AASD4001 Applied Mathematical Concepts for Machine Learning

DIABETES CLASSIFICATION

Group 6

- Ahmed Ali
- Nichapat Boonprasertsri
- Anuphap Chansatit
- Karthikeyan Jeyabalasuntharam
- Yat Chit Law
- Halari Shanpru
- Vitchaya Siripoppohn
- Chotiros Srisiam

01 DATASET OVERVIEW

Id	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1

SKIN THICKNESS
Triceps skinfold thickness measured in millimeters (mm).

INSULIN
2-Hour serum insulin levels measured in micro international units per milliliter (mu U/ml).

BLOOD PRESSURE
Diastolic blood pressure measured in millimeters of mercury (mm Hg)

BMI
Body mass index, calculated as weight in kilograms divided by height in meters squared (kg/m²)

2,768 SAMPLES
WITHIN
8 FEATURES

GLUCOSE
Plasma glucose concentration measured over 2 hours in an oral glucose tolerance test.

DIABETES PEDIGREE FUNCTION
A genetic score representing the likelihood of diabetes.

PREGNANCIES
Number of times the individual has been pregnant.

AGE
Age of the individual in years.

Data source: Healthcare Diabetes Dataset - https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes

01 DATASET OVERVIEW

EDA

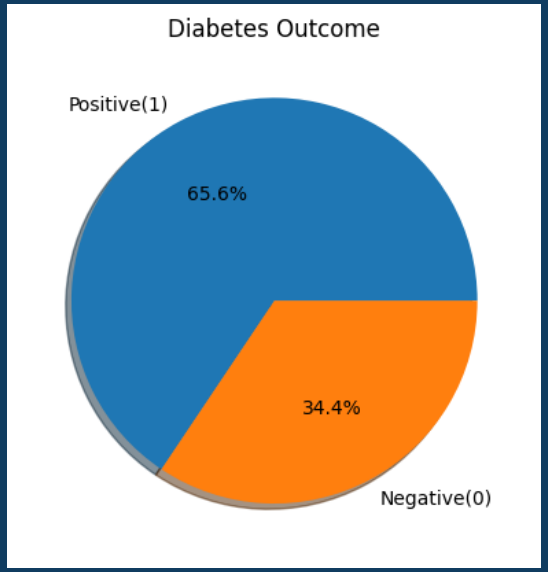
HOW MANY DUPLICATED SAMPLES?

```
0
```

HOW MANY MISSING VALUES?

```
Id                          0
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

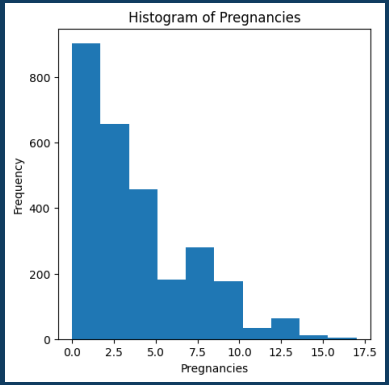
DIABETES OUTCOME:

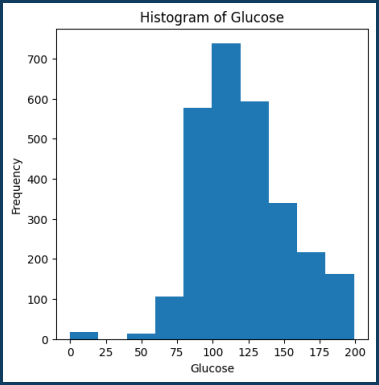


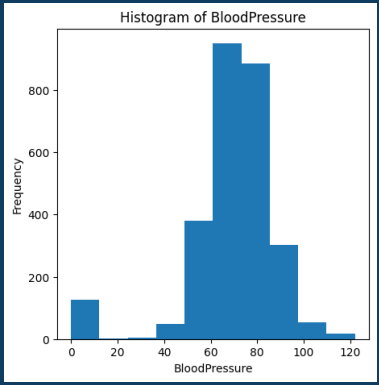
01 DATASET OVERVIEW

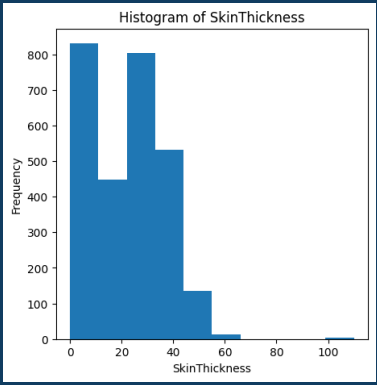
EDA

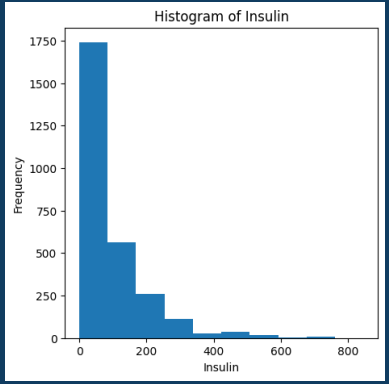
FEATURE HISTOGRAMS:

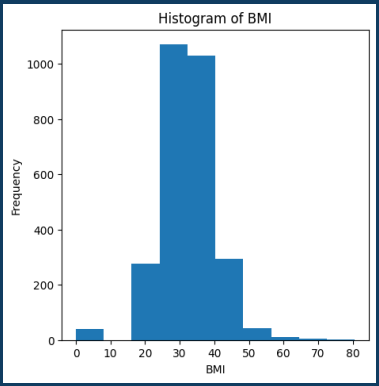


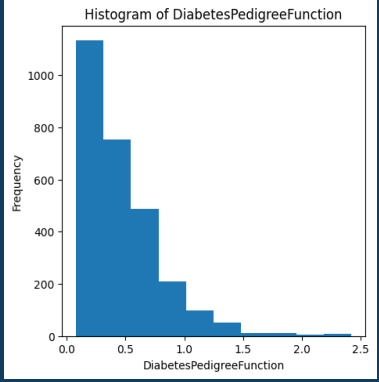


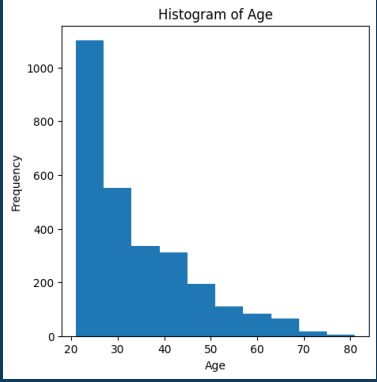












02 DATA PREPROCESSING

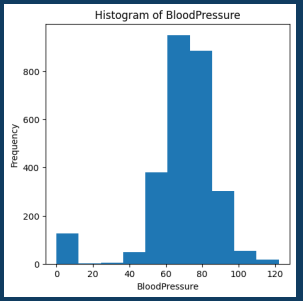
REMOVING DATA

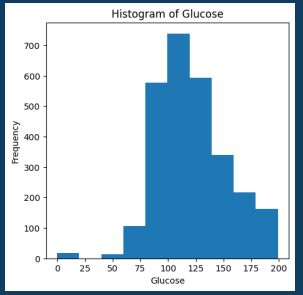
1.1) Blood Pressure (missing 125)
Diastolic BP is the latter number when you take a blood pressure test.
Recommend measure for healthy people is more than 80 [1].

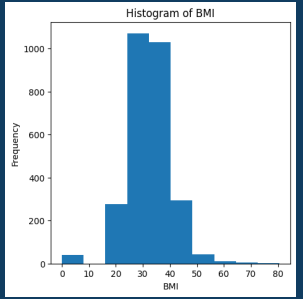
1.2) Glucose Tolerance Test (missing 18)
The recommend measure is the range between 70-140
Lower than the range considered as Hypoglycemia
More than the range considered impaired glucose tolerance, or diabetes [2].

1.3) BMI (missing 39)
Recommend measure is the range between 18.5 to 25 [3]
Lower than average means underweight
Greater than average means overweight, or obesity

The missing data in these features were very small compared to the total number (182 vs 2768). We considered removing it as the first approach.



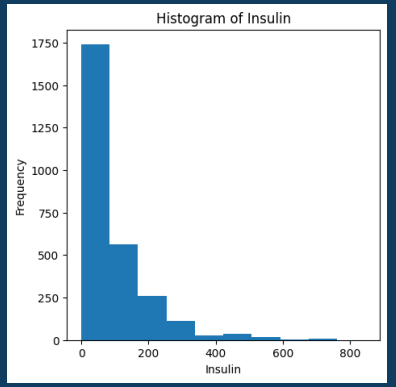




[1] NIH National Institute on Aging (NIA). *High Blood Pressure and Older Adults*. U.S. Department of Health and Human Services, National Institutes of Health. Retrieved September 22, 2023, from www.nia.nih.gov
[2] Rao SS, Disraeli P, McGregor T. *Impaired glucose tolerance and impaired fasting glucose. Am Fam Physician*. 2004;69(8):1961-1968.
[3] Centers for disease control and prevention (CDC). *Healthy Weight, Nutrition, and Physical Activity*. Division of Nutrition, Physical Activity, and Obesity, National Center for Chronic Disease Prevention and Health Promotion. Retrieved September 22, 2023, from www.cdc.gov

02 DATA PREPROCESSING

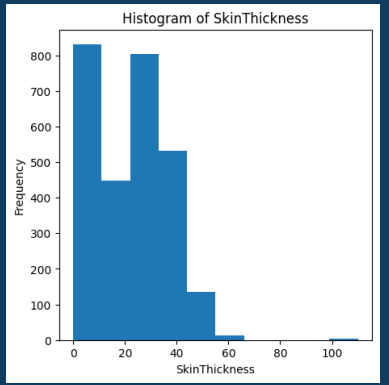
IMPUTATION



2.1) Insulin Test (missing 1330)

Insulin moves glucose from the blood into cells, particularly into muscles and fat tissue

The average value of serum insulin metabolic parameter is 14.3±6.4 μU/ml [1].



2.2) Tricept Skinfold Thickness (missing 800)

The average TSF thickness is 14.3 - 31.7 mm.

The world smallest TSF is 10 mm [2].

