Analyzing the Effects of Missing Data in Time Series with Machine Learning Algorithms

Jazmin Quezada[1], Maria C. Mariani[2], Claire McKay Bowen[3], Joanne Wendelberger[3]
[1] Graduate Student, Department of Mathematical Sciences, University of Texas at El Paso
[2] Supervisor, Chair of the Department of Mathematical Sciences, University of Texas at El Paso
[3] Supervisor, Statistical Sciences Group, Los Alamos National Laboratory

(LANL review: LA-UR-19-29640)

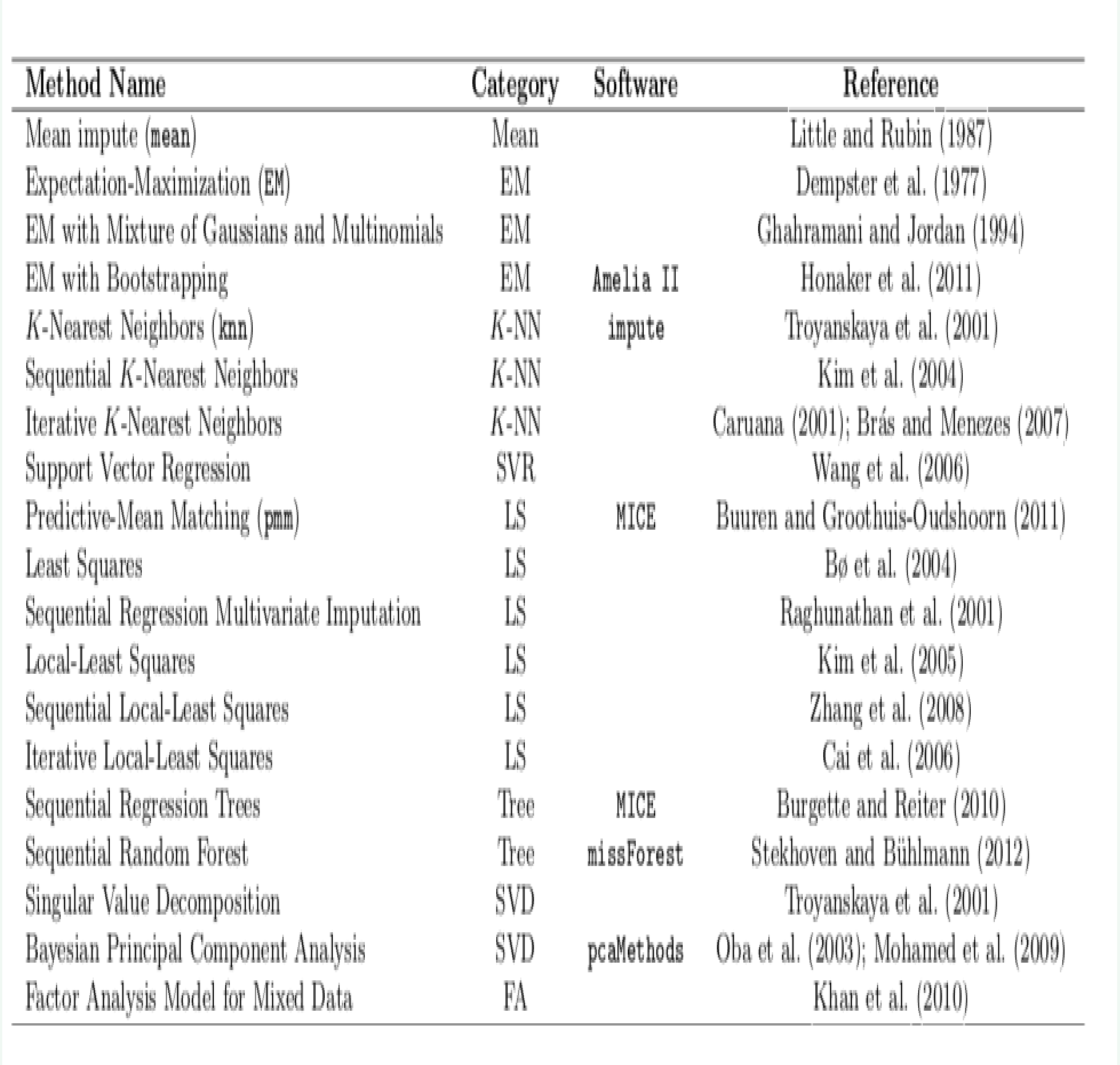
1 Abstract

In statistics, imputation is the process of replacing missing data with substituted values. When substituting for a data point, it is known as "unit imputation", when replacing a component of a data point, it is known as "item imputation". The topic of using machine learning algorithms along with imputation methods has been rapidly growing due to the vast amount of data available and the increase in resources to analyze and or predict an outcome. In this study, we will combine different machine learning methodologies to forecast stock prices and impute missing educational data. The educational data used in this work was synthetically generated by using a machine learning approach.

3 Methods

In this work the following subjects were considered:

1. Multivariate Time Series.
2. Machine learning algorithms.
   (a) Neural Networks.
      i. Deep Neural Networks.
      ii. Convolutional Neural Networks.
      iii. Recurrent Neural Networks.
   (b) Logistic Regression.
   (c) Naive Bayes.
   (d) K-Nearest Neighbors.
   (e) Support Vector Machine.
   (f) Decision Tree.
3. R packages: Amelia, missForest, MI, MICE.



Method Name	Category	Software	Reference
Mean impute (`mean`)	Mean		Little and Rubin (1987)
Expectation-Maximization (`EM`)	EM		Dempster et al. (1977)
EM with Mixture of Gaussians and Multinomials	EM		Ghahramani and Jordan (1994)
EM with Bootstrapping	EM	`Amelia II`	Honaker et al. (2011)
K-Nearest Neighbors (`knn`)	*K*-NN	`impute`	Troyanskaya et al. (2001)
Sequential *K*-Nearest Neighbors	*K*-NN		Kim et al. (2004)
Iterative *K*-Nearest Neighbors	*K*-NN		Caruana (2001); Brás and Menezes (2007)
Support Vector Regression	SVR		Wang et al. (2006)
Predictive-Mean Matching (`pmm`)	LS	`MICE`	Buuren and Groothuis-Oudshoorn (2011)
Least Squares	LS		Bø et al. (2004)
Sequential Regression Multivariate Imputation	LS		Raghunathan et al. (2001)
Local-Least Squares	LS		Kim et al. (2005)
Sequential Local-Least Squares	LS		Zhang et al. (2008)
Iterative Local-Least Squares	LS		Cai et al. (2006)
Sequential Regression Trees	Tree	`MICE`	Burgette and Reiter (2010)
Sequential Random Forest	Tree	`missForest`	Stekhoven and Bühlmann (2012)
Singular Value Decomposition	SVD		Troyanskaya et al. (2001)
Bayesian Principal Component Analysis	SVD	`pcaMethods`	Oba et al. (2003); Mohamed et al. (2009)
Factor Analysis Model for Mixed Data	FA		Khan et al. (2010)

Figure 5: Table of Imputation Methods.

7 Future Research

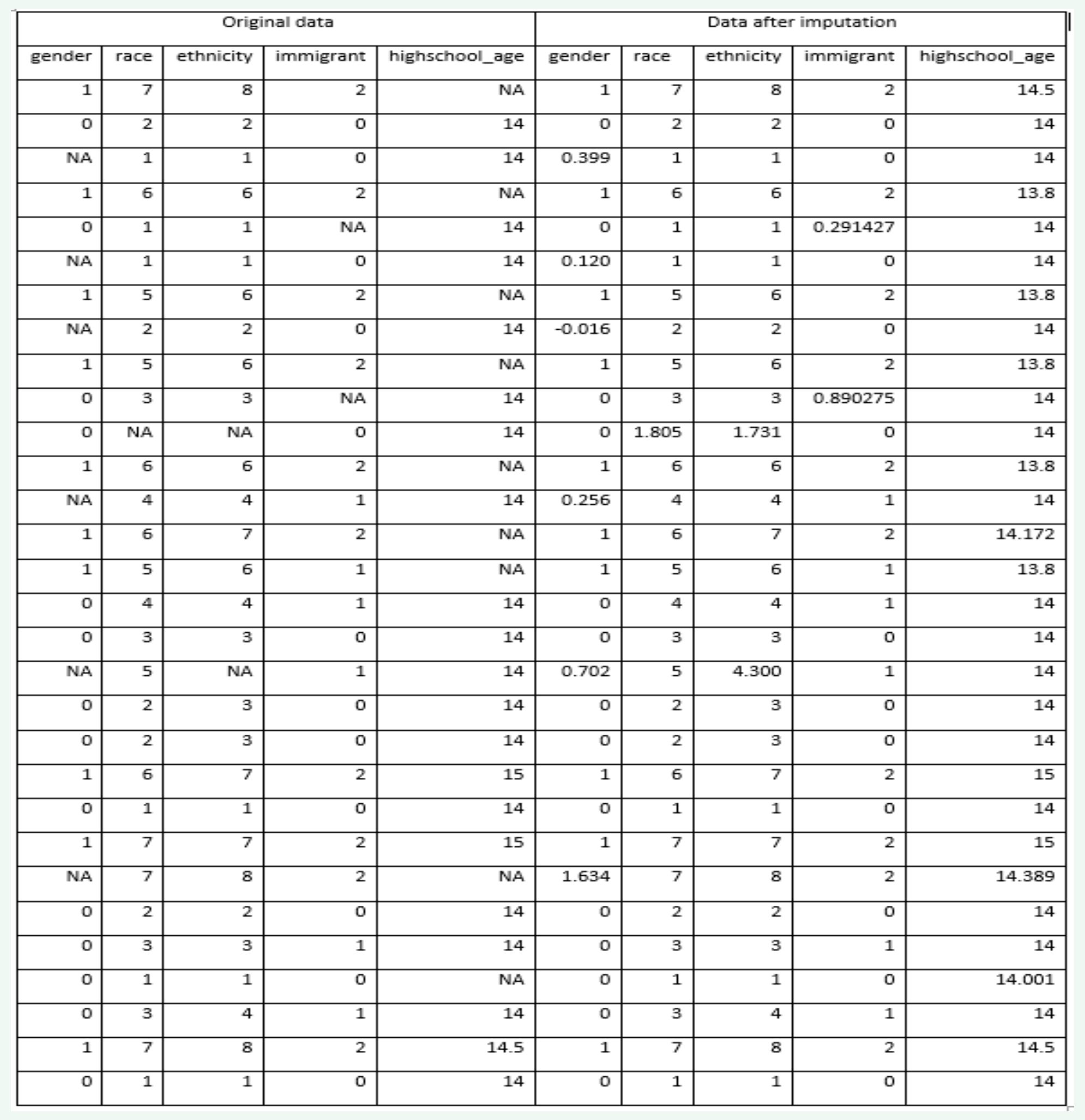
- I plan to evaluate the effectiveness of different existing statistical tools (MICE, missForest, MI etc.) for forecasting missing data, including appropriate visualizations.
- I will also work on different neural networks and other machine learning techniques for predicting missing data and apply the models to solve problems in education, economy, and engineering.

2 Motivation

- Applications of modern methods for analyzing data with missing values, based primarily on multiple imputation, have in the last half-decade become common in many fields. For example educational data, data from American politics and political behavior among others.
- The increase of historical data throughout time along with its heavily missing values is a great topic of many literatures. In this work, we will combine different machine learning algorithms to forecast stock prices and impute missing educational data.

5 Results 2- Impute missing data

Finding missing data in educational data. The educational data used in this work was synthetically generated by using a machine learning approach.



Original data					Data after imputation				
gender	race	ethnicity	immigrant	highschool_age	gender	race	ethnicity	immigrant	highschool_age
1	7	8	2	NA	1	7	8	2	14.5
0	2	2	0	14	0	2	2	0	14
NA	1	1	0	14	0.399	1	1	0	14
1	6	6	2	NA	1	6	6	2	13.8
0	1	1	NA	14	0	1	1	0.291427	14
NA	1	1	0	14	0.120	1	1	0	14
1	5	6	2	NA	1	5	6	2	13.8
NA	2	2	0	14	-0.016	2	2	0	14
1	5	6	2	NA	1	5	6	2	13.8
0	3	3	NA	14	0	3	3	0.890275	14
0	NA	NA	0	14	0	1.805	1.731	0	14
1	6	6	2	NA	1	6	6	2	13.8
NA	4	4	1	14	0.256	4	4	1	14
1	6	7	2	NA	1	6	7	2	14.172
1	5	6	1	NA	1	5	6	1	13.8
0	4	4	1	14	0	4	4	1	14
0	3	3	0	14	0	3	3	0	14
NA	5	NA	1	14	0.702	5	4.300	1	14
0	2	3	0	14	0	2	3	0	14
0	2	3	0	14	0	2	3	0	14
1	6	7	2	15	1	6	7	2	15
0	1	1	0	14	0	1	1	0	14
1	7	7	2	15	1	7	7	2	15
NA	7	8	2	NA	1.634	7	8	2	14.389
0	2	2	0	14	0	2	2	0	14
0	3	3	1	14	0	3	3	1	14
0	1	1	0	NA	0	1	1	0	14.001
0	3	4	1	14	0	3	4	1	14
1	7	8	2	14.5	1	7	8	2	14.5
0	1	1	0	14	0	1	1	0	14

Figure 6: Sample missing data in education.

