.

Rapport

Réalisé par :
DAHMANI Zineb

Table de figures

Table des figures

Table des matières

1 Chapitre 1 : Réalisation

Ce chapitre présente la réalisation de projets data science décomposé en trois parties pour les trois differents cas de problémes : Classification binaire, multiclassification et puis régression .

1.1 Classification Binaire

1.1.1 Buisness case

Les indicateurs clés de performance (KPIs) qui seront influencés positivement :
- Amélioration de la réactivité des systèmes de santé .
- Amélioration de la qualité de vie .

Ceci en déterminant les personnes à risque de développer des maladies coronariennes .

1.1.2 Data

1.1.2.1 Data source Le jeu de données est disponible publiquement sur le site Kaggle et provient d'une étude cardiovasculaire en cours sur les résidents de la ville de Framingham, Massachusetts. L'objectif de la classification est de prédire si le patient présente un risque de maladie coronarienne à 10 ans.Le jeu de données fournit des informations sur les patients. Il comprend plus de 4 000 enregistrements et 15 attributs,chaque attribut est un facteur de risque potentiel. Il existe des facteurs de risque démographiques, comportementaux et médicaux.

1.1.2.2 Data description Une description des 15 variables du dataset :

Demographique :
• Sex : « M » pour le sexe masculin et « F » pour le sexe féminin)
• Age : Age du patient au moment de l'étude .
Comportemental :
• is smoking : si le patient est un « YES » ou un non « NO » fumeur actuel .
• Cigs Per Day : le nombre de cigarettes que la personne a fumé en moyenne en une journée. (Peut être considéré comme continu car on peut avoir n'importe quel nombre de cigarettes, même une demi-cigarette).
Medicale(history) :
• BP Meds : si le patient prenait ou non des médicaments contre l'hypertension (nominale) .
• Prevalent Stroke : si oui ou non le patient a déjà eu un AVC (nominal) .
• Prevalent Hyp : si le patient était ou non hypertendu (nominal) .
• Diabetes : si le patient était diabétique ou non (nominal) .
Medical(current) :
• Tot Chol : taux de cholestérol total (continu) .
• Sys BP : pression artérielle systolique (continue) .

• Dia BP : pression artérielle diastolique (continue) .
• BMI : indice de masse corporelle (continu) .
• Heart Rate : fréquence cardiaque (continue) .
• Glucose : taux de glucose (continu) .
Variable à prédire(desired target) :
• TenYearCHD : Risque à 10 ans de maladie coronarienne CHD(binaire : "1", signifie "Oui", "0" signifie "Non") .

1.1.3 Data exploration

1.1.3.1 Principe : L'exploration des données est l'étape initiale de l'analyse des données, au cours de laquelle on peut explorer un grand ensemble de données de manière non structurée afin de découvrir les premiers modèles, caractéristiques et points d'intérêt. Ce processus n'a pas pour but de révéler toutes les informations contenues dans un ensemble de données, mais plutôt d'aider à créer une image générale des tendances importantes et des points majeurs à étudier plus en détails.

```
id                1667
age                 49
education          NaN
sex                  F
is_smoking          NO
cigsPerDay         0.0
BPMeds             0.0
prevalentStroke      0
prevalentHyp         0
diabetes             0
totChol          246.0
sysBP            107.0
diaBP             73.0
BMI              29.36
heartRate         79.0
glucose           80.0
TenYearCHD           0
Name: 1667, dtype: object
```

```
Data columns (total 17 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   id               3390 non-null   int64
 1   age              3390 non-null   int64
 2   education        3303 non-null   float64
 3   sex              3390 non-null   object
 4   is_smoking       3390 non-null   object
 5   cigsPerDay       3368 non-null   float64
 6   BPMeds           3346 non-null   float64
 7   prevalentStroke  3390 non-null   int64
 8   prevalentHyp     3390 non-null   int64
 9   diabetes         3390 non-null   int64
 10  totChol          3352 non-null   float64
 11  sysBP            3390 non-null   float64
 12  diaBP            3390 non-null   float64
 13  BMI              3376 non-null   float64
 14  heartRate        3389 non-null   float64
 15  glucose          3086 non-null   float64
 16  TenYearCHD       3390 non-null   int64
dtypes: float64(9), int64(6), object(2)
```

FIGURE 1 – Affichage d'un exemple aléatoire et des caractéristiques des features

1.1.3.2 Résultats : On remarque :

— La présence de valeurs manquantes :

- education (87).
- cigsPerDay (4).
- BPMeds (44).
- totChol (38).
- BMI (14).

- heartRate (1).
- glucose (304).

— 43.75% des variables ont des valeurs manquantes .
— 2 variables de type object (nécessitent un traitement pour qu'elles soient compréhensibles par la machine) .

1.1.4 Data preprocessing

1.1.4.1 Principe : Le prétraitement des données comprend les étapes que nous devons suivre pour transformer ou coder les données afin qu'elles puissent être facilement analysées par la machine.

Pour qu'un modèle soit exact et précis dans ses prédictions, l'algorithme doit être capable d'interpréter facilement les caractéristiques des données. Ceci à travers :

Data cleaning

Se fait notamment dans le cadre du prétraitement des données pour nettoyer les données en remplissant les valeurs manquantes(handling missing values), en lissant les données bruyantes, en résolvant les incohérences et en supprimant les valeurs aberrantes(Removing outliers)...

Data transformation

Une fois le nettoyage des données effectué, nous devons consolider les données de qualité sous d'autres formes en modifiant la valeur, la structure ou le format des données à l'aide des stratégies de transformation des données comme la normalisation.

1.1.4.2 Résultats : *Handling missing values :*

Puisque Les valeurs manquates semblent etre des MARs (Missing at random) et l'abandon de ces enregistrements entraînerait une perte de données, nous devons donc les remplacer .
Pour ceci ,on remplacera pour :

- *education , heartRate* par la valeur la plus fréquente .

- *cigsPerDay , BPMeds , totChol , BMI , glucose* par la moyenne de la variable correspondante .

Encoding :

Un meilleur codage conduit à un meilleur modèle et la plupart des algorithmes ne peuvent pas traiter les variables catégorielles à moins qu'elles ne soient converties en une valeur numérique. C'est pour cela la transformation de *sex* , *is smoking* est nécessaire .

Normalization :

La normalisation est la transformation des données de manière à ce qu'elles soient sans dimension et / ou aient des distributions similaires.

— Utilité :

Améliore considérablement la précision du modèle en donnant des poids / importance égaux à chaque variable de sorte qu'aucune variable unique ne dirige les performances du modèle dans une seule direction simplement parce qu'il s'agit de nombres plus grands.

— Outils et techniques :

Une des techniques les plus connues qu'on utilisera lors de notre application est la normalisation du score z ou bien standardisation :

$$z_{ij} = \frac{x_{ij} - \bar{x_j}}{\sigma_j} \tag{1}$$

où :

$$\begin{cases} \bar{x_j} : moyennedelavariable \\ \sigma_j : carttypedelavariablej \end{cases}$$

—

Comparaison entre données brutes et transformées :

— Données brutes : distance augmente avec présence de variables dispersées .

— Données transformées : distance augmente si pour une variable l'écart au carré entre individus est important par rapport à la variance de cette variable .

1.1.5 Data analytics

1.1.5.1 Principe : Le Data Analytics, abrégé par DA, est une science consistant à examiner des données brutes, dans le but de tirer des conclusions à partir de ces informations.

1.1.5.2 Résultats :

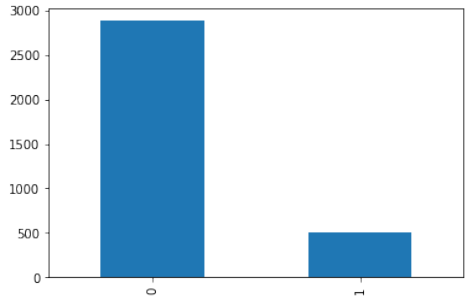


FIGURE 2 – Distribution de 'Target variable'

— La distribution est fortement déséquilibrée. En effet, le nombre de cas négatifs est supérieur au nombre de cas positifs. Cela conduirait à un problème de déséquilibre de classe lors de l'ajustement de nos modèles. Par conséquent, ce problème doit être traité et pris en charge (Class balance).

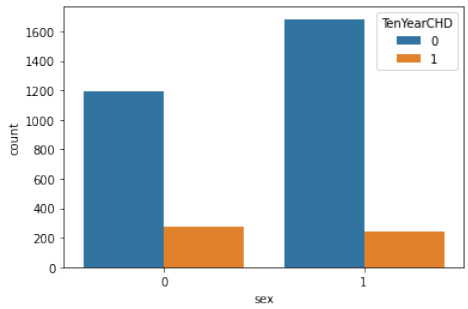


FIGURE 3 – Distribution de 'Target variable' selon le sexe - M :0 / F :1

— -Les hommes présentent plus de risque à avoir ces maladies.

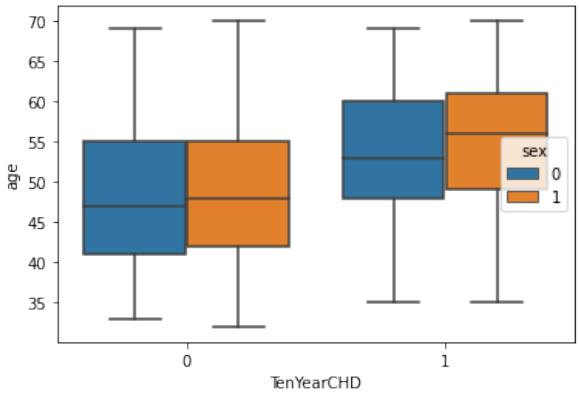


FIGURE 4 – Box plot age par rapport à la variable visée

— Chez les sujets présentant un risque , il s'avère que la moyenne d'age pour :
* le sexe féminin est : 56 , majoritairement entre [49,63] .
* le sexe masculin est : 53 , majoritairement entre [48,60] .
Chez les sujets ne présentant pas de risque , il s'avère que la moyenne d'age pour :

* le sexe féminin est : 48 , majoritairement entre [45,55] .
* le sexe masculin est : 47 , majoritairement entre [41,50] .
Le risque augmente avec l'age (généralement lorsqu'on dépasse les cinquantaines pour les deux sexes.

— On remarque aussi la présence des outliers ,par exemple la présente d'une femme à risque à l'age de 35 ans .

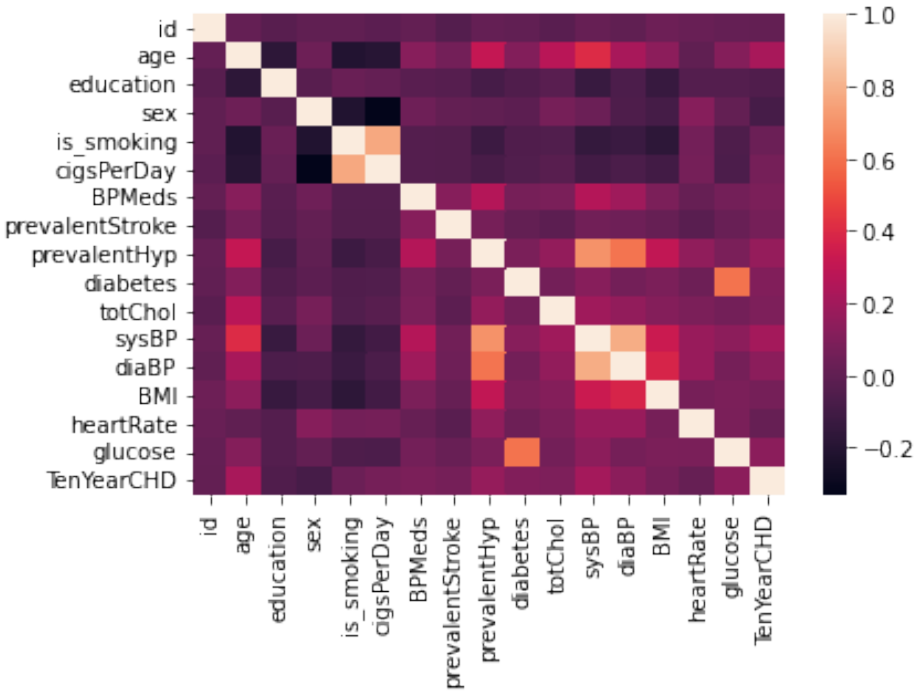


FIGURE 5 – Matrice de corrélation

— Variables fortement corrélées :
- is smoking et cigsPerDay .
- sysBP et diaBP .
- sysBP et prevakentStroke .

— Parmi les variables numériques continues :
-totChol, sysBP, diaBPet BMI ont des distributions uniformes et les autres sont distribuées de manière inégale.
-cigsPerDay a une distribution très inégale avec la plupart des données présentes dans 0.
-cigsPerDay et sysBP présentent respectivement une forte et une légère asymétrie vers la droite.

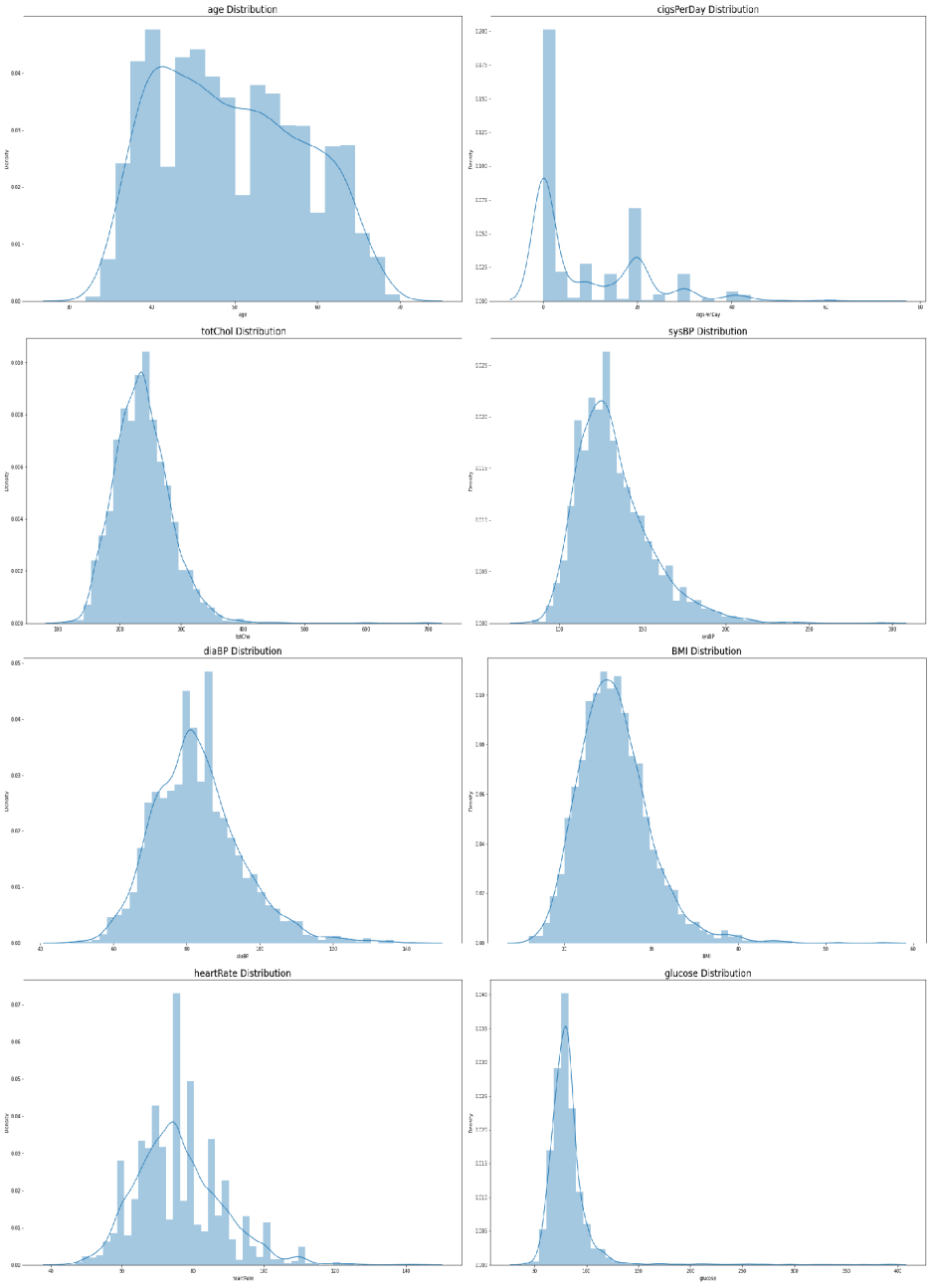


FIGURE 6 – Distribution des variables numériques continues

— Le graphique de la proportion du risque de maladie coronarienne à 10 ans par niveau d'éducation montre les distributions du risque de maladie coronarienne à 10 ans dans chaque groupe d'éducation. Il n'y a pas de grande différence entre ces distributions, mais le risque le plus élevé se trouve dans le groupe d'éducation 1 et le plus faible dans le groupe 2.

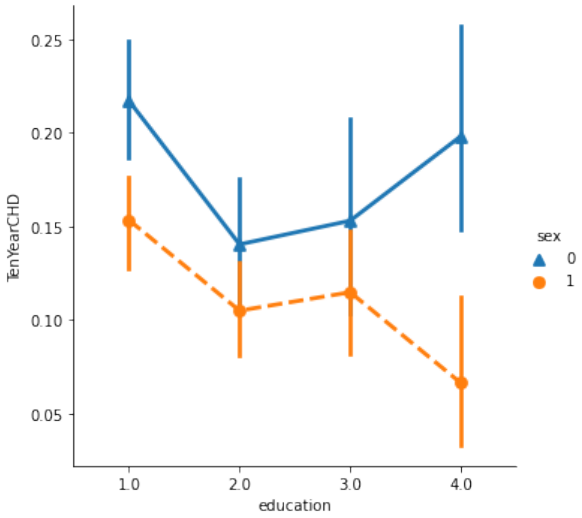


FIGURE 7 – Distribution du 'target variable' pour chaque groupe d'education

— De 22 personnes ayant déja fait un avc 10 personnes présentent un risque de developper une maladie coronarienne , c'est à dire un pourcentage de 45% qui reste largement important .On peut dire que les gens ayant déja fait un avc présente un risque très modéré ,or qu'une population de 22 personnes seulement n'est pas fiable .

