**CS5331: Special Problems in CS:**

**Data Analysis and Machine Learning with R**

**Sasanka Abeysinghe**

**11/16/17**

**Briefly describe the mechanisms of operation of four different data imputation methods.**

Missing data mechanisms can be classified as missing completely at Random (MCAR), missing at Random (MAR) and not missing at Random (NMAR). In MCAR, it is a probability that a data point is missing is completely independent of any data or parameters. In MAR, the probability that point is missing is independent can be modeled by other variables. In NMAR, the probability of data point is missing depends on that value of the variable.

There are number of imputation methodologies.

* Model based
  + Linear regression
  + Decision trees
  + Random forest
* Multivariate imputation
  + Imputation based on expectation-maximization algorithm
  + Iterative random forest imputation
* Donor imputation
  + K-nearest neighbor
  + Sequential hotdeck
  + Predictive mean matching
* Other
  + Mean imputation
  + Proxy imputation where copy another variable to compute imputed values.

KNN (K Nearest Neighbor) is another method of imputation that each sample or individual a set of K-nearest neighbors and then replaces the missing data for a given variable by averaging (non-missing) values of its neighbors.

Random Forest is another method of imputation, and impute missing values in predictor data using proximity from randomForest.

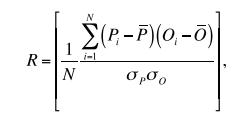
Mean and median imputation method that is used in the respective column in the data set where the mean or the median at a particular point. For each of the column with missing values, the corresponding mean or median of the observed values are calculated and this values can be used to replace missing values.

**Using the iris dataset, randomly create missing values to occupy x% of the data. Use R to apply 3 different methods of imputation and compare their performance when x=2, 5, 10 , 15, 20, 25 (All code snippets must be included in the report appendices and on github). Use the following methods to evaluate the performance of the methods:**

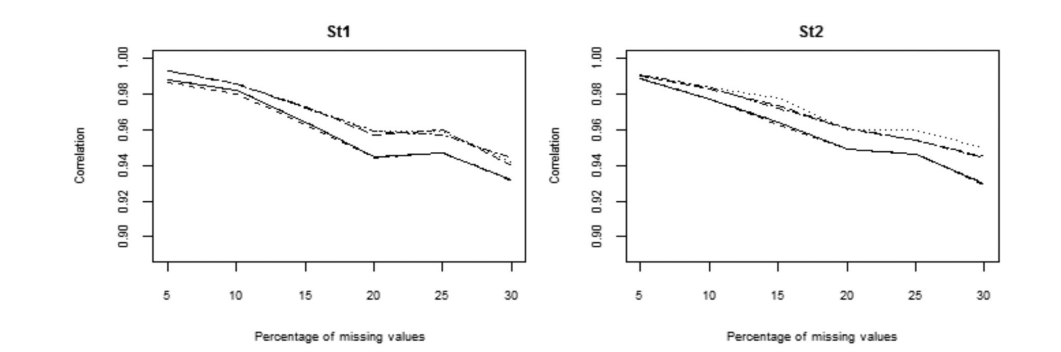
According to my first reference, A comparison of Varioud Imputation Methods for Missing Values in Air Quality Data, to evaluate the imputation methods, there are three performance indicators are used such as the correlation coefficient (R), the index of agreement (d) and the mean absolute error (MAE). So that, these methods are taking consideration between predicted and their corresponding observed values to select best method for predicting missing data.

* The correlation coefficient (R)

This explains about the variability in the imputed data and how much they are related to the observed data.



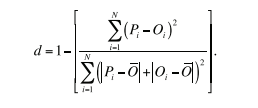
N – Number of imputations, Oi – Observed data point, Pi – Imputed data points, O’- the average of observed data, P’-the average of imputed data,  - standard deviation of imputed data, and  - standard deviation of observe data.



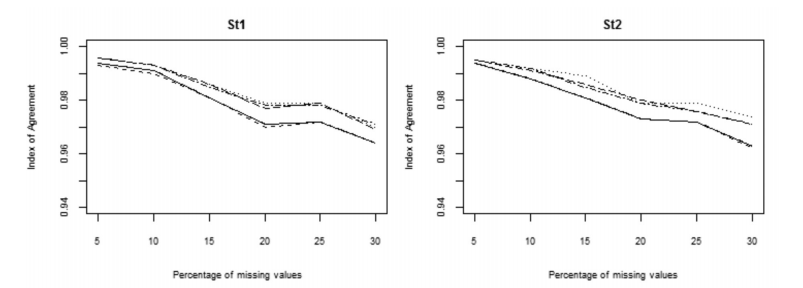


EM, KNN, and SKNN have superior compared to mean, median, and SVD.

