Diabetes Detection

Team Name: Kustlers

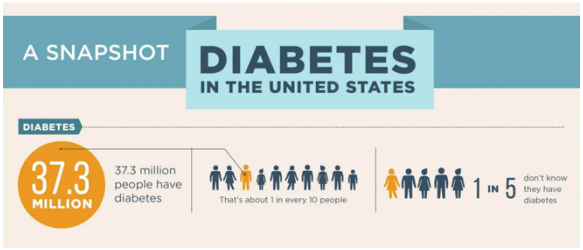
Byungkwon Min
Technology, Cybersecurity and Policy
University of Colorado Boulder
Boulder, CO, USA
byumgkwon.min@colorado.edu

Jongbae Yoon
Applied Mathematics
University of Colorado Boulder
Boulder, CO, USA
jongbae.yoon@colorado.edu

KEYWORDS
Diabetes Mellitus, WHtR, BMI, Machine Learning, Supervised Learning, Binary Classification, Logistics Regression, SGD Classifier, K Nearest Neighbors, Linear Discriminant Analysis, Support Vector Classifier, Decision Tree, Random Forest, Gradient Boosting Classifier, Light GBM Classifier

1. Problem Space

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Diabetes is very serious and costly because it can damage the heart, blood vessels, eyes, kidneys, and nerves. Early diagnosis is a key factor to lower the risk of diabetes-related disease and the cost as well. However, 37.3 million Americans have diabetes, but about 1 in 5 people with diabetes don't even know they have it [1]. This fact motivates us to develop a machine learning model that everybody can easily diagnose diabetes at home for free.



<Figure 1. CDC 2022 Diabetes Report>

The main goal of this project is to find the best model for a binary classification problem that whether a user has a diabetes or not using only basic information of the user.

Our solution for this problem is unique for the following reasons.

First, most of the works, if not all, use BMI (Body Mass Index) as one of the key features of their models. However, BMI was first introduced in the 19th century and proved that it is not an accurate predictor for the following reasons:

- BMI does not measure body fat percentage
- BMI does not account for different demographics.
- BMI does not measure body fat distribution.

Various studies claim that Waist to Height Ratio (WHtR) is more valid indicator than BMI to predict type-2 diabetes. Fatemeh claims that WHtR is a more accurate tool for predicting hypertension than BMI in patients With Type 2 Diabetes [2]. Thus, for this project instead of using BMI, we use WHtR for the feature of our model.

Second, it is widely known that family history and obesity are two of the most critical and impactful factors in diabetes. This makes us assume that if someone whose family member has diabetes is more susceptible to diabetes if he/she is obese than someone who is obese but do not have any of family members diagnosed with diabetes. So, our hypothesis is that if we categorize the dataset into sub datasets based on 'family history' and 'obesity', the same model would perform differently on each dataset, thus each dataset would have different best model for itself. Hence, these more specified models would perform better for people in the specified category than the one generally built on not specified, entire dataset. Thus, we divide the whole dataset into 4 different categories and find the best model for each. Then we compare the results with the one built on the whole dataset to find out in which case we can find the best performing models.

2. Approach

The approach to solve this problem is categorized as below 4 steps.

2.1. Data preprocessing

First, data gathering and pre-processing. We have two datasets, one for diabetes, and the other for body measurements. The samples with missing values are removed from the dataset or replaced with median values. There are some categorical features, and they are transformed into binary or one hot encoded values and normalization (Standard-scaling) is applied. As mentioned in the section 1. Problem space, we use WHtR instead of BMI as an indicator of physical status. However, the diabetes dataset we use do not contain WHtR or body measurement that we can compute the WHtR. So, we use a separate dataset which has individual body measurements. We calculate the WHtR and BMI from the body measurements dataset and find the formula converting BMI to WHtR. Eun-Gyong Yoo claims that WHtR cutoff of 0.5 can be used in different sex and ethnic groups and is generally accepted as a universal cutoff for central obesity [3]. We use the same cutoff value 0.5 and convert WHtR values to binary value, 1 for obesity and 0 for normal.

