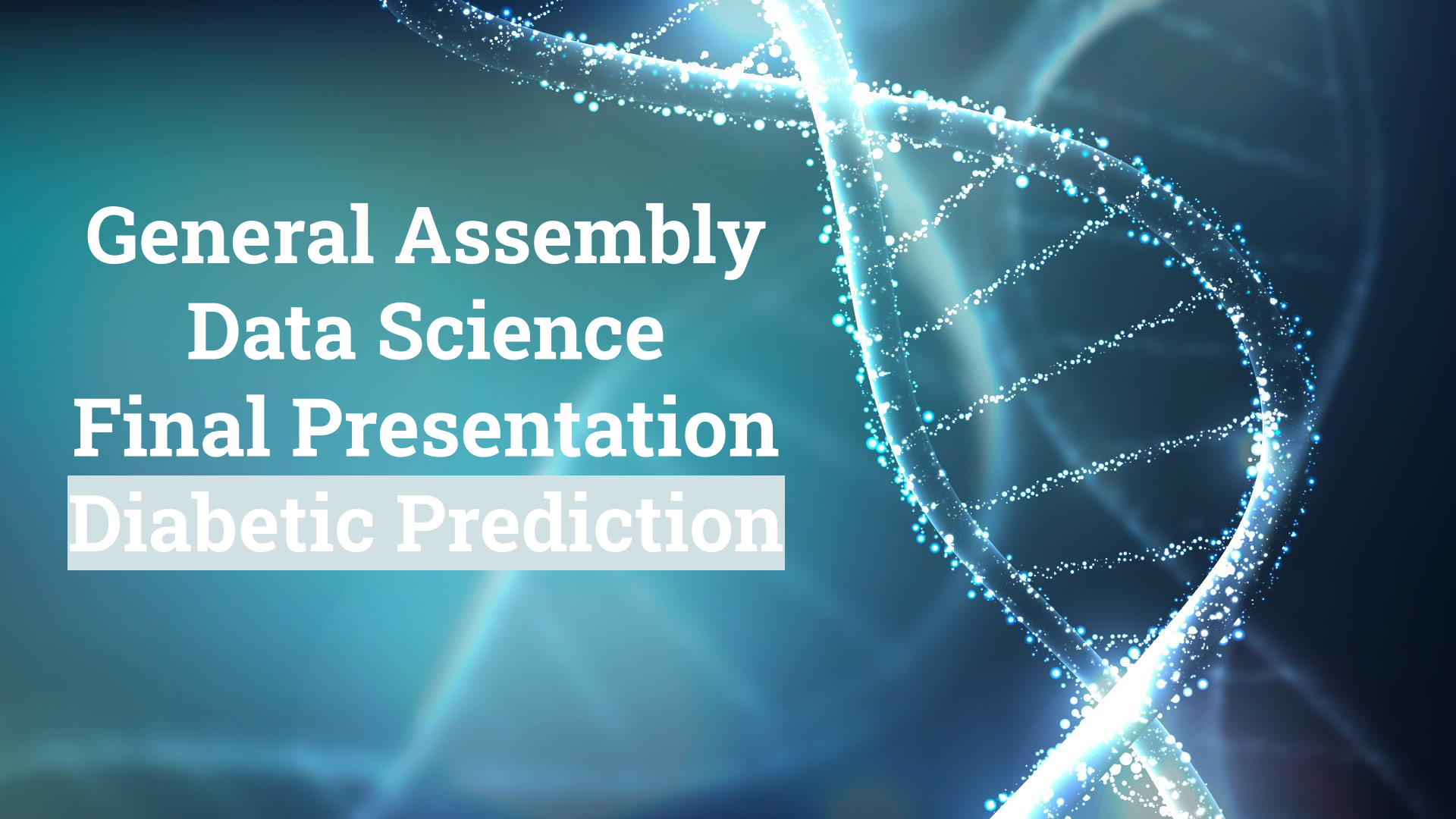
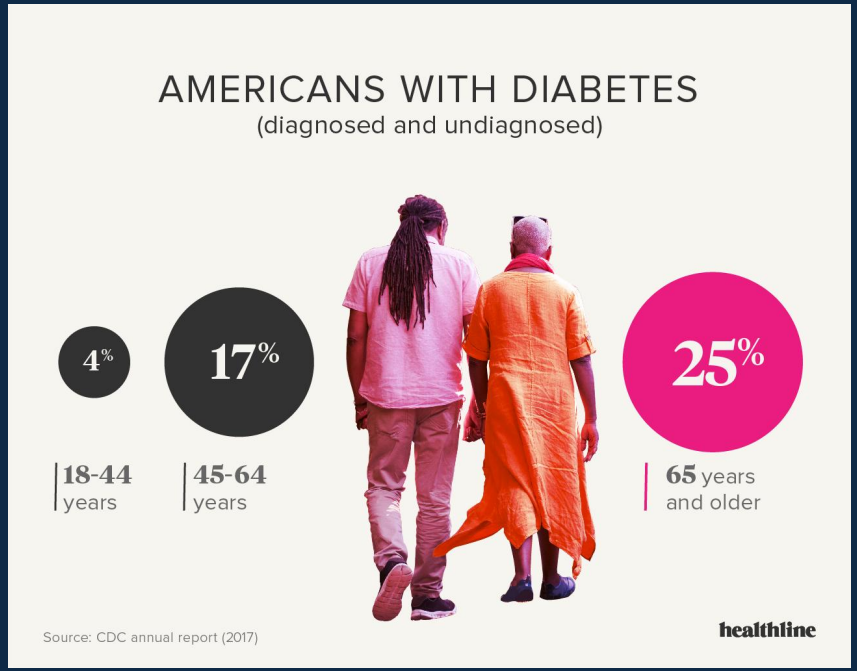