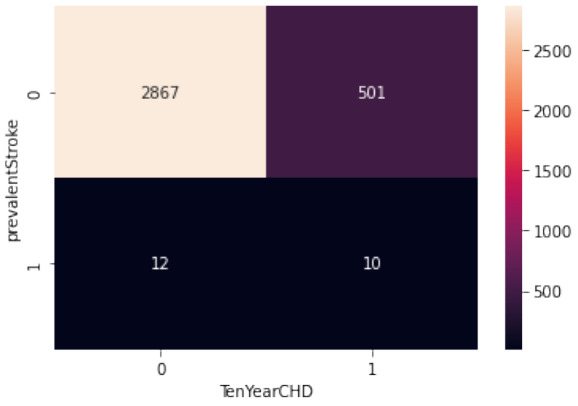


FIGURE 8 – Relation entre risque è developper une maladie coronarienne et avc antécédant

1.1.6 Feature engineering

1.1.6.1 Principe :

De meilleures variables signifient flexibilité, modèles plus simples et meilleurs résultats.
Le Feature Engineering est un processus qui consiste à transformer les données brutes en caractéristiques représentant plus précisément le problème sous-jacent au modèle prédictif. Pour faire simple,il s'agit d'appliquer une connaissance du domaine pour extraire des représentations analytiques à partir des données brutes et de les préparer pour le Machine Learning.
Le Feature Engineering repose sur un ensemble de procédures et de méthodes bien définies. Les procédures à utiliser varient en fonction des données, et c'est avec l'expérience et la pratique que l'on apprend lesquelles utiliser en fonction du contexte.
On distingue differents procédés tel que :
-Feature Extraction : La construction automatique de nouvelles caractéristiques à partir de données brutes .
-Feature Selection : De nombreuses fonctionnalités à quelques-unes qui sont utiles.
-Feature Construction : La construction manuelle de nouvelles caractéristiques à partir de données brutes.

1.1.6.2 Résultats :

— "id" ne servira à aucun point pour augmenter la qualité des modèles , elle sera donc supprimée .
— On a déja signalé que ces Variables sont fortement corrélées :
* is smoking et cigsPerDay .
* sysBP et diaBP .
* sysBP et prevakentStroke .

–> is smoking est une variable binaire indiquant si l'exemple est un current fumeur ou nn , par contre cigsPerDay est une variable quantitative indiquant la moyenne du nombre de cigarettes fumées par l'exemple par jour.Cette variable est donc plus informative et englobe la première . On peut donc supprimer la variable " is smoking" .
–> sysBP est fortement corélée avec diaBP et prevalentStroke or ces dernieres ne sont pas fortement corélees ,le plus juste donc est de supprimer la variable sysBP .

1.1.7 Application du modele ML "SVC"

Cas linéairement Séparable

1.1.7.1 Hard Margin SVC :

Dans un espace vectoriel de dimension finie n, un hyperplan est un sous-espace vectoriel de dimension n1.
Soit un espace vectoriel E de dimension n. L'équation caractéristique d'un hyperplan est de la forme : $w_1x_1 + w_2x_2 + ... + w_nx_n$ où $w_1, w_2 \ldots\ldots, w_n$sont des scalaires.
Comme on peut le constater, un hyperplan vectoriel passe toujours par 0. C'est pour cette raison qu'on utilisera un hyperplan affine, qui n'a pas quant à lui obligation de passer par l'origine.

Ainsi, si l'on se place dans $\mathbb{R}^{\mathbb{N}}$,pendant son entraînement le SVM calculera un hyperplan vectoriel d'équation $w_1x_1 + w_2x_2 + ... + w_nx_n = 0$ ainsi qu'un scalaire (un nombre réel) b. C'est ce scalaire b qui va nous permettre de travailler avec un hyperplan affine, comme nous allons le voir.

Formulations et notations :

Le vecteur $w=\begin{pmatrix} w_1 \\ : \\ w_n \end{pmatrix}$ est appelé vecteur de poids et le scalaire b biais.

Une fois l'entraînement terminé, pour classer une nouvelle entrée $x=\begin{pmatrix} x_1 \\ : \\ x_n \end{pmatrix}$,le SVM regardera le signe de :

$$h(x) = w_1x_1 + w_2x_2 + ... + w_nx_n = w^n.b \tag{2}$$

pour un point x, s'il se trouve d'un côté ou de l'autre de l'hyperplan. La fonction h permet de répondre à cette question, grâce à la classification suivante : $\begin{cases} h(x) \geq 0 \Rightarrow x \in categorie1 \\ h(x) \leq 0 \Rightarrow x \in categorie2 \end{cases}$

On est face à une minimisation avec contraintes , on peut donc résoudre ce problème par la méthode classique des multiplicateurs de Lagrange, le Lagrangien de ce problème est donné par :

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum \alpha_i[y_i(w^T x_i + b) - 1] où \forall \alpha_i \geq 0 \tag{3}$$

Conditions K.K.T : $\begin{cases} \frac{\partial L}{\partial w} = w - \sum \alpha_i y_i x_i = 0 \Rightarrow w = \sum \alpha_i y_i x_i \\ \frac{\partial L}{\partial b} = -\sum \alpha_i y_i = 0 \end{cases}$

Dans (3) , on remplace :

$$L(\alpha) = \frac{1}{2}(\sum \alpha_i y_i x_i)(\sum \alpha_j y_j x_j) - (\sum \alpha_i y_i x_i)(\sum \alpha_j y_j x_j) - \sum \alpha_i y_i + \sum \alpha_i \tag{4}$$

$$L(\alpha) = \sum \alpha_i - \frac{1}{2}\sum\sum \alpha_i \alpha_j y_i y_j x_i x_j \tag{5}$$

Cas linéairement non Séparable

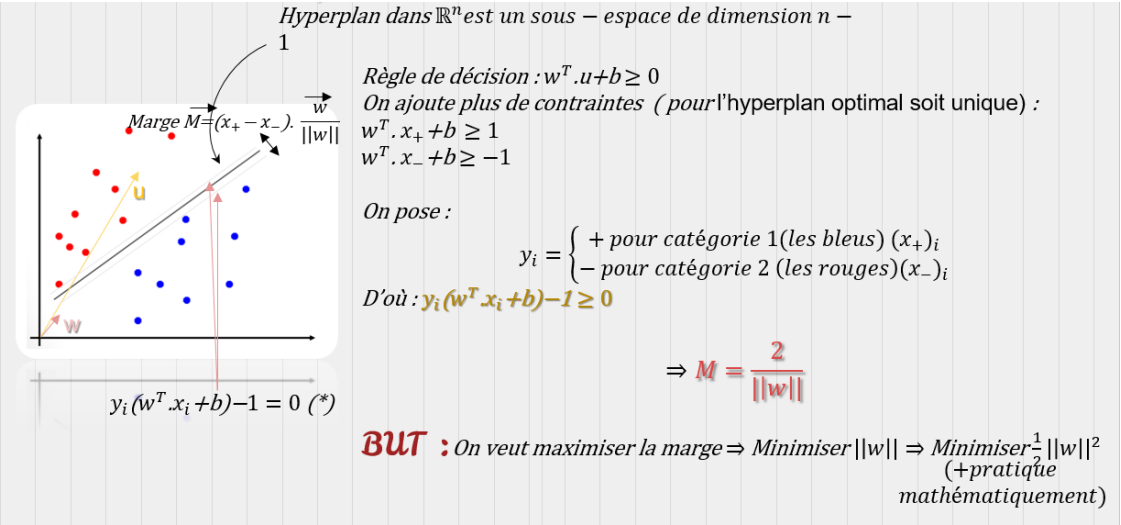


FIGURE 9 – Figure illustrative du principe de SVC

1.1.7.2 Soft Margin SVC :

Formulations et notations :

— ξ est un vecteur de taille n .

— $\xi_i \geq 0$ matérialise l'erreur de classement pour chaque observation .

— $\xi_i = 0$ elle est nulle lorsque l'observation est du bon côté de la droite «marge» associée à sa classe.

— $\xi_i < 1$ le point est du bon côté de la frontière, mais déborde de la droite «marge» associée à sa classe .

— $\xi_i > 1$ l'individu est mal classé .

La tolérance aux erreurs est plus ou moins accentuée avec le paramètre C(''cost" parameter)
* C trop élevé, danger de sur-apprentissage .
* C trop faible, sous-apprentissage .
Le choix de C constituera un enjeu important en pratique .

Explication :

Dans le cas où les données ne sont pas linéairement séparables, ou contiennent du bruit (outliers : données mal étiquetées) les contraintes du problème ne peuvent être vérifiées, et il y a nécessite de les relaxer un peu. Ceci peut être fait en admettant une certaine erreur de classification des données $\xi_i \geq 0$. On aura pour la classification :

$$\begin{cases} y_i(w^T x_i + b) \geq 1 - \xi_i \\ w \in \mathbb{R}^{dim(x)}, b \in \mathbb{R} \end{cases}$$ Le hard marging essaye donc de résoudre le problème d'op-

timisation suivant : $\begin{cases} min\frac{1}{2}||w||^2 + C\sum \xi_i \\ y_i(w^T x_i + b) \geq 1 - \xi_i, i \in 1, 2, .., n \\ w \in \mathbb{R}^{dim(x)}, b \in \mathbb{R}, \xi_i \geq 0 \end{cases}$

Multiplicateur de Lagrange, le lagrangien de ce problème est donné par :

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}||w||^2 + C\sum \xi_i \sum \alpha_i[y_i(w^T x_i + b) - 1 + \xi_i] - \sum \beta_i \xi_i \quad (6)$$

Conditions K.K.T : $\begin{cases} \frac{\partial L}{\partial w} = w - \sum \alpha_i y_i x_i \Rightarrow w = \sum \alpha_i y_i x_i \\ \frac{\partial L}{\partial \xi_i} = 0 \Rightarrow 0 \leq \alpha_i \leq C \end{cases}$

En remplaçant ces valeurs dans l'équation (6) On obtient : $L(\alpha) = \sum \alpha_i - \frac{1}{2}\sum\sum \alpha_i \alpha_j y_i y_j x_i x_j$ (Forme duale)

Cas non linéairement Séparable De la même manière ,on est face à une minimisa-

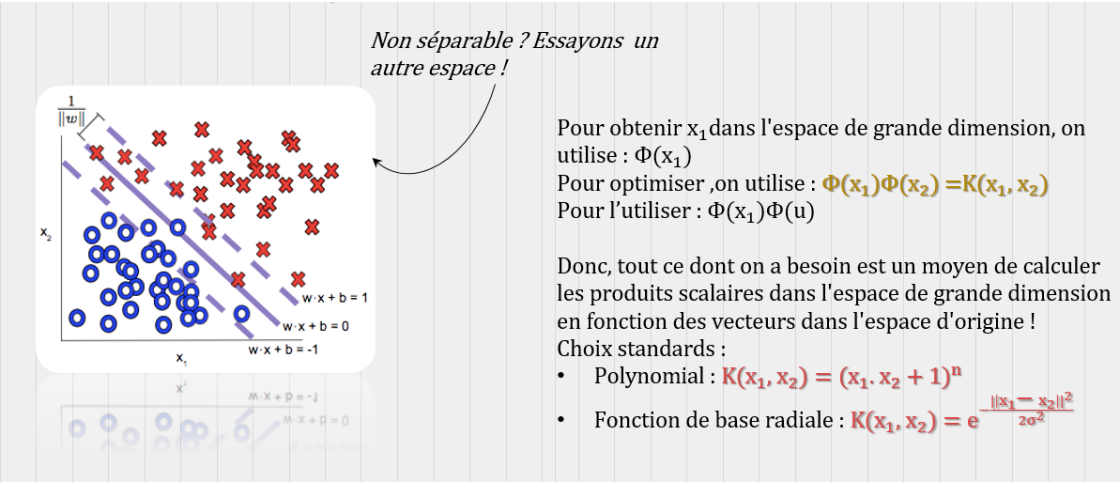


FIGURE 10 – Figure illustrative du principe de SVC-non lineairement séparable

tion avec contraintes , on peut donc résoudre ce problème par la méthode classique des multiplicateurs de Lagrange, le Lagrangien de ce problème est donné par :

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum \alpha_i[y_i(w^T \Phi(x_i) + b) - 1] o\grave{u} \forall \alpha_i \geq 0 \quad (7)$$

Conditions K.K.T : $\begin{cases} \frac{\partial L}{\partial w} = w - \sum \alpha_i y_i \Phi(x_i) = 0 \Rightarrow w = \sum \alpha_i y_i \Phi(x_i) \\ \frac{\partial L}{\partial b} = -\sum \alpha_i y_i = 0 \end{cases}$

W somme linéaire des tranformations de vecteurs dans l'ensemble d'échantillons

Dans (7) , on remplace :

$$L(\alpha) = \frac{1}{2}(\sum \alpha_i y_i x_i)(\sum \alpha_j y_j \Phi(x_j)) - (\sum \alpha_i y_i \Phi(x_i))(\sum \alpha_j y_j \Phi(x_j)) - \sum \alpha_i y_i + \sum \alpha_i \quad (8)$$

$$L(\alpha) = \sum \alpha_i - \frac{1}{2}\sum\sum \alpha_i \alpha_j y_i y_j \Phi(x_i)\Phi(x_j) \quad (9)$$

1.1.7.3 Résultas : En appliquant SVC , on obtient une accuracy de 84% : Ce qui indique un bon exemple .

1.2 Multiclassification

1.2.1 Buisness case

Les indicateurs clés de performance (KPIs) qui seront influencés positivement
- Amélioration de la réactivité des systèmes de santé .
- Amélioration de la qualité de vie .
Ceci en déterminant si une lésion est begnine ou maligne

1.2.2 Data

1.2.2.1 Data source Le jeu de données est disponible publiquement sur le site Kaggle ,il s'agit d'une collection de 10015 images dermatoscopiques de différentes populations. Les cas comprennent une collection représentative de toutes les catégories diagnostiques importantes dans le domaine des lésions pigmentées.

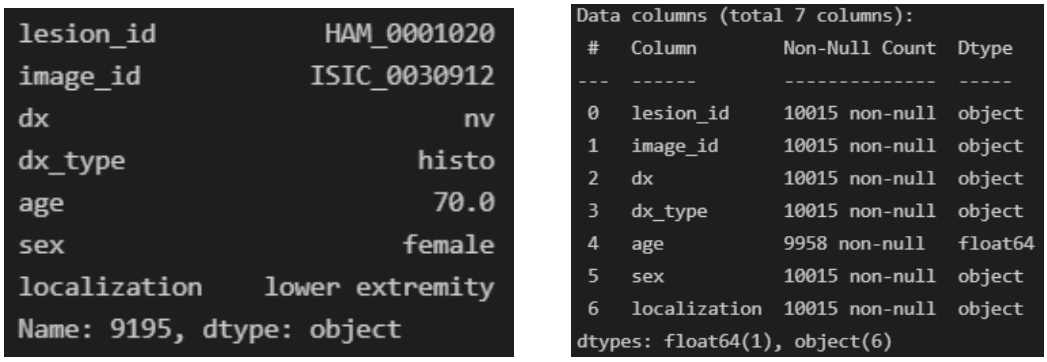
1.2.2.2 Data description
-Description des variables :

- lesion_id
- image_id
- dx : Types de lésions :
  - Maladie de Bowen (akiec). Malin
  - Carcinome basocellulaire (bcc).Malin
  - Lésions vasculaires (vasc) .Malin
  - Le mélanome (mel) .Malin
  - Les naevus mélanocytaires (nv) .Bénin
  - Lésions bénignes de type kératose (bkl). Bénin
  - Le dermatofibrome(df) . Bénin
- dx_type Outils de diagnostique :
  - Histopathologie (histo) .
  - Suivi du diagnostic (follow_up) .
  - Diagnostic consensuel (consensus) .
  - Microscope confocal (confocal) .
- Localisation :
  - Visage (face) .
  - Oreille (ear) .
  - Scalp (cuir chevelu) .
  - Cou (neck) .
  - Tête (head) .
  - Poitrine (chest) .
  - Main (hand) .
  - Pied (foot) .

— Membre inférieur (lower extremity) .
— Membre supérieur (upper extremity) .
— Dos (back) .
— Tronc (trunk)
— Abdomen (abdomen) .
— Organes génitaux (genital) .

— Acrales (acral) .

1.2.3 Data exploration



```
lesion_id            HAM_0001020
image_id            ISIC_0030912
dx                            nv
dx_type                    histo
age                         70.0
sex                       female
localization    lower extremity
Name: 9195, dtype: object
```

```
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   lesion_id     10015 non-null  object
 1   image_id      10015 non-null  object
 2   dx            10015 non-null  object
 3   dx_type       10015 non-null  object
 4   age           9958 non-null   float64
 5   sex           10015 non-null  object
 6   localization  10015 non-null  object
dtypes: float64(1), object(6)
```

FIGURE 11 – Affichage d'un exemple aléatoire et des caractéristiques des features

On remarque :

— La présence de valeurs manquantes : Age (57).

— 6 variables de type object (nécessitent un traitement pour qu'elles soient compréhensibles par la machine) .

1.2.4 Data Analytics

Vu que la plupart des variables ne sont pas numériques , on avancera l'étape de la visualisation cette fois - ci aprés avoir se débarasser des valeurs manquantes dans 'age'.

Distributions des features :

— La maladie la plus répandue est le naevus mélanocytaire, tandis que la moins répandue est le dermatofibrome.

— Les maladies de peau les plus fréquentes sont ceux bénines .

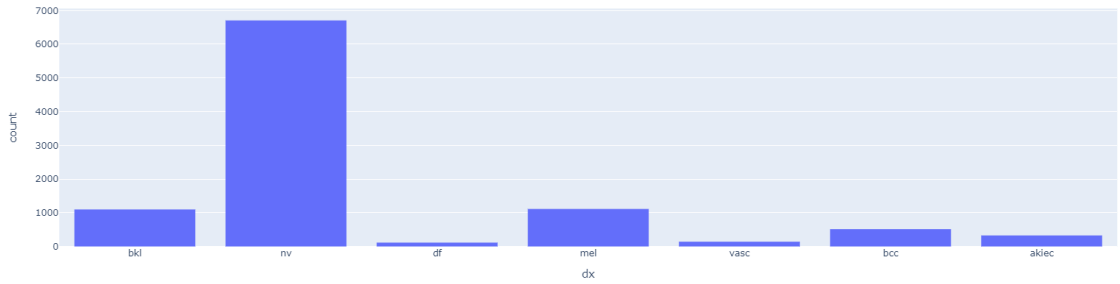


FIGURE 12 – Distribution des types de lésions

— Les maladies de peau sont plus visibles sur le "dos" du corps et moins sur les "surfaces acrales".

FIGURE 13 – Distribution des localisations

— Les maladies de peau sont plus fréquentes chez les hommes que chez les femmes. On remarque aussi la présence de valeurs manquantes .

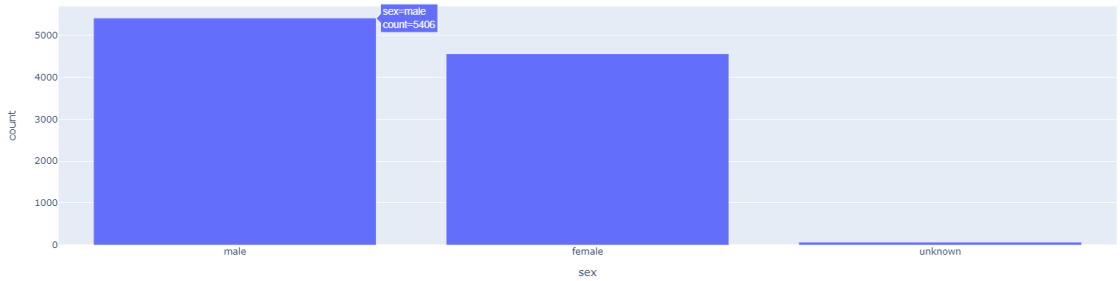


FIGURE 14 – Distribution du sexe

— Les maladies de la peau ont été découvert le plus avec histophatologie et de moins par un microscope confocal.

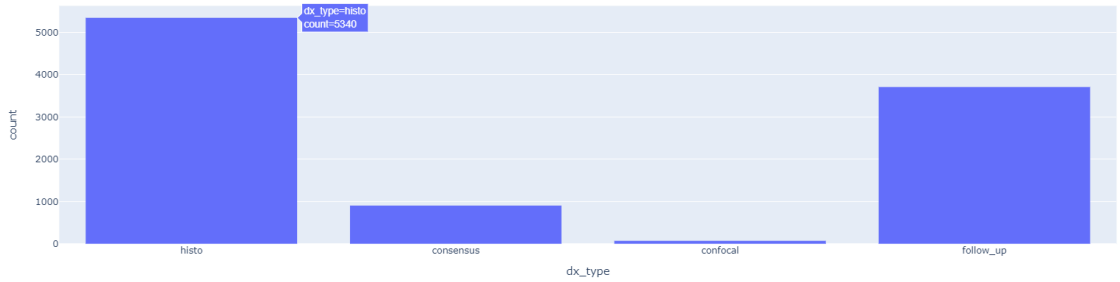


FIGURE 15 – Distribution du type de diagnostique

Relation d'age avec les autres variables :

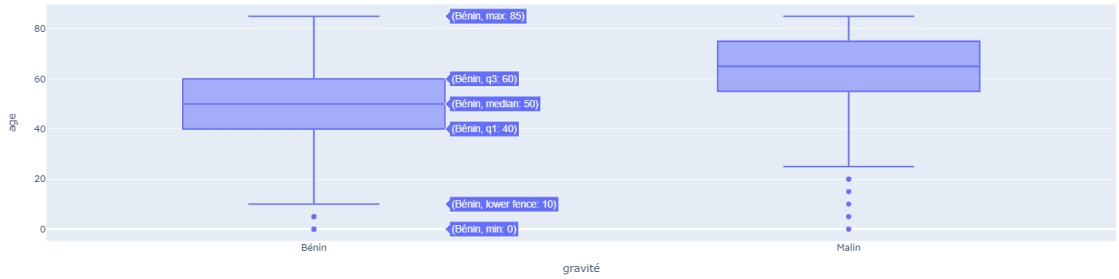


FIGURE 16 – Relation d'age avec type de lésion

— On observe que la probabilité d'avoir une maladie de la peau augmente avec l'âge .
Or que (bkl) qui est bénine représente une exception , elle est présente chez les individus les plus agés

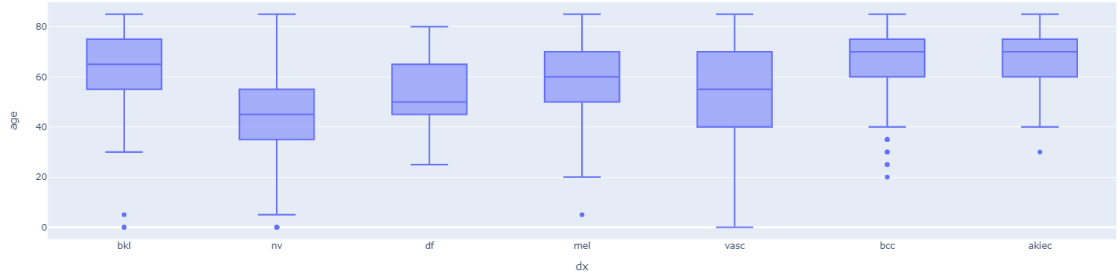


FIGURE 17 – Relation d'age avec type de maladie

— Il apparait que chez les personnes plus agés , on a recours de plus au microscope confocal .
— On remarque que :

— La moyenne d'age est élévée lorsqu'on parle de lésions localisés sur le visage , le cuir chevelure et les oreilles.

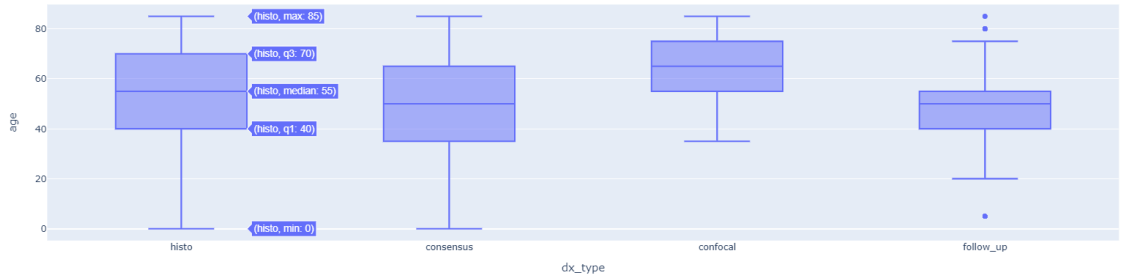


FIGURE 18 – Relation d'age avec type de diagnostique

— La moyenne d'age est basse losque les lésions sont au niveau des acrales, abdomen, parties genitales,pieds et le tronc .

— La présence d'outliers pour un peu près toutes les lésions sauf la potrine , le cou , les mains et les acrales .

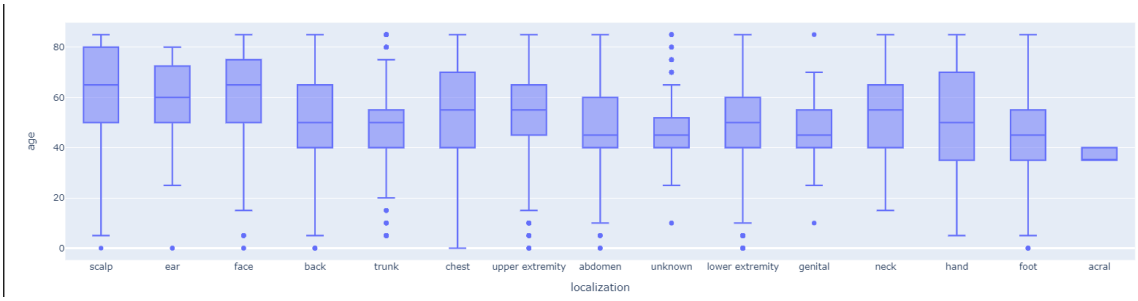


FIGURE 19 – Relation d'age avec localisation

Relation des types de lésions avec les autres variables :

Localisation :

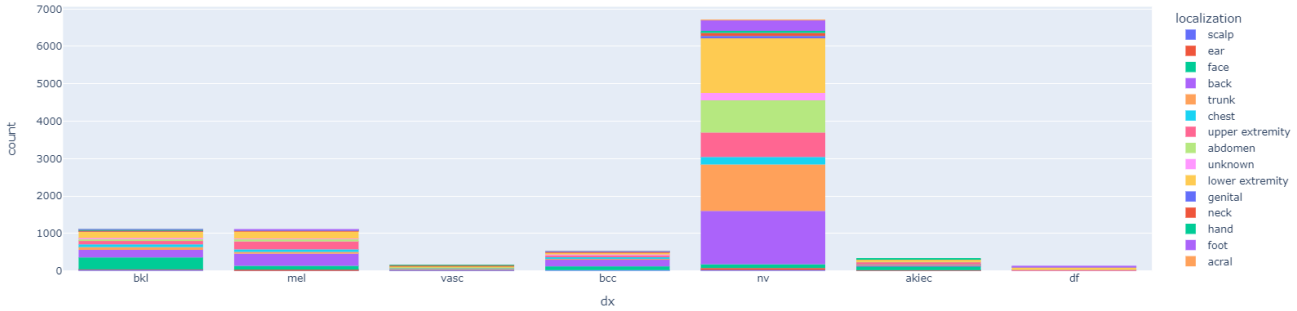


FIGURE 20 – Relation entre type de lésion et localisation

— Lésions vasculaires sont plus présentes au tronc .

— Maladies de Brown sont plus présentes au visage .

— Les dermatofibromes sont plus présentes au membres inférieurs .

— Carcinome basocellulaires sont plus présentes au dos .

— Les mélanomes sont plus présentes au dos .

— Lésions bénignes de type kératos sont plus présentes au visage .

— Les naevus mélanocytaires sont plus présentes au membres inférieurs .
On remarque qu'au niveau des arcales et des parties génitales , seules les naevus mélanocytaires sont présentes qui sont bénines .

Outils de diagnostique :

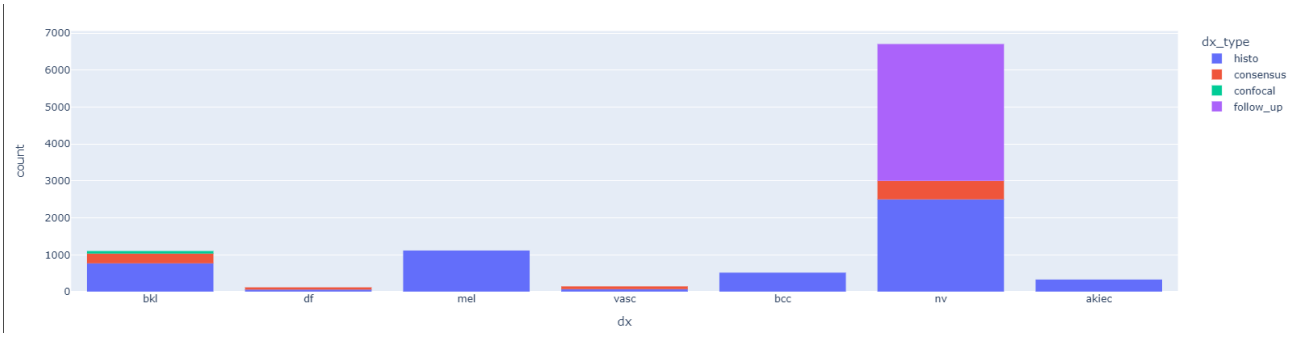


FIGURE 21 – Relation entre type de lésion et type de diagnostique

— Lésions vasculaires et les dermatofibromessont sont détectés par deux types de diagnostique : Diagnostic consensuel puis par histopathologie.

— Maladies de Brown , les carcinomes basocellulairessont et les mélanomes sont détectés uniquement par histopathologie .

— Lésions bénignes de type kératos sont détectés par trois types de diagnostique : histopathologie suivie de diagnostic consensuel puis par microscope confocal .

— Les naevus mélanocytaires sont sont détectés par trois types de diagnostique :un Suivi du diagnostic , de histopathologie puis par diagnostic consensuel.

Relation du sexe avec les autres variables :

Localisation :

— Seules les femmes developpent des lésions au niveau des acrales .

— Les femmes developpent plus de lésions au niveau des parties génitales,membres inférieurs,oreilles,les mains,pieds et sur des parties inconnues .

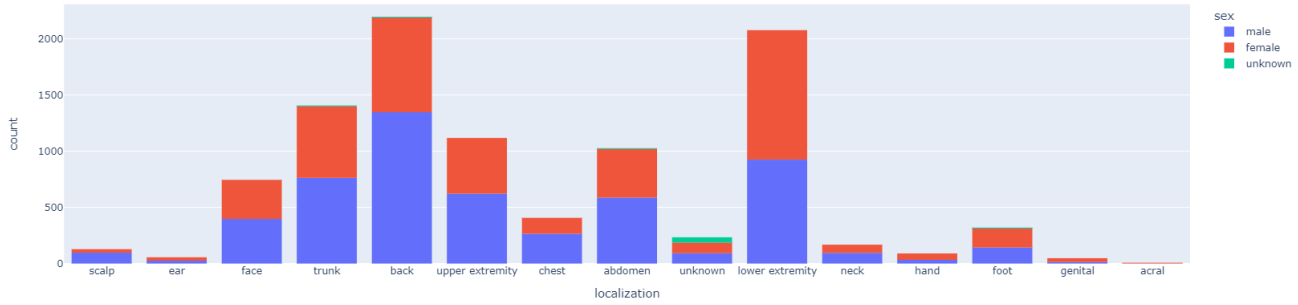


FIGURE 22 – Partition de localisation 'acrale' par sexe

Type de lésions :
Les femmes developpent plus de Lésions vasculaires que les hommes .Par contres ces derniers developpent plus les autres lésions .

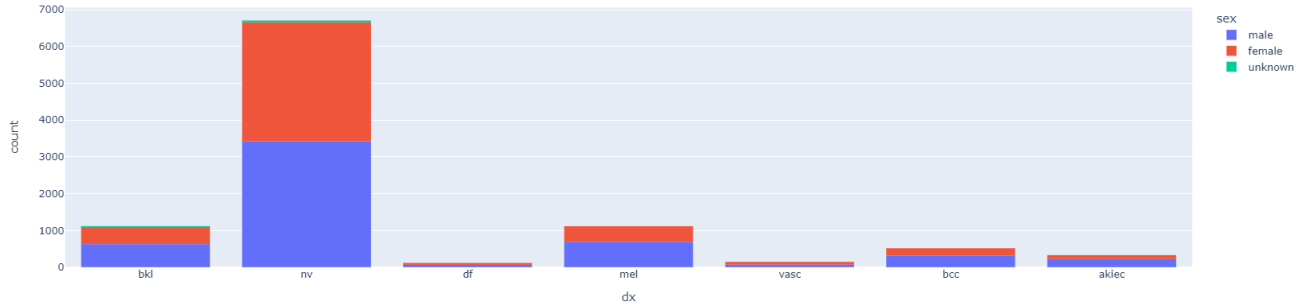


FIGURE 23 – Distribution des types de lésions par sexe

Valeurs manquantes :

Il se peut qu'il existe une relation entre les valeurs manquantes dans la variable localisation et celle du sexe .⇒ il se peut donc qu'il s'agit de MNAR's (Missing not at random) .

1.2.5 Data preprocessing

On remarque la présence de 'unknown' dans notre dataset , on doit les remplacer en NaN afin d'obtenir tous les valeurs manquantes avant l'étape de l 'encoding.

Encoding :
Pour cette étape ,j'ai transformé *sex,localization,dx et dx_type* en des valeurs numériques.

Manipulation des données manquantes :
J'ai beau essayé de comprendre et de chercher une raison qui peut justifier mon hypothèse (qu'il s'agit de MNAR) mais en vain , j'ai fint donc par les remplacer avec les valeurs les plus fréquentes .

FIGURE 24 – Partition des localisations pour sexe manquant

Redimensionnemnt des images :
Redimensionnement des images car les dimensions originales de 600 * 450 prennent beaucoup de temps à traiter .

1.2.6 Application du model ML"SVC"

Accuracy :
La précision est l'une des mesures permettant d'évaluer les modèles de classification. De manière informelle, la précision est la fraction des prédictions que notre modèle a réussi à obtenir. Formellement, la précision a la définition suivante :

$$\text{Accuracy} = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre total de prédictions}}$$

En appliquant SVC , on obtient une accuracy de 71% :
Ce qui indique un bon exemple .

1.3 Regression

1.3.1 Buisness case

Les indicateurs clés de performance (KPIs) qui seront influencés positivement
- Adaptation des coûts du traitement .
- Mettre l'accent sur l'impact de la médecine sur notre porte-monnaie .

1.3.2 Data

1.3.2.1 Data source Le jeu de données est disponible publiquement sur le site Kaggle qui vise nous donner une approximation de ce que seront les frais et les charges médicales des patients .

1.3.2.2 Data description Ce jeu de données est composé de 7 colonnes :
- age : Age du patient .
- sex : Sexe du patient .
- bmi : Indice de masse corporel du patient .
- children : Nombre d'enfants du patient .
- smoker : Si le patient est un fumeur ou pas
- region : La région ou vit le patient .
- charges : Charges médicales du patient .

1.3.3 Data exploration

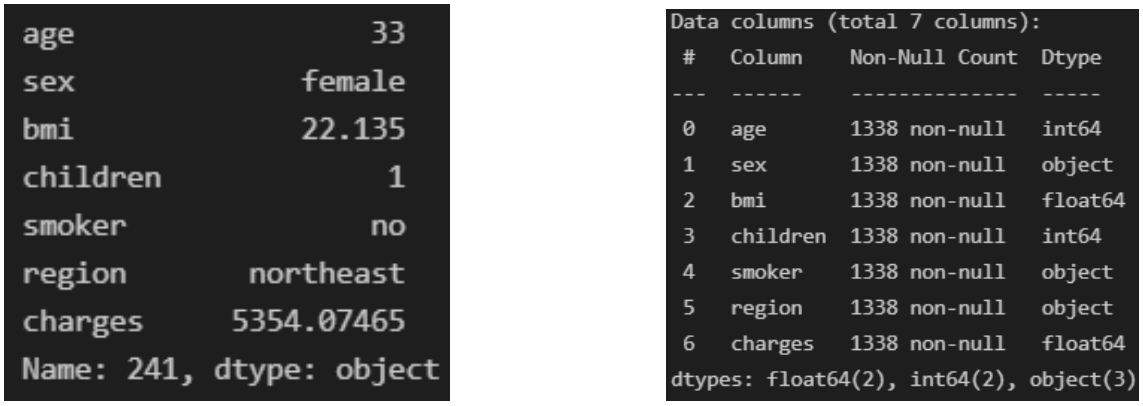


FIGURE 25 – Affichage d'un exemple aléatoire et des caractéristiques des features

— Aucune valeur manquante dans ces données .

— 3 variables de type object (nécessitent un traitement pour qu'elles soient compréhensibles par la machine) .

