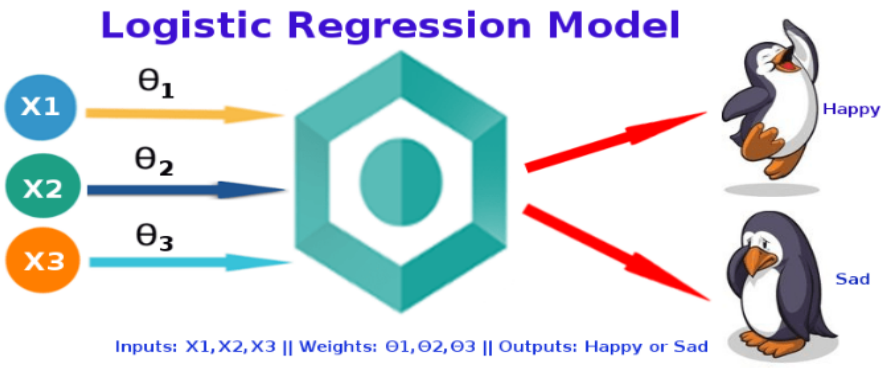

FIGURE 2.3 – Régression logistique

Par exemple,pour prédire si un e-mail est du spam (1) ou (0) ou bien par exemple si la tumeur soit maligne (1) ou non (0) selon les comportements des individus etc...

CHAPITRE 3

ANALYSE EXLORATOIRE

3.1 Source de données

Le projet de régression logistique consiste à réaliser un modèle de régression logistique à partir d'une table de données.
La table de données provient du site Kaggle et contient des informations de santé concernant des individus.
Ici, le projet consiste à réaliser un modèle de régression logistique censé déterminer si un individu quelconque est diabétique ou non.

Notre table de données contient 768 individus et 9 variables :
- Pregnancies : le nombre de grossesses
- Glucose : le taux de glucose
- BloodPressure : le niveau de pression sanguine
- SkinThickness : l'épaisseur de la peau
- Insulin : le taux d'insuline
- BMI : l'indice de masse corporelle
- DiabetesPedigreeFunction

- Age
- Outcome : vaut 1 si l'individu est déclaré diabétique, 0 sinon

L'ensemble de ces variables sont des variables quantitatives, à l'exception de la variable "Outcome" qui est une variable categorielle(binaire) mais aussi la variable cible.

Avant de procéder aux analyses statistiques sur notre table, nous nous sommes interrogés sur la présence ou non de valeurs aberrantes voire manquantes.
Lors de la lecture de la table, nous avons constaté la présence de plusieurs zéros qui semblent être des valeurs aberrantes dans le cas de ces variables :
- Glucose => absence de glucose ?
- BloodPressure => pas de pression sanguine chez l'individu ?
- SkinThickness => absence de peau ?
- BMI => l'individu pèserait 0 kilogrammes ?

La première idée à laquelle nous avons pensé était de supprimer les individus concernées par ces valeurs, ce qui serait revenu à supprimer environ un quart des données de la table.
Comme cela revient à supprimer un nombre trop important de données, et donc d'informations, nous avons choisi d'émettre l'hypothèse qu'il s'agit de valeurs manquantes.
Nous avons finalement choisi de conserver ces valeurs manquantes car nous avons pensé que les informations en question n'ont pas pu être obtenues lors de l'examen de l'individu.

Comme précisé précédemment, nous allons procéder à réaliser deux analyses statistiques :
- Une analyse univariée
- Une analyse bivariée

Mais avant de procéder à l'analyse univariée, nous vous présentons un tableau de statistiques descriptives de l'ensemble des variables (à l'exception de la variable "Outcome" car il s'agit pour rappel d'une variable catégorielle). Pour

cela, on va construire deux variables X et Y qui contiennent respectivement les variables explicatives et la variable cible

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.0
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000

FIGURE 3.1 – Tableau descriptive de données

D'après ce tableau, on peut constater par exemple que l'âge moyen des individus est de 33 ans.

3.2 Analyse univariée

L'analyse univariée consiste à analyser les données d'une seule variable.

Cependant, il faut savoir à quel type de données nous avons à faire avant de faire cette analyse.

On peut representer la boite à moustache des variables quantitatives pour savoir la distribution de chacunes de variables .