Index of Agreement (d) is a measure of relative error between imputed and observed data.



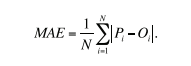
According to the above journal, the best performance with highest index agreement are EM, KNN,

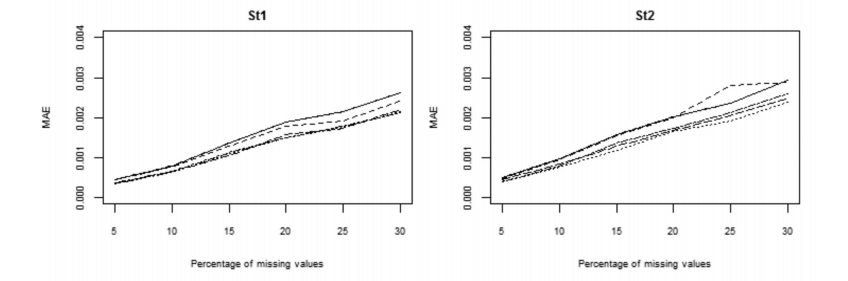




SKNN compared to mean, median, and SVD.

Mean absolute error (MAE) is the average difference between imputed and observed data point.

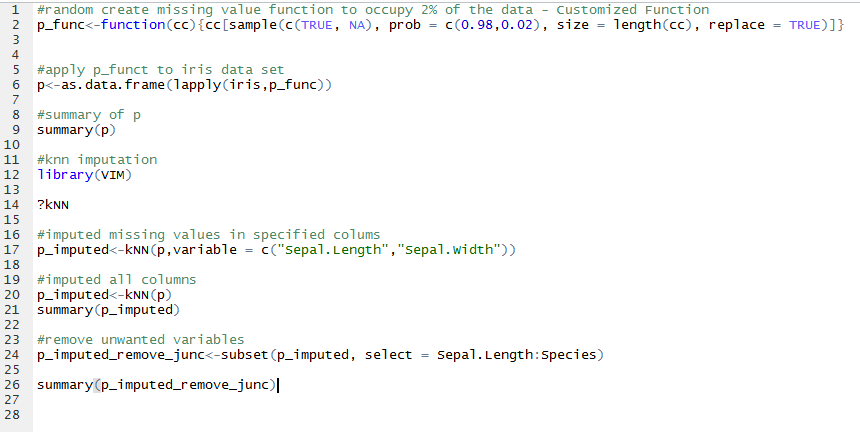


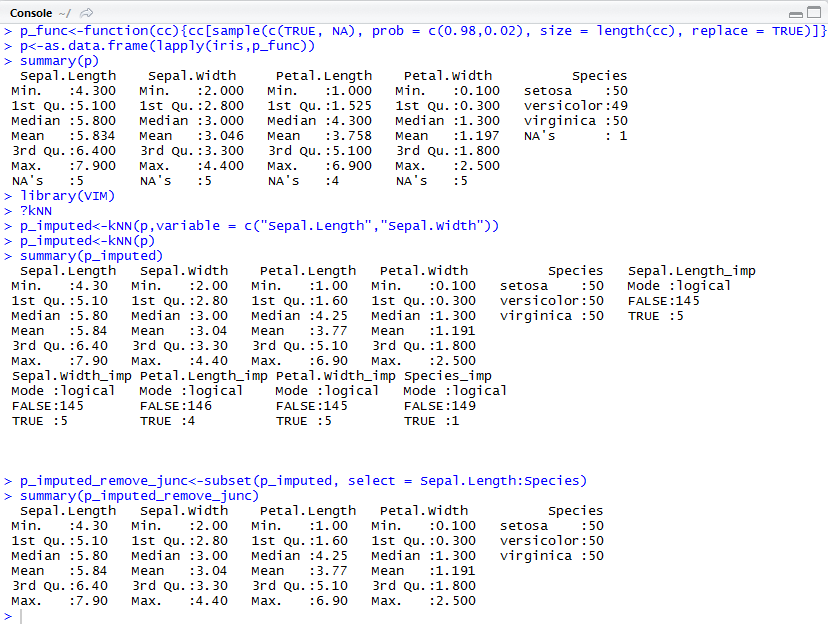




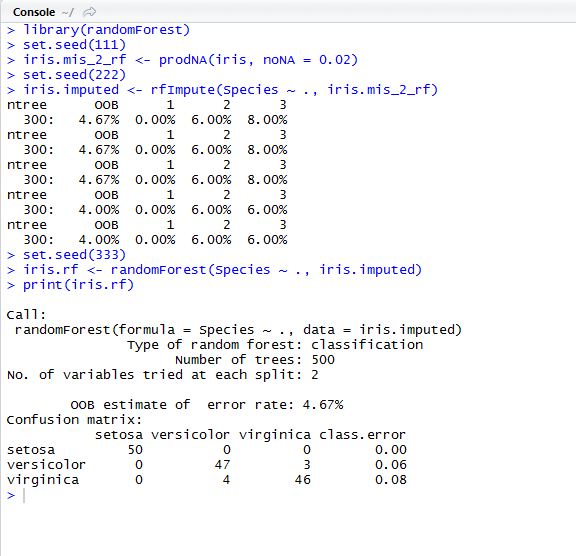
For 2% missing values

KNN (K Nearest Neighbor)

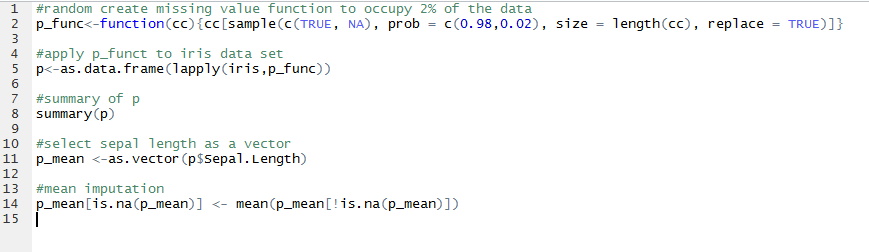


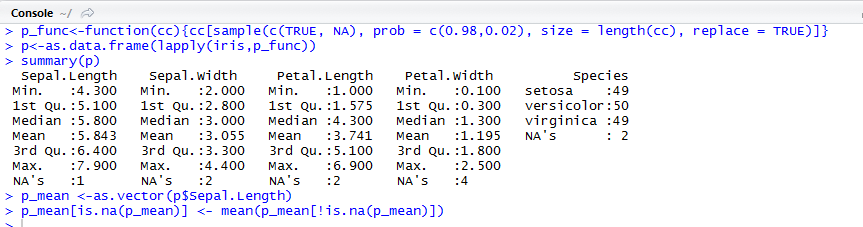


Random Forest

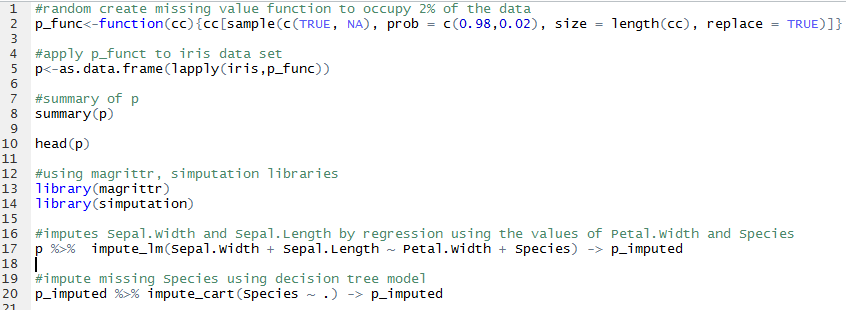


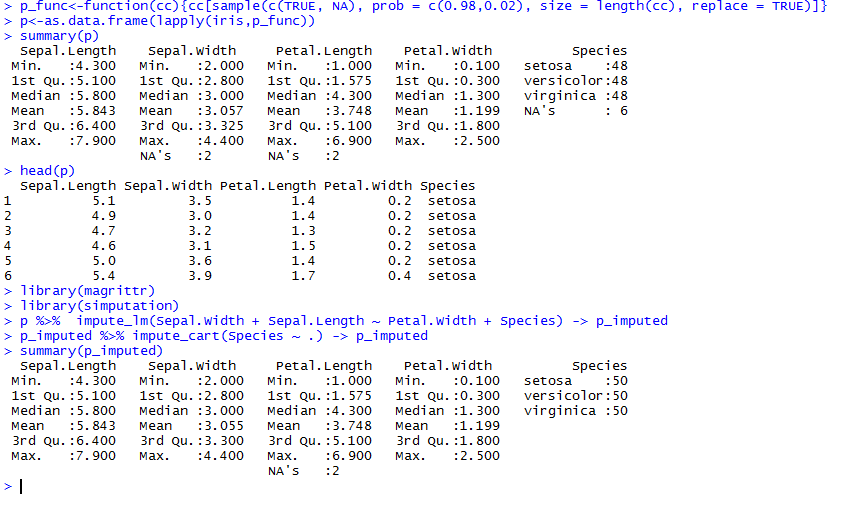
Mean



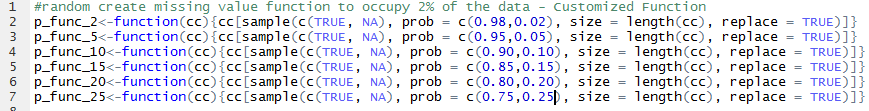


Regression

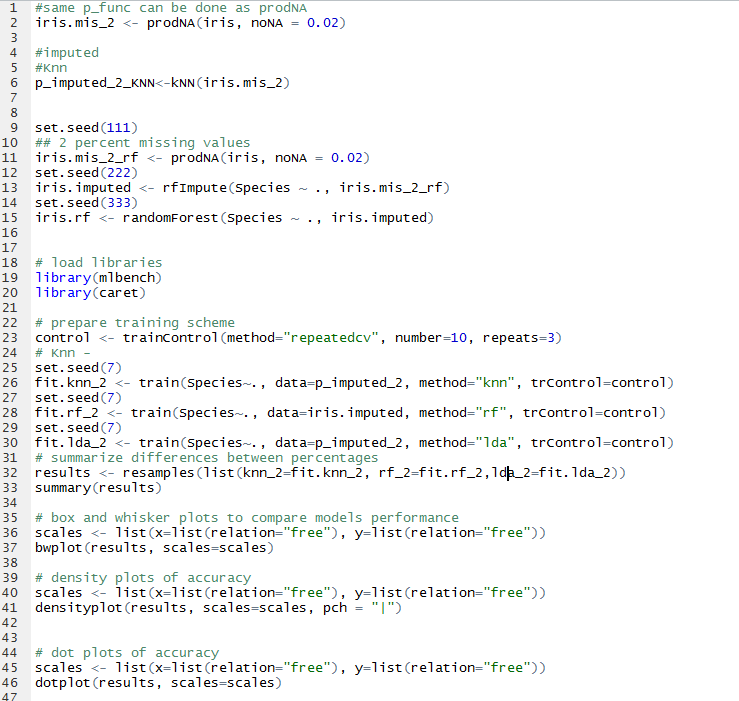


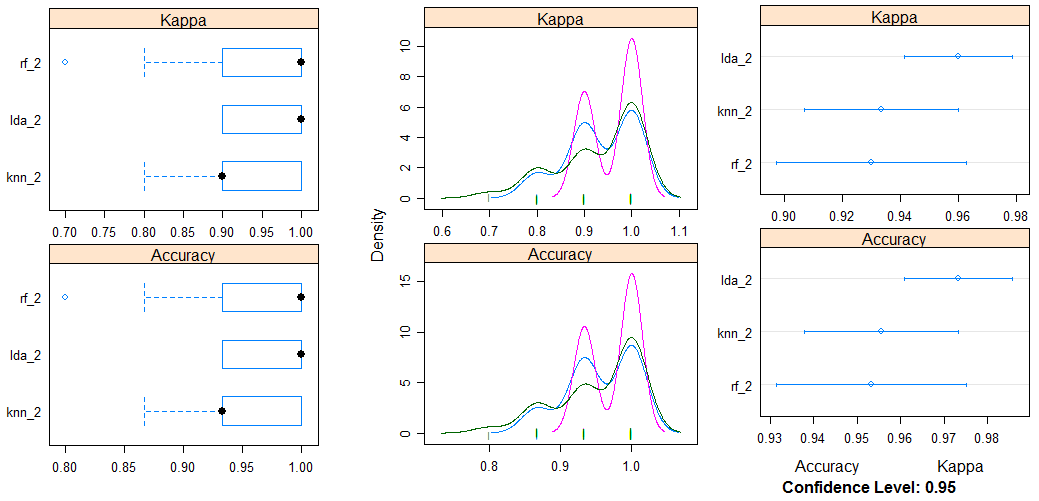


This can be extended to 2%-25% as including these customize functions



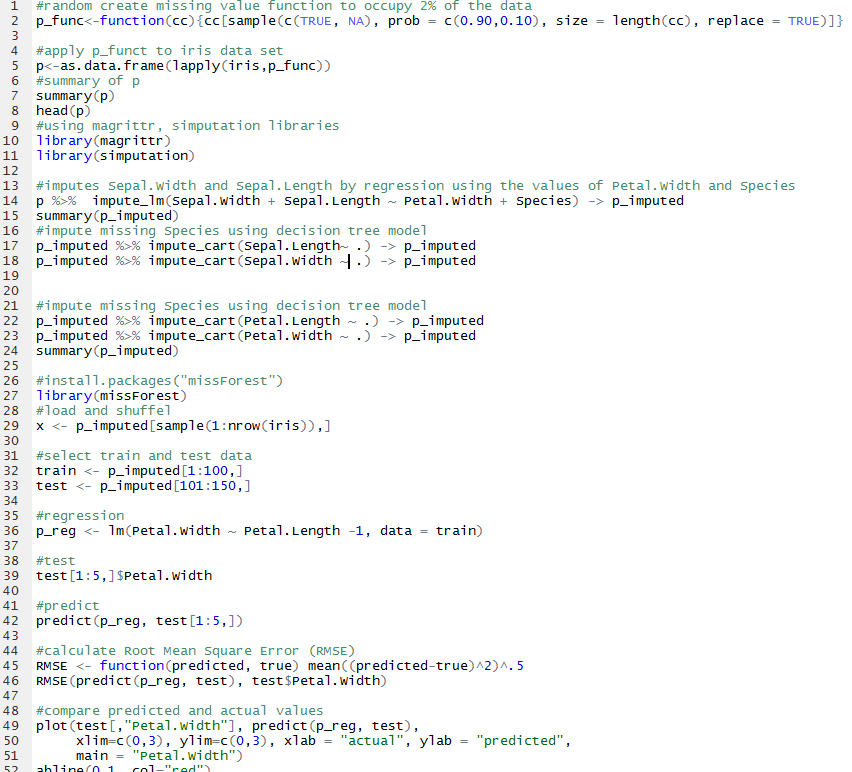
Performance

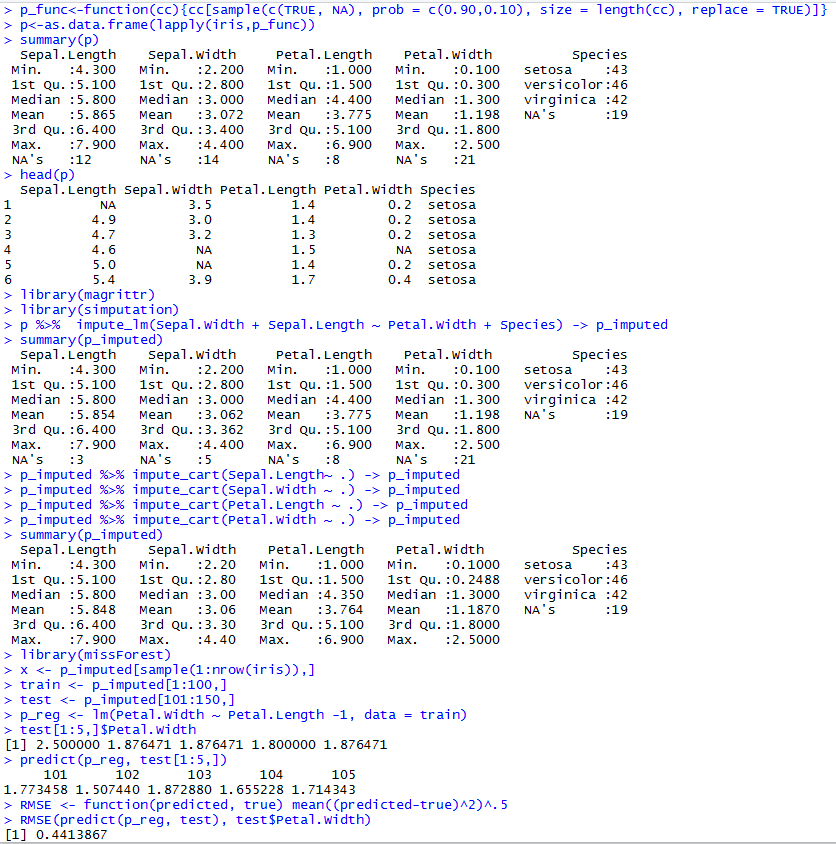


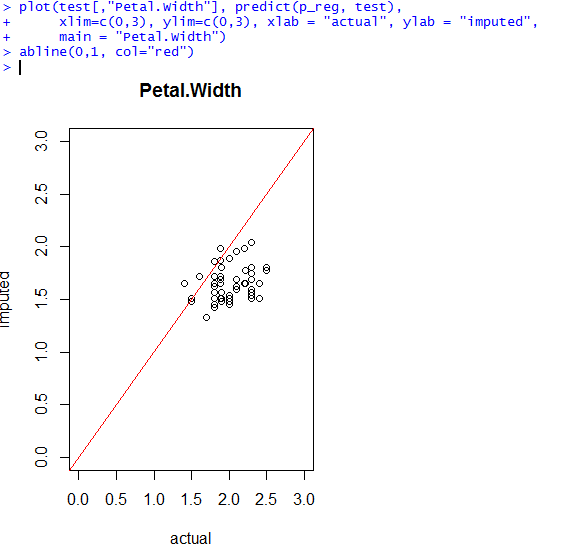


Performance can be investigated by comparing models by differentiation of visualization techniques. For that, we need to prepare iris dataset, train models, and the compare models after imputation methods applied. Here I have used Linear Discriminant Analysis, K-nearest neighbor and Random Forest methods. After these models are trained, then it is added to a list and resamples on the list of models. The function compare each models and they used same training scheme to evaluate metrics for each fold and each repeat for above three algorithm to be evaluated. Using Box and Wisker Plots, Density plots, Dot plots help to identify performance of each algorithms used.

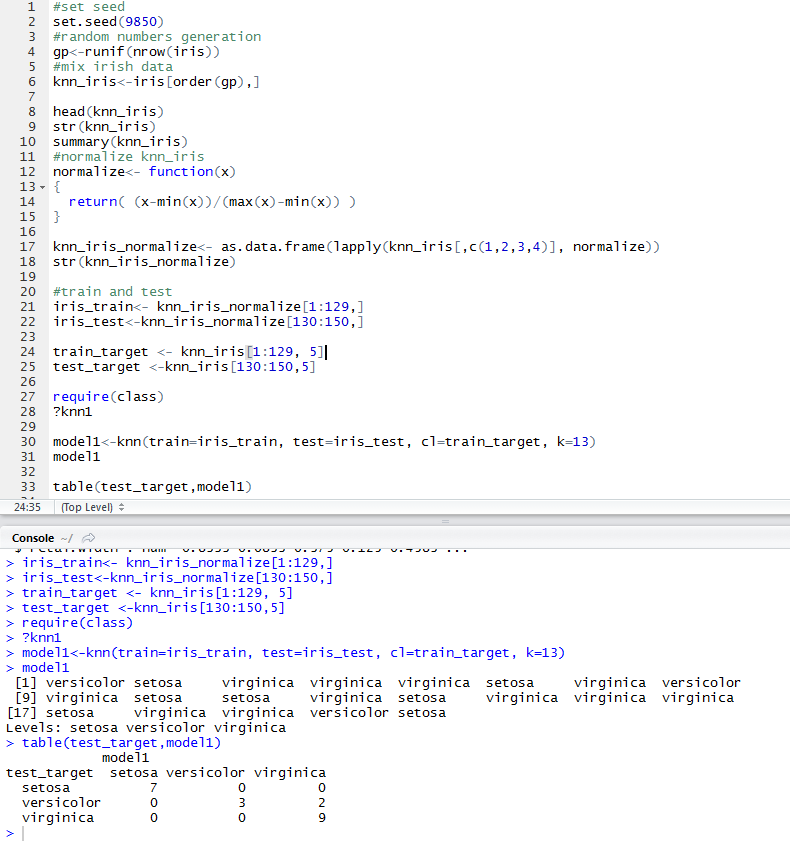
**Root mean square error (RMSE) between imputed and true values**







