# DiabeteStudy

## Progetto di Ingegneria della Conoscenza


### Componenti del gruppo:

• **D**onatello Scigliuto
Matricola: 707052, email: d.scigliuto@studenti.uniba.it

• **L**uigi Vulcano
Matricola: 698958, email: l.vulcano@studenti.uniba.it

• **C**ristiana Sorrenti
Matricola: 682718, email: c.sorrenti3@studenti.uniba.it


Link del repository (GitHub):
https://github.com/DLC-Adventure/DiabeteStudy

Link diretto del progetto, su Google Colab:
https://colab.research.google.com/drive/1K2E_z7aH304hTbYnFeFt3WGoXnvnwSgi

### Tecnologie utilizzate:

Il campo di lavoro utilizzato è **Google Colaboratory**, uno strumento web che si è rivelato molto efficace per la programmazione in **Jupyter Notebook** *(.ipynb)*.


Il progetto è stato anche convertito in **Python** *(.py)*, nell'eventualità che debba essere eseguito su una macchina in locale.

## Introduzione:

Il progetto consiste in un sistema di classificazione che, sulla base di misurazioni diagnostiche, prevede se un paziente ha il diabete.

In particolare, si occupa di:
- Classificare secondo vari algoritmi e valutarne le prestazioni.
- Calcolare la probabilità con cui un soggetto possa avere il diabete.

## Dataset utilizzato:

Il dataset utilizzato contiene osservazioni su 768 pazienti di origine "Pima Indian" (indiani d'America/nativi americani).

Link sorgente: https://www.kaggle.com/mathchi/diabetes-data-set

Le feature presenti nel dataset sono:
- Pregnancies (Gravidanze): Numero di volte in gravidanza.
- Glucose (Glucosio): Concentrazione di glucosio a 2 ore da un test orale.
- BloodPressure (Pressione sanguigna): Pressione sanguigna minima.
- SkinThickness (Spessore della pelle): Spessore della piega cutanea del tricipite.
- Insulin (Insulina): 2 ore da un siero di insulina.
- BMI (Body Mass Index=Indice di massa corporea): [peso/(altezza^2)]
- DiabetesPedigreeFunction (Predisposizione genealogica al diabete)
- Age (Età)

Ed una variabile target:
- Outcome (Esito): 0=non diabetico, 1=diabetico.

# 1. Lettura dei dati ed analisi esplorativa

## 1.1) Import del dataset e breve panoramica dei dati

Tramite la stampa di una tabella, mostriamo a schermo una breve panoramica dei dati per verificare l'effettivo import del dataset.

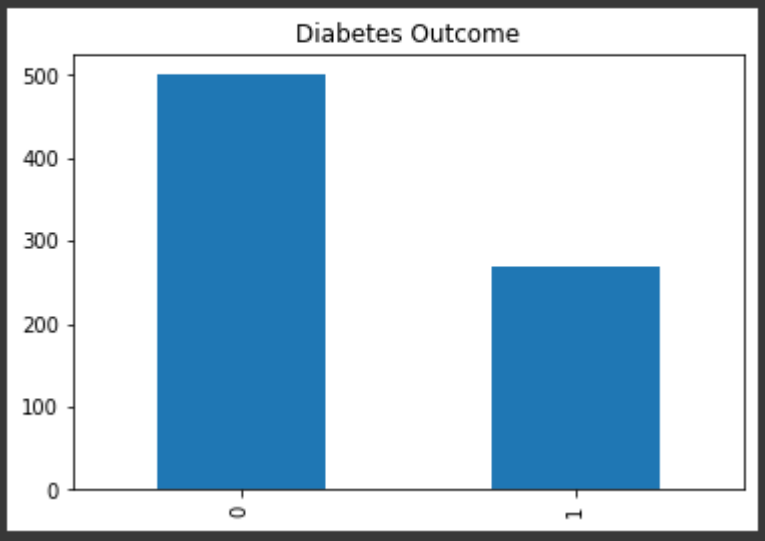
```
Righe (pazienti) : 768
Colonne (features) : 9
```

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| **1** | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| **2** | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| **3** | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| **4** | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

## 1.2) Bilanciamento delle classi

Attraverso un grafico, visualizziamo la proporzione dei pazienti non diabetici (0) e di quelli diabetici (1).

```
Pazienti non diabetici: 500 (% 65.10)
Pazienti diabetici: 268 (% 34.90)
```



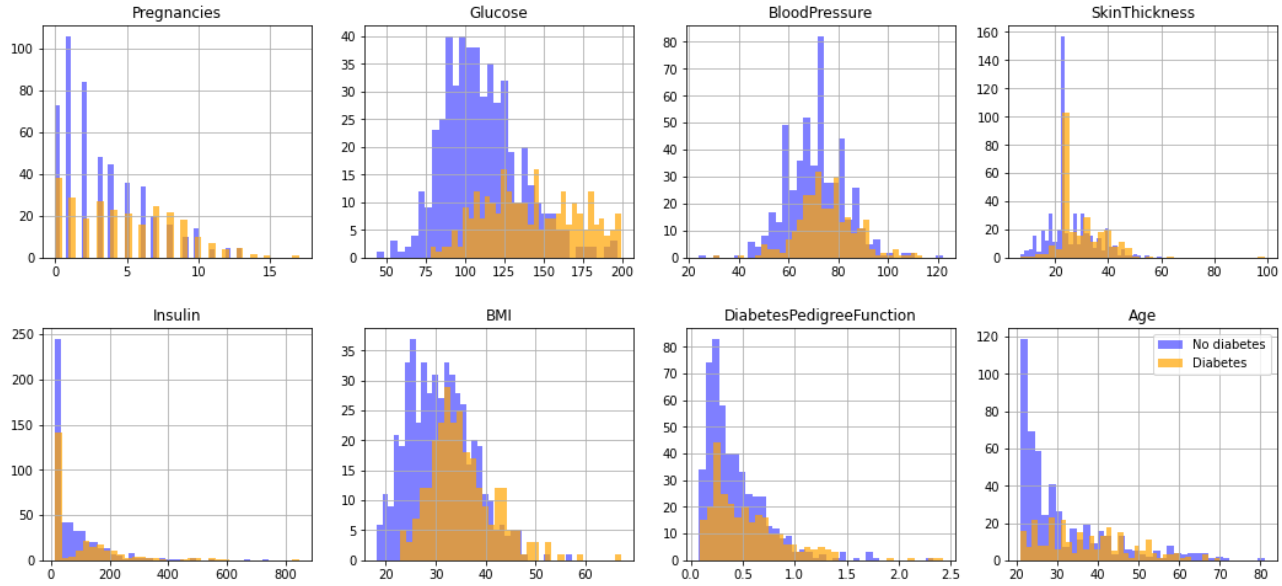

Possiamo notare che le classi sono leggermente sbilanciate.

## 1.3) Sostituzione degli zero con la mediana

Ci sono dei valori zero nel dataset (in particolare in Glucose, BloodPressure, SkinThickness, Insulin e BMI) che influenzeranno sull'accuratezza del training, quindi li sostituiamo con il valore mediano.

