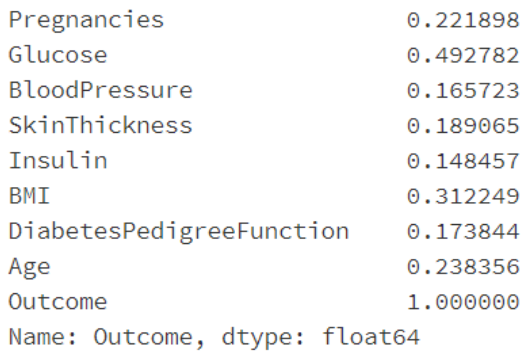
**Service-Oriented Software Engineering (Project Proposal)**

# Nice Title

*Group Members (5): ZijiaHe, ChangXu, HongyiHuang, WangzhihuiMei*

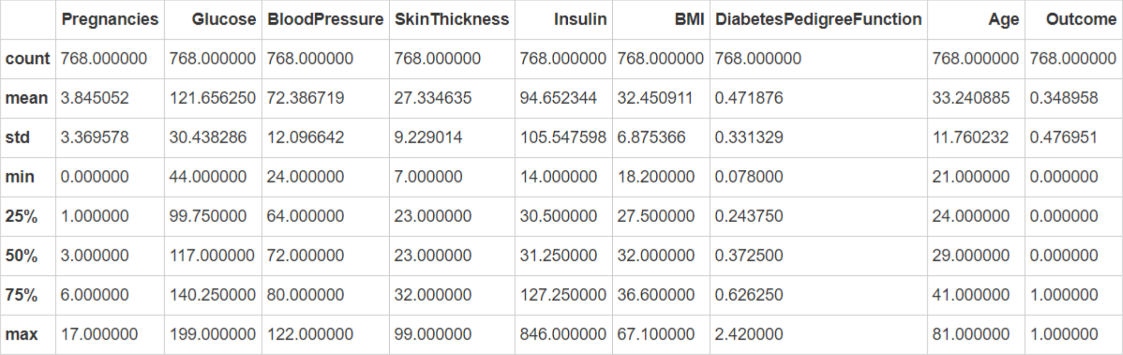
## 1 Data preparation

There are 8 features (explanatory variables) and 1 label (response variable). These data are collected from actual patients and represent a task, usually performed by a human doctor, with the purpose of identifying the patients most likely to have diabetes in order to propose preventive measures.Some data values are 0, which is impossible, because these physical quantities cannot be 0 (for living people). Therefore, this has told us that we need to estimate these five columns. The scope of other variables seems to be reasonable.Next, we can calculate the relevant values to see the relationship between the characteristics and the results. Of course, correlation does not mean causality, but because we are building a linear model, correlation features may be useful for learning the mapping between patient information and whether they have diabetes.In problems with a large number of features, we can use relevant thresholds to delete variables. In this case, we may want to keep all variables and let the model decide which ones are related.In this brief exploratory data analysis, we learned about the two main aspects of data sets that can be used for modeling. First, we need to enter missing values in several columns, because these values are physically impossible. We can use the median method as a simple and effective way to fill the value of 0. We also learned that there is a correlation between features and responses, although the correlation is not strong. In addition, all features are at least slightly positively correlated with the result (whether or not the patient has diabetes).Next, in order to facilitate testing and training, we randomly select 75% of the data set for training and 25% of the data set for testing.



```
Pregnancies                 0.221898
Glucose                     0.492782
BloodPressure               0.165723
SkinThickness               0.189065
Insulin                     0.148457
BMI                         0.312249
DiabetesPedigreeFunction    0.173844
Age                         0.238356
Outcome                     1.000000
Name: Outcome, dtype: float64
```

**Fig. 1.** Correlation



| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 |
| **mean** | 3.845052 | 121.656250 | 72.386719 | 27.334635 | 94.652344 | 32.450911 | 0.471876 | 33.240885 | 0.348958 |
| **std** | 3.369578 | 30.438286 | 12.096642 | 9.229014 | 105.547598 | 6.875366 | 0.331329 | 11.760232 | 0.476951 |
| **min** | 0.000000 | 44.000000 | 24.000000 | 7.000000 | 14.000000 | 18.200000 | 0.078000 | 21.000000 | 0.000000 |
| **25%** | 1.000000 | 99.750000 | 64.000000 | 23.000000 | 30.500000 | 27.500000 | 0.243750 | 24.000000 | 0.000000 |
| **50%** | 3.000000 | 117.000000 | 72.000000 | 23.000000 | 31.250000 | 32.000000 | 0.372500 | 29.000000 | 0.000000 |
| **75%** | 6.000000 | 140.250000 | 80.000000 | 32.000000 | 127.250000 | 36.600000 | 0.626250 | 41.000000 | 1.000000 |
| **max** | 17.000000 | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 | 2.420000 | 81.000000 | 1.000000 |

**Fig. 2.** Data filled with missing values

## 2 Classifiers