PREDICTION OF DIABETES BY CLASSIFICATION MODELS

TABLE OF CONTENTS

01

BACKGROUND AND MOTIVATION

DIABETES AROUND THE WORLD IN 2021

DIABETES AROUND THE WORLD IN 2021:

537
million
adults are living with diabetes

3 in 4
adults with diabetes
live in low- and middle-income countries

6.7
million
deaths due to diabetes in 2021

537 million adults (20-79 years) are living with diabetes - 1 in 10. This number is predicted to rise to 643 million by 2030 and 783 million by 2045.

Over 3 in 4 adults with diabetes live in low- and middle-income countries.

Diabetes is responsible for 6.7 million deaths in 2021 - 1 every 5 seconds.

Diabetes caused at least U.S. 966 billion dollars in health expenditure – a 316% increase over the last 15 years.

541 million adults have Impaired Glucose Tolerance (IGT), which places them at high risk of type 2 diabetes.

02

EXPLORATORY DATA ANALYSIS

DATA INFORMATION

A dataset of **253,680** survey responses to the CDC's *The Behavioral Risk Factor Surveillance System* (BRFSS) 2015.

- **Target variable: Diabetes**
- **21 Feature Variables:**
    - **Categorical:**
        - 'HighBP', 'HighChol', 'CholCheck', 'Smoker', 'Stroke', 'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'Veggies', 'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost', 'GenHlth', 'DiffWalk', 'Sex', 'Age', 'Education', 'Income'
    - **Numerical:**
        - 'BMI', 'MentHlth', 'PhysHlth'

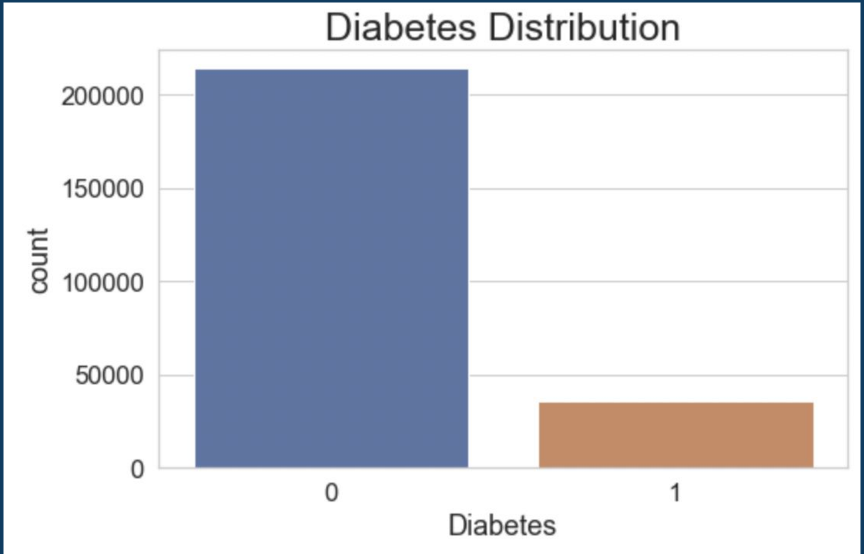
DATA CLEANING

1. CHECK DATA TYPES
2. HANDLE MISSING VALUES
3. DROP UNNECESSARY FEATURES

HighBP	float64
HighChol	float64
CholCheck	float64
BMI	float64
Smoker	float64
Stroke	float64
HeartDiseaseorAttack	float64
PhysActivity	float64
Fruits	float64
Veggies	float64
HvyAlcoholConsump	float64
AnyHealthcare	float64
NoDocbcCost	float64
GenHlth	float64
MentHlth	float64
PhysHlth	float64
DiffWalk	float64
Sex	float64
Age	float64
Education	float64
Income	float64

DATASET VISUALIZATION

TARGET VARIABLE: DIABETES



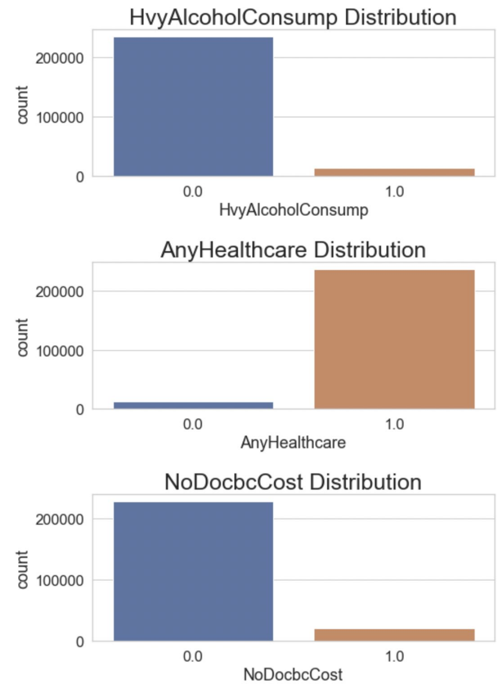
- **1** for diabetes
- **0** for no diabetes or only during pregnancy

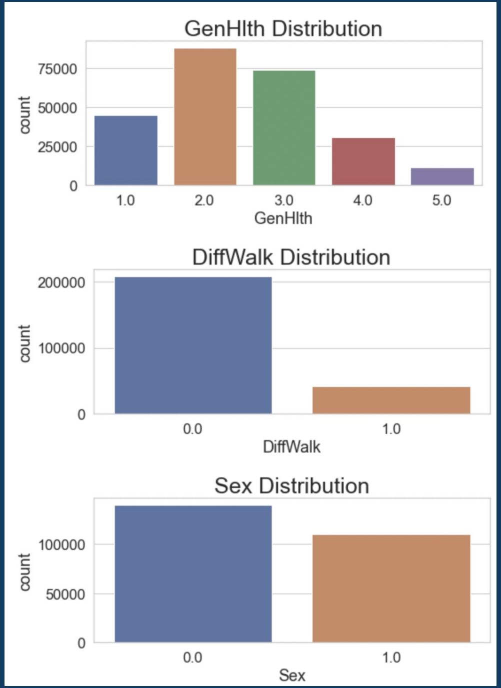
"There is class imbalance in this dataset."