2.2. Split Dataset

After data preprocessing, we match each sample's BMI of the diabetes dataset with the obesity category we just computed. As mentioned in the section 1, since family history and physical status are the most impactful factors in type-2 diabetes, the whole dataset is split into 4 different datasets based on the family history and obesity. So, we have total 5 datasets, the entire dataset, and 4 sub-categorized datasets. Each dataset including the entire dataset is split into training, and test dataset by 75% and 25% respectively.

2.3. Testing Machine Learning Models

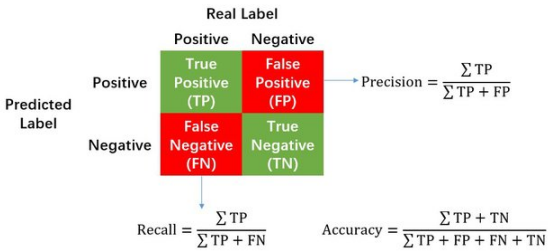
Most widely and commonly used 9 classification machine learning models are considered as follows:

- Logistics Regression
- Stochastic Gradient Descent (SGD) Classifier
- K Nearest Neighbors
- Linear Discriminant Analysis
- Support Vector Classifier
- Decision Tree
- Random Forest
- Gradient Boosting Classifier
- Light Gradient Boosting Classifier

These 9 different machine learning models are trained on the training datasets and tested on the test datasets to find the best performing model for each dataset split. In addition, we trained with Stratified KFold set to 10, and found the optimal parameters using RandomizedSearchCV. Among the results of these tests, the highest performance was output as the result of the model.

2.4. Evaluation

For the evaluation, we measured accuracy, precision, recall, and F-1 scores for each model we built. Since this is a medical problem, diagnosing 'people with diabetes' as 'having no-diabetes' is more severe than diagnosing 'people without diabetes' as 'having diabetes'. Thus, the model minimizing the 'False Negative' is the best performing model. In other words, the model with the highest 'Recall' score is the best model for this problem.



<Figure 2. Evaluations in Confusion Matrix >

Once we find the best model for each dataset, we compare the performance. If the models tested on the sub-categorized dataset perform better than the model tested on the entire dataset, then the result proves our hypothesis which is that separate model on each sub-categorized dataset returns more accurate and reliable results. Thus, when people diagnose diabetes, they should use different models based on their family history and WHtR value.

3. DATA

We use two different datasets for this project. The first dataset is diabetes dataset. The dataset is from Kaggle [4]. 952 individuals' data have been collected through an online and offline questionnaire with 18

features related to age, health, lifestyle, family background, etc.

diabetes_dataset_2019

Age	Gender	Family_Diabetes	highBP	PhysicallyActive	BMI	Smoking	Alcohol	Sleep	SoundSleep	RegularMedicine	JunkFood	Stress	BPLevel	Pregancies	Pdiabetes	UrinationFreq	Diabetic
50-59	Male	no	yes	one hr or more	39	no	no	8	6	no	occasionally	sometimes	high	0	0	not much	no
50-59	Male	no	yes	less than half an hr	28	no	no	8	6	yes	very often	sometimes	normal	0	0	not much	no
40-49	Male	no	no	one hr or more	24	no	no	6	6	no	occasionally	sometimes	normal	0	0	not much	no
50-59	Male	no	no	one hr or more	23	no	no	8	6	no	occasionally	sometimes	normal	0	0	not much	no
40-49	Male	no	no	less than half an hr	27	no	no	8	8	no	occasionally	sometimes	normal	0	0	not much	no
40-49	Male	no	yes	none	21	no	yes	10	10	no	occasionally	sometimes	high	0	0	not much	yes
less than 40	Male	no	no	one hr or more	24	no	no	8	8	no	occasionally	sometimes	normal	0	0	not much	no
less than 40	Male	no	no	less than half an hr	20	no	no	7	7	yes	occasionally	sometimes	low	0	0	not much	no

<Figure 3. 'Diabetes' dataset>

The below features are categorical datasets and transformed to either binary or one hot encoded.

- Age: less than 40, 40-49, 50-59, 60 or older
- Gender: Male / Female
- Family_Diabetes: yes / no
- highBP(Blood Pressure): yes / no
- PhysicallyActive: none, less than half an hour, more than half an hour
- Smoking: yes / no
- RegularMedicine: yes / no
- JunkFood: Occasionally, Often, Very often, always
- Stress: not at all, sometimes, very often, always
- BPLevel: low, normal, high
- Pregnancies (Number of pregnancies): 0, 1, 2, 3, 4
- Pdiabetes (Diabetes during pregnancies): yes / no
- UritationFreq: not much, quite often

The second dataset is body measurement dataset or 'bodyfat' dataset, and this is also from Kaggle [5]. It has 251 individuals' body measurements including height, weight, abdomen circumference, etc. (15 features)

bodyfat

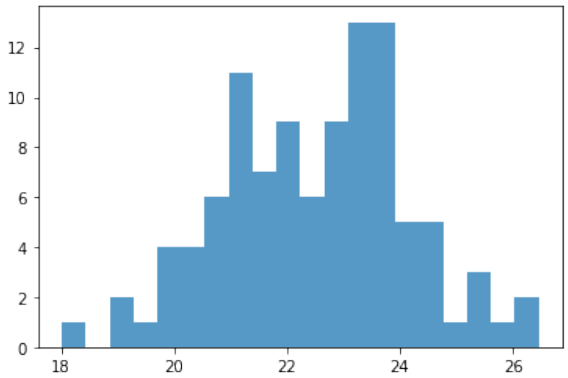
Density	BodyFat	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
1.0708	12.3	23	154.25	67.75	36.2	93.1	85.2	94.5	59.0	37.3	21.9	32.0	27.4	17.1
1.0853	6.1	22	173.25	72.25	38.5	93.6	83.0	98.7	58.7	37.3	23.4	30.5	28.9	18.2
1.0414	25.3	22	154.00	66.25	34.0	95.8	87.9	99.2	59.6	38.9	24.0	28.8	25.2	16.6
1.0751	10.4	26	184.75	72.25	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2
1.0340	28.7	24	184.25	71.25	34.4	97.3	100.0	101.9	63.2	42.2	24.0	32.2	27.7	17.7
1.0502	20.9	24	210.25	74.75	39.0	104.5	94.4	107.8	66.0	42.0	25.6	35.7	30.6	18.8

<Figure 4. 'bodyfat' dataset>

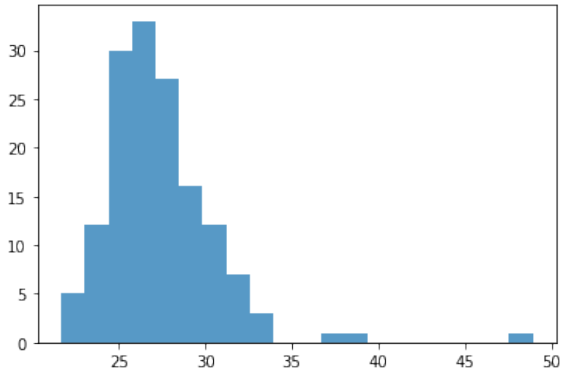
Since the diabetes dataset does not contain WHtR values, we use bodyfat dataset to compute WHtR. All we need from this dataset is Height, Weight, and Abdomen. The formula for WHtR and BMI are as follows:

- BMI = weight (lb) / [height (in)]^2 x 703
- WHtR = Waist(Abdomen) / Height

As briefly introduced in the section 2, we use 0.5 for the cutoff value, thus people with WHtR value less than 0.5 are labeled as 'Normal' while people with WHtR value larger than 0.5 are labels as 'Obese'. After that, we calculate the mean of BMI of each group, and find the distribution of each group. The mean BMI for normal people is 22.5 and the mean BMI for obese people is 27.5. Thus, we conclude that BMI value 25 can be used as the cutoff value for the central obesity.



<Figure 5. Histogram for Normal>



<Figure 6. Histogram for Obesity>

With this BMI based cut-off value along with 'family history', we divide the entire 'diabetes' dataset into 4 sub-categorized datasets as below.

		Family History	
		O	X
Obesity	O	225	216
	X	229	282

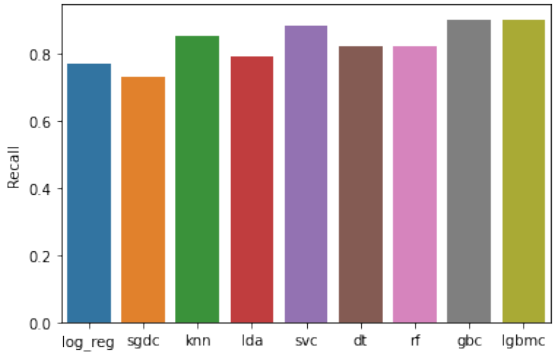
<Figure 7. Sub-categories with number of samples>

4. Results

We've trained and tested 9 different models on 5 different datasets, and the results are as follows:

4.1. Whole dataset

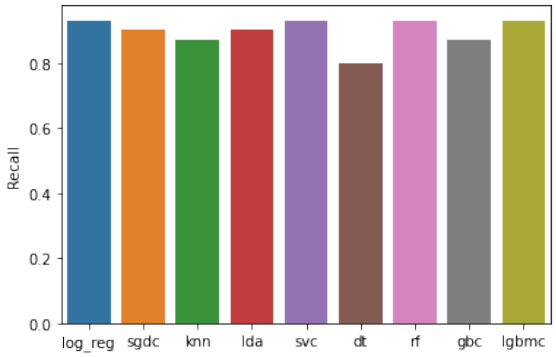
The best performing models on the whole dataset in terms of 'Recall' are Gradient Boosting Classifier and Light Gradient Boosting Classifier. Both return 0.9 of recall for the entire dataset.



<Figure 8. Recall values per model on whole dataset>

4.2. Family history: yes, Obesity: yes

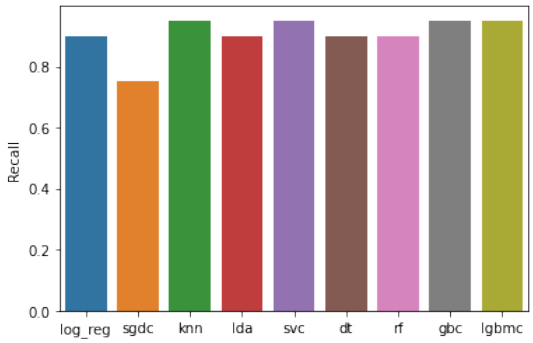
The best performing models on the first sub-dataset, 'family history: yes, obesity: yes', are Logistics Regression, Support Vector Classifier, Random Forest, and Light Gradient Boosting Classifier, and the recall value is 0.93.



<Figure 9. Recall values per model on 1st sub-dataset>

4.3. Family history: yes, Obesity: no

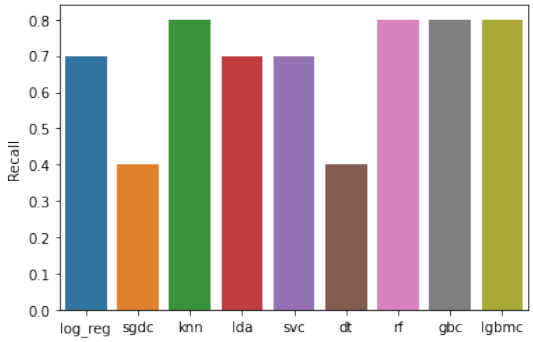
The best performing models on the second sub-dataset, 'family history: yes, obesity: no', are K Nearest Neighbors, Support Vector Classifier, Gradient Boosting Classifier, and Light Gradient Boosting Classifier, and the returned recall value is 0.95.