**Supervised classification error (Use k-NN classifier)**



Classification Error = 2/21 = 0.09523

This can be extended and create classification error matrix by increasing/ decreasing K value and by selecting different sets of train/test of data samples.

Reference

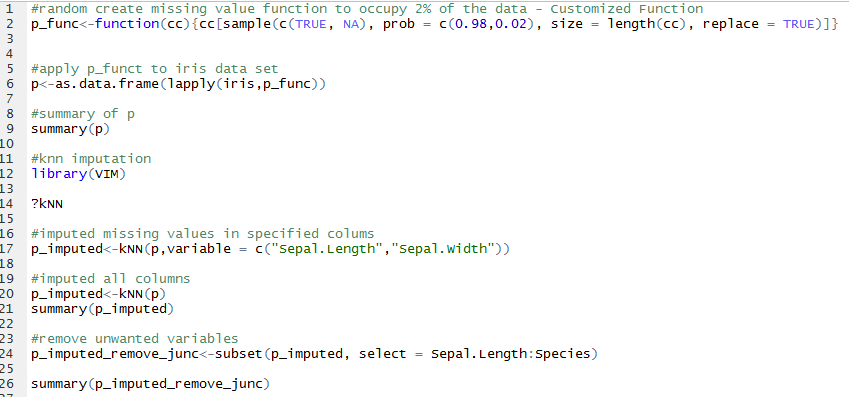
[1] NURYAZMIN AHMAT ZAINURI\*, ABDUL AZIZ JEMAIN & NORA MUDA, A Comparison of Various Imputation Methods for Missing Values in Air Quality Data, Sains Malaysiana 44(3)(2015): 449–456

[2] R Documentation, Missing Value Imputations by randomForest, retrieved on 11/16/2017

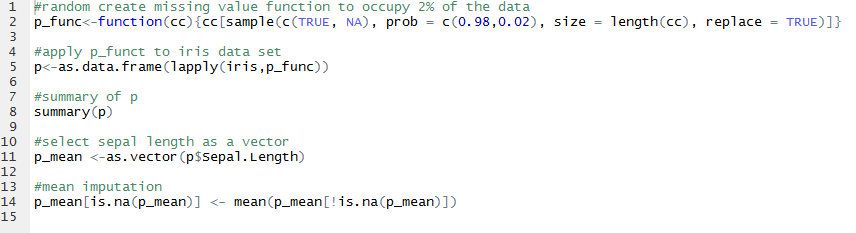
<http://math.furman.edu/~dcs/courses/math47/R/library/randomForest/html/rfImpute.html>