In order to study effectiveness of different computational, mathematical, and statistical methods it is possible to adaptively create different test cases in an autonomous way.

```
library(missForest)
library(mice)
MyDataInput <- read.csv(file="data.csv",header=TRUE,sep=",")
MyDataInput.mis <- prodNA(MyDataInput, noNA = 0.1)
write.csv(MyDataInput.mis,'data-education.csv')
```

Figure 7: Generation of missing data in R.

8 References

[1] Jazmin Quezada, Analyzing the Effects of Missing Data in Time Series with Machine Learning Algorithms, Scientific Report, The Computer, Computational, and Statistical Sciences (CCS) Division, Los Alamos National Lab, 2019

[2] Jazmin Quezada, Predicting Crashes in Stock Market and Computing Credit Card Default by Using Neural Networks. Master's Thesis, University of Texas at El Paso, 2019

4 Results 1- Forecasting of stock prices

In theory, it is possible to use arbitrary number of input files which increase accuracy of the predictions, and allows us to take into account a lot of dependency in the input data, which is not available for the calculations with single data file.

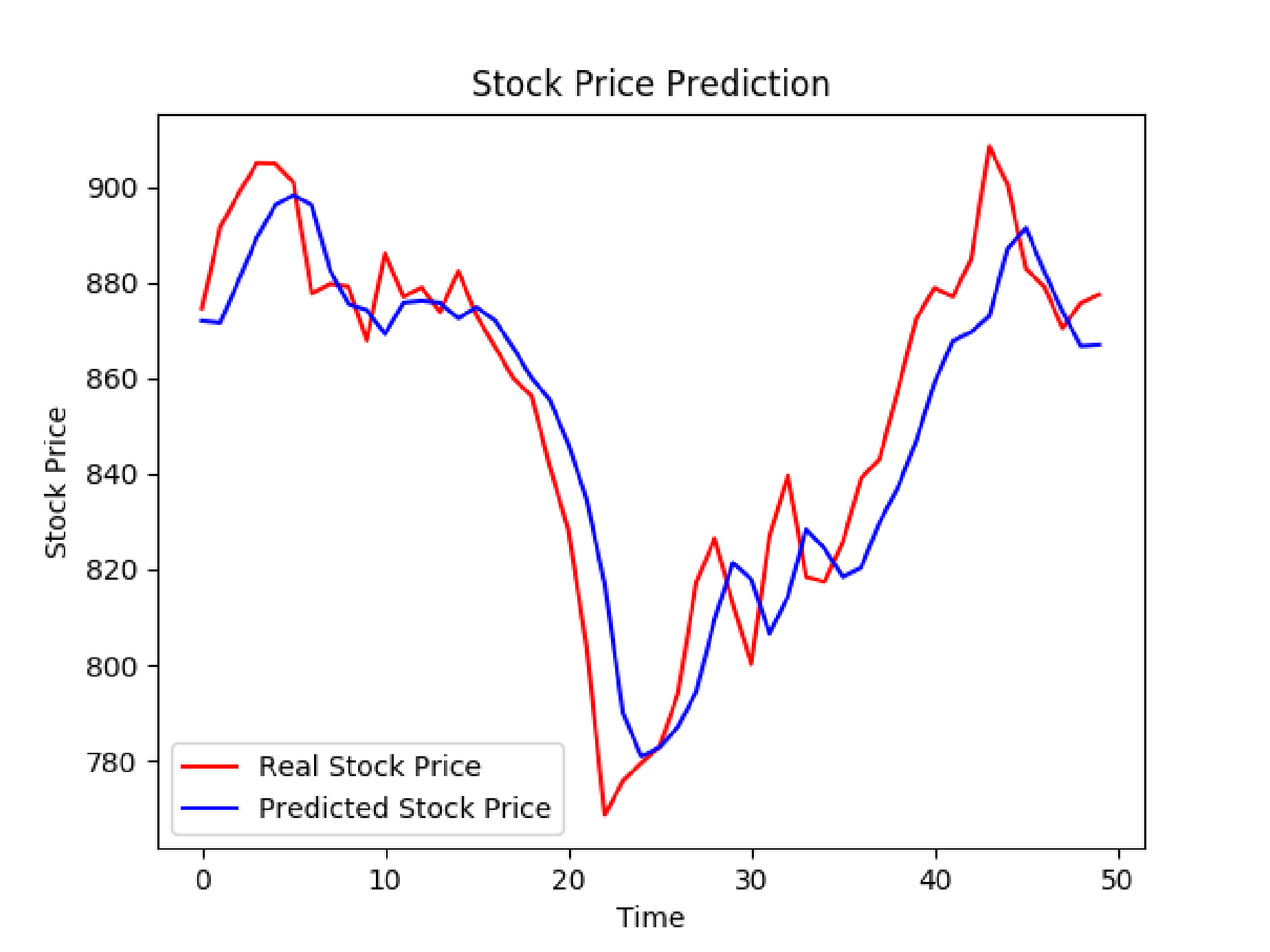


Figure 1: Training the neural network for SP500

An argument can be made that it is possible to create one mathematical model for all stocks that are available on the market. It is also possible to include additional information: Natural disasters, international and national political situation, foreign exchange etc.

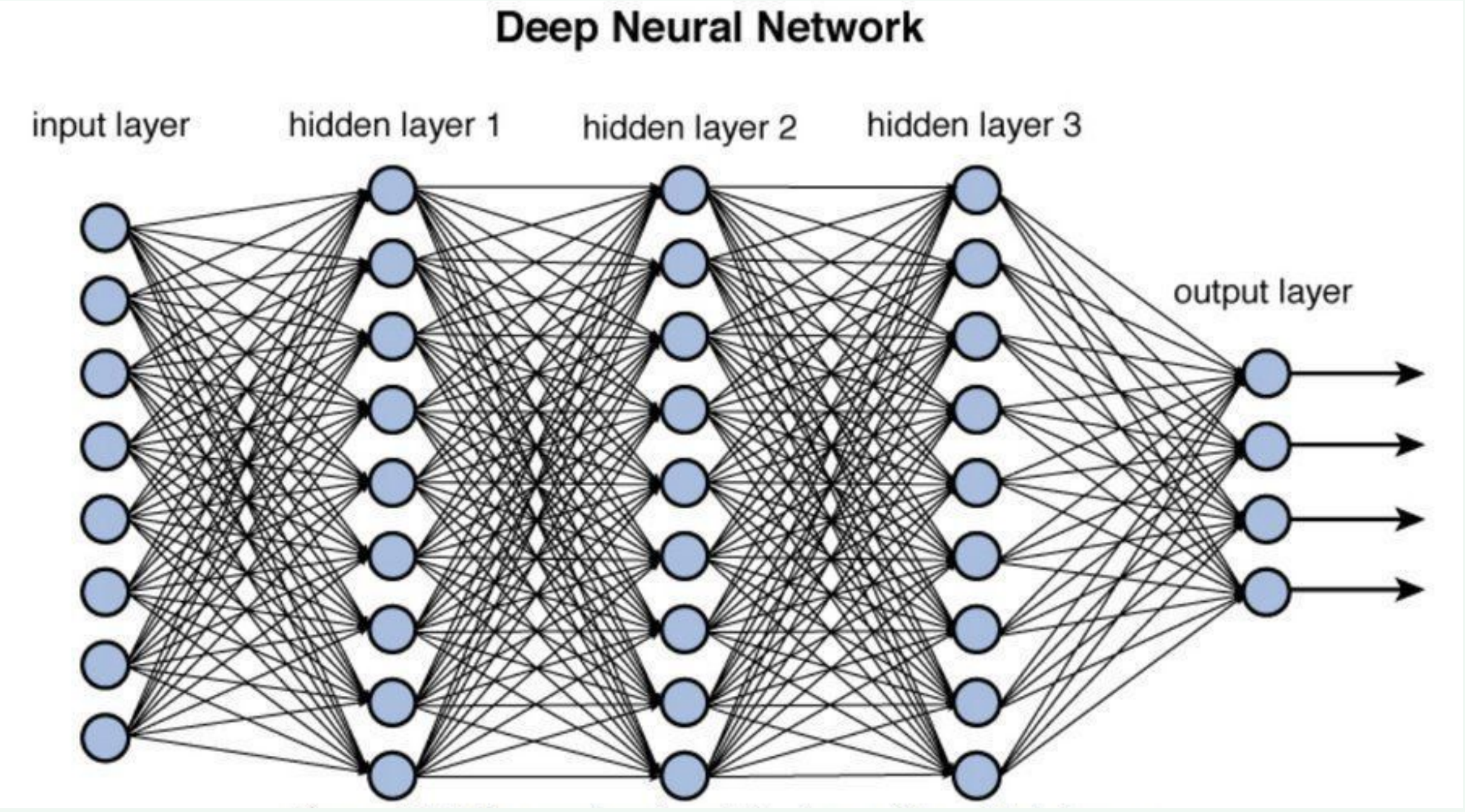


Figure 2: Sample Neural Network.

6 Conclusion

- In this study, we adapt the neural networks algorithm to process incomplete data.
- The experiment performed confirms its practical usefulness in various tasks and for diverse network architectures. In particular, it gives comparable results to other methods which require complete data in training.
