**Summary of PID References**

**Data & Features**

The Pima Indian Diabetes (PID) dataset is originally from the national institute of diabetes and digestive and kidney diseases (NIDDK). PID dataset consists of 768 female patients’ record from a near Phoenix, Arizona, USA population who were examined for diabetes, and this dataset is composed of diabetic (268, 34.9%) and non-diabetic (500, 65.1%). Except the diabetes identifier, PID is comprised of 8 numeric attributes which contains personal health data and medical examination results. **Table 1** shows the description and statistic of the attributes of PID.



**Table 1**: Attributes of PID dataset

The PID dataset doesn’t have blank value, but some variables, such as Glucose and Diastolic Blood Pressure, show value as 0, which is not reasonable. Data quality is important, so we need to fix the issue of missing value. Based on the domain knowledge, the above 8 attributes’ value is supposed to related to if the patient is diabetic or not, which means we should assign value by diabetes identifier. Also, mean is a measure impacted by the extreme value, which needs to be avoided. So, in this research, median by diabetes (or not) is assigned to each variable with missing value; if median is 0, mean value is applied.

Correlation is calculated by Pearson’s correlation method.

Chart

Description automatically generated with low confidence

Normalization – improve speed and reduce runtime complexity. More: PCA & K-means

**Prelim Results – Logistic Regression**

Text

Description automatically generated with low confidence

**References**

1. T.M. Alam, et al., Informatics in medicine unlocked a model for early prediction of diabetes, Inform. Med. Unlocked 16 (2019) 100204.
   1. ANN – 75.7%
2. D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, Procedia Comput. Sci. 132 (2018) 1578–1585.
   1. SVM
   2. NB – 76.30%
   3. DT
3. N.P. Tigga, S. Garg, Predicting type 2 Diabetes using Logistic Regression accepted to publish in: Lecture Notes of Electrical Engineering, Springer.
   1. LR – 75.32%
   2. # of pregnancies, BMI & glucose level are the most significant variables
4. Salim Amour Diwani, Anael Sam, Diabetes forecasting using supervised learning techniques, Adv. Comput. Sci.: Int. J. [S.l.] (ISSN: 2322-5157) (2014) 10–18, Available at: <http://www.acsij.org/acsij/article/view/156>.
   1. 10 cross-validations
   2. NB – 76.3021%
   3. DT
5. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting Diabetes Mellitus with Machine Learning Techniques, Vol. 9, Frontiers in genetics, 2018, p. 515, <http://dx.doi.org/10.3389/fgene.2018.00515>.
   1. RF – 77.21%
   2. DT
   3. ANN
   4. PCA
   5. Minimum Redundancy Maximum Relevance (mRMR)
6. Bokhare, Anuja and Vandan Raj, N. 2023 International Conference for Advancement in Technology (ICONAT) Advancement in Technology (ICONAT), 2023 International Conference for. :1-5 Jan, 2023.
   1. LR – 76.66%
   2. RF – 74%