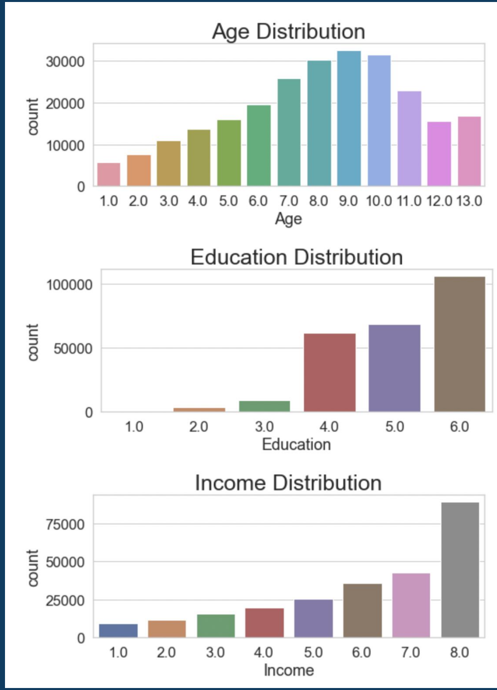
DATASET VISUALIZATION

CATEGORICAL VARIABLES

IMBALANCED





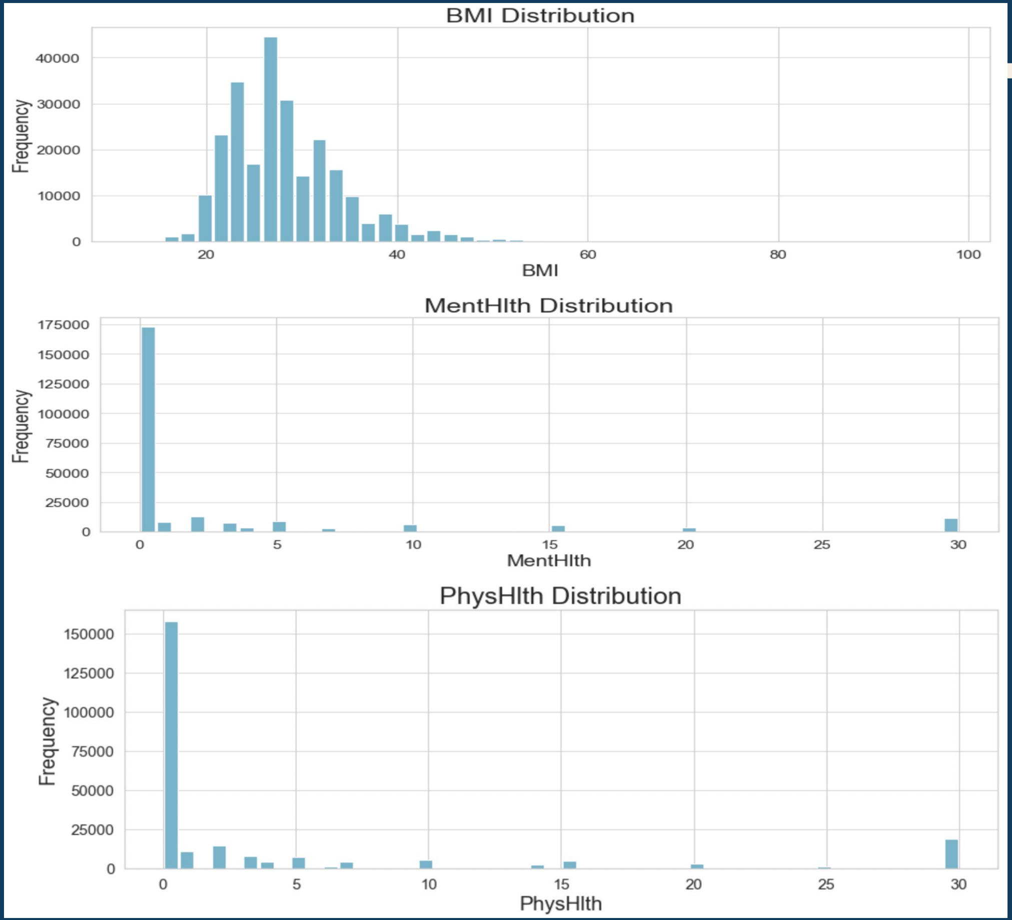


DATASET VISUALIZATION

NUMERICAL VARIABLES

Right-skewed?

Log Transformation



EXAMINE RELATIONSHIP BETWEEN FEATURES AND DIABETES

Diabetes	0	1	count	0_rate	1_rate
HighBP					
0.0	134391	8742	143133	0.938924	0.061076
1.0	79312	26604	105916	0.748820	0.251180

Diabetes	0	1	count	0_rate	1_rate
HighChol					
0.0	132673	11660	144333	0.919215	0.080785
1.0	81030	23686	104716	0.773807	0.226193

Diabetes	0	1	count	0_rate	1_rate
CholCheck					
0.0	9167	241	9408	0.974384	0.025616
1.0	204536	35105	239641	0.853510	0.146490

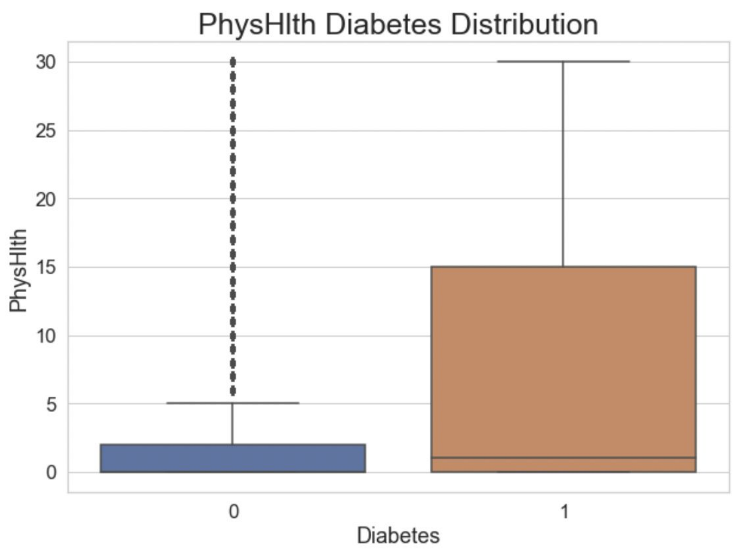
For most categorical features, the positive rates are **significantly** different for different classes

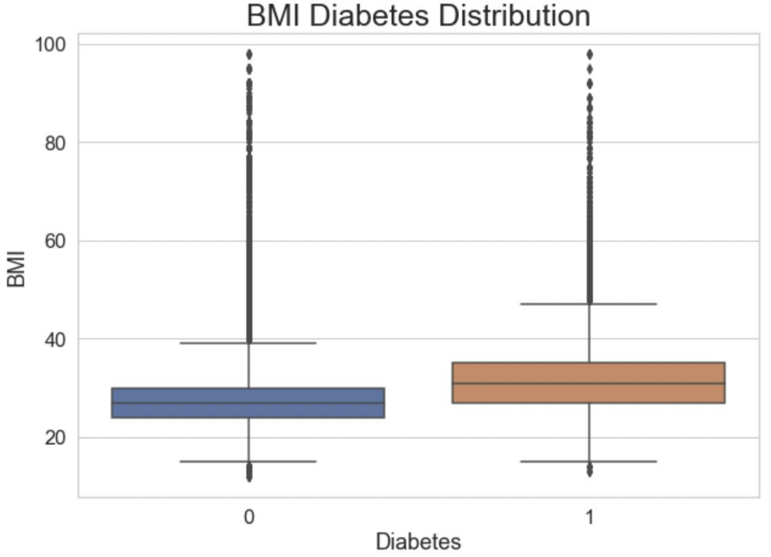
Diabetes	0	1	count	0_rate	1_rate
GenHlth					
1.0	43846	1140	44986	0.974659	0.025341
2.0	81489	6381	87870	0.927381	0.072619
3.0	60461	13457	73918	0.817947	0.182053
4.0	20755	9790	30545	0.679489	0.320511
5.0	7152	4578	11730	0.609719	0.390281

EXAMINE RELATIONSHIP BETWEEN FEATURES AND DIABETES

Similarly, for numerical features,

the distributions for ‘diabetes’ and ‘no diabetes’ are different.

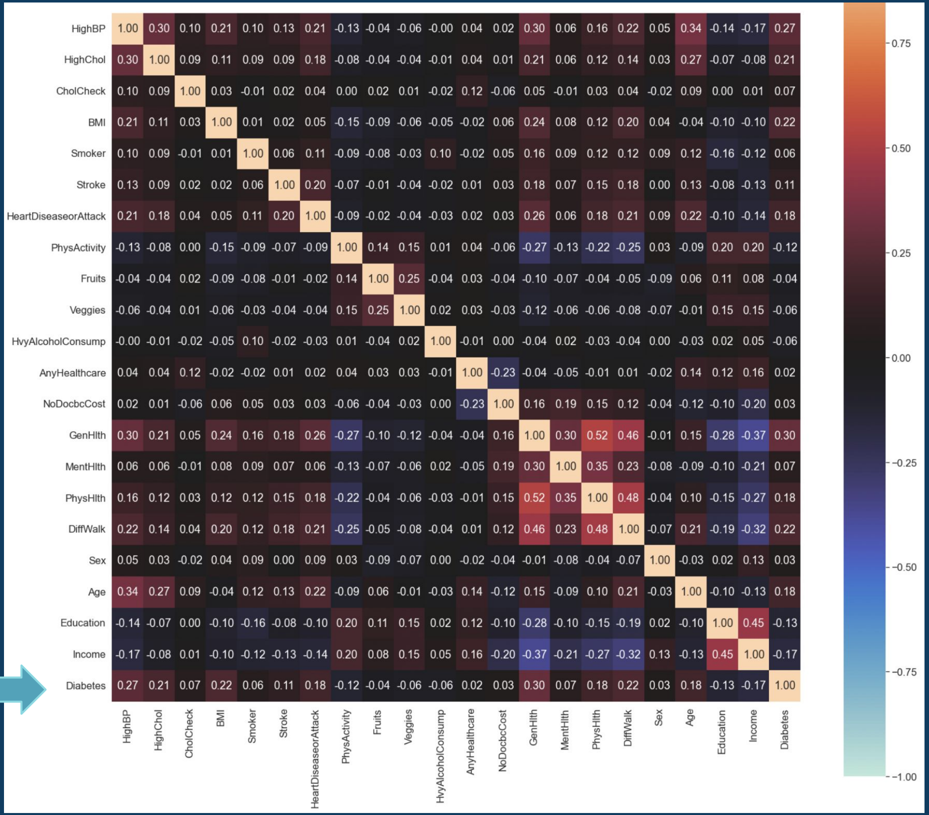




COMPUTE CORRELATION

If abs(correlation)< 0.05, the feature is nearly uncorrelated to diabetes.

Thus we drop those features to increase the efficiency of our model.



03

CLASSIFICATION MODELS

BACKGROUND INFORMATION

- Why we use classification?

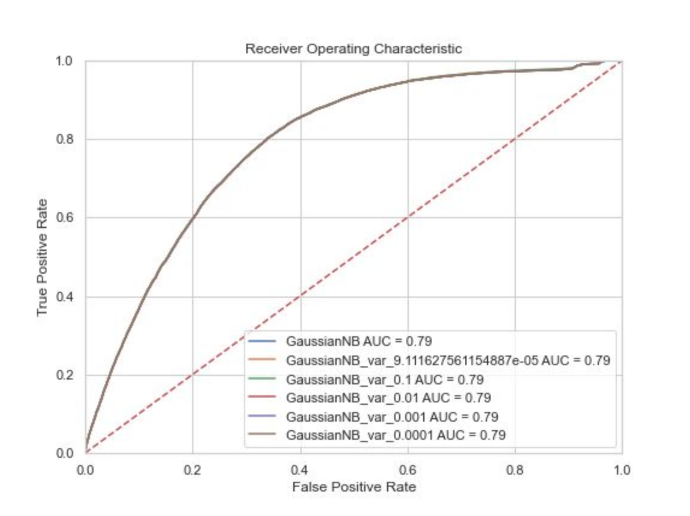
  Categorical data

- What is the result we should focus on?

  Recall instead of Accuracy

- What we do next for finding the best fitting
  - Imbalanced dataset $\xrightarrow{\text{Under sample}}$ balanced dataset
  - Hyperparameter tuning

NAIVE BAYES (GAUSSIAN)



- **Select Parameter:**

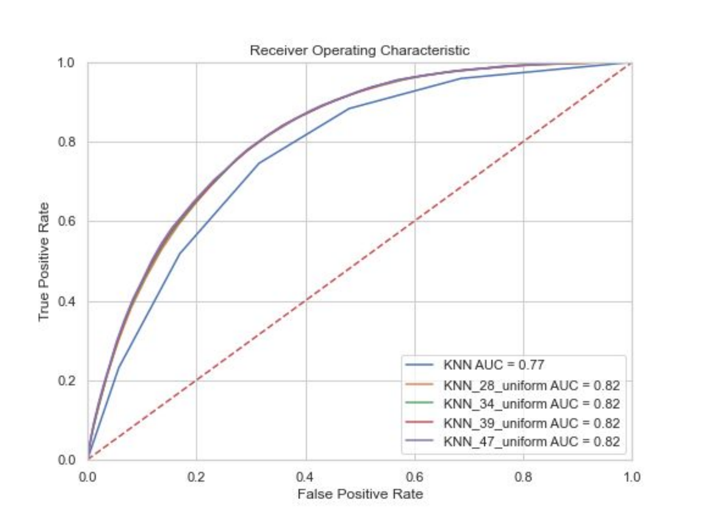
  var_smoothing

- **Best fitting:**

  var_smoothing = 9.11*10^(-5)

	acc	precision	recall	f1	auc
GaussianNB	0.77544000535367700	0.33311709373986400	0.581101471142965	0.42347604975603100	0.78879516188977700
GaussianNB_var_9.111627561154887e-05	0.77544000535367700	0.33311709373986400	0.581101471142965	0.42347604975603100	0.78879622538618500
GaussianNB_var_0.1	0.78412634678444800	0.339808640185561	0.55262165220671400	0.42084096376889700	0.78999600672544000
GaussianNB_var_0.01	0.77640366726895500	0.33389590592334500	0.57836665409279500	0.42337429242026800	0.78891976086411000
GaussianNB_var_0.001	0.77552031051328400	0.3331529971867560	0.58072425499811400	0.42340484048404800	0.78880774733274100
GaussianNB_var_0.0001	0.77542662116041000	0.333081040168676	0.58100716710675200	0.42342187553692300	0.78879630628840100

K NEAREST NEIGHBOR



- **Select Parameter:**

  N_neighbors, weighting , leaf_size

- **Best fitting:**

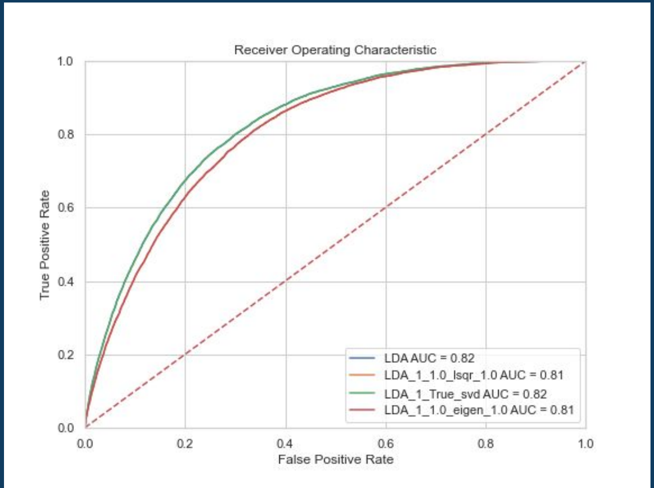
  N_neighbors = 47

  weighting = ’uniform’

  leaf_size = 30

	acc	precision	recall	f1	auc
KNN	0.6934082848156330	0.2812488887308420	0.745850622406639	0.4084699806326660	0.7721542776317300
KNN_28_uniform	0.7160008030515960	0.3044039063939560	0.7789513391173140	0.4377434484220570	0.8153723688370760
KNN_34_uniform	0.7133105802047780	0.3028292256088670	0.7832893247831010	0.4367900715187210	0.8166289633900170
KNN_39_uniform	0.7030315197751460	0.2966149308237940	0.7965861938890990	0.4322706105112330	0.8174521456476890
KNN_47_uniform	0.7036873452452650	0.2973257897881010	0.797906450396077	0.4332198356416890	0.8188863352749880

LDA



- **Select Parameter:**

  Solver, shrinkage, n_components, tol

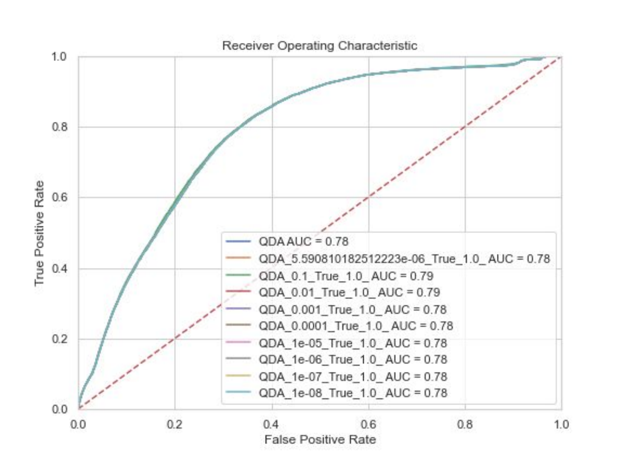
- **Best fitting:**

  Solver = ‘lsqr’ or ‘eigen’ Tol = 1

  shrinkage = 1 n_components = 1

	acc	precision	recall	f1	auc
LDA	0.859533	0.511876	0.22152	0.309221	0.824735
LDA_1_1.0_lsqr_1.0	0.812956	0.37861	0.495756	0.429336	0.807283
LDA_1_True_svd	0.859533	0.511876	0.22152	0.309221	0.824735
LDA_1_1.0_eigen_1.0	0.812956	0.37861	0.495756	0.429336	0.807283