General Assembly
Data Science
Final Presentation
Diabetic Prediction

PROBLEM STATEMENT

1. In general ⅓ of diabetes cases (type 2) are undiagnosed.
2. Diabetes can lead to several other complications throughout the body (e.g., kidney and heart diseases), and compromise immune system
3. Doctor's often make assumptions based off age, weight, and blood sugar levels, leaving room for misdiagnosis.



USING MACHINE LEARNING

Using a Machine Learning Model can Help us:

- Diagnosis patients who are not showing symptoms or not considered high risk
- Provide cheaper alternatives than performing in depth tests
    - Utilizing data about your patient that is readily available
- Saving us time - type 2 can take a long time to diagnose

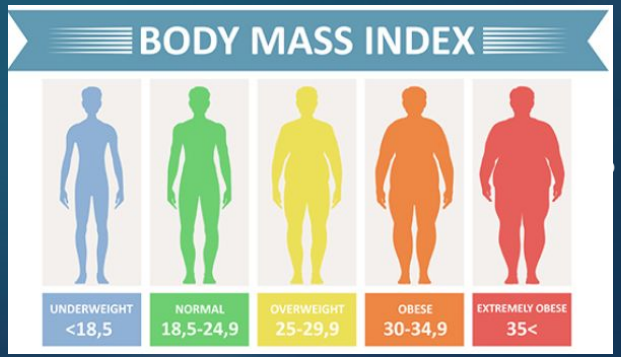
Overall Approach:

- Prepare and Understand the Data
- Set Metrics for success
- Finalize and Improve Models
- Determine next steps based on Business goals and model performance

LOOKING AT OUR DATA

FEATURES

- Glucose
- BMI
- Insulin
- Diabetes Pedigree Function
- Age
- Pregnancies
- Skin Thickness
- Blood Pressure



TARGET

- Diabetic
- Not Diabetic

DATA OVERVIEW

- Several of our columns had “zero” values including Glucose, BMI, Insulin, and Skin Thickness
- Wanted to avoid dropping zero values. I used the median value to replace the zero values.
- This helped us see a more normal distribution for our variables
- Overall we saw highest correlation with Glucose, Pregnancies, and Age

