**PRÉSENTATION DES RÉSULTATS D’ANALYSE**

1. **PRÉSENTATION DES DONNÉES**

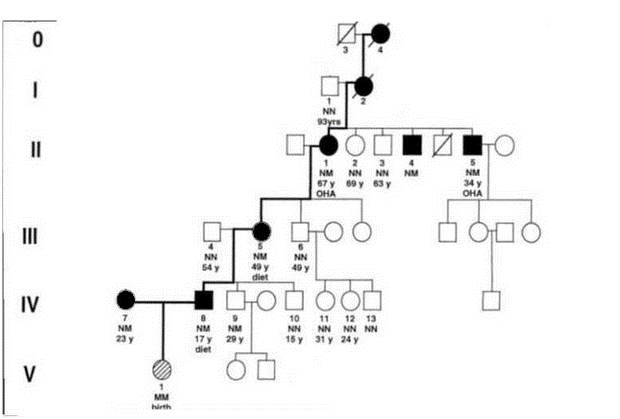
Nous disposons d’une base de données provenant de l'Institut national du diabète et des maladies digestives et rénales.

Dans le cadre de cette analyse, l’étude a été porté sur 768 individus de sexe féminin.

Sur chacun des individus, 8 mesures ont été prélevées. A savoir :

* Pregnancies (Grossesses) : le nombre de grossesse
* Glucose : Concentration plasmatique de glucose à 2 heures dans un test oral de tolérance au glucose.
* BloodPressure (pression artérielle) : Pression artérielle diastolique (mm Hg)
* SkinThickness (Épaisseur de peau) : Épaisseur du pli cutané du triceps (mm)
* Insulin (insuline) : Insuline sérique 2 heures (mu U/ml)
* BMI ou Body mass index (IMC ou indice de masse corporelle) : c’est le rapport entre le poids (en Kg) et le carré de la taille (en m).
* DiabetesPedigreeFunction (Fonction généalogique du diabète) : Ici on retrace l’arbre généalogique du diabète dans la famille.

Comme illustration, nous avons :

Figure 1: Évolution du diabète sur plusieurs générations

On essaie aussi de voir les gènes en cause. (source : <http://acces.ens-lyon.fr/acces/thematiques/evolution/dossiers-thematiques/epigenetique/anomalies-genetiques-et-epigenetiques-dans-le-diabete-mody-4-et-le-diabete-de-type-2>).

* Age : l’âge est pris en année révolue.

1. **APPROCHE DESCRIPTIVE DE LA MALADIE DE DIABÈTE**

Il est difficile d’avancer dans l’analyse sans pour autant donner une description claire sur la population sous étude. Dans cette section, sans entrer en détail ou sans énumérer les formules théoriques, les résultats descriptifs seront obtenus à l’aide des sorties du logiciel R > 4, SPSS-26, PYTHON > 3.9, etc.

En ce qui concerne la statistique descriptive, nous allons décrire notre phénomène sur le plan univarié et bivarié. Partant de l’approche univarié, nous allons ressortir la structure de chaque variable (Fréquence, moyenne, écart-type, médiane, etc.), dans l’approche bivarié, nous essayerons de ressortir les liaisons entre différentes variables sous étude.

* 1. **Structure de la population selon la présence ou l’absence de la maladie**

Tableau 1 : Répartition de la population selon la maladie

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Fréquence** | **Pourcentage** | **Pourcentage cumulé** |
| **absence** | 500 | 65.1 | 65.1 |
| **présence** | 268 | 34.9 | 100.0 |
| **Total** | 768 | 100.0 |  |

*Source : Auteur, par l’intermédiaire du logiciel SPSS-26*

Aux vus des résultats fournis par le tableau 1, nous remarquons que 500 patients sur 768, soit près de 65% de la population sous étude n’est pas atteint du diabète, contre 35% de la population qui est atteint de la maladie du diabète.

Cette information nous sera très importante lors de l’élaboration du modèle, car cela cause généralement le problème de *l’imbalence*[[1]](#footnote-1).

* 1. **Structure de la population selon l’âge**

Tableau 2 : description de la variable Age

|  |  |
| --- | --- |
|  | **Age** |
| **N** | 768 |
| **Moyenne** | 33.24 |
| **Mode** | 22.00 |
| **Médiane** | 29.00 |
| **std** | 11.76 |
| **Min** | 21.00 |
| **N si âge < 25** | 219 |
| **N Si 25 ≤ âge <65** | 533 |
| **N Si âge ≥ 65** | 16 |
| **Max** | 81.00 |
| **Q1** | 23.90 |
| **Q2** | 29.10 |
| **Q3** | 40.49 |

*Source : Auteur, par l’intermédiaire du langage de programmation Python 3.9.6*

Les résultats du tableau 2 montrent que dans cette population :

* L’âge des individus est en moyenne de 33 ans
* L’âge le plus fréquent est de 22 ans
* La plus jeune des femmes a 21 ans
* La plus vielle a 81 ans
* Près de 219 femmes sur 768, soit 28% des femmes ont moins de 25 ans
* Près de 533 femmes sur 768, soit 69% des femmes ont un âge compris entre 25 et 64 ans

Nous remarquerons sans doute que cette population est constituée des jeunes, car près de 75% des femmes dans cette population ont moins de 41 ans. En plus, les adultes sont majoritaires.

* 1. **Structure de la population selon le nombre des grossesses**

Tableau 3 : Structure de la population par rapport au nombre au nombre des grossesses

|  |  |  |  |
| --- | --- | --- | --- |
| **Pregnancies** | **Fréquence** | **Pourcentage** | **Pourcentage cumulé** |
| **0** | 111 | 14.5 | 14.5 |
| **1** | 135 | 17.6 | 32.0 |
| **2** | 103 | 13.4 | 45.4 |
| **3** | 75 | 9.8 | 55.2 |
| **4** | 68 | 8.9 | 64.1 |
| **5** | 57 | 7.4 | 71.5 |
| **6** | 50 | 6.5 | 78.0 |
| **7** | 45 | 5.9 | 83.9 |
| **8** | 38 | 4.9 | 88.8 |
| **9** | 28 | 3.6 | 92.4 |
| **10** | 24 | 3.1 | 95.6 |
| **11** | 11 | 1.4 | 97.0 |
| **12** | 9 | 1.2 | 98.2 |
| **13** | 10 | 1.3 | 99.5 |
| **14** | 2 | 0.3 | 99.7 |
| **15** | 1 | 0.1 | 99.9 |
| **17** | 1 | 0.1 | **100.0** |
| **Total** | **768** | **100.0** |  |

*Source : Auteur, par l’intermédiaire SPSS-26*

Les résultats de l’analyse montrent que dans cette population :

* Près de 45% des femmes ont porté aux plus deux grossesses
* La majorité des femmes n’ont porté qu’une grossesse lors de l’étude
* Chose étonnante, il y a 14 femmes dans cette population qui ont eu à porter au moins 14 grossesses.

Les informations ci-dessous pourrait mettre certaines personnes en doute, et la question suivante peut-être posée :

*Comment expliquer que dans une population dont 69% des femmes ont un âge compris entre 25 et 69 ans, il n’y a que près des 45% des femmes ayant porté aux plus deux grossesses ?*

La réponse à cette sera donnée dans la suite des analyses.

* 1. **Structure de la population par rapport à L’IMC**

L’information par rapport à l’IMC (indice de masse corporelle) nous sera cruciale, car non seulement nous cherchons à savoir si notre population est diabétique ou pas, mais en plus de cela, nous aurons une information complémentaire en rapport avec l’état nutritionnel des femmes sous étude.

