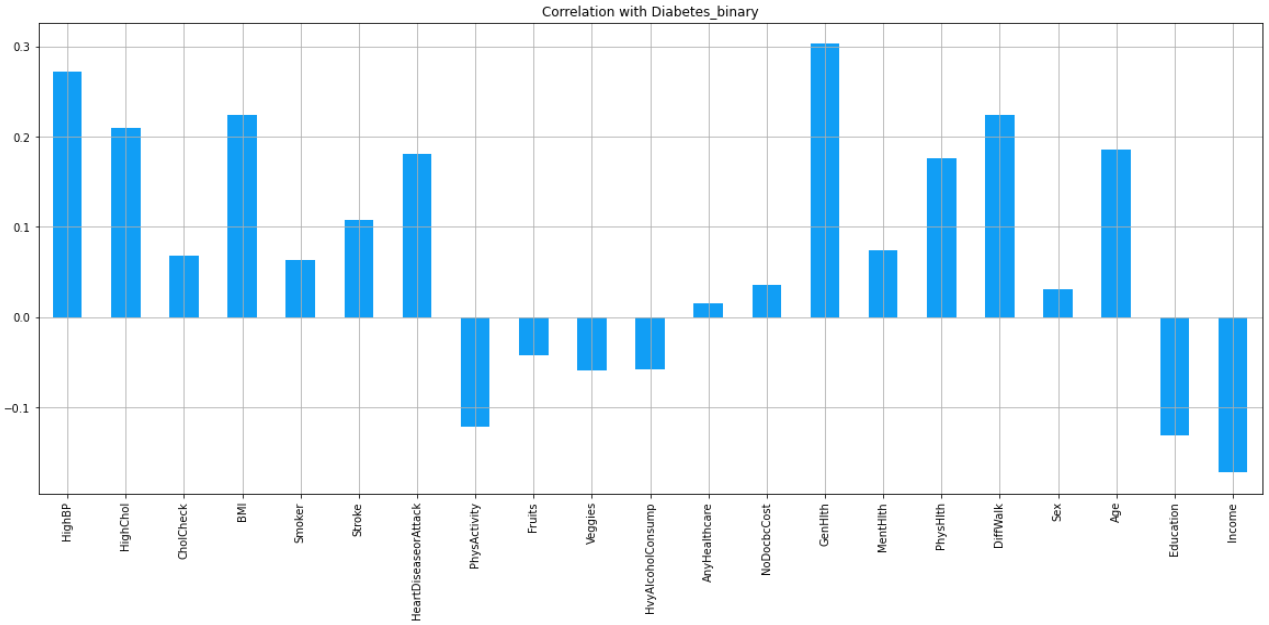
Artificial Intelligence of Things
Programming Practice 1

a.Print out <dataset>.info() after removing low correlation columns.
- Please provide the reason(s) for the exclusion. (15%)



根據上圖不同欄位與「是否有糖尿病」此 label 的相關程度圖表，可發現有幾個欄位的相關程度較低。在這裡我選擇移除最不相關的四種欄位，分別是：'AnyHealthcare', 'Sex', 'NoDocbcCost', 'Fruits'。

截圖如下：

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 18 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_012          253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Veggies               253680 non-null  float64
 10  HvyAlcoholConsump     253680 non-null  float64
 11  GenHlth               253680 non-null  float64
 12  MentHlth              253680 non-null  float64
 13  PhysHlth              253680 non-null  float64
 14  DiffWalk              253680 non-null  float64
 15  Age                   253680 non-null  float64
 16  Education             253680 non-null  float64
 17  Income                253680 non-null  float64
dtypes: float64(18)
memory usage: 34.8 MB
```

b.Print out model.score() of training data and test data using Decision Tree method (20%)

```
---DecisionTreeClassifier---
train acc: 0.9801798060387689
test acc: 0.9674785334924312
```

c.Print out model.score() of training data and test data using Random Forest method (20%)

```
---RandomForestClassifier---
train acc: 0.9948058632912578
test acc: 0.9939519430991319
```

d.Rationally select (or design) an ML model for the analysis of diabetes prediction. Discuss your reason(s).
- Print out each value of the accuracy of training data and test data using your selected method (10%)

前面使用的 Decision Tree 跟 Random Forest 都是基於邏輯判斷、樹狀的模型，也因此我想到 XGboost 這個同樣是建立樹，而且透過其他演算法改良更加強大的模型。我想知道這樣的模型是否能達成更好的結果。

參數的部分我參考了前兩個模型範例程式的設定，將 max_depth 與 n_estimators 設一樣值，沒有再多做調整，分數如下：

```
---xgboostModel---
train acc: 0.9947824662790562
test acc: 0.990945462202569
```

此外我也曾嘗試使用如：LogisticRegression、KNN、SVM、Naive Bayes 等方法，不過它們要不是成果較差(準確度約 60%)，要不就是因為數據數量過大而非常費時，因此最後沒有選擇使用。

比較三種模型的結果，Random Forest 與 XGboost 表現優於 Decision Tree，我推測原因是 Decision Tree 只建一顆決策樹，然而 Random Forest 跟 XGboost 會生成多棵樹讓結果更好。此外Random Forest 跟 XGboost 的結果並沒有明顯差異，我認為是這個資料集較為單純，像是只有兩種類別、欄位維度不高，因此 XGboost 的優化技巧沒有展現出來。

Decision Tree	Random Forest	XGboost
array([[40015, 2656], [124, 42687]], dtype=int64)	array([[42290, 489], [28, 42675]], dtype=int64)	array([[41958, 731], [43, 42750]], dtype=int64)

最後，除了準確率以外，我也有查看他們的 confusion matrix，整理成上方表格。整體來說，同樣是 Decision Tree 比較無法準確的預測是否有糖尿病。另外我也發現這三種模型都是在 type 1 error 表現沒那麼好，推測是資料集本身的緣故，也許更細緻的處理資料能有更進一步的提升。