SAVE MARSHMALLOW

START

INSULIN PANDAS TO THE RESCUE!

ONLY YOU CAN MONITOR YOUR GLUCOSE LEVEL

PREDICT DIABETES

YOU COULD BE NEXT...

Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year. According to the CDC survey in 2018, 34 million Americans have diabetes and 88 million Americans are prediabetic.

Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, and can lead to reduced quality of life and life expectancy.

MARSHMALLOW'S PROