Mini Projet

NBA dataset , source : https://www.kaggle.com/drgilermo/nba-players-stats
Diabetes dataset , source : https://www.kaggle.com/mathchi/diabetes-data-set

---

Présentation des datasets :

NBA Dataset :

J'ai trois documents CSV de Joueurs avec leurs Stats , Positions , Age , Education Universitaires , matches jouées. Et toutes sont distribuées dans les trois Fichiers.

Diabetes Dataset :

J'ai un document CSV avec des paramètres des patients et leur résultats de diabètes.

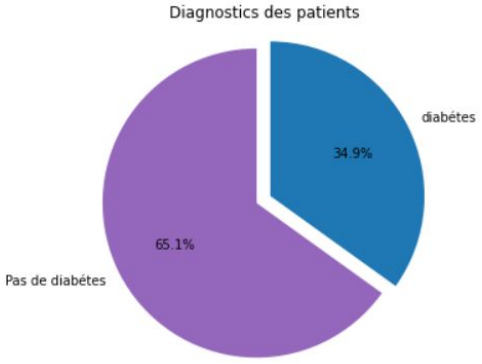
Partie 1 : Analyse graphique des données , " dataset Diabetes "

- Vérification du nombre de données, si plusieurs données sont peu représentées (<3%) alors regrouper dans une seule et même catégorie, 1 pie chart avant/après.
    - Mes données intéressants étaient le nombre de patients qui ont les diabétiques ou non , ce que j'ai pu montré) partir de Pie chart.

Alors il y a 500 personnes sans diabetes et 268 avec diabetes

Présentation fo Pie chart with categorical types

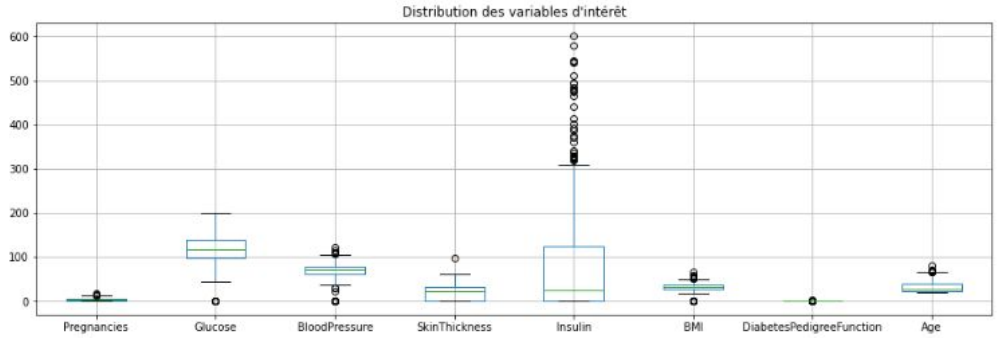
trée [123]:
```
Outcome_proportion=df.Outcome.value_counts()/df.Outcome.value_counts(

fig =plt.figure(figsize=(17,5))
ax2 = fig.add_subplot(1,2,2)
ax2.pie(x=Outcome_proportion, explode=(0,0.1), labels=['Pas de diabét
        colors=('tab:purple','tab:blue'))
ax2.axis('equal')
ax2.set_title("Diagnostics des patients")

plt.show()
```



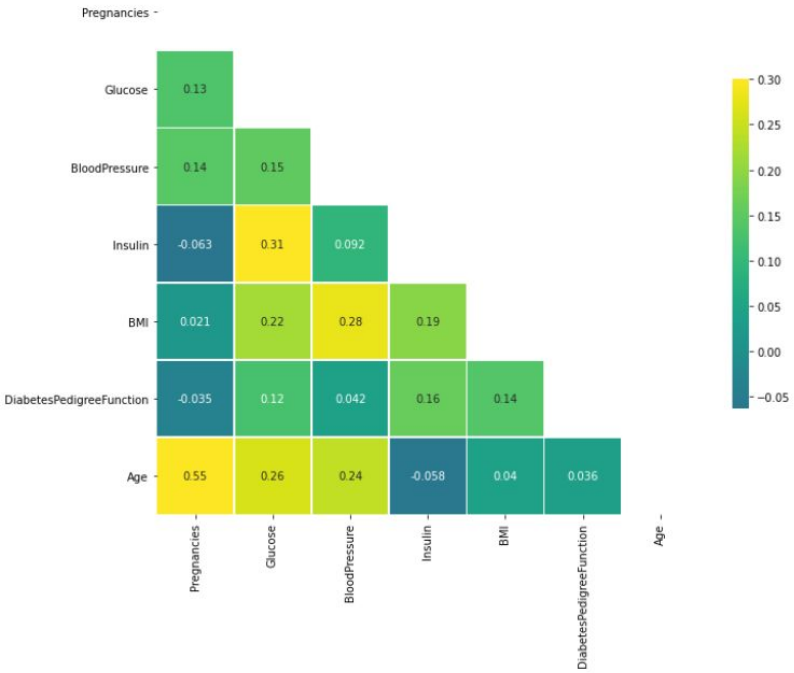
- Nettoyage des données manquantes, encodage
    - La Nettoyage est fait sur le Dataset est plutôt fait de séparer colone : “Outcome” du dataframe puisque c’est une colonne du type de donnée catégoriale.
    - Il y a aussi le nettoyage qui est faire sur NBA dataset , la suppression des colonnes inutiles en plus de merger et combiner les colonnes vues utiles pour faire une exemple de régression linéaire
- Boîtes à moustache avec données extrêmes
    - J’ai visualisé mon dataset pour voir les extrêmes des points s’il y avait des grandes valeurs qui sont pas nécessaires mais quand même pas négligeable afin d’avoir une meilleur résultat.

Entrée [120]:
```
data=df
data=data.drop(['Outcome'],axis=1)
data = data.drop(data[data.Insulin > 600 ].index)
```

Entrée [121]:
```
plt.figure(figsize=(16,5))
data.boxplot()
plt.title("Distribution des variables d'intérêt")
plt.show()
```



- Heatmap + observations sur les corrélations (Diabetes Dataset)
    - On observe ici que les parties qui ont moins foncées alors où c'est susceptible que le plus de patient aura de diabètes mais la forme de



Heatmap indique les différents patients qui peut y avoir diabètes , il y a pas une corrélation significative.

Partie 2: Model Building (NBA , Diabetes Datasets)

- 2 algorithmes avec 2 paramètres différents que vous expliquerez en commentaire
    - Pour (Diabetes Dataset) :

C'est le variable d' "Outcome" , " Résultats si quelqu'un a de diabètes ou non" , et les autres paramètres de Data ['Pregnancies', 'Glucose', 'Blood Pressure', 'Insulin', 'BMI' , 'Diabetes Pedigree Function', 'Age'] et le score eu est 24% accuracy , ce qui est bien parce qu'ils nous montre que la maladie vient pas directement de tous ces paramètres mais que peut être plutôt un variable génétique.

    - NBA Dataset :

Important les Matchs Jouées : (G_x) , Les Points gagnées par les Joueurs : (PTS_x) And The Players. La corrélation était simple et vraie puisque par rapport aux matchs jouées c'est sûre que les Joueurs auront des points en totalité de plus.

- Affichage des coefficients/ accuracy
    - Diabetes Dataset :

```
Entrée [133]: regr.score(X_test,y_test)
Out[133]: 0.24843386287506763
```

    - NBA Dataset :

```
Entrée [302]: regr.score(X_test,y_test)
Out[302]: 0.7483469964248707
```