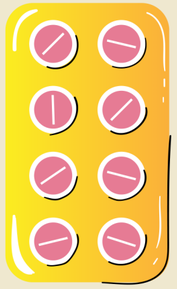
Diabetes Classification

STAT451 Group 8

Ramona Liu, Yueming Xu, Ling Zhang,
Meishu Zhao, Yingqi Zhao

Presentation Outline

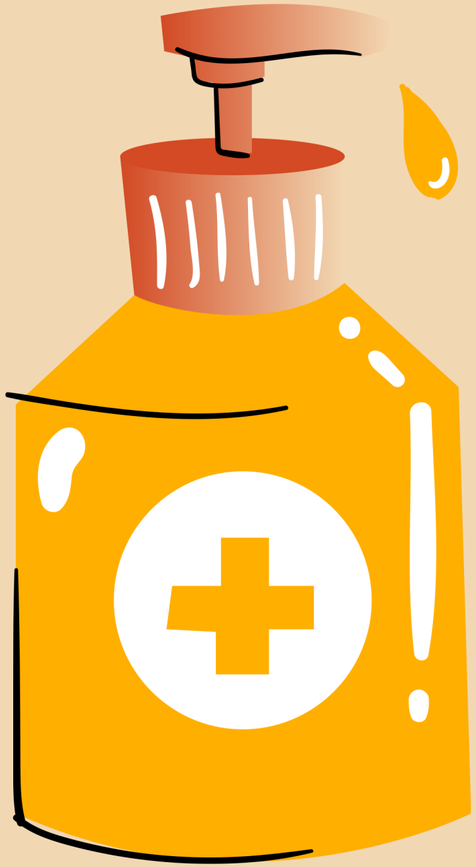
MOTIVATION
&
OBJECTIVE



MODELS

ENSEMBLE METHODS

RESULTS



Motivation

“How is diabetes different for women than it is for men? Diabetes increases the risk of **heart disease** (the most common diabetes complication) by about **four times in women but only about two times in men**, and women have worse outcomes after a heart attack. Women are also at higher risk of other diabetes-related complications such as blindness, kidney disease, and depression.”

------ *"Diabetes and Women", Centers for Disease Control and Prevention*

Objective

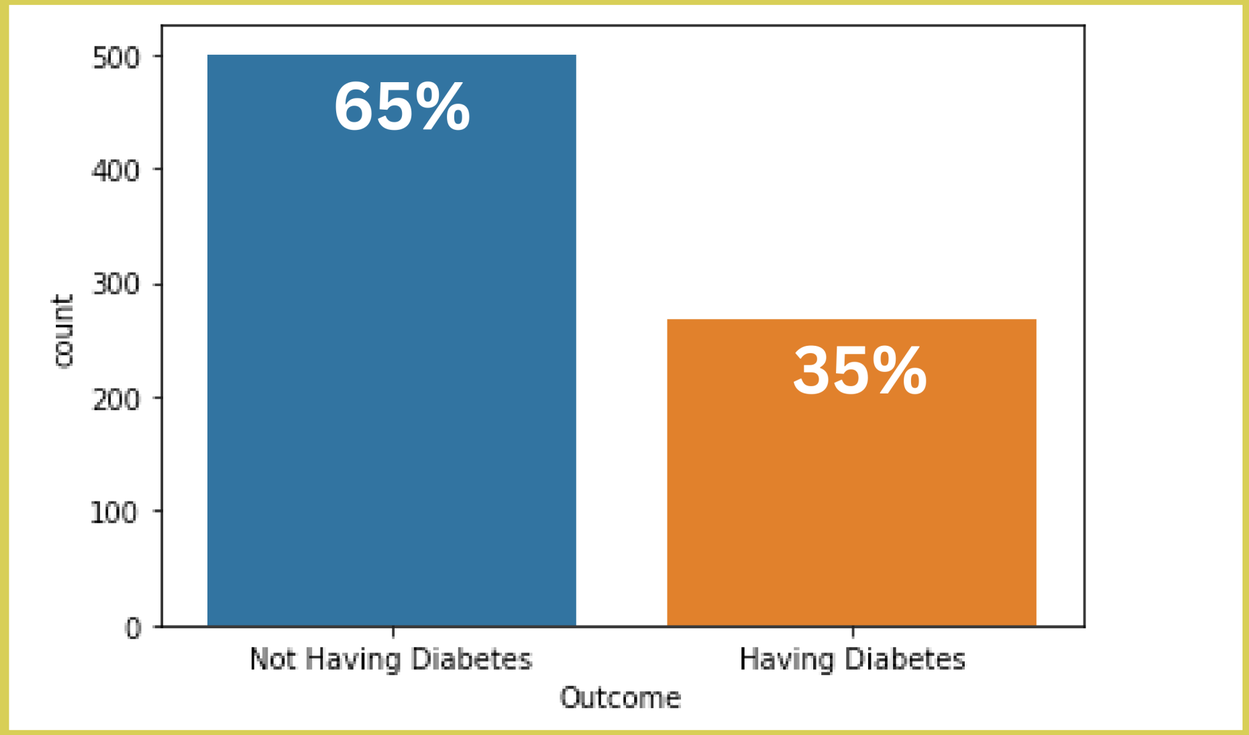
Find the best machine-learning model of determining whether a female has diabetes based on her clinical data.