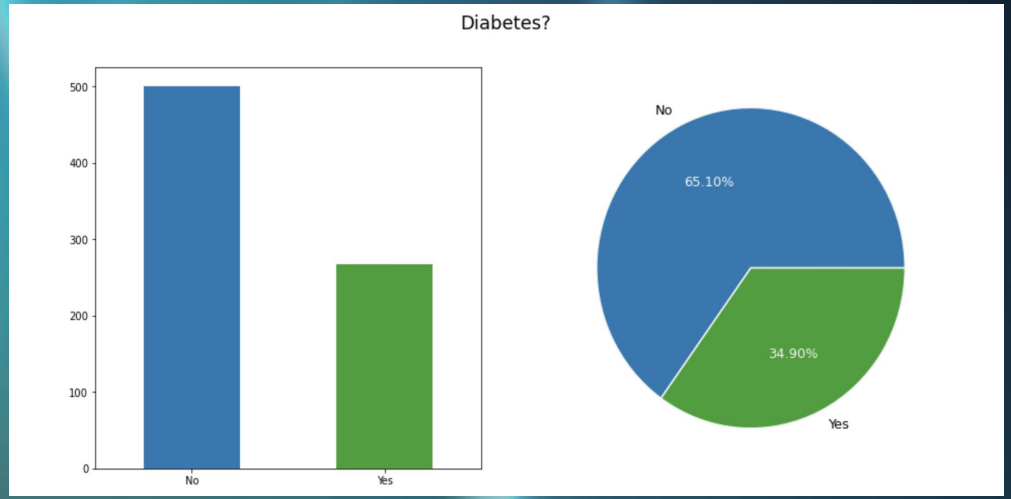
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	121.656250	72.386719	27.334635	94.652344	32.450911	0.471876	33.240885	0.348958
std	3.369578	30.438286	12.096642	9.229014	105.547598	6.875366	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.750000	64.000000	23.000000	30.500000	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	31.250000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

DATA OVERVIEW CONT'D

Total Data: 768
-Diabetic: 268
-Not Diabetic:500

Training Set Data:576
-Diabetic:374
-Not Diabetic:202

Test Set Data: 192
-Diabetic:126
-Not Diabetic:66



IDEAL SITUATION

- Our biggest priority is to avoid informing patients who have diabetes that they don't have diabetes
  - (False Negatives)
- According to the National Academy of Science, incorrect diagnosis is expensive
  - False Negatives are more expensive than False Positives.
- We ideally want to create a model that has high overall accuracy but is also cost efficient for the patient and the physician

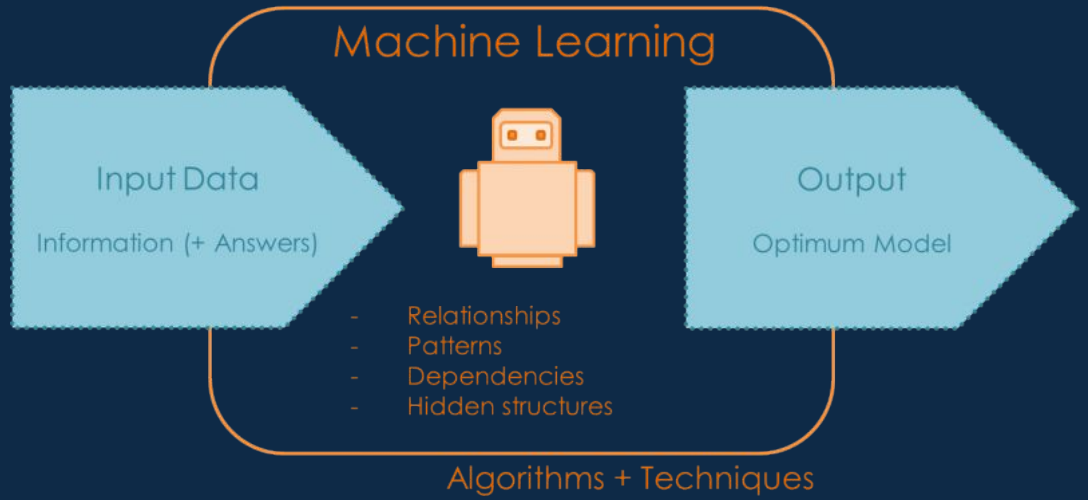
MODELS USED

MODELS USED:

- LOGISTIC REGRESSION
- DECISION TREE
- RANDOM FOREST

METRICS OF SUCCESS:

- Fbeta 2 (PRIMARY)
- PREDICTION ACCURACY (SECONDARY)



RESULTS

192
Test size

576
Training Size

	ACCURACY SCORE	FBETA2	FALSE NEGATIVES
LOGISTIC REGRESSION	61%	78%	3
DECISION TREE	81%	74%	15
RANDOM FOREST	80%	69%	21

ANALYSIS

NEXT STEPS:

- Increase the overall score for our model by tuning our model and using alternate ML model types
- Working with a larger dataset that includes more features like ethnicity, gender, etc.
- Tune our training and test data set to work with a more balanced class

BUSINESS STRATEGY:

- Determine which groups of people are the least likely to develop diabetes and utilize the top performing model on that group to avoid misdiagnosing them.
- Our Model can also be used to determine if a patient is likely to develop diabetes. Developing a preventative model will help us be even more cost efficient.