The missing values of these two features were approximately 50% of the dataset. Removing all of these values may affect the model performance. Thus, we suggested imputation method instead, and replaced 0 with median values.

[1] Chevenne D, Trivin F, Parquet D. Insulin assays and reference values. *Diabetes Metab*. 1999;25(6):459-476.

[2] Li W, Yin H, Chen Y, et al. Associations between adult triceps skinfold thickness and all-cause, cardiovascular and cerebrovascular mortality in nhanes 1999-2010: a retrospective national study. *Front Cardiovasc Med*. 2022;9:858994. doi:10.3389/fcvm.2022.858994

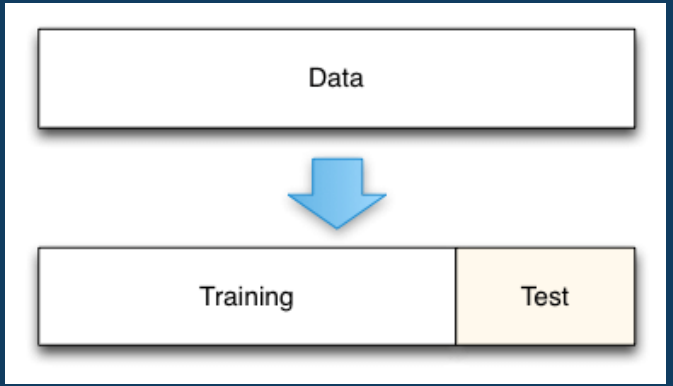
02 DATA PREPROCESSING

DATA SPLITTING AND FEATURE SCALING

Dataset is separated into Feature Variable (X) and Target Variable (y)

Data Splitting

- Train Data – 80%
- Test Data – 20%



Feature Scaling

- Features were scaled down with "StandardScaler" from scikit-learn.
- Standardize the features to 0 mean and 1 standard deviation

03 DATASET CLASSIFICATION

LOGISTIC REGRESSION

DEFAULT	
Parameter	max_iter =100
Accuracy	76.02%
APPLYING GRIDSEARCHCV HYPERPARAMETER	
Best parameter	max_iter =1000
Accuracy	79.21% (↑)

DIFFERENCES

- The maximum tier has been increased to 1000, and the accuracy has been improved
- Using this model with GridSearchCV is recommended

03 DATASET CLASSIFICATION

DECISION TREE

DEFAULT	
Parameter	criterion = gini and splitter = best
Accuracy	98.90%
APPLYING GRIDSEARCHCV HYPERPARAMETER	
Best parameter	criterion = entropy and splitter = random
Accuracy	95.67% (↓)

DIFFERENCES

- The criteria and splitter were changed to entropy and random, resulting in lower accuracy
- Using this model without GridSearchCV is recommended

03 DATASET CLASSIFICATION

RANDOM FOREST

DEFAULT	
Parameter	criterion = gini and n_estimators = 100
Accuracy	99.53%
APPLYING GRIDSEARCHCV HYPERPARAMETER	
Best parameter	criterion = gini and n_estimators = 200
Accuracy	96.06%(↓)
DIFFERENCES	