Data Summary

1. **Pregnancies**: Number of times of pregnancies
2. **Glucose**: 2-hour Plasma Glucose Concentration in an oral glucose tolerance test (mg/dl)
3. **Blood Pressure**: Diastolic Blood Pressure(mmHg)
4. **Skin Thickness**: Triceps Skin Fold Thickness (mm)
5. **Insulin**: 2-hour Serum Insulin (mu U/ml) in blood
6. **BMI**: Body Mass Index (weight in kg/ height in m^2)
7. **Diabetes Pedigree Function**: Score of diabetes likelihood based on history in relatives of the patient
8. **Age**: Age of the patient
9. **Outcome**: binary values, 1 = has diabetes, 0 = does not have diabetes

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.00	768.00	768.00	768.00	768.00	768.00	768.00	768.00	768.00
mean	3.85	121.66	72.39	29.11	140.67	32.46	0.47	33.24	0.35
std	3.37	30.44	12.10	8.79	86.38	6.88	0.33	11.76	0.48
min	0.00	44.00	24.00	7.00	14.00	18.20	0.08	21.00	0.00
25%	1.00	99.75	64.00	25.00	121.50	27.50	0.24	24.00	0.00
50%	3.00	117.00	72.00	29.00	125.00	32.30	0.37	29.00	0.00
75%	6.00	140.25	80.00	32.00	127.25	36.60	0.63	41.00	1.00
max	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00	1.00

Data


500
400
300
200
100
0
count
65%
35%
Not Having Diabetes
Having Diabetes
Outcome

Models

Linear SVM

RBF Kernel SVM

Polynomial Kernal SVM

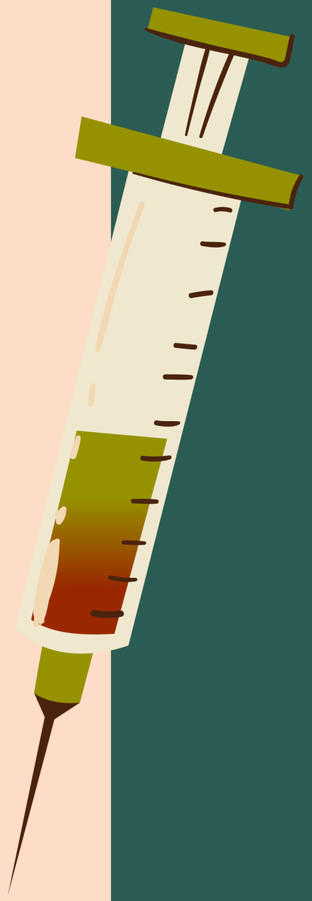
Logistic Regression

KNN

Decision Tree

Accuracy

- Use when both classes are equally important.
- Easy to interpret
- Tells you how many times the ML model was correct overall.



Recall

- Estimates how many of the Actual Positives our model captures.
- Use when False Negative has a large cost.