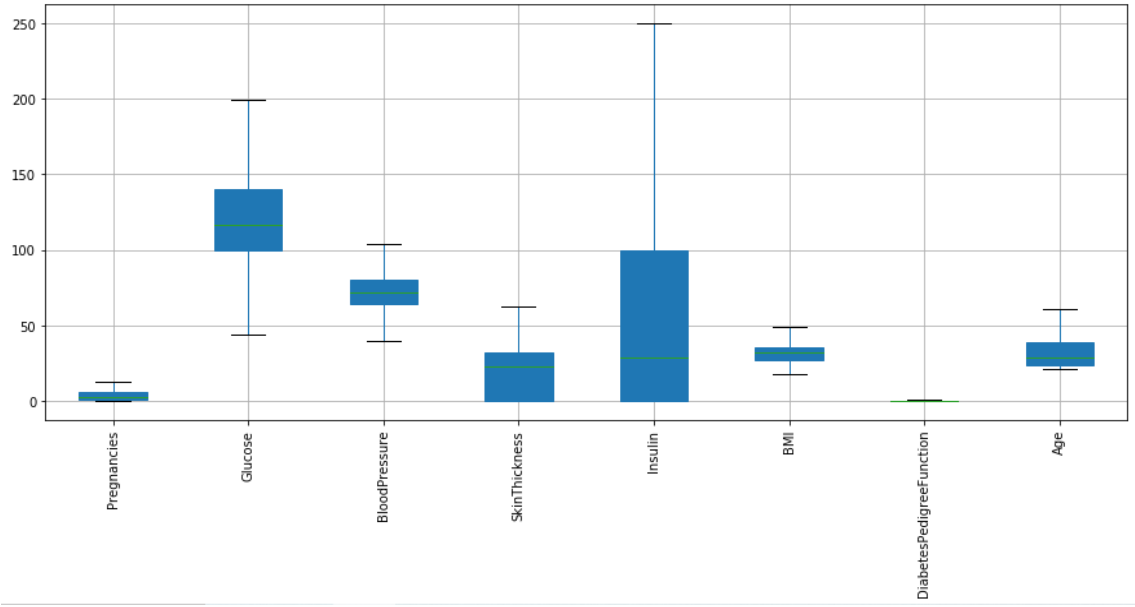

FIGURE 3.2 – La boite à moustache des variables

On voit bien que la majorité des variables ont des distributions normales et qu'elle nous facilite à repondre l'une des contraintes pour faire une analyse de l'ANOVA sur les variances pour la suite de l'analyse bivariée.

Pour l'ensemble des variables, l'analyse est déjà faite à l'exception des variables Pregnancies, Age et Outcome. En effet, ces trois variables ont la particularité de présenter en soi les nombres entiers : on peut donc y faire pour chacune d'entre elles un tableau recensant les occurrences de leurs valeurs présentes (voir annexes) puis les illustrer dans les graphiques ci-dessous.

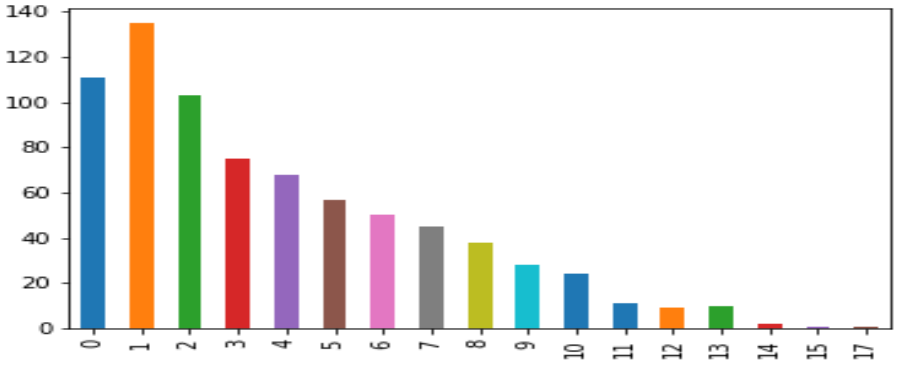

FIGURE 3.3 – Variable Pregnancies

Sans trop de surprise, on constate que, globalement, plus le nombre de grossesses est élevé, moins il y a de femmes concernées.

Ici, on a 14,45% des individus qui n'ont connu aucune grossesse, 17,57% qui ont connu une grossesse et 13,41% qui ont connu deux grossesses. Ce qui fait qu'on a plus de 45% des femmes qui ont connu deux grossesses maximum. Par conséquent, plus de la moitié des femmes a connu au moins trois grossesses et cela monte à dix-sept grossesses.

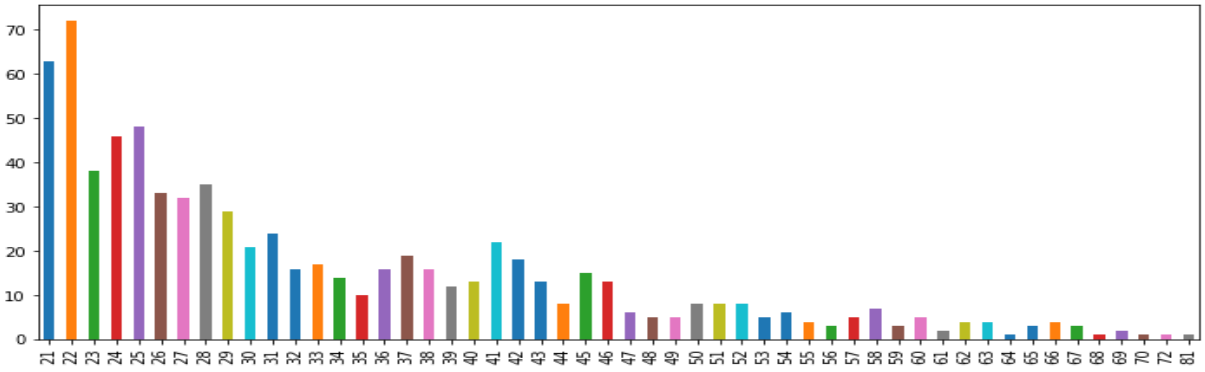

FIGURE 3.4 – Variable Age

Comme pour le cas des grossesses, on constate dans l'ensemble que plus l'âge

est élevé, plus le nombre d'individus diminue. D'après ce graphique, les individus les plus nombreux sont âgés de 21 à 29 et chacun de ces âges représentent au moins 5% de l'échantillon et cela monte à plus de 9% pour les individus âgés de 22 ans.

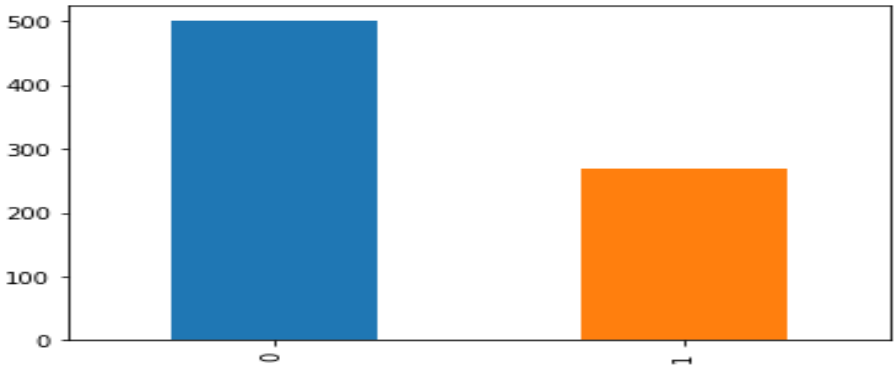

FIGURE 3.5 – Variable Outcome

D'après le tableau, on a 500 individus déclarés non diabétiques, ce qui représente près de deux tiers des données de l'échantillon, contre 268 diabétiques.

3.3 Analyse Bivariée

Maintenant que nous avons fait l'analyse univariée, on peut procéder à l'analyse bivariée.

Pour cela, nous avons construit une matrice de corrélations qui renseigne les coefficients de corrélation entre chaque variable. Ici, nous avons d'abord fait une matrice "classique" qui recense les coefficients de corrélation, puis nous en avons réalisé une autre version avec un code couleur afin de mieux mettre en avant les niveaux de corrélation.

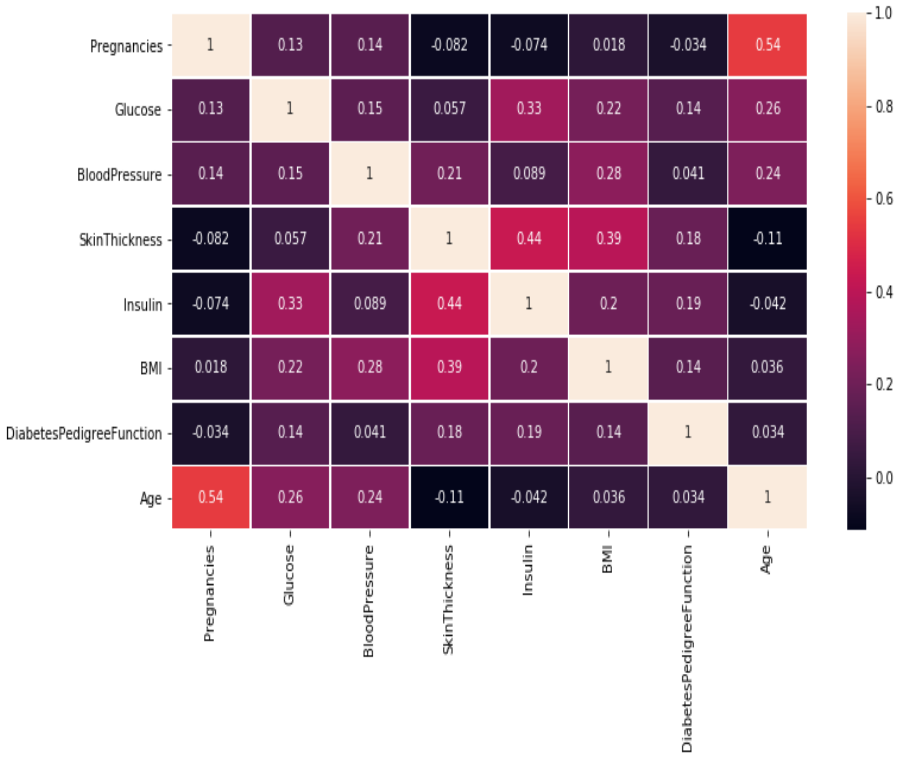

FIGURE 3.6 – Matrice de corrélations des variables

Dans l'ensemble, les coefficients de corrélation entre variables explicatives sont proches de 0, ce qui signifie qu'il n'y a pas vraiment de corrélation entre elles. Le coefficient de corrélation le plus élevé est égal à 0.56 et on l'obtient pour les variables Age et Pregnancies. On obtient aussi un coefficient de 0.42 entre Insulin et SkinThickNess et 0.39 entre SkinThickNess et BMI.

	df	sum_sq	mean_sq	F	PR(>F)
Pregnancies	1.0	8.591143	8.591143	53.638189	6.164705e-13
Glucose	1.0	34.020758	34.020758	212.406175	1.327103e-42
BloodPressure	1.0	0.123476	0.123476	0.770911	3.802132e-01
SkinThickness	1.0	0.863789	0.863789	5.393003	2.048132e-02
Insulin	1.0	0.255349	0.255349	1.594251	2.071077e-01
BMI	1.0	6.780158	6.780158	42.331432	1.398272e-10
DiabetesPedigreeFunction	1.0	1.817752	1.817752	11.349002	7.929848e-04
Age	1.0	0.458924	0.458924	2.865257	9.092163e-02
Residual	759.0	121.567819	0.160168	NaN	NaN

FIGURE 3.7 – Test ANOVA à plusieurs facteurs

Après la matrice de corrélations, on décide de faire une analyse de la variance ANOVA afin de savoir quelle(s) variable(s) explicative(s) aurait une influence sur le diagnostic du patient.

D'après l'analyse de la variance (ANOVA) entre la variable dépendante et les variables explicatives, on observe que la variable cible semble dépendante de certaines variables explicatives car leur p-value est inférieure à 5%. Il s'agit des variables Pregnancies, Glucose, SkinThickNess, BMI et DiabetesPedigreeFunction.

Les variables pour lesquels il y aurait indépendance sont BloodPressure (38%), Insulin (20,7%) et Age (9,1%).

Le résultat sur l'insuline est le plus surprenant puisque d'une part, l'insuline est utilisée pour le traitement du diabète, et d'autre part, il y aurait indépendance entre le taux d'insuline et le diagnostic de l'individu d'après l'ANOVA.