• Increasing n_estimators to 200 did not improve accuracy • Using this model without GridSearchCV is recommended	

03 DATASET CLASSIFICATION

SUPPORT VECTOR MACHINES

DEFAULT	
Parameter	kernel = rbf
Accuracy	74.61%
APPLYING GRIDSEARCHCV HYPERPARAMETER	
Best parameter	kernel = rbf
Accuracy	84.54% (↑)
DIFFERENCES	

- The parameter does not change, but the accuracy increases
- Using this model with GridSearchCV is recommended

STOCHASTIC GRADIENT DESCENT

DEFAULT	
Parameter	alpha = 0.0001, loss = hinge, penalty = l2, max_tier = 1000
Accuracy	82.76%
APPLYING GRIDSEARCHCV HYPERPARAMETER	
Best parameter	alpha = 0.01, loss = modified_huber, penalty = None, max_tier = 5000
Accuracy	80.21% (↓)

DIFFERENCES

- The alpha increased to 0.01, loss changed to modified_huber, the penalty changed to none, and max_tier increased to 5000, but the accuracy decreased
- Using this model with GridSearchCV is recommended

03 DATASET CLASSIFICATION

RANDOMLY REMOVE FEATURE

	LOGISTIC REGRESSION	DECISION TREE	RANDOM FOREST	SGD	SVM
PREGNANCIES	76.97%	97.85%	98.37%	70.61%	77.31%
GLUCOSE	70.23%	97.32%	97.70%	67.60%	70.23%
BLOOD PRESSURE	77.32%	97.37%	98.23%	70.42%	77.12%
SKIN THICKNESS	77.70%	97.65%	98.13%	74.77%	77.02%
INSULIN	77.36%	97.61%	98.13%	74.77%	77.02%
BMI	77.02%	97.56%	97.89%	70.56%	77.26%
DIABETES PEDIGREE FUNCTION	77.12%	97.56%	98.42%	71.04%	77.17%
AGE	77.26%	97.08%	97.99%	69.70%	77.02%

CONCLUSIONS

- Diabetes Prediction Dataset is the dataset of our choice
- Pre-processing was done ensuring minimal drops
- Classification on the dataset was implemented using various approaches
- Default parameter values were first obtained
- Using GridSearchCV, the optimal results were obtained and compared
- Features were then randomly dropped to see their effect on the model
- Based on the obtained results, Random Forest is found to be the best approach

THANKS!

Do you have any questions?
Feel free to ask us..