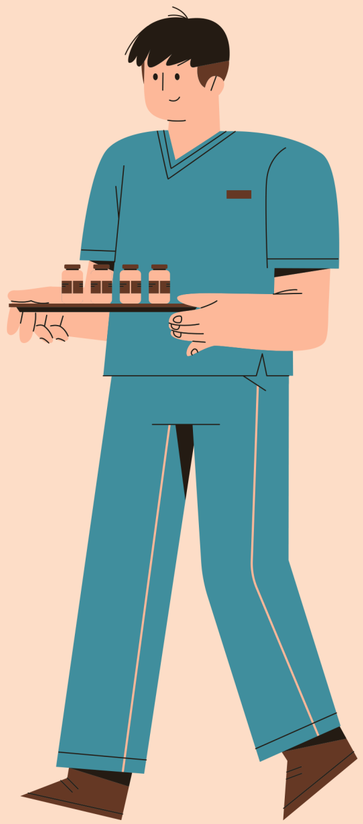
Best hyperparameter & Best Recall

RandomSearch: 5-Fold, 200 iterations

Model	Validation Recall	Best Parameters
Linear SVM	0.52	'C': 3.29
RBF Kernel SVM	0.59	'C': 9.22
Polynomial Kernel SVM	0.41	'C': 49.93, 'degree': 3
Logistic Regression	0.52	'C': 1.52, 'max_iter': 5000, 'penalty': 'l2'
Decision Tree	0.37	'criterion': 'entropy', 'max_depth': 22
KNN	0.52	'n_neighbors': 9

Ensemble Learning

- Performed on all models
  - Bagging
  - Adaptive Boosting
- Performed on Decision Tree
  - Gradient Boosting
  - Random Forest



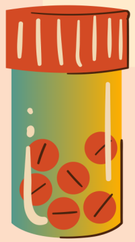
Model Performance

- Accuracy
- Recall=(TP/(TP+FN))
- Precision=(TP/(TP+FP))
- F1=(2*Precision*Recall)/(Precision + Recall)
- ROC(AUC)
- Runtime

Best hyperparameter & Best Recall

RandomSearch: 5-Fold, 200 iterations

Model	Validation Recall	Best Parameters
Gradient Boosting	0.30	'learning_rate': 1.92, 'max_depth': 2, 'n_estimators': 183
Random Forest	0.52	'criterion': 'entropy', 'max_depth': 18, 'n_estimators': 17
Bagging-Decision Tree	0.56	'base_estimator': DecisionTreeClassifier(criterion='entropy', max_depth=22), 'n_estimators': 153
Adaptive Boosting -Logistic Regression	0.48	'algorithm': 'SAMME', 'base_estimator': LogisticRegression(C=1.52 max_iter=5000, solver='liblinear'), 'n_estimators': 186



Test Results--Recall

Model	Gradient Boosting	Random Forest	Bagging-Decision Tree	Adaptive Boosting-Logistic Regression
Test Recall	0.63	0.59	0.74	0.67

Model	Linear SVM	RBF Kernel SVM	Polynomial Kernel SVM	Logistic Regression	Decision Tree	KNN
Test Recall	0.56	0.81	0.37	0.56	0.41	0.63

Test Results -- All

Model	Recall	Accuracy	Precision	F1	AUC
Linear SVM	0.56	0.79	0.65	0.79	-
RBF Kernel SVM	0.81	0.81	0.81	0.87	-
Polynomial Kernel SVM	0.37	0.77	0.50	0.74	-
Logistic Regression	0.56	0.79	0.65	0.79	0.9
Decision Tree	0.41	0.61	0.49	0.70	0.84
KNN	0.63	0.73	0.68	0.79	0.85
Gradient Boosting	0.63	0.59	0.61	0.71	0.72
Random Forest	0.59	0.70	0.64	0.77	0.86
Bagging- Decision Tree	0.74	0.83	0.78	0.86	0.93
Adaptive Boosting - Logistic Regression	0.67	0.69	0.68	0.78	0.82

Run Time

Model	Training Time (s)	Prediction Time (s)
Linear SVM	0.0150	0.0021
RBF Kernel SVM	0.0125	0.0034
Polynomial Kernel SVM	0.0608	0.0020
Logistic Regression	0.0013	0.0010
Decision Tree	0.0035	0.0009
KNN	0.0012	0.0054
Gradient Boosting	0.1838	0.0016
Random Forest	0.0359	0.0036
Bagging- Decision Tree	0.5192	0.0202
Adaptive Boosting - Logistic Regression	0.0102	0.0014

Conclusion

RBF Kernel SVM

- Highest recall
- Fairly high accuracy
- Low training & prediction time

Citations

- https://www.cdc.gov/diabetes/library/features/diabetes-and-women.html
- https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression
- https://www.analyticssteps.com/blogs/what-precision-recall-f1-score-statistics


Q
A