QDA



- **Select Parameter:**

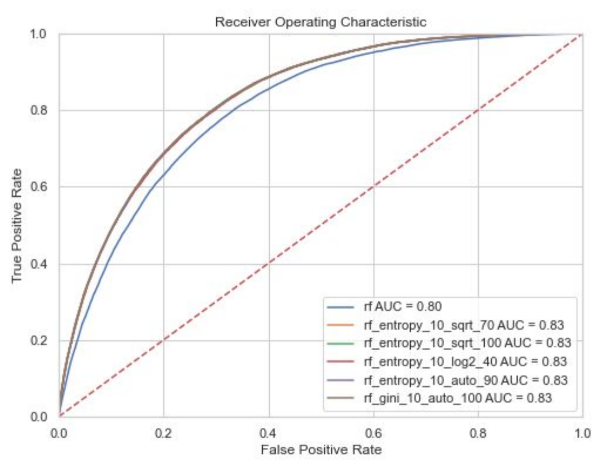
  Reg_param, store_covariance, tol

- **Best fitting:**

  None

	acc	precision	recall	f1	auc
QDA	0.772121	0.324995	0.562335	0.411923	0.784898
QDA_0.1_True_1.0_	0.782935	0.334979	0.537344	0.412689	0.787304
QDA_0.01_True_1.0_	0.773258	0.32605	0.560072	0.412159	0.785137
QDA_0.001_True_1.0_	0.772268	0.325151	0.562146	0.411998	0.784922
QDA_0.0001_True_1.0_	0.772134	0.325012	0.562335	0.411937	0.7849
QDA_1e-05_True_1.0_	0.772121	0.324995	0.562335	0.411923	0.784898

RANDOM FOREST CLASSIFICATION



- **Select Parameter:**

  Criterion, Max_depth, Max_features, N_estimators

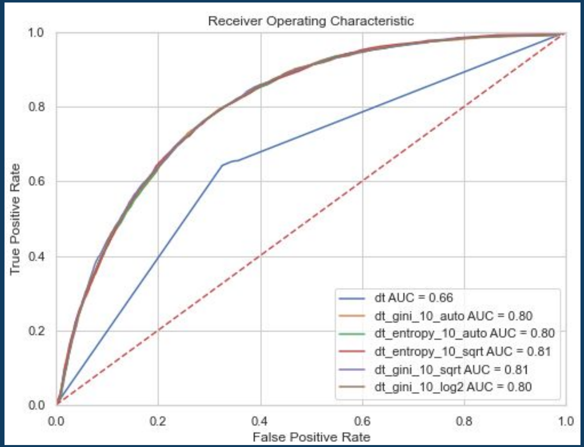
- **Best fitting:**

  Criterion = 'gini' Max_depth = 10

  Max_features = 'auto' N_estimators = 100

	acc	precision	recall	f1	auc
rf	0.707970287	0.29577157	0.765843078	0.426736029	0.804858346
rf_entropy_10_sqrt_70	0.724365924	0.313271028	0.790267823	0.448680195	0.831782696
rf_entropy_10_sqrt_100	0.725102054	0.314153167	0.791870992	0.449843302	0.832079969
rf_entropy_10_log2_40	0.724620224	0.313284147	0.788853263	0.448465353	0.831270703
rf_entropy_10_auto_90	0.723536104	0.31289554	0.792625424	0.448673464	0.831951313
rf_gini_10_auto_100	0.72352272	0.313314763	0.795548849	0.449572331	0.832108292

DECISION TREE CLASSIFICATION



- **Select Parameter:**

  Criterion, Max_depth, Max_features

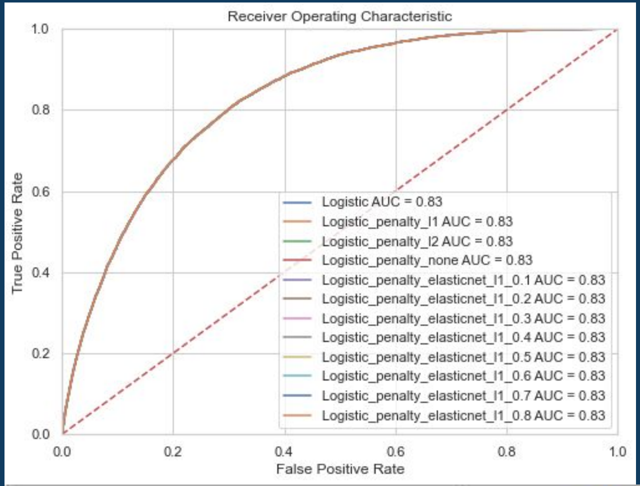
- **Best fitting:**

  Criterion = 'entropy'   Max_depth = 10

  Max_features = 'auto'

	acc	precision	recall	f1	auc
dt	0.668152312	0.245095926	0.643342135	0.354961236	0.656998475
dt_gini_10_auto	0.707823061	0.297364621	0.776782346	0.430085631	0.804287408
dt_entropy_10_auto	0.697771532	0.291115839	0.787061486	0.425024826	0.802898349
dt_entropy_10_sqrt	0.708452118	0.297092242	0.771784232	0.429032004	0.805195642
dt_gini_10_sqrt	0.712052466	0.299536969	0.768672199	0.43108737	0.805426072
dt_gini_10_log2	0.718409958	0.30322459	0.758204451	0.433201325	0.803765153

LOGISTIC REGRESSION



- **Select Parameter:**

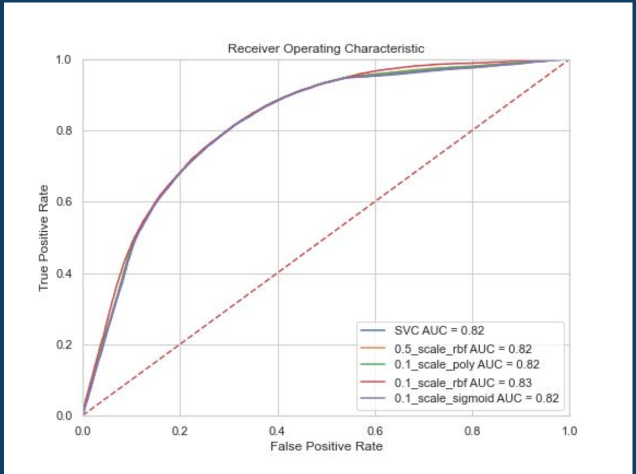
  Several regularization

- **Best fitting:**

  None

	acc	precision	recall	f1	auc
Logistic	0.862344	0.547655	0.172859	0.262777	0.827302
Logistic_penalty_l1	0.862344	0.547683	0.172765	0.262671	0.827302
Logistic_penalty_l2	0.862357	0.547818	0.172859	0.262796	0.827302
Logistic_penalty_none	0.862344	0.547655	0.172859	0.262777	0.827302
Logistic_penalty_elasticnet_l1_0.1	0.862344	0.547683	0.172765	0.262671	0.827302

SVC



- **Select Parameter:**

  C, Gamma, Kernel

- **Best fitting:**

  C = 0.1 Gamma = 'scale'

  Kernel = 'rbf'

	acc	precision	recall	f1	auc
SVC	0.7109683464	0.3045210401	0.807336854	0.4422346773	0.8180997922
0.5_scale_rbf	0.7105534364	0.3043524567	0.8084685025	0.4422263489	0.8207670206
0.1_scale_poly	0.7095228535	0.3034670822	0.8080912863	0.4412347777	0.8197418291
0.1_scale_rbf	0.7079569029	0.3029238122	0.8129007922	0.4413722478	0.8258717651
0.1_scale_sigmoid	0.7108077361	0.3043772002	0.80724255	0.4420688409	0.8169414269

04

CONCLUSIONS

MODEL SELECTION

- Choose recall score as our primary criterion. Because we want to know the number of accurate predictions within all people with diabetes.
- SVC is the best in terms of recall scores.
- Considering training time and performance, we prefer random forest

	acc	precision	recall	f1	auc	train_time	predict_time
SVC_best	0.711397	0.304938	0.807808	0.442745	0.818521	674.642445	382.396422
KNN_best	0.703152	0.297165	0.799604	0.433298	0.818462	0.003280	97.878076
QDA_best	0.686783	0.283978	0.793286	0.418237	0.784127	0.036154	0.029693
RandomForest_best	0.72446	0.313757	0.793003	0.449619	0.831344	1.638924	0.793561
DecisionTree_best	0.706873	0.297014	0.779423	0.430122	0.807748	0.027780	0.012291
Logistic_best	0.736131	0.321346	0.772727	0.453923	0.827844	0.936724	0.004051
GaussianNB_best	0.724875	0.302289	0.717465	0.425361	0.788483	0.013829	0.022791
LDA_best	0.751482	0.323305	0.687099	0.43971	0.807442	0.076563	0.002063

IN REAL WORLD

- Using ML techniques to predict diabetes risk have received wide attention（Quan Zou, 2018).
- ML models are efficient and effective in prevention.

FURTHER WORK

- We can further explore time series data to analyze the accumulation effects.
- ML models are powerful in predictions. But we need causation effects for diagnosis.
- Need cross section data to improve the generalization ability.

THANKS!

Any questions?

Citation

1. *Diabetes Around the World in 2021*, https://diabetesatlas.org/
2. *Predicting Diabetes Mellitus With Machine Learning*, Quan Zou, 2018, https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full
3. Data Resources: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
4. PPT Template : created by Slidesgo, including icons by Flaticon and infographics & images by Freepik