Results

Training sets with complete data (T)

```
Logistic regression score T original: 0.967
Tree clf score T original: 0.9485
KNN score T original: 0.9705
```

Training sets with calculated data (T')

```
Logistic regression score T'1: 0.969 (mean)
Logistic regression score T'2: 0.958 (logistic)
Logistic regression score T'3 : 0.969 (knn)

Tree clf score T'1: 0.9525 (mean)
Tree clf score T'2: 0.9485 (logistic)
Tree clf score T'3 : 0.9465 (knn)

KNN score T'1: 0.9705 (mean)
KNN score T'2: 0.969 (logistic)
KNN score T'3 : 0.9705 (knn)
```

Observations

What we did in the lab, was delete data from the training set, then we filled those missing values with 3 different techniques.

1. Using the mean
2. Using logistic regression
3. Using KNN k=20

I was actually very impressed of the results, because even thought we removed data, the results for the T' datasets are very precise and also very similar to the original. In some cases, precision is even slightly greater.

Reflection

As seen in class, in this lab we were able to confirm that even if you have an incomplete dataset or incomplete data, we can calculate the missing information and still have very good precision.