1.3.4 Data preprocessing

Encoding :
Pour cette étape ,j'ai transformé *sex,smoker et region* en des valeurs numériques.

Manipulation des données manquantes :
Pour la suite :

- Sex : '0' : male , '1' : female .
- Smoker : '0' : NO , '1' : yes .
- Region : '0' :Northeast ,'1' :Northwest ,'2' :Southwest ,'3' :Southeast .

Aucune valeur manquante pour les variables de type objet aussi .

1.3.5 Data analytics

Variable 'smoker' et ses relations avec les autres variables

- Les non fumeurs dominent les données .

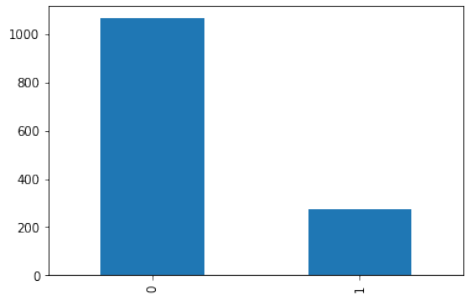


FIGURE 26 – Distribution de 'smoker'

- Le nombre d'hommes fumeurs dépasse celui des femmes fumeuses .

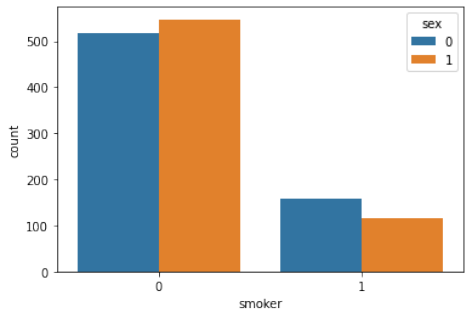


FIGURE 27 – Distribution de 'smoker' par sexe

— Les patients fumeurs dépensent plus pour leur traitement.

— Les hommes fumeurs ont les charges les plus élevés .

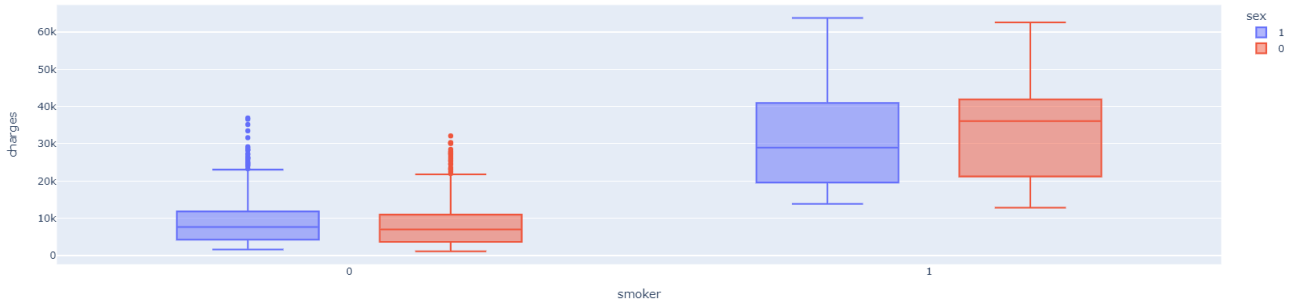


FIGURE 28 – Distribution de charges pour les fumeurs et les non fumeurs

— Les patients fumeurs dépensent plus pour leur traitement pour toutes les tranches d'age meme chez les sujets les plus jeunes .

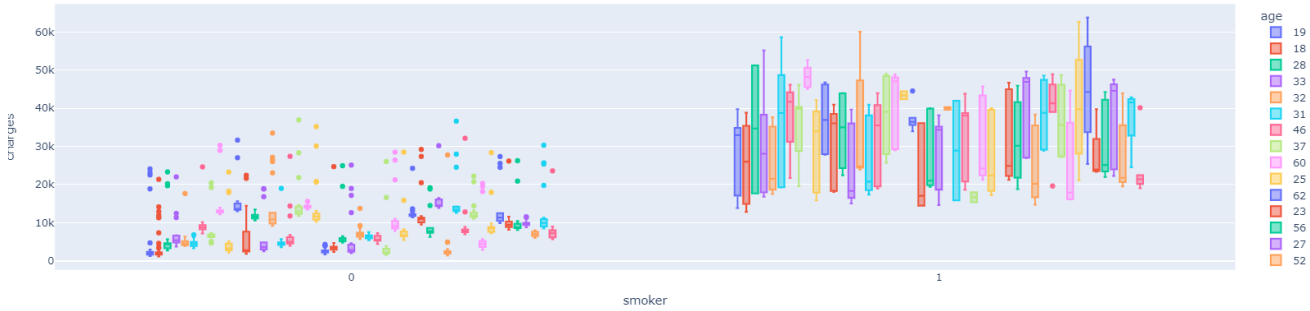


FIGURE 29 – Distribution de charges pour les fumeurs et les non fumeurs

— La distribution des fumeurs par régions est un peu la meme , sauf pour la région 'southwest'.

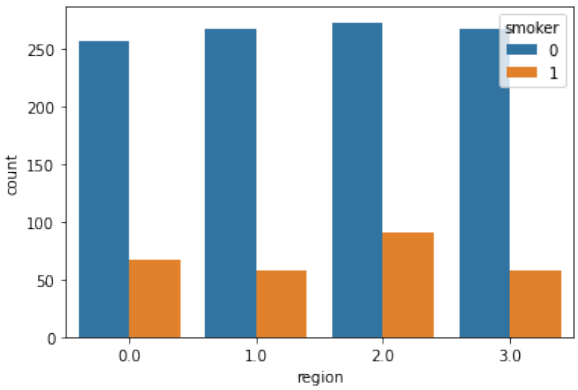


FIGURE 30 – Distribution des fumeurs par région

Variable 'age' et ses relations avec les autres variables

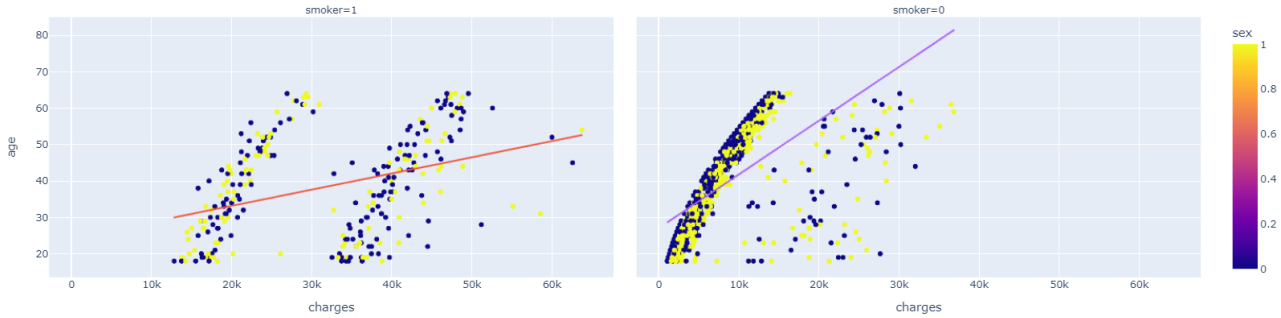


FIGURE 31 – Relation age-charges-sexe

— Clairement, les charges augmentent avec l'age .

— Cette augmentation est proportionnelle au cas fumeur ou pas , elle est plus importante chez les sujets non fumeurs ce qui met aussi l'accent sur l'important dépense des charges médicales par l'habitude de fumer .

Variable 'bmi' et ses relations avec les autres variables

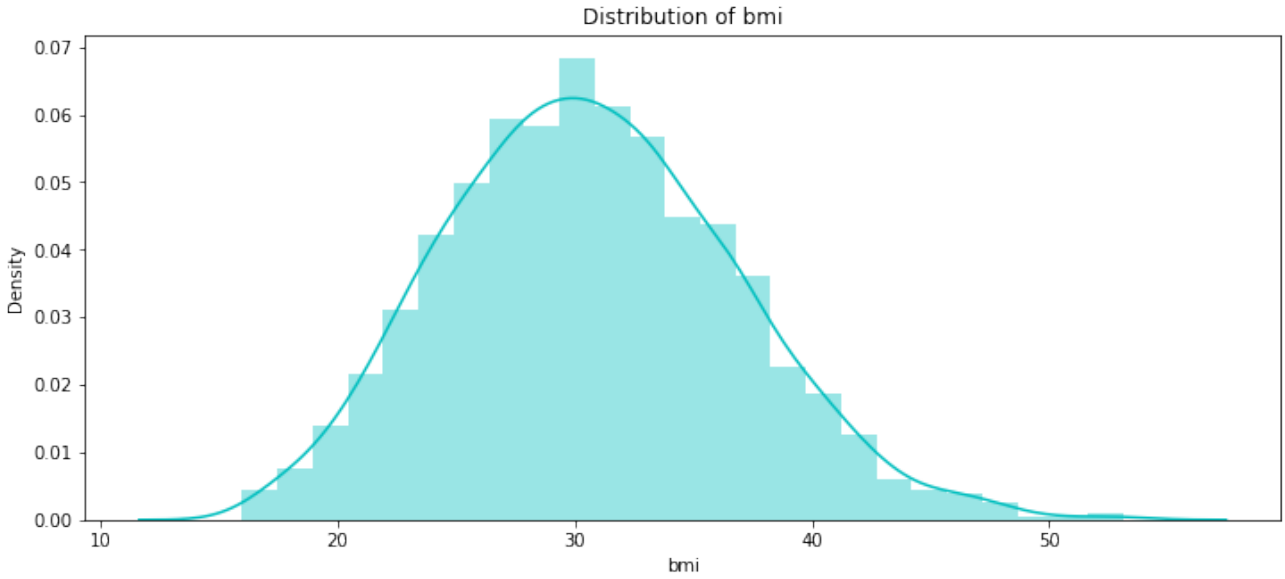


FIGURE 32 – Distribution bmi

— L'IMC moyen des patients est de 30.

— BMI=30 indique l'obésité ,c'est à dire la plupart de sujets souffrent d'obésité . Essayons donc de voir si l'augmentation ou bien la baisse de ce indicateur influencera les charges .

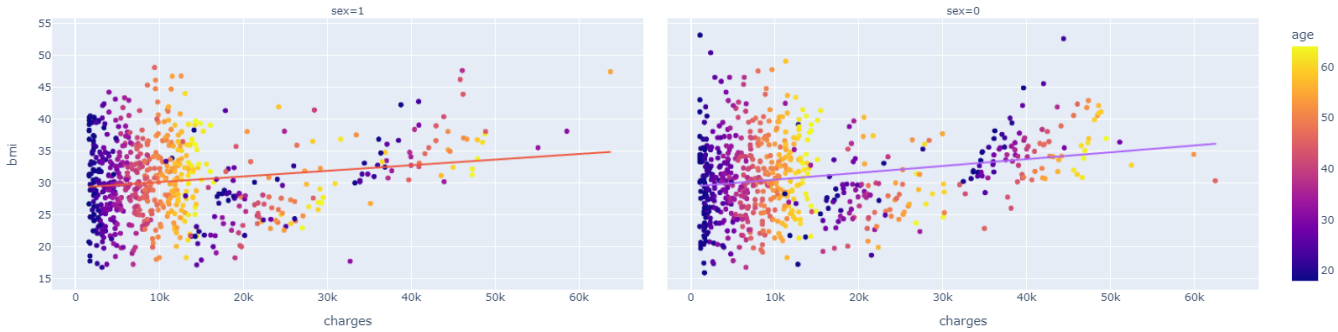


FIGURE 33 – Relation age-charges-sexe-bmi

— Cette figure très informative montre bien que les charges augmentent avec l'age et indice de mace corporelle élevé pour les deux sexes .

Conclusion

Nous pouvons observer les dépenses médicales s' élevent lorsque les gens vieillissent, ce qui est prévisible. Mais, quel que soit l'âge, les fumeurs ont des dépenses médicales plus élevées que les non-fumeurs, comme on l'a déduit précédemment. Il semble vraiment que 'smoker' est la variable la plus importante pour prédire les frais médicaux.

1.3.6 Application du model ML : SVR

1.3.6.1 Principe :

Support Vector Regression (SVR) utilise le même principe que le SVM, mais pour les problèmes de régression. Prenons quelques minutes pour comprendre l'idée derrière SVR .

Le problème de la régression est de trouver une fonction qui se rapproche de la cartographie d'un domaine d'entrée en nombres réels sur la base d'un échantillon d'entraînement. Plongeons maintenant en profondeur et comprenons comment fonctionne réellement le SVR.

Considérez ces deux lignes rouges comme la limite de décision et la ligne verte comme

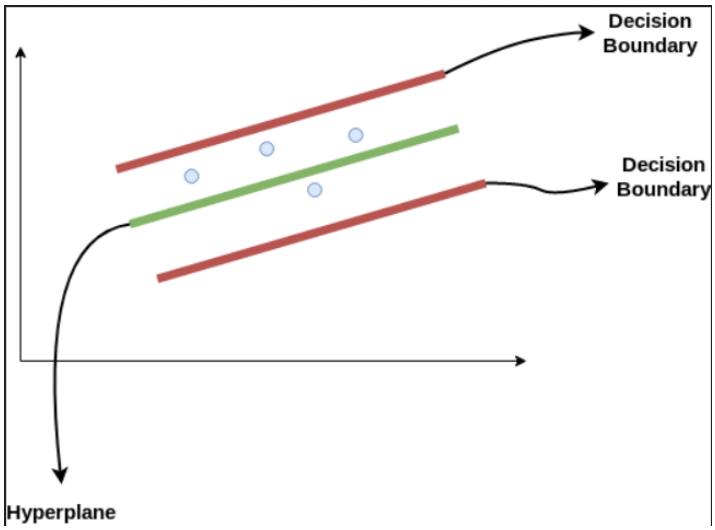


FIGURE 34 – SVR

l'hyperplan. Notre objectif, lorsque nous avançons avec le SVR, est de considérer essentiellement les points qui sont à l'intérieur de la ligne de limite de décision. Notre ligne de meilleur ajustement est l'hyperplan qui a un nombre maximum de points.

La première chose que nous allons comprendre est ce qu'est la limite de décision (la ligne rouge de danger ci-dessus !). Considérez ces lignes comme étant à n'importe quelle distance, disons 'a', de l'hyperplan. Ce sont donc les lignes que nous traçons à la distance '+a' et '-a' de l'hyperplan. Dans le texte, ce "a" est appelé epsilon.

En supposant que l'équation de l'hyperplan soit la suivante :

$$Y = wx + b(equationofhyperplane) \tag{10}$$

Then the equations of decision boundary become :

$$wx + b = +a \tag{11}$$

$$wx + b = -a \tag{12}$$

Ainsi, tout hyperplan qui satisfait notre SVR devrait satisfaire :

$$-a < Y - wx + b < +a \tag{13}$$

Notre objectif principal ici est de décider d'une limite de décision à une distance "a" de l'hyperplan original, de sorte que les points de données les plus proches de l'hyperplan ou des vecteurs de support se trouvent à l'intérieur de cette ligne de limite.

Par conséquent, nous allons prendre uniquement les points qui se trouvent à l'intérieur de la limite de décision et qui ont le taux d'erreur le plus faible, ou qui se trouvent dans la marge de tolérance. Nous obtenons ainsi un modèle mieux adapté.

1.3.6.2 Résultas :

Il existe trois mesures d'erreur qui sont couramment utilisées pour évaluer et rendre compte des performances d'un modèle de régression :

— Mean Squared Error (MSE).

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(y_i - \tilde{y}_i)^2 \tag{14}$$

— Root Mean Squared Error (RMSE).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(y_i - \tilde{y}_i)^2} \tag{15}$$

$$RMSE = \sqrt{MSE} \tag{16}$$

— Mean Absolute Error (MAE)

$$MAE = \frac{1}{N}\sum_{i=0}^{N}|y_i - \tilde{y}_i| \tag{17}$$

où :
-N : le nombre d'observations .
-les y_i : valeurs réelles .
-les $\tilde{y}_i$: valeurs prédites .

Une MSE,MRSE,MAE parfaite est de 0.0, ce qui signifie que toutes les prédictions correspondent exactement aux valeurs attendues.
Fonction de score de régression de la variance expliquée.

Pour notre cas , on a appliqué le SVR pour des données brutes et pour des données transformées pour mettre l'accent sur le rôle du centrage et réduction des données et on a obtenu les résultats suivants :

— MSE = 0.0085 (Données normalisées) ,173353842.8 (Données brutes) .
— RMSE = 0.0927(Données normalisées) ,13166.3 (Données brutes) .
— MAE = 0.0699 (Données normalisées) ,6664.9 (Données brutes) .

Ce chapitre comporte l'élaboration des points essentiels à la réalisation de d'un projet data science : Buisness case , data ,data exploration, data analysis ,date preprocessing , feature engineering et puis l'application du SVM. Et ceci pour trois différents types de problémes :Classification binaire, multiclassification et puis régression . Dans le chapitre prochain , on abordera les moyens et les outils avec laquelle s'est fondue cette réalisation .

2 Chapitre 2 : Environnement logiciel

Ce chapitre comporte tous les outils employés pour la réalisation d'un travail bien structuré et complet .On citera en premier la méthode employée pour la construction d'une structure normalisée et flexible , suivie des languages utilisés , les bibliothéques employées et finalement une brève comparaison entre les languages utilisés .

2.1 Structure

La structure du projet a été produite à l'aide de : cookiecutter Data science qui est une structure de projet logique, raisonnablement normalisée, mais flexible, pour réaliser et partager des travaux de science des données.

2.2 Etude technique

2.2.1 Languages de programmation

2.2.1.1 Python

Python est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions ,populaire chez les data scientist .

FIGURE 35 – Logo de python

2.2.2 Bibliothéques

2.2.2.1 Scikit-learn

Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria .

2.2.2.2 Plotly

La bibliothèque graphique Python de Plotly permet de créer des graphiques interactifs de qualité professionnelle. Exemples de création de diagrammes linéaires, de diagrammes de dispersion, de diagrammes de surface, de diagrammes à barres, de barres d'erreur,

de diagrammes en boîte, d'histogrammes, de cartes thermiques, de sous-points, d'axes multiples, de diagrammes polaires et de diagrammes à bulles.

2.2.2.3 Seaborn

Seaborn est une bibliothèque Python de visualisation de données basée sur matplotlib. Elle fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs.

2.2.2.4 Matplotlib

Matplotlib est une bibliothèque complète permettant de créer des visualisations statiques, animées et interactives en Python. Matplotlib rend les choses faciles faciles et les choses difficiles possibles.