***Assessment 2***

**Assessment type:** Individual

**Max. Marks:** 10

**Last date to submit the assessment:** 20th Sept 2018, 3:00 PM

**Question1:**

**Data Description**

A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases.

These data frames contains the following columns:

1. Npreg: number of pregnancies.
2. Glu : plasma glucose concentration in an oral glucose tolerance test.
3. Bp : diastolic blood pressure (mm Hg).
4. Skin : triceps skin fold thickness (mm).
5. Bmi : body mass index (weight in kg/(height in m)\^2).
6. Ped : diabetes pedigree function.
7. Age : age in years.
8. Type : Yes or No, for diabetic according to WHO criteria.

**TASK:**

1. What is the probability for someone age 35 with bmi of 32, to be diabetic?
2. According to your model what is the probability that a woman in your sample is diabetic given age 55, a bmi 37, bp 68 and npreg of 2?
3. For a woman aged 35 and mother of 2 children, by how much does the probability of diabetes increase, if her bmi was 35 instead of 25 according to the model?
4. What is the accuracy of Models?

NOTE: All answers must be rounded to 2 decimal places.

**Question 2**

**Data Description**

Use data set Titanic

**TASK:**

1. Apply the k-means algorithm to get 4 clusters. Remember to set the seed to 1234 so the results are comparable. Note that you can remove few column, which are not used in K-means.
2. Describe, in words, each of the clusters.
3. Now, 4 clusters was an arbitrary choice. What seems like a reasonable number of clusters?

**Question 3**

**Data Description**

Use the NFL Play by Play 2009-2017 (v4).csv

**Acknowledgements**

This dataset was compiled by Ron Yurko, Sam Ventura, and myself. Special shout-out to Ron for improving our current expected points and win probability models and compiling this dataset. All three of us are proud founders of the Carnegie Mellon Sports Analytics Club.

**TASK**

1. Get the number of missing data points per column.
2. Percent of data that is missing in terms of cells.
3. Find the number of rows after removing all the rows that contain a missing value.
4. Find the number of columns after removing all columns with at least one missing value.
5. Replace all missing value by the mean value of that column.

NOTE: Please submit this assessment in form of an R-markdown file. Remember to set the seed to 1234 so the results are comparable.