<Figure 10. Recall values per model on 2nd sub-dataset>

4.4. Family history: no, Obesity: yes

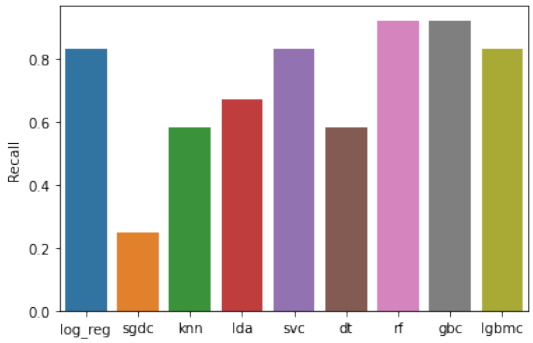
The best performing models on the third sub-dataset, 'family history: no, obesity: yes', are K Nearest Neighbors, Random Forest, Gradient Boosting Classifier, and Light Gradient Boosting Classifier, and the returned recall value is 0.8.



<Figure 11. Recall values per model on 3rd sub-dataset>

4.5. Family history: no, Obesity: no

The best performing models on the last sub-dataset, 'family history: no, obesity: no', are Random Forest, and Gradient Boosting Classifier models, and the returned recall value is 0.92.



<Figure 12. Recall values per model on 4th sub-dataset>

4.6. Summary

The best performing models and their recall values per each dataset are summarized as below table.

Dataset	Model	Recall
Whole dataset	GBC, LGBMC	0.9
Family History: yes, Obesity: yes	LR, SVC, GBC, LGBMC	0.93
Family History: yes, Obesity: no	KNN, SVC, GBC, LGBMC	0.95
Family History: no, Obesity: yes	KNN, RF, GBC, LGBMC	0.8
Family History: no, Obesity: no	RF, GBC	0.92

<Figure 13. Summary Table>

The recall value of the best performing models on the whole dataset is 0.9, while the same values of the best performing models on the separate datasets are 0.93, 0.95, 0.8, and 0.92 respectively.

5. Discussion

5.1. Conclusion

As shown in the figure 13. Summary Table, except for the dataset 'family history: no, obesity: yes', all the recall values of the best models from separate datasets are larger than the recall value of the model on the whole dataset. This proves that separate model on each sub-categorized dataset returns more accurate and reliable results. Thus, we should use different models on different group of people based on their family history and WHtR values.

5.2. Limitations

One of the biggest limitations of this work is its dataset. Since the diabetes dataset has no body measurements of each sample, we use another dataset to compute WHtR and find the relationship between WHtR and BMI for conversion. There should be some errors due to this conversion process. In addition, the number of samples are relatively small. The diabetes dataset has only 952 samples, which is considered not enough for machine learning. Moreover, we divide 952 datasets into 4 sub-datasets, so each has less than 300 samples. This is not even close to enough number of samples for machine learning problem. Thus, the result could be affected by this small number of samples. If we had larger datasets, the result would have been more reliable.

5.3. Potential Future Works

Our potential future work is the same diabetes detection, but this time instead of inputting all the information about a user, we input images of the user's body and return the result whether the user has a diabetes. For this work, we consider using neural network algorithms, potentially Convolutional Neural Network. Since this is supervised learning, we would need lots of body images with labels. With this model, people easily detect whether they have a diabetes or not.

REFERENCES

[1] CDC(Center for Disease Control and Prevention), National diabetesstatistica l report, 2022, https://www.cdc.gov/diabetes/health-equity/

[2] Fatemeh et al., Waist-To-Height Ratio Is a More Accurate Tool for Predicting Hypertension Than Waist-To-Hip Circumference and BMI in Patients With Type 2 Diabetes: A Prospective Study, https://pubmed.ncbi.nlm.nih.gov/34692623/

[3] Eun-Gyong Yoo, Waist-to-height ratio as a screening tool for obesity and cardiometabolic risk https://pubmed.ncbi.nlm.nih.gov/27895689/

[4] NEHA PRERNA TIGGA ·, Diabetes Dataset 2019, https://www.kaggle.com/datasets/tigganeha4/diabetes-dataset-2019

[5] FEDESORIANO, Body Fat Prediction Dataset https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset