IDMPF: An Intelligent Diabetes Mellitus Prediction Framework using Machine Learning

Table S1: Evaluation of Past Works on DT-based RF and SVM Classification Models for Diabetes Prediction.

Work	Dataset			Accuracy	Precision	Recall	F-measure	ROC curve	AUC	Execution time
	# features	Features	# observations							
	RF									
[1]	8	Age, body mass index (BMI), diabetes pedigree function, diastolic blood pressure, number of times pregnant, plasma glucose concentration, triceps skin fold thickness and 2-Hour serum insulin	768	74.8%	0.656^{+} 0.791^{-} 0.744^{A}	0.59^{+} 0.834^{-} 0.749^{A}	0.621^{+} 0.812^{-} 0.745^{A}	×	×	×
[2]	11	Age, alcohol consumption, BMI, diastolic blood pressure, family history of diabetes, family history of hypertension, gender, smoking, systolic blood pressure, waist and weight	30122	85.09%	×	×	×	×	0.883	×
[3]	14	Age, breathe, fasting glucose, height, HDL, left diastolic pressure, left systolic pressure, low density lipoprotein (LDL), physique index, pulse rate, right diastolic pressure, right systolic pressure, waistline and weight	68994	80.84% 75.08%*	×	0.849^{A} 0.833^{A*}	×	×	×	×
	8	Age, BMI, diabetes pedigree function, diastolic blood pressure, number of times pregnant, plasma glucose concentration, triceps skin fold thickness and 2-Hour serum insulin	392	76.04% 77.21%*	×	0.757^{A} 0.745^{A*}	×	×	×	×
	SVM (Linear, Polynomial, RBF and Sigmoid kernels)									
[4]	14	Age, alcohol use, BMI, education, family history of diabetes, gender, height, household income, hypertension, physical activity, race and ethnicity, smoking, waist circumference and weight	6314	×	×	0.776	×	✓	0.824	×
	SVM (×)									
[5]	13	Age, BMI, diastolic blood pressure, family history of diabetes, gender, height, history of aborted baby, history of gestational diabetes, history of high blood pressure,	2536	SVM - 81.19%	×	SVM - 0.830^{A}	×	×	×	×

		history of pregnancy, history of use of drugs for high blood pressure, systolic blood pressure and weight								
[6]	8	Age, BMI, diabetes pedigree function, diastolic blood pressure, number of times pregnant, plasma glucose concentration, triceps skin fold thickness and 2-Hour serum insulin	768	SVM - 65.10%	SVM – 0.424^A	SVM – 0.651^A	SVM – 0.513^A	✗	SVM – 0.5	✗
[7]	9	Gender, age, plasma glucose fasting, plasma glucose postprandial, serum urea, serum creatinine, serum sodium, serum potassium and HBAIC	545	SVM – 65.04%	SVM – 0.499^A	SVM – 0.65^A	SVM – 0.52^A	✗	SVM - 0.497	✗
	RF and SVM (Linear kernel)									
[8]	✗	✗	✗	✗	✗	✗	✗	✓	RF - 0.879 SVM - 0.841	✗
	RF and SVM (✗)									
[9]	8	Age, BMI, diabetes pedigree function, diastolic blood pressure, number of times pregnant, plasma glucose concentration, triceps skin fold thickness and 2-Hour serum insulin	✗	RF - 100% SVM - 77.73%	✗	RF – 1.00^A SVM - 0.513^A	✗	✗	✗	✗
[10]	15	Age, blood glucose level, blood pressure, BMI, diabetes pedigree function, gender, HBAIC, insulin, plasma glucose fasting, plasma glucose postprandial, pregnancy, serum creatinine, serum potassium, serum sodium and skin thickness	2500	RF – 75.39% RF – 79%* SVM – 77.73% SVM – 80%*	✗	RF - 0.332^A SVM – 0.570^A	✗	✓	✗	✗
	RF and SVM (Linear, Polynomial, RBF and Sigmoid kernels)									
This paper	11	Age, alcohol consumption, blood pressure, blurred vision, cholesterol, gender, heart disease, obesity, pregnancy, race and uric acid	65,839	✓	✓	✓	✓	✓	✓	✓

✗ → Not reported

$^+$Value for diabetic class

$^-$Value for non-diabetic class

^{A}Weighted average value of both diabetic and non-diabetic classes values

*Computed for the dataset after feature selection

REFERENCES

1. Tamilvanan B, Bhaskaran VM (2017) An Experimental Study of Diabetes Disease Prediction System Using Classification Techniques. IOSR J Comput Eng 19:39–44. https://doi.org/10.9790/0661-1901043944
2. Nai-arun N, Moungmai R (2015) Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Comput Sci 69:132–142. https://doi.org/10.1016/j.procs.2015.10.014
3. Zou Q, Qu K, Luo Y, et al (2018) Predicting Diabetes Mellitus With Machine Learning Techniques. Front Genet 9:. https://doi.org/10.3389/fgene.2018.00515
4. Yu W, Liu T, Valdez R, et al (2010) Application of support vector machine modeling for prediction of common diseases:

the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak 10:. https://doi.org/10.1186/1472-6947-10-16
5. Heydari M, Teimouri M, Heshmati Z, Alavinia SM (2016) Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. Int J Diabetes Dev Ctries 36:167–173. https://doi.org/10.1007/s13410-015-0374-4
6. Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. Procedia Comput Sci 132:1578–1585. https://doi.org/10.1016/j.procs.2018.05.122
7. Saravananathan K, Vekmurugan T (2016) Analyzing Diabetic Data using Classification Algorithms in Data Mining. Indian J Sci Technol 9:1–6
8. Khalilia M, Chakraborty S, Popescu M (2011) Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 11:. https://doi.org/10.1186/1472-6947-11-51
9. Kandhasamy JP, Balamurali S (2015) Performance Analysis of Classifier Models to Predict Diabetes Mellitus. Procedia Comput Sci 47:45–51. https://doi.org/10.1016/j.procs.2015.03.182
10. Sneha N, Gangil T (2019) Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data 6: