Journal of Systemics, Cybernetics and Informatics (2025) 23(7), 39-44
https://doi.org/10.54808/JSCI.23.07.39

Comparison of Three Methods to Generate Synthetic Datasets for Social Science

Li-jing Arthur CHANG, PhD (CDS&E[1]), PhD (Journalism)
Department of Journalism and Media Studies, Jackson State University
Jackson, MS 39204, USA

ABSTRACT

Many researchers often have difficulties finding enough data to test their hypotheses [1][2]. This study explores three different ways to create "synthetic"[2] (i.e., artificial) data that mimics real-world data in statistical traits like correlations (i.e., relationships between the variables). To see how well these methods perform, the study compares the patterns of synthetic data to their real-world counterparts and sees how closely the data maintain the correlations. Additionally, the study uses seven machine learning[3] prediction methods to see how these synthetic data perform. The findings indicate that two methods more effectively preserve the original correlation structure, while the third method yields better predictive performance.

Keywords: Synthetic Data, Generation Methods, Machine Learning, Performance Comparison, Social Science.

1. INTRODUCTION

Generating synthetic data has a few research applications, including facilitating theory building, testing machine learning models, mitigating data imbalance, and addressing privacy concerns [3][4][5][6]. Synthetic data is instrumental in social science, where there is a need to alleviate data scarcity, reduce information biases, complement privacy restrictions, and facilitate hypothesis testing [1][2][6][7].

The present study used a correlation matrix about journalists' job satisfaction and its predictors obtained from a meta-analysis to generate synthetic datasets [8]. During the data generation process, the study also employed means, standard deviations, a minimum value, and a maximum value from another research to provide further statistical constraints [9]. In addition, the datasets are also constrained to be integers consistent with real-life survey data using the Likert scale [2]. Three data-generation techniques were applied using these statistical traits: multivariate normal, Cholesky decomposition, and Gaussian copula with custom margins [10][11][12][13]. After that, the study compared the synthetic correlations with their original counterparts to assess the quality of the generated synthetic datasets. In addition, the study also applied a few machine learning techniques to see the prediction performance of the datasets. The study aims to explore a few synthetic generating techniques regarding their utilities for social science as a possible solution to the field's data shortage challenge.

2. LITERATURE REVIEW

Researchers have been trying to generate synthetic datasets for decades to ease data access issues deriving from theory testing needs, machine learning requirements, and privacy concerns [3][4][6][14].

The availability of large amounts of data is essential for theory testing because it enhances the accuracy of the theoretical models built [3]. For example, larger data can more easily account for the complex relationships among the independent variables [3]. Additionally, bigger sample sizes are more likely to reduce Type I errors (false positive) and Type II errors (false negative) [2] [15][16]. All these factors enhance theory testing [17][18][19].

Besides theory testing, the rising popularity of machine learning, a branch of AI (artificial intelligence), has been calling for more available data [4][14]. For instance, the grid search with 10-fold cross-validation, a machine learning technique to fine-tune model parameters, benefits from large data because the data needs to be split into 10 sub-subsamples for the model training [20]. As another example, deep learning, a sub-branch of machine learning, needs large data to train multiple hidden layers of neurons [3][21]. Such structural requirements necessitate large data quantities for model training to take place [3][21].

Another case of data needs is privacy concerns during the data collection process. The presence of sensitive questions could discourage survey respondents from completing the questionnaires or, in some cases, lead them to finish only part of the questions, resulting in missing data for later analysis [22].

Although the data shortage challenges many research fields, social scientists face a significant need for data, especially from human subjects [1]. Given the complexity and nuances of the human population, collecting privacy-related data could be challenging [22].

Mimicking real data's key traits such as correlation matrices, synthetic datasets enable researchers to conduct analysis, build models, and test theories [23]. Among the most widely used synthetic generating methods are multivariate normal

[1] CDS&E stands for Computational and Data-Enabled Science and Engineering.
[2] Synthetic datasets are data generated for the training, testing, or validating of models, mimicking real-world data while addressing privacy or bias issues [6][23].
[3] Computational programs that enable systems to learn patterns from data to develop prediction models without explicit programming [14].

distribution, Cholesky decompositions, and Gaussian copula with custom margins [10][11][12][13][23].

The multivariate normal distribution expands the univariate normal distribution to multiple dimensions [24]. This method is suitable for generating synthetic data because it often approximates the "true" population distribution [24]. The limitation of this approach is its assumption of normality, which may not be the case for real-world data [24].

Cholesky decomposition can be also used in the context of generating synthetic datasets—including multivariate normal data [25]. Cholesky decomposes the original correlation matrix and means and standard deviations for data generation [11]. This approach is helpful because it reduces computation complexity during data generation [26].

Different from the other two methods, Gaussian copula with custom margins models dependencies between variables using a multivariate normal distribution while allowing each variable to have its marginal distribution [13]. This method generates synthetic data with pre-specified marginal distributions while preserving the correlation structure [13]. The allowance of marginal distributions for individual variables makes this method suitable for capturing real-world data patterns, often non-Gaussian (e.g., skewed or heavy-tailed) [24].

Synthetic data can be helpful for machine learning model testing. Since the present study aims to use a meta-analysis correlation matrix of journalists' job satisfaction and its key predictors, along with other statistical traits like means and standard deviations, as the basis to generate the data, the target variable (i.e., job satisfaction) of the synthetic dataset is of the continuous type [2]. Due to the nature of the data, regression, rather than classification, is suitable for machine learning model building [14]. Among the popular regression machine learning algorithms are ridge, lasso, elastic net, gradient boosting, random forest, Support Vector Machine, and k-nearest neighbors [27][28][29][30].

Ridge, lasso, and elastic net use penalties like L2, L1, and both L1 and L2, respectively, to constrain coefficients and/or choose features to control multicollinearity and overfitting [27][31][32][33][34]. Gradient boosting and random forest are tree-based algorithms [27][28]. The first tree algorithm optimizes its performance sequentially by correcting the errors of the last tree until the process converges [28]. The second tree algorithm leverages bootstrapping to generate sub-samples for training and obtain the mean of the predictions [35]. Support Vector Machine can predict a target variable by searching for the best-fitting "line" (or hyperplane in higher dimensions) [29][36]. K-nearest neighbors predicts a target value by computing a weighted average of the target values of its k nearest data points. [37][38].

RQ1: Among the three synthetic data generating methods—multivariate normal, Cholesky decomposition, and Gaussian copula with custom margins—which one(s) generate(s) a dataset or datasets with a correlation matrix that closely approximates its original counterpart?

RQ2: Among the three synthetically generated datasets, which one or ones yield the best results when evaluated using the seven machine learning models?

3. METHODS

The present study employs three popular methods to generate synthetic datasets. The data generation is based on statistical properties such as a correlation matrix of job satisfaction and its predictors, variable means, standard deviations, a minimum value and a maximum value. The statistical traits are derived from journalist surveys [8][9]. This data type is chosen for two reasons: (1) most of the public's information comes from journalists, and (2) job satisfaction is a strong predictor of turnover intentions [39]. The correlation matrix is from a meta-analysis of journalists' satisfaction and its key predictors[8]. The selection of a meta-analysis, rather than a single study, is based on the idea that a meta-analysis can capture intervariable relations that are more generalizable to the population[40]. As the meta-analysis used for the study does not report means, standard deviations, or minimum and maximum values, these statistics come from a single national study of American journalists [9].

To evaluate the synthetic datasets, the study used Fisher's z-test to compare each correlation pair between the synthetic and original correlations [41]. A correlation shows how strong the relationship is between two things, like a person's height and weight [42]. Fisher's z-test helps explain the relationship by turning the correlation numbers (i.e., coefficients) into a standard form for easier comparison [41]. The study then calculates the difference between transformed numbers to see if the difference is large enough to matter (i.e., statistically significant) [41][43].

As another means of assessment, the datasets generated by the three methods underwent evaluations using seven machine learning regression algorithms: Ridge, lasso, elastic net, gradient boosting, random forest, Support Vector Machine, and k-nearest neighbors. Before the evaluations, each dataset is randomly divided into training and test datasets with an 80:20 ratio.

Afterward, each regression model underwent a grid search on the training dataset. Grid search is a method that finds optimal hyperparameters by thoroughly assessing preset combinations using techniques like cross-validation [20][44]. Cross-validation divides the training dataset into subsets for iterative training/validation to reduce overfitting risks and provide more accurate results [20]. Although requiring repeated computation, grid search is reliable for obtaining elevated performance metrics [44]. After the grid search, the trained model is evaluated on the test dataset using performance metrics like R^2 (R-squared) and RMSE (root mean squared error). R^2 measures the variance of the target variable explained by its predictors (i.e., features) [45]. RMSE measures average prediction errors between predicted and actual values of the target variable, with lower scores meaning better model performance [46][47][48].

4. RESULTS

The study is designed to explore three popular synthetic data-generating methods to see which one(s) can more effectively reproduce the pre-specified correlations. The created synthetic datasets also underwent further evaluations via machine learning model building to see which synthetic data generation method creates a dataset with the highest prediction performance.

When the average absolute correlation difference is considered, the correlation matrices of all the synthetic datasets closely approximate the original correlation matrix.

The lowest average absolute difference between the synthetic and original correlations is 0.014 for the multivariate normal dataset, 0.016 for the Cholesky decomposition dataset, and 0.047 for the Gaussian copula dataset.

However, when the Fisher's z transformation is used to test the difference between the synthetic and original correlations, only two methods strongly preserve the original correlation matrix.

Fisher's z transformation was used to statistically compare each synthetic dataset's correlation matrix with the original correlation matrix, comprising eight variables (28 unique pairwise correlations). The proportion of non-significant differences between the synthetic and original correlation coefficients was calculated to assess structural fidelity.

For the multivariate normal-based synthetic dataset, 78.6% (22 out of 28) of the pairwise correlations were not significantly different ($p \geq .05$) from the original correlations. Similarly, the Cholesky-based dataset also preserved 78.6% (22 out of 28) of the correlations without significant differences ($p \geq 0.5$). In contrast, the copula-based synthetic dataset preserved only 3.6% (1 out of 28) of the original correlations ($p \geq 0.5$), indicating a substantial deviation in inter-variable relationships.

As a visual demonstration of similarity, Figure 1 shows a comparison of the heatmaps between the multivariate normal and original correlations. The heatmap follows a color scheme that shows positive correlations in warm colors and negative correlations in cool colors. The similar color hues between the heatmaps of the two correlation matrices show that the two matrices are not too different.

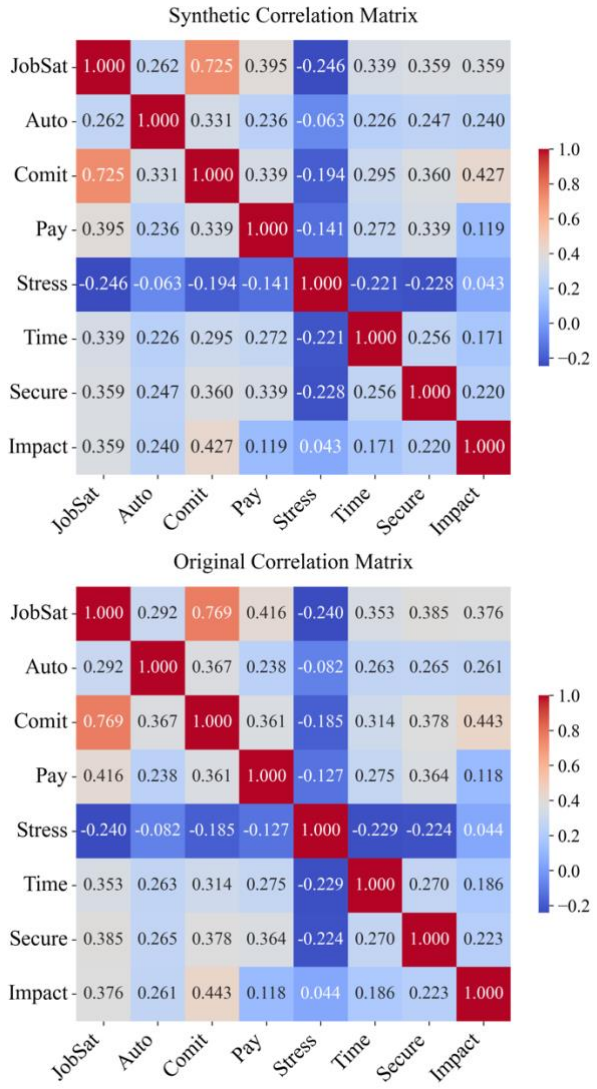
Besides comparing the synthetic and original correlations, the study used machine learning regressors to test how the synthetic datasets perform in predictions. The evaluations involved examining the R^2 and RMSE metrics.

Table 1 shows that the copula-based dataset has the largest R^2 score of 0.624 when the ridge, lasso, and elastic net regressors were tested. The three regressors are also top performers for the Cholesky-based dataset ($R^2 = 0.577$) and the multivariate normal dataset ($R^2 = 0.572$). For the ridge model tested on multivariate normal and Cholesky decomposition datasets, all features are statistically significant ($p < 0.5$). Besides these cases, most features for the three algorithms tested on the datasets are statistically significant. The R^2 scores are acceptable in social science as they exceed 0.5 and most features are statistically significant [49].

Table 2 shows that the Gaussian copula dataset scored the least RMSE error of 0.872 for ridge, lasso, and elastic net. The three algorithms are also top performers for multivariate normal and Cholesky decomposition datasets (RMSE = 0.923). The RMSE scores are less than the standard deviation of the target variable of the test data for the three datasets (1.423, 1.411 and 1.419 for the Gaussian copula, multivariate normal, and Cholesky decomposition datasets, respectively), indicating that the models perform well. This interpretation is based on two key ideas: (1) the standard deviation of the target variable represents the natural variability or "noise" in the data, and a model with RMSE lower than this threshold is performing better than a naive baseline (e.g., predicting the mean); (2) RMSE can be viewed as the standard error of estimate computed on the validation dataset, quantifying the average magnitude of prediction errors [50][51].

Based on the R^2 and RMSE metrics, the Gaussian copula with custom margins data has the best regression models among the three datasets. To get more details about the seven algorithms tested on the copula dataset, the study also examined Q-Q plots and histograms of residuals plots. The Q-Q plots showed that all seven regression models have residuals along the 45-degree line, revealing that residuals are normally distributed (see Figure 2) [41]. Also, the histogram of residuals plots shows that except for k-nearest neighbors, the models have residuals plots following the bell-shaped normal curve (see Figure 3) [41]. Overall, visual inspection of the Q-Q plots and histograms showed that residuals from the seven regression models, except for the histogram of k-nearest neighbors residuals, approximated normality. This normality indicates effective data capture without major issues like outliers.

Figure 1. Synthetic vs. Original Correlations*



The synthetic correlation comes from the multivariate normal dataset.
Note: Full variable names—JobSat (job satisfaction); Auto (autonomy); Comit (organizational commitment); Pay (feeling about pay); Stress (work stress); Time (work schedule); Secure (job security); Impact (impact on the community).

Table 1. Synthetic Datasets R^2 Performance Comparison

	Multivariate Normal	Cholesky Decomp*	Gaussian Copula**
Ridge	0.572	0.577	0.624
Lasso	0.572	0.577	0.624
Elastic Net	0.572	0.577	0.624
SVM	0.570	0.574	0.623
Gradient Boosting	0.566	0.572	0.618
Random Forest	0.559	0.568	0.608
KNN	0.516	0.530	0.577

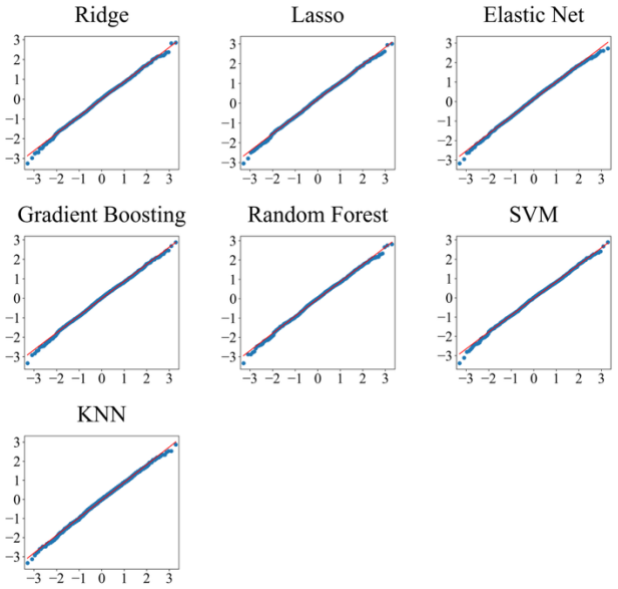
* Cholesky decomposition.
** Gaussian copula with custom margins.
Note: Abbreviations—SVM (Support Vector Machine), KNN (k-nearest neighbors).

Table 2. Synthetic Datasets RMSE Performance Comparison

	Multivariate Normal	Cholesky Decomp*	Gaussian Copula**
Ridge	0.923	0.923	0.872
Lasso	0.923	0.923	0.872
Elastic Net	0.923	0.923	0.872
SVM	0.926	0.926	0.874
Gradient Boosting	0.929	0.928	0.879
Random Forest	0.937	0.933	0.891
KNN	0.982	0.972	0.926

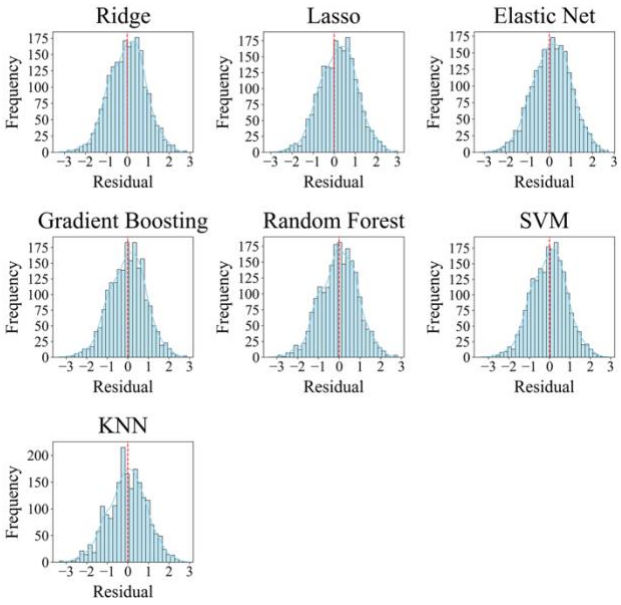
* Cholesky decomposition.
** Gaussian copula with custom margins.
Note: Abbreviations—SVM (Support Vector Machine), KNN (k-nearest neighbors).

Figure 2. Q-Q Plot for Gaussian Copula* Dataset



* Gaussian copula with custom margins.
Note: Abbreviations—SVM (Support Vector Machine), KNN (k-nearest neighbors).

Figure 3. Histogram of Residuals for Copula* Dataset



* Gaussian copula with custom margins.
Note: Abbreviations—SVM (Support Vector Machine), KNN (k-nearest neighbors).

5. DISCUSSION

In social science, synthetic data can address privacy concerns and speed up research. Synthetic data also helps ease data shortages and meet the significant data requirements for machine learning. Findings from the present study showed that synthetic procedures can produce data for model and theory testing [17][23].

The present study found that based on Fisher's z tests, the multivariate normal and Cholesky approaches perform better than the Gaussian copula technique in terms of approximating the original correlations.

When tested with the machine learning regressors, the datasets perform well because their top three performers (ridge, lasso and elastic net) have an R^2 score above 0.5, deemed acceptable in social science when most of the features' regression coefficients are statistically significant, which was confirmed by examining these three regressors.

The data from the Gaussian copula with custom margins shows the best prediction performance among the three synthetic datasets. Of the seven algorithms tested on the Gaussian copula dataset, ridge, lasso, and elastic net have the best R^2 and RMSE scores. These regressors are also top models when tested on the multivariate normal and Cholesky decomposition datasets.

Further analysis using the Q-Q plots on the Gaussian copula showed that all regressors produced normally distributed residuals. Also, the histograms of residuals plots showed that except for k-nearest neighbors, all the other algorithms have a bell-shaped normal residuals curve.

To further understand which predictor(s) most influenced the regression outcome (i.e., the target variable), a post hoc feature importance analysis was conducted on the best-performing

Gaussian copula data (R^2 = 0.624; RMSE = 0.872, see Tables 1 and 2). Among the seven features, organizational commitment demonstrated a disproportionately high importance score of 0.574, while the remaining six features ranged from 0.064 to 0.077 (see Table 3). This shows organizational commitment is a strong predictor of the target variable job satisfaction, which is in line with past research [52][53]. This finding supports the robustness of the model and reinforces empirical evidence found in the literature.

Table 3. Feature Importance Analysis* on Copula Data**

Feature	Importance
Comit	0.574
Auto	0.077
Stress	0.075
Impact	0.073
Secure	0.070
Time	0.067
Pay	0.064

* The analysis was conducted using the random forest algorithm.
** Gaussian copula with custom margins.
Note: Full variable names—Auto (autonomy); Comit (organizational commitment); Pay (feeling about pay); Stress (work stress); Time (work schedule); Secure (job security); Impact (impact on the community).

6. CONCLUSION

Multivariate normal, Cholesky decomposition, and Gaussian copula methods generate synthetic data through Monte Carlo-style sampling from estimated probabilistic models [54]. This study uses these methods as well as a correlation matrix, means, standard deviations, a minimum value and a maximum value to generate synthetic datasets. The findings show benefits and drawbacks of each method in preserving the original statistical traits and in machine learning predictions. Fisher's z tests showed that the multivariate normal and Cholesky-based techniques strongly preserved the original correlations, while the copula-based method did not. However, regarding machine learning model prediction performance, the copula dataset produces better results than the other two methods. Nevertheless, the choice of method should depend on the specific application, as this study shows trade-offs may exist between prediction performance and fidelity to the original data distribution. More work should be conducted to retest the conclusion of the present study. Future work could explore more complex dependency structures or validate these techniques on real-world datasets to further assess their robustness.

7. REFERENCES

[1] S. E. Maxwell, "The persistence of underpowered studies in psychological research: causes, consequences, and remedies.," *Psychol Methods*, vol. 9, no. 2, p. 147, 2004.

[2] E. R. Babbie, *The Practice of Social Research*. Cengage Au, 2020.

[3] M. Sordo and Q. Zeng, "On sample size and classification accuracy: A performance comparison," in *International Symposium on Biological and Medical Data Analysis*, 2005, pp. 193–201.

[4] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intell Syst*, vol. 24, no. 2, pp. 8–12, 2009.

[5] M. Ye-Bin, N. Hyeon-Woo, W. Choi, N. Kim, S. Kwak, and T.-H. Oh, "SYNAuG: Exploiting Synthetic Data for Data Imbalance Problems," *arXiv preprint arXiv:*2308.00994, 2023.

[6] J. P. Reiter, "Satisfying disclosure restrictions with synthetic data sets," *J Off Stat*, vol. 18, no. 4, p. 531, 2002.

[7] F.-J. Rodrigo-Ginés, J. Carrillo-de-Albornoz, and L. Plaza, "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it," *Expert Syst Appl*, vol. 237, p. 121641, 2024.

[8] L. A. Chang and B. L. Massey, "Factors Related to Journalist Job Satisfaction: Meta-Analysis and Path Model," in *AEJMC annual conference*, Aug. 2008, pp. 1–29.

[9] L. A. Chang and B. L. Massey, "Work motivation and journalists in Taiwan and the US: An integration of theory and culture," *Asian J Commun*, vol. 20, no. 1, pp. 51–68, 2010.

[10] J. Heine, E. E. E. Fowler, A. Berglund, M. J. Schell, and S. Eschrich, "Techniques to produce and evaluate realistic multivariate synthetic data," *Sci Rep*, vol. 13, no. 1, p. 12266, 2023.

[11] A. Marchev Jr and V. Marchev, "Automated algorithm for multi-variate data synthesis with Cholesky decomposition," in *Proceedings of the 7th International Conference on Algorithms, Computing and Systems*, 2023, pp. 1–6.

[12] Y. Wei, "Using Gaussian Copulas to Generate a Synthetic Population," Oct. 2018. [Online]. Available: https://www.washstat.org/presentations/20181024/Wei.pdf

[13] I. Žežula, "On multivariate Gaussian copulas," *J Stat Plan Inference*, vol. 139, no. 11, pp. 3942–3946, 2009.

[14] T. M. Mitchell, *Machine Learning*, vol. 1, no. 9. McGraw-hill New York, 1997.

[15] D. V Knudson and C. Lindsey, "Type I and Type II errors in correlations of various sample sizes," *Comprehensive Psychology*, vol. 3, pp. 03–CP, 2014.

[16] K. J. Rothman, "Curbing type I and type II errors," *Eur J Epidemiol*, vol. 25, pp. 223–224, 2010.

[17] G. Shmueli, "To explain or to predict?," 2010.

[18] M. Kuhn, K. Johnson, and others, *Applied Predictive Modeling*, vol. 26. Springer, 2013.

[19] P. J. G. Teunissen, *Testing Theory: An Introduction*. TU Delft Open, 2024.

[20] A. Géron, "Hands-on machine learning with scikit-learn, keras, and tensorflow: concepts," Aurélien *Géron-Google Kitaplar, yy https://books. google. com. tr/books*, 2019.

[21] S. I. Nikolenko and others, *Synthetic Data for Deep Learning*, vol. 174. Springer, 2021.

[22] B. Howe, J. Stoyanovich, H. Ping, B. Herman, and M. Gee, "Synthetic data for social good," *arXiv preprint arXiv:1710.08874*, 2017.

[23] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in 2016 *IEEE international conference on data science and advanced analytics (DSAA)*, 2016, pp. 399–410.

[24] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. Upper Saddle River, NJ: Pearson, 2007.
[25] J. Wang and C. Liu, "Generating multivariate mixture of normal distributions using a modified Cholesky decomposition," in *Proceedings of the 2006 Winter Simulation Conference*, 2006, pp. 342–347.
[26] GeeksforGeeks, "Cholesky Decomposition: Matrix Decomposition," 2025. [Online]. Available: https://www.geeksforgeeks.org/cholesky-decomposition-matrix-decomposition/
[27] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2. Springer, 2009.
[28] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann Stat*, pp. 1189–1232, 2001.
[29] R. G. Brereton and G. R. Lloyd, "Support vector machines for classification and regression," *Analyst*, vol. 135, no. 2, pp. 230–267, 2010.
[30] H. Dubey, "Efficient and accurate kNN based classification and regression," *A Master Thesis Presented to the Center for Data Engineering, International Institute of Information Technology*, vol. 500, no. 32, 2013.
[31] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J R Stat Soc Series B Stat Methodol*, vol. 67, no. 2, pp. 301–320, 2005.
[32] A. B. Owen, "A robust hybrid of lasso and ridge regression," *Contemporary Mathematics*, vol. 443, no. 7, pp. 59–72, 2007.
[33] S. Altelbany, "Evaluation of ridge, elastic net and lasso regression methods in precedence of multicollinearity problem: a simulation study," *Journal of Applied Economics and Business Studies*, vol. 5, no. 1, pp. 131–142, 2021.
[34] J. Y.-L. Chan et al., "Mitigating the multicollinearity problem and its machine learning approach: a review," *Mathematics*, vol. 10, no. 8, p. 1283, 2022.
[35] L. Breiman, "Random forests," *Mach Learn*, vol. 45, pp. 5–32, 2001.
[36] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat Comput*, vol. 14, pp. 199–222, 2004.
[37] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor," *Am Stat*, vol. 46, no. 3, pp. 175–185, 1992.
[38] scikit-learn.org, "KNeighborsRegressor," 2025.
[39] E. Lambert and N. Hogan, "The importance of job satisfaction and organizational commitment in shaping turnover intent: A test of a causal model," *Crim Justice Rev*, vol. 34, no. 1, pp. 96–118, 2009.
[40] D. Jackson, R. Riley, and I. R. White, "Multivariate meta-analysis: potential and promise," *Stat Med*, vol. 30, no. 20, pp. 2481–2498, 2011.
[41] A. Field, Discovering statistics using IBM SPSS statistics. Sage publications limited, 2017.
[42] A. G. Asuero, A. Sayago, and A. G. González, "The correlation coefficient: An overview," *Crit Rev Anal Chem*, vol. 36, no. 1, pp. 41–59, 2006.
[43] N. J. Cox, "Speaking Stata: Correlation with confidence, or Fisher's z revisited," *Stata J*, vol. 8, no. 3, pp. 413–439, 2008.
[44] M. R. Hossain and D. Timmer, "Machine learning model optimization with hyper parameter tuning approach," *Glob. J. Comput. Sci. Technol. D Neural Artif. Intell*, vol. 21, no. 2, p. 31, 2021.
[45] D. B. Figueiredo Filho, J. A. S. Júnior, and E. C. Rocha, "What is R2 all about?," *Leviathan (São Paulo)*, no. 3, pp. 60–68, 2011.
[46] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," *Geosci Model Dev*, vol. 7, no. 3, pp. 1247–1250, 2014.
[47] T. O. Hodson, "Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not," *Geoscientific Model Development Discussions*, vol. 2022, pp. 1–10, 2022.
[48] G. Casella and R. Berger, *Statistical Inference*. CRC press, 2024.
[49] P. K. Ozili, "The acceptable R-square in empirical modelling for social science research," in *Social Research Methodology and Publishing Results: A Guide to Non-native English Speakers*, IGI global, 2023, pp. 134–143.
[50] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*. McGraw-hill, 2005.
[51] G. Shmueli, P. C. Bruce, I. Yahav, N. R. Patel, and K. C. Lichtendahl Jr, *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. John Wiley & Sons, 2017.
[52] R. J. Vandenberg, C. E. Lance, "Examining the causal order of job satisfaction and organizational commitment," *Journal of Management*, vol. 18, no. 1, pp. 153-167, 1992.
[53] R. Saleem, "Organizational commitment as a predictor of job satisfaction among private sector employees," *International Journal of Indian Psychology*, vol. 4, no. 4, pp. 2349-3429, 2017.
[54] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2004.