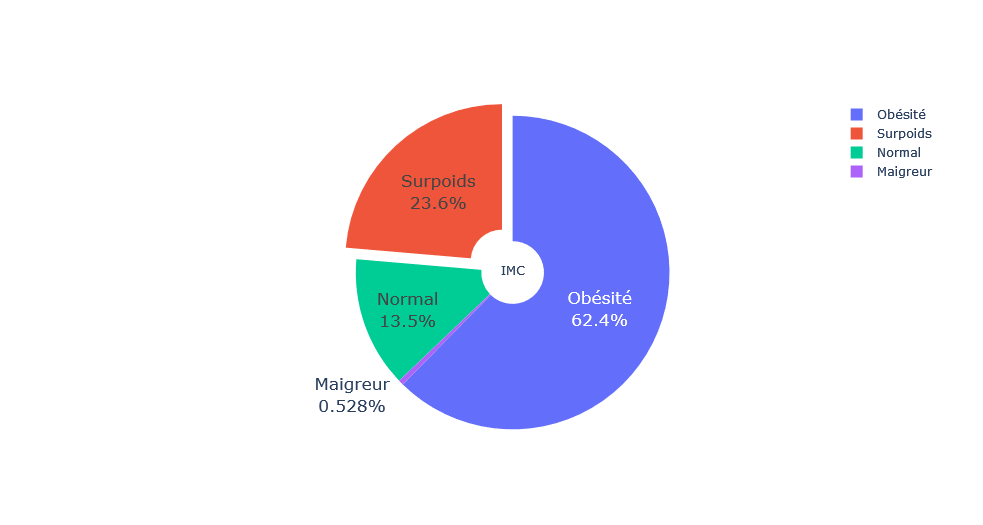
Rappelons que selon l’OMS (organisation mondiale de la santé), la classification en rapport à l’IMC peut être résumer par rapport aux types de grades (sévérité). Mais cela peut être résumé en :

Tableau 4 : Classification de l'IMC selon l'OMS

|  |  |
| --- | --- |
| **IMC** | **Classiffication** |
| < 18,5 | **Maigreur** |
| 18,5 - 24,9 | **Normal** |
| 25 - 29,9 | **Surpoids** |
| ≥ 30 | **Obésité** |

*Source : Organisation mondiale de la santé*

En ce qui nous concerne, nous avons :

Graphique 1 : Indice de masse corporelle de la population

*Source : Auteur, par l’intermédiaire du langage de programmation Python 3.9.6*

La majorité des femmes (soit 62.4%) de cette population sont obèses, 23.6% des femmes sont en état de surpoids, 0.53% sont dans un état de maigreur et 13,5% des femmes sont en état normal.

En effet, cette information nous est très crucial et dont nous ne devons pas prendre à la légère lors de la modélisation.

Ceci pourrait être la réponse à la question que nous nous sommes posés au point 2.3.

1. **CLASSIFICATION NAIVE BAYÉSIENNE**

**3.1 Introduction**

Pour dire qu’une personne est diabétique (diabète de type II) ou non, nous nous servirons de la classification de Naïve Bayes, car ce classifieur a eu à faire ses preuves en ce qui concerne le problème de classification.

En plus, le classifieur naïve Bayes se base sur le théorème de Bayes. Ce dernier est un classique de la théorie des probabilités, et il est fondé sur les probabilités conditionnelles.

Cette classification pose une forte condition majeure se basant sur l’indépendance des variables exogènes avant son application, car si cette condition ne pas validée, nous n’aurons pas des bonnes performances lors de la classification.

Dans le cadre de cette analyse, nous utiliserons la méthode de 80/20 proposée par Pareto, càd 20% des données vont nous servir de test pour performance du modèle.

Pour tester la performance de la prédiction, nous nous servirons de mesure comme : la matrice de confusion, l’accuracy, l’aire sous la courbe ROC (Receiver Operating Curve) et quelques mesures sur la spécificité du modèle.

