نامه به سردبیر

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره هفتم شماره سوم. ص ۳۵۳-۳۵۱، ۱۳۹۹

وجود داده‌های گمشده در مجموعه داده PIMA و عدم توجه به آن در مقاله پذیرفته شده در آن مجله

فاطمه آهوز[۱]، امین گلاب‌پور[۲]*

• دریافت مقاله: ۹۸/۶/۹ • پذیرش مقاله: ۹۸/۶/۲۳

مجموعه داده دیابت PIMA در مقالات بسیاری مورد بررسی پژوهشگران قرار گرفته است. این مجموعه داده که از پایگاه داده یادگیری ماشین ایروین دانشگاه کالیفرنیا (The University of California, Irvine) UCI گرفته شده است، شامل اطلاعات ۷۶۸ بیمار خانم، حداقل ۲۱ ساله و با تبار سرخپوستان PIMA است که از این تعداد ۲۶۸ فرد دارای دیابت و ۵۰۰ فرد فاقد دیابت هستند.

اولین کاری که روی دیتاست PIMA صورت گرفته است به سال ۱۹۸۸ برمی‌گردد جایی که «Smith و همکاران» [۱] با استفاده از مدل شبکه عصبی ADAP ارزیابی را جهت پیش‌بینی داشتن دیابت در جمعیت سرخپوستان PIMA انجام دادند [۲]. Smith و همکاران مشخص کردند که مجموعه داده دارای مقادیر ناشناخته در ویژگی‌های قندخون دو ساعت بعد صبحانه، فشارخون دیاستولیک، ضخامت عضله سر بازو، انسولین سرم و شاخص توده بدنی است (جدول ۱).

کلیدواژه‌ها: دیابت، مقادیر گمشده، مجموعه داده PIMA، داده‌کاوی

•**ارجاع:** آهوز فاطمه، گلاب‌پور امین. وجود داده‌های گمشده در مجموعه داده PIMA و عدم توجه به آن در مقاله پذیرفته شده در آن مجله. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۸؛ ۷(۳):

۱. مربی، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه صنعتی خاتم‌الانبیاء بهبهان، بهبهان، ایران
۲. استادیار، گروه فناوری اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شاهرود، شاهرود، ایران
* **نویسنده مسئول:** امین گلاب‌پور
آدرس: شاهرود، میدان هفت تیر، دانشگاه علوم پزشکی شاهرود، گروه فناوری اطلاعات سلامت
• **شماره تماس:** ۰۲۳۳۲۳۹۵۰۵۴
• **Email:**a.golabpour@shmu.ac.ir

جدول ۱: دسته‌های موجود برای هر یک از متغیرهای ورودی مجموعه داده PIMA

Input variable	# of categories	Categories
Number of times pregnant	۳	{۰ ،۱ ،۲},{۳ ،۴ ،۵ ،۶}, {>۷}
۲ Hr glucose tolerance	۶	{۸۹-۰/۱},{۸۹/۲ – ۱۰۷/۱} {blank,۱۰۷/۲-۱۲۳/۱} {۱۲۳/۲-۱۴۳/۱} {۱۴۳/۲-۱۶۵/۱} {>۱۶۵/۲}
Diastolic BP	۴	{blank}{۱-۷۶/۱} {۷۶/۲-۹۸/۱} {>۹۸/۲}
Triceps skin fold	۴	{blank} {۱-۲۵} {۲۶-۳۲} {>۳۳}
۲ Hr serum insulin	۵	{blank} {۱-۱۱۰} {۱۱۱-۱۵۰} {۱۵۱-۲۴۰} {>۲۴۱}
Body mess inbox	۵	{۱-۲۲/۸۱۴},{۲۲/۸۱۵-۲۶/۳۴},{blank,۲۶/۳۴۱-۳۳/۵۵},{۳۳/۵۵۱-۳۵/۵۶۳} {>۳۵/۵۶۴}
Diabetes Pedigree Function	۵	{۰-۰/۲۴۴} {۰/۲۴۵-۰/۵۲۵} {۰/۵۲۶-۰/۹۰۵} {۰/۹۰۶-۱/۱۱} {>۱/۱۱}
Age	۵	{۲۱-۲۴} {۲۵-۳۰} {۳۱-۴۰} {۴۱-۵۵} {>۵۵}

در این جدول مقادیر ناشناخته با کلمه **blank** مشخص شده‌اند [۱].

این مقادیر ناشناخته که در مجموعه داده با مقدار صفر مشخص شده‌اند، در مقالات بسیاری از جمله [۲۵-۲] تحت نام مقادیر گمشده ذکر شده‌اند. جدول ۲ تعداد مقادیر معتبر و گمشده در این مجموعه داده را نشان می‌دهد که با نرم‌افزار SPSS نسخه ۱۶ محاسبه شده است. در حقیقت ویژگی‌هایی ذکر شده به لحاظ پزشکی نمی‌توانند مقدار صفر داشته باشند؛ به عنوان مثال منطقی نیست که میزان قندخون در یک انسان زنده صفر باشد [۷].

این در حالی است که در مقاله صباغ‌گل [۲۶]، بدون در نظر گرفتن مقادیر گمشده و راهی برای اداره کردن آن به اجرای روش‌های داده کاوی پرداخته است. در این مقاله جدولی (با نام جدول ۱: متغیرهای اطلاعاتی مورد استفاده) ذکر شده که حاوی اطلاعات مجموعه داده PIMA است و در آن بازه مجاز ویژگی‌های دارای مقدار گمشده را شامل صفر دانسته که خلاف قوانین پزشکی و کارهای پژوهشی صورت گرفته بر روی این مجموعه داده است.

جدول ۱ مقادیر گمشده ویژگی‌ها در مجموعه داده PIMA [۲]

ردیف	ویژگی	گمشده		تعداد مقادیر
		تعداد	درصد	معتبر
۱	Pregnancies	۰	-	۷۶۸
۲	PG Concentration	۵	٪۰/۶۵	۷۶۳
۳	Diastolic BP	۳۵	٪۴/۵۶	۷۳۳
۴	Tri Fold Thick	۲۲۷	٪۲۹/۵۶	۵۴۱
۵	Serum Ins	۳۷۴	٪۴۸/۷	۳۹۴
۶	BMI	۱۱	٪۱/۴۳	۷۵۷
۷	DP Function	۰	-	۷۶۸
۸	Age	۰	-	۷۶۸
۹	Diabetes	۰	-	۷۶۸

References

1. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Proc Annu Symp Comput Appl Med Care 1988;261-5.

2. Barhate R, Kulkarni P. Analysis of classifiers for prediction of type ii diabetes mellitus. Fourth International Conference on Computing Communication Control and Automation (ICCUBEA); 2018 Aug 16-18; Pune, India, India: IEEE; 2018. p. 1-6. doi: 10.1109/ICCUBEA.2018.8697856

3. Santhanam T, Padmavathi MS. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Computer Science 2015;47:76-83. https://doi.org/10.1016/j.procs.2015.03.185

4. Alirezaei M, Niaki ST, Niaki SA. A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines. Expert Systems with Applications 2019;127:47-57. https://doi.org/10.1016/j.eswa.2019.02.037

5. Bashir S, Qamar U, Khan FH, Naseem L. HMV: A medical decision support framework using multi-layer classifiers for disease prediction. Journal of Computational Science 2016;13:10-25. doi: 10.1016/j.jocs.2016.01.001

6. Farid DM, Zhang L, Rahman CM, Hossain MA, Strachan R. Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert Systems with Applications 2014;41(4):1937-46. doi: 10.1016/j.eswa.2013.08.089

7. Patil BM, Joshi RC, Toshniwal D. Hybrid prediction model for type-2 diabetic patients. Expert Systems with Applications 2010;37(12):8102-8. doi: 10.1016/j.eswa.2010.05.078

8. Rubaiat SY, Rahman MM, Hasan MK. Important feature selection & accuracy comparisons of different machine learning models for early diabetes detection. International Conference on Innovation in Engineering and Technology (ICIET); 2018 Dec 27-28; Dhaka, Bangladesh, Bangladesh: IEEE; 2018. p. 1-6. doi: 10.1109/CIET.2018.8660831

9. Zhu C, Idemudia CU, Feng W. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked 2019;17:100179. https://doi.org/10.1016/j.imu.2019.100179

10. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics 2020; 1-6. doi: 10.1016/j.aci.2018.12.004

11. Xie J, Gao H, Xie W, Liu X, Grant PW. Robust clustering by detecting density peaks and assigning

points based on fuzzy weighted K-nearest neighbors. Information Sciences 2016;354:19-40. https://doi.org/10.1016/j.ins.2016.03.011

12. Maniruzzaman M, Rahman MJ, Al-MehediHasan M, Suri HS, Abedin MM, El-Baz A, Suri JS. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. Journal of Medical Systems 2018;42(5):92. doi: 10.1007/s10916-018-0940-7

13. Bhat VH, Rao PG, Shenoy PD, Venugopal KR, Patnaik LM. An efficient prediction model for diabetic database using soft computing techniques. Proceedings of the 12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing; 2009 Dec 15; Berlin, Heidelberg: Springer; 2009. p. 328-35. https://doi.org/10.1007/978-3-642-10646-0_40

14. Dzulkalnine MF, Sallehuddin R. Missing data imputation with fuzzy feature selection for diabetes dataset. SN Applied Sciences 2019;1(4):362. doi: 10.1007/s42452-019-0383-x

15. Ramezani R, Maadi M, Khatami SM. A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. Alexandria Engineering Journal 2018;57(3):1883-91.

16. Gautam C, Ravi V. Counter propagation auto-associative neural network based data imputation. Information Sciences 2015;325:288-99. https://doi.org/10.1016/j.ins.2015.07.016

17. Karadogan SG, Marchegiani L, Hansen LK, Larsen J. How efficient is estimation with missing data?. International Conference on Acoustics, Speech, and Signal Processing (ICASSP); 2011 May 22-27; Prague, Czech: IEEE; 2011. p. 2260-3. doi: 10.1109/ICASSP.2011.5946932

18. Wei Y, Tang Y, McNicholas PD. Mixtures of generalized hyperbolic distributions and mixtures of skew-t distributions for model-based clustering with incomplete data. Computational Statistics & Data Analysis 2019;130:18-41. doi: 10.1016/j.csda.2018.08.016

19. Cheng CH, Chan CP, Sheu YJ. A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. Engineering Applications of Artificial Intelligence 2019;81:283-99. https://doi.org/10.1016/j.engappai.2019.03.003

20. Yadav M, Ravi V. Quantile Regression Random Forest Hybrids Based Data Imputation. 17th International Conference on Cognitive Informatics & Cognitive Computing; 2018 Jul 16-18; Berkeley, CA, USA: IEEE. p. 195-201. doi: 10.1109/ICCI-CC.2018.8482040

21. Kumar MA, Aroquiaraj IL. Adaptive Divergence Weight Firefly Algorithm (ADWFA) with Improved K-Means Algorithm and Adaptive Neuro Fuzzy Inference System (ANFIS) for Type 2 Diabetes Mellitus Prediction. Journal of Advanced Research in Dynamical and Control Systems 2009; 11(6): 18-31.

22. Sutanto DH, Ghani MK. A Benchmark of Classification Framework for Non-Communicable Disease Prediction: A Review. ARPN Journal of Engineering and Applied Sciences 2015; 10(20): 9941-55.

23. Marion R, Bibal A, Frénay B. BIR: A method for selecting the best interpretable multidimensional scaling rotation using external variables. Neurocomputing 2019;342:83-96. https://doi.org/10.1016/j.neucom.2018.11.093

24. Maniruzzaman M, Kumar N, Abedin MM, Islam MS, Suri HS, El-Baz AS, et al. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Comput Methods Programs Biomed 2017;152:23-34. doi: 10.1016/j.cmpb.2017.09.004

25. Shahid AH, Singh MP. Computational intelligence techniques for medical diagnosis and prognosis: Problems and current developments. Biocybernetics and Biomedical Engineering 2019;39(3):638-72.

26. Sabbagh Gol H. A detection of type2 diabetes using C4. 5 decision Tree. Journal of Health and Biomedical Informatics 2018;5(2):293-303. [In Persian]