## 1.4) Analisi esplorativa dei dati

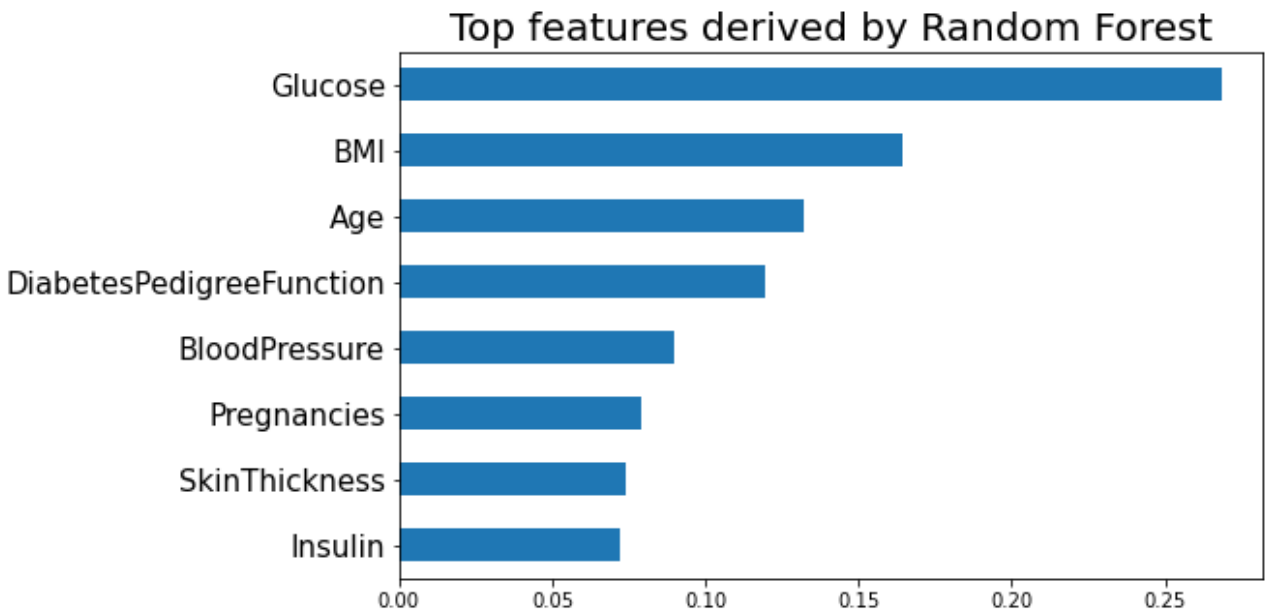
Analizziamo la distribuzione dei risultati delle feature.




Mostriamo a grafico la quantità di valori per ogni feature, così da avere una panoramica generale.

## 1.5) Verifica dell'importanza delle feature

Tramite il classificatore Random Forest, esaminiamo l'importanza di ogni singola feature.




Possiamo notare che Glucosio e BMI sono le caratteristiche mediche predittive più importanti per la diagnostica del diabete.

## 2. Model Selection:

Per il Model Selection utilizziamo la K-Fold Cross Validation. Dal momento che le classi sono in squilibrio, utilizziamo in particolare lo Stratified K-Fold.

I modelli valutati sono:
- KNN (K-Nearest Neighbors)
- Decision Tree
- Random Forest
- SVC (Support-Vector Classification)
- Logistic Regression

Le metriche di performance utilizzate nella valutazione sono:
Accuratezza, Precisione, Richiamo, F1-score.

I risultati della valutazione sono:

| | model | accuracy | precision | recall | f1score |
|---|---|---|---|---|---|
| 0 | KNN | 0.679739 | 0.534483 | 0.584906 | 0.558559 |
| 1 | DecisionTree | 0.725490 | 0.627907 | 0.509434 | 0.558559 |
| 2 | RandomForest | 0.758170 | 0.690476 | 0.547170 | 0.610526 |
| 3 | SVC | 0.771242 | 0.750000 | 0.509434 | 0.606742 |
| 4 | LogisticRegression | 0.758170 | 0.710526 | 0.509434 | 0.593407 |

Basandoci sulle performance dell'F1-score, possiamo dire che il modello migliore è il Random Forest.

## 3. Rete bayesiana:

### 3.1) Creazione della rete bayesiana

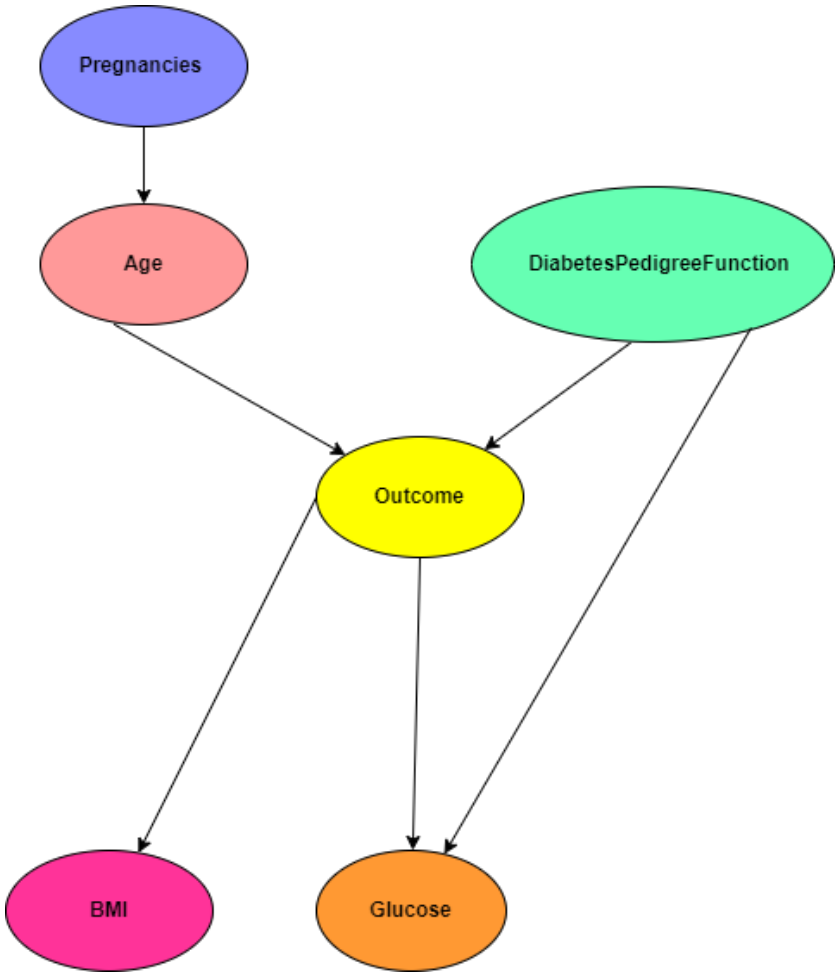
Per poter successivamente effettuare il calcolo della probabilità, abbiamo creato una rete bayesiana. Così la predizione del diabete avverrà in base ai nostri fattori, ovvero le feature.

Per accertarci dell'effettiva creazione della rete, visualizziamo a schermo i nodi e gli archi della rete.

```
Nodi della rete:
 ['Pregnancies', 'Age', 'Insulin', 'SkinThickness', 'DiabetesPedigreeFunction', 'Outcome', 'Glucose', 'BMI']

Archi della rete:
 [('Pregnancies', 'Age'), ('Age', 'Outcome'), ('Insulin', 'SkinThickness'),
  ('DiabetesPedigreeFunction', 'Outcome'), ('DiabetesPedigreeFunction', 'Glucose'),
  ('Outcome', 'Glucose'), ('Outcome', 'BMI')]
```

La rete bayesiana costruita è la seguente:



Possiamo quindi scrivere la funzione di probabilità.

Data la formula:

$$\boldsymbol{P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i))}$$

La nostra funzione di probabilità è:

$$P(Pregnancies, Age, DiabetesPedigreeFunction, Outcome, BMI, Glucose) = P(Pregnancies) * P(Age \mid Pregnancies) * P(DiabetesPedigreeFunction) * P(Outcome \mid Age, DiabetesPedigreeFunction) * P(BMI \mid Outcome) * P(Glucose \mid Outcome)$$

### 3.2) Calcolo della probabilità

Sfruttando la rete bayesiana precedentemente creata, calcoliamo la probabilità per un soggetto presumibilmente non diabetico (0) ed uno diabetico (1) di avere il diabete.

```
Probabilità per un soggetto potenzialmente non diabetico:
+------------+----------------+
| Outcome    |   phi(Outcome) |
+============+================+
| Outcome(0) |         0.8344 |
+------------+----------------+
| Outcome(1) |         0.1656 |
+------------+----------------+


Probabilità per un soggetto potenzialmente diabetico:
+------------+----------------+
| Outcome    |   phi(Outcome) |
+============+================+
| Outcome(0) |         0.4286 |
+------------+----------------+
| Outcome(1) |         0.5714 |
+------------+----------------+
```

Per un soggetto potenzialmente non diabetico, ovvero un paziente con dei valori diagnostici tali che precludono la possibilità che abbia il diabete, la probabilità che possa avere il diabete è del 16%. La probabilità che effettivamente non abbia il diabete è quindi dell'83%.

Al contrario, per un soggetto potenzialmente diabetico, ovvero un paziente con dei valori diagnostici tali da incrementare la possibilità che abbia il diabete, la probabilità che effettivamente abbia il diabete è del 57%. La probabilità che invece non abbia il diabete è quindi del 42%.