Appendix

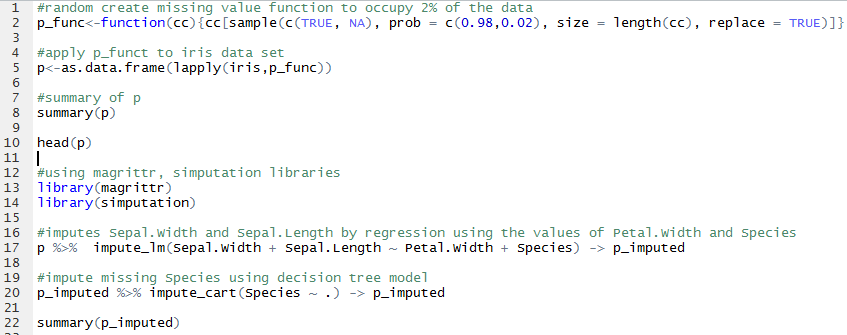
Knn



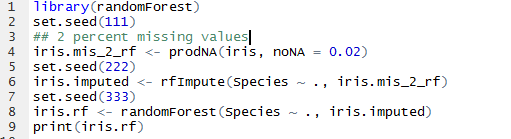
Mean



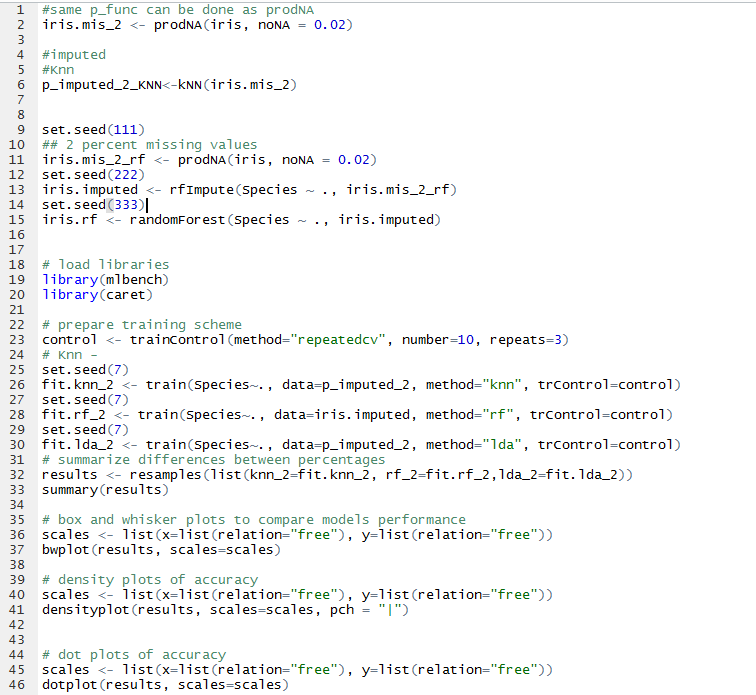
Regression



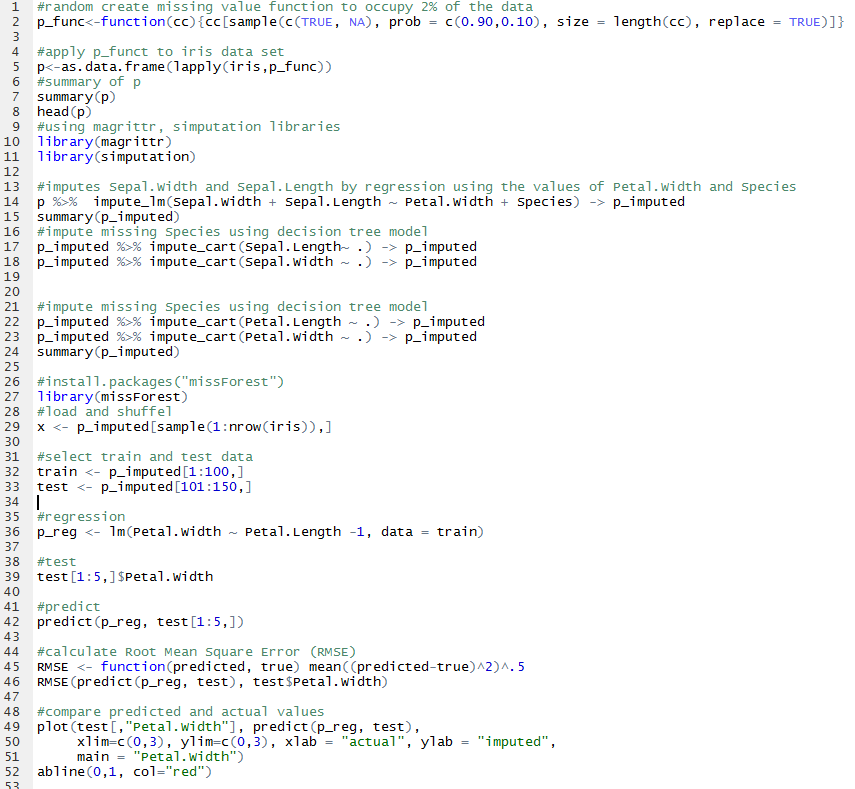
Random Forest



Performance



RMSE



Supervised Classification Knn