- The algorithms studied can be applied to several applications in education, economy, and engineering problems.
- The forecasting techniques used in this work can be extended to different problems with hidden information or to complete data sets.

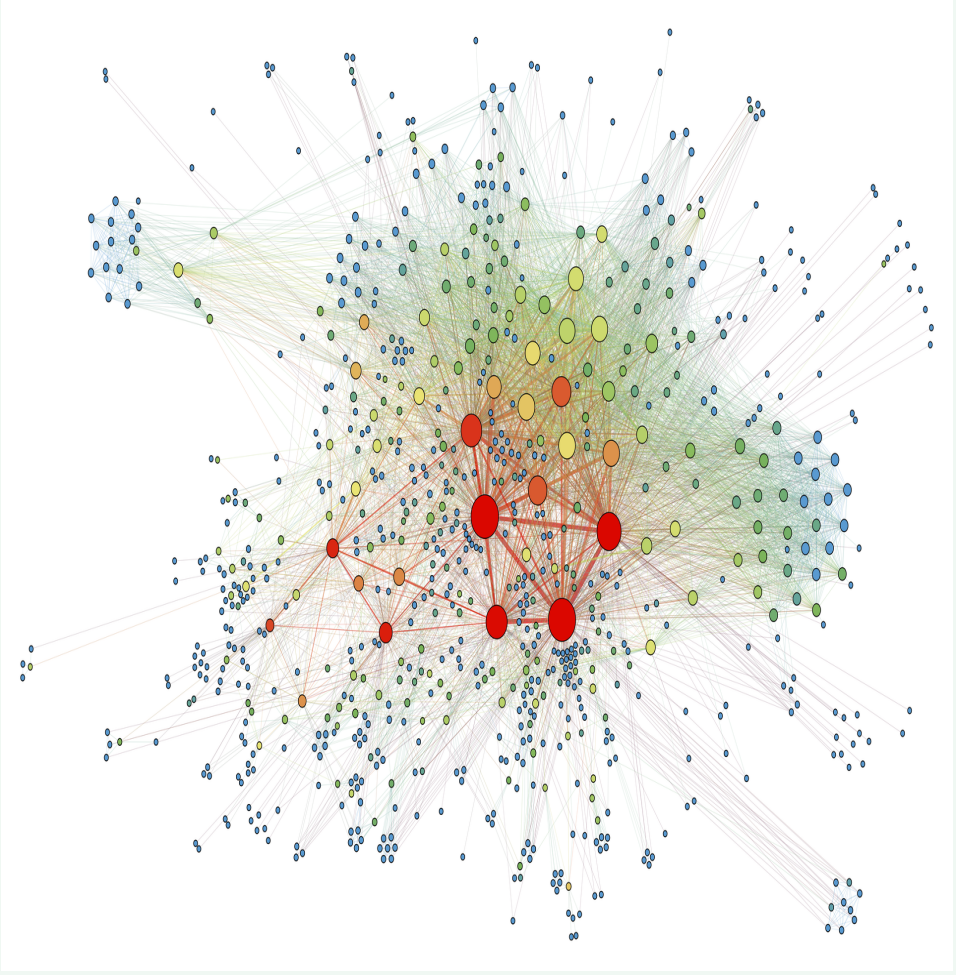


Figure 4: Relations between information in a dataset.

9 Contact Information

Web http://utminers.utep.edu/jquezada5/
Email jquezada5@miners.utep.edu

Figure 3: Researchers from Los Alamos National Lab.