Dans notre table, on constate que plusieurs individus présentent un taux

d'insuline nul, qu'ils soient diagnostiqués diabétiques ou non. L'indépendance s'expliquerait par le fait que des individus diabétiques ne se sont pas forcément injecté d'insuline lors des analyses.

Ces résultats peuvent s'expliquer par le fait qu'un individu non diabétique n'a pas besoin de s'injecter d'insuline. Quant aux diabétiques, ces derniers peuvent prendre des médicaments au lieu de prendre de l'insuline.

CHAPITRE 4

MODELISATION

La première chose à faire est de diviser notre échantillon en deux : un ensemble d'entraînement pour 80% de nos données et un ensemble test pour 20% de nos données .

A la fin de la modélisation, on va essayer de voir si les valeurs prédites par notre modèle se rapprochent des vraies valeurs de l'ensemble de test.

4.1 Matrice de confusion

La première façon de regarder si le modèle a marché est la matrice de confusion.

Dans l'ensemble d'entraînement par exemple, 125 individus sur 154 ont bien été classés (86 vrais positifs et 29 vrais négatifs)

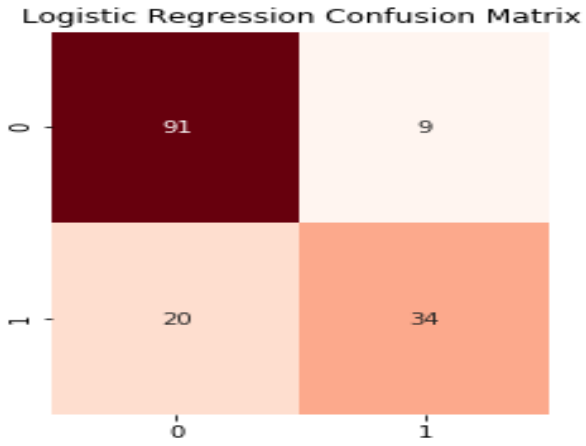

FIGURE 4.1 – Matrice de confusion de la regression logistique

4.2 Interprétation des coefficients et Odds Ratio

La régression logistique nous donne des paramètres ou coefficients qui permettent d'affirmer à quel point les variables explicatives influencent la variable cible .
Par définition ,le rapport de côte est une mesure statistique qui indique par exemple le rapport de côte d'un individu jeune d'avoir le diabète avec celle d'un individu plus âgé .
Un rapport des côtes de 1 correspond à l'absence d'effet. En cas d'effet bénéfique, le rapport des cotes est inférieur à 1 et il est supérieur à 1 en cas d'effet délétère.

Plus le rapport des cotes est éloigné de 1, plus l'effet est important. Il est calculé par **Odds =Exp(coefficient)** Dans le tableau ci-dessous regroupe les coefficients de variables explicatives et leurs rapports de cotes.

	Coefficient	Rapport de cote
Pregnancies	0.095385	1.100082
Glucose	0.023287	1.023561
BloodPressure	-0.014128	0.985971
SkinThickness	-0.002634	0.997370
Insulin	-0.000599	0.999401
BMI	0.058262	1.059992
DiabetesPedigreeFunction	0.705564	2.024989
Age	0.013249	1.013337

FIGURE 4.2 – Coefficients et Odds Ratio des variables explicatives

On observe que les variables Age,Glucose et BMI,Pregnancies ont des **coefficients positifs** .

On peut dire les individus jeunes ont tendance à ne pas etre detecté diabetique . Plus l'IMC de l'individu est grand ,moins il a de chance d'etre detecté diabetique La quantité de glucose dans le sang augmente avec le fait d'etre detecté diabetique. Le nombre de fois que le patient soit enceinte augmente la detection de la maladie de diabete

Par contre les variables Insulin,BloodPressure dimunient avec la detection d'un individu d'etre diabetique car leurs **coefficients sont negatifs.**

Gardez à l'esprit que notre variable dépendante, ce que nous essayons de classer, est de savoir si un individu souffrira ou non de diabète. Voici quelques exemples de la façon dont vous pouvez interpréter les rapports de cotes ci-dessus :

Âge : Pour chaque année supplémentaire, les chances de souffrir de diabète sont 1,007 fois plus importantes.

IMC : Pour chaque augmentation de 1 de l'IMC, les chances d'avoir le diabète sont 1,11 fois plus importantes.

4.3 Le score et la probabilité de risque

Par définition , le score est une note attribuée a individu qui mesure le risque ou la probabilité de sinistre de cet individu. Il ne s'agit pas seulement de classer les clients potentiels en deux catégories, les bons et les mauvais assurés, mais de produire un indicateur quantitatif du risque individuel que présente chaque client.

Dans la plupart de temps ,on préfère plus la probabilité de risque que le score attribuée à un assuré.

	Probabilité de risque	Score	Diabetique
103	0.012736	1.020648	1
27	0.018629	-1.901013	0
84	0.052233	0.172571	1
148	0.054382	-1.546612	0
79	0.064831	-0.659328	0
69	0.073332	-2.115929	0
142	0.078440	-0.851255	0
80	0.080525	-1.595072	0
74	0.092355	-0.427948	1

FIGURE 4.3 – Score et probabilité de risque

Pour essayer de bien savoir le degré de risque d'etre diabetique ,On attribue chaque patient à un score et un probabilité de risque.
Le score ou la probabilité permet de classer les patients du moins risqué d'avoir le diabète au plus risqué.
On voit bien que le 103eme patient presente la probabilité la plus faible c'est à dire qu'il est le plus risqué d'avoir le diabète .

Pour essayer de bien observer la visualisation de deux classes d'etudes par

leurs distributions de probabilités,on a commencé par représenter sur le graphique ci-dessous . On peut ainsi constater l'efficacité ou la précision de notre modèle construit à partir de la régression logistique .

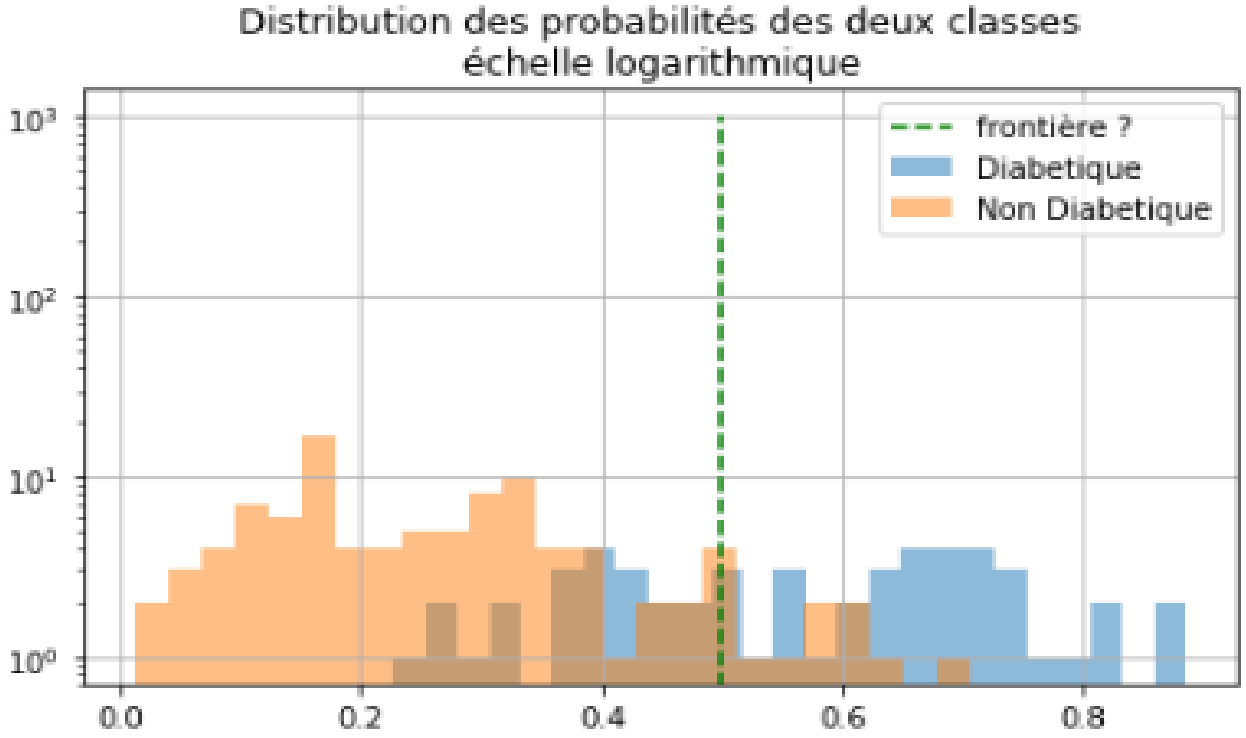

FIGURE 4.4 – Visualisation de deus classes

On remarque la distribution des probabilités de deux classes (Diabétique et Non Diabétique) est cette fois-ci bien séparée par la frontière de couleur verte .L'histogramme colorié rouge représente les individus non diabétique et l'histogramme bleue représente les individus diabétiques.

4.4 La courbe ROC

La courbe ROC est une mesure de performance pour un problème de classification à différents paramètres de seuils. ROC est une courbe de probabilité et l'AUC représente le degré ou la mesure de la séparabilité. Il indique dans quelle mesure le modèle est capable de distinguer les classes. Plus l'AUC est élevée, mieux le modèle est de prédire 0 comme 0 et 1 comme 1. Par analogie, plus l'AUC est élevée, mieux le modèle consiste à faire la distinction entre les patients atteints de maladie de diabète et ceux qui n'en ont pas. La courbe ROC est tracée avec TPR contre le FPR où TPR est sur l'axe des y et FPR est sur l'axe des x. Dans notre cas ,l'aire sous la courbe ou AUC est égale à 0.77.

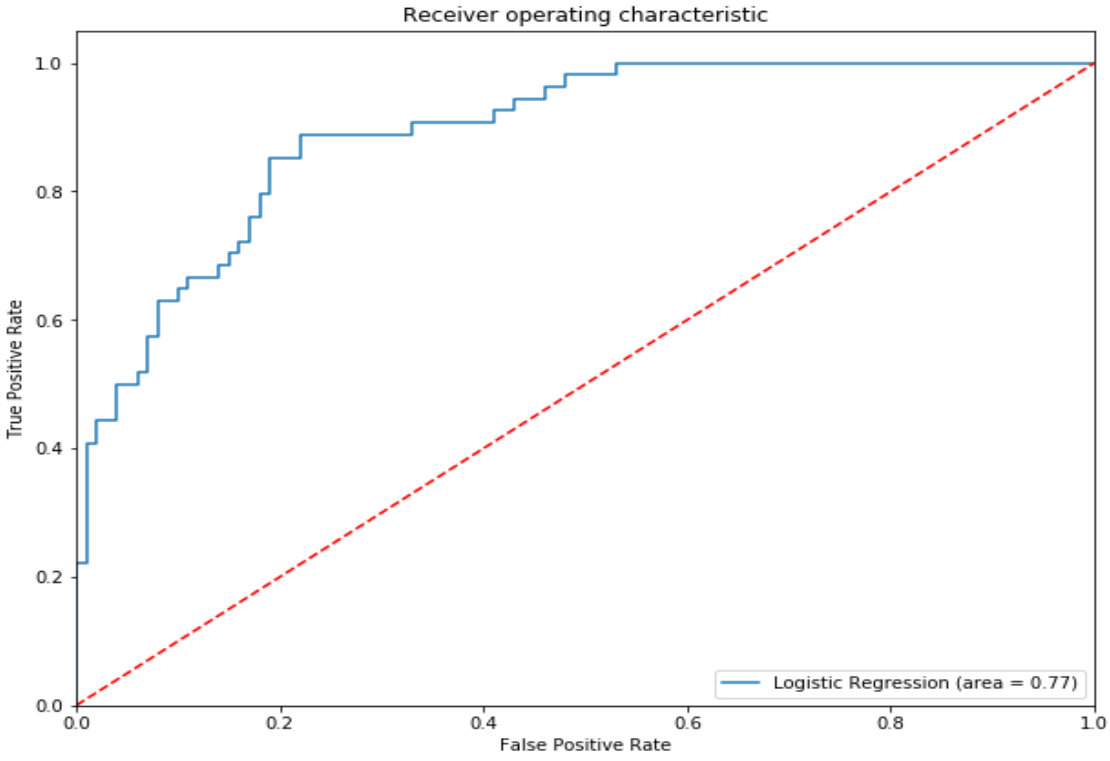

FIGURE 4.5 – La courbe ROC du modèle