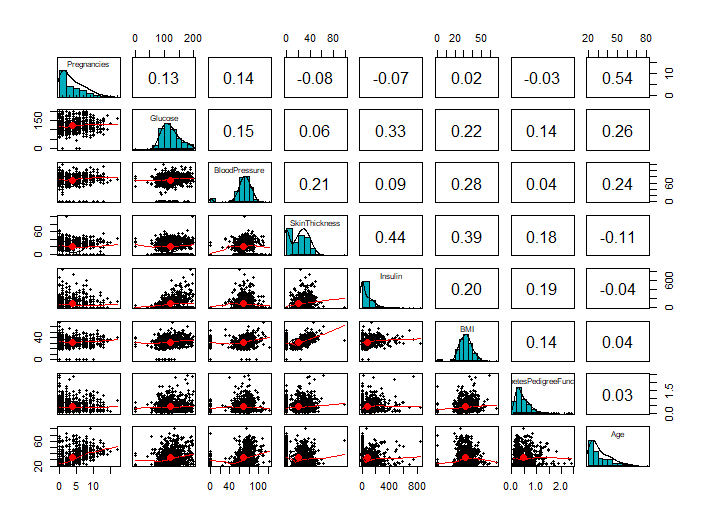
3.2 **Corrélation des Facteurs**

Comme nous l’avons évoqué au point précédent, la classification basée sur les probabilités pose des sérieux problèmes. Ce type de classification sont extrêmement sensibles aux problèmes de multicolinéarité[[2]](#footnote-2). Il sied de signaler que la majorité des maladies comme celui des diabètes sont extrêmement liées à d’autres types de pathologie, ce qui pourrait conduire aux problèmes de multicolinéarité.

Or, l’intégration de variables corrélées fausse les résultats du modèle et doit être évitée (Ooghe et al. 1994). Mais Il y a un vrai risque de passer à côté d'une variable exogène importante tout simplement parce qu'elle est redondante avec une autre. La colinéarité[[3]](#footnote-3) entre variables exogènes rend illusoire la lecture des résultats sur la base des valeurs et de la significativité des coefficients. Il est indiqué de la détecter et de la traiter avant toute interprétation approfondie.

Les variables colinéaires seront éliminées du modèle lors de la sélection des variables par le critère AIC-Schwartz.

Graphique 2 : Corrélation des variables exogènes



*Source : Auteur, à l’aide du logiciel R*

Au regard du graphique, nous remarquons que les corrélations entre les variables sont très faibles et même sans pour autant passer par un test statistique, cela confirme bien l’indépendance entre les variables exogènes. D’où l’application du classifieur de Bayes ne donnera pas des mauvaises performances.

3.3 **Analyse du modèle de Naïve Bayes**

A ce point, nous allons mettre en place le modèle prédictif permettant de dire si une personne est atteinte du diabète de type II ou pas. Pour cela, le modèle dont nous utiliserons est ce lui de ***Naïve Bayes Classifier.***

Enfin de compte, nous allons réaliser une analyse approfondie sur le modèle de prédiction, cela nous permettra de savoir si le modèle est fiable ou pas.

3.3.1 **Résultat de l’analyse du classifieur de Bayes**

Dans les sections qui suivent, nous étudierons la qualité de ce modèle selon les règles et hypothèses d’ajustement par la méthode de maximum de vraisemblance.

3.3.1.1 **Evaluation du modèle**

Tableau 5 : Evaluation du modèle sur les données d'entrainement et sur les données de test

|  |  |
| --- | --- |
| **Train** | **Test** |
| **0.7525** | **0.7662** |
| *Source : A l'aide du langage de programmation Python 3.11* | |

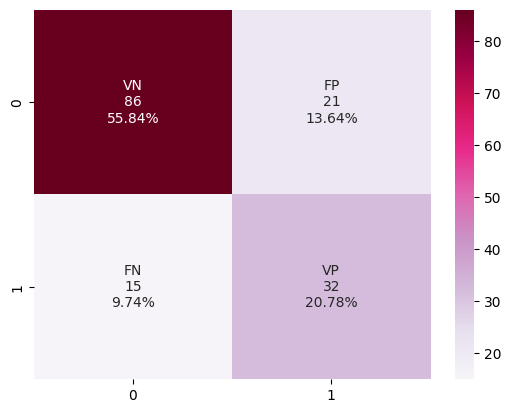
Globalement le modèle est de très bonne qualité. Le modèle est capable de prédire à 75% sur les données qui ont servi de création du modèle et sur les données de test, le modèle est capable de prédire à 77%. Tout ce qui précède signifie que si nous présentons une nouvelle information au modèle, notre modèle aura une probabilité de 0.77 de réussir sa prédiction. Ce qui conduit à dire que la confiance au modèle ne sera pas décevante.

3.3.1.1 **Performance du modèle**

Pour mesurer la performance du modèle, il serait question de se servir de la matrice de confusion.

Cette matrice de confusion se présente comme :

Figure 2 : Matrice de confusion



*Source : Auteur, à l’aide du langage de programmation Python 3.11*

Avec de la matrice de confusion, certaines informations restent cachées. D’où, pour avoir plusieurs informations, nous devons nous référer au rapport de classification[[4]](#footnote-4). Car ce dernier nous fournit plusieurs informations en rapport avec : le rappel, la précision, le score F1 et le support du modèle entrainé.

En ce qui concerne cette analyse, le rapport de classification est :

Tableau 6 : Rapport de classification du model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **Absence** | 0.85 | 0.8 | 0.83 | 107 |
| **Présence** | 0.6 | 0.68 | 0.64 | 47 |
|  |  |  |  |  |
| **accuracy** |  |  | 0.77 | 154 |
| **macro avg** | 0.73 | 0.74 | 0.73 | 154 |
| **weighted avg** | 0.78 | 0.77 | 0.77 | 154 |
| *Source : Auteur, à l'aide du langage de programmation Python 3.11* | | | | |

**3.3.1.2 Analyse de la performance du modèle**

**1. Sensibilité**

Rappelons que la sensibilité (ou rappel) qui est, en effet, le taux des vrais positifs (TVP), représente la capacité du modèle à retrouver les positifs, càd la capacité d’un examen diagnostique à fournir un résultat positif en présence de la maladie.

Elle se calcul de la manière suivante :



Pour ce modèle, la capacité d’un examen diagnostique à fournir un résultat positif en présence de la maladie est estimée à 68%.

2. **Taux d’erreur**

Le taux d'erreur est égal au nombre de mauvais classement rapporté à l'effectif total.



La probabilité de mauvais classement est estimée à 0.23, soit 23%. Cela signifie que pour un test réalisé, la probabilité pour que le résultat soit faux est estimée à 23%.

3. **Taux de succès (Accuracy)**

Le taux de succès est égal au nombre de bon classement rapporté à l'effectif total, ou même le complémentaire à l’unité du taux d’erreur.



La probabilité de bon classement est estimée à 0.77, soit 77%. Cela signifie que pour un test réalisé, la probabilité pour que le résultat soit vrai est estimée à 77%.

4. **La spécificité**

La spécificité, à l’inverse de la sensibilité, indique la proportion des négatifs détectés.



Pour un test réalisé sur une personne qui n’a pas le diabète, ce modèle a une capacité de 80% de confirmer cela.

Comparativement au cas où nous sommes en présence de la maladie, le modèle a une forte capacité de dire que la personne n’est pas diabétique.

Cette grande différence en terme de la capacité de prédiction est due au problème de l’imbalence (soulevé au point 3.2).

5. **La Précision**

La précision indique la proportion de vrais positifs parmi les individus qui ont été classés positifs.



La proportion des vrais diabétiques parmi les patients prédit diabétiques par le modèle est de 60%.

**CONCLUSION**

1. Imbalance : Cas du déséquilibre en termes de pourcentage au niveau de la variable cible. [↑](#footnote-ref-1)
2. Multicolinéarité : la multicolinéarité est la forte corrélation entre plus de deux variables explicatives.

   On parle de multicolinéarité parfaite lorsqu'une des variables explicatives d'un modèle est une combinaison linéaire d'une ou plusieurs autres. [↑](#footnote-ref-2)
3. Colinéarité : Étymologiquement, **colinéaire signifie sur une même ligne.** En géométrie classique, deux vecteurs sont colinéaires si on peut en trouver deux représentants situés sur une même droite. [↑](#footnote-ref-3)
4. Rapport de classification : Un rapport de classification est une mesure d’évaluation des performances dans l’apprentissage automatique. Il est utilisé pour montrer la précision, le rappel, le score F1 et le support de votre modèle de classification entraîné. [↑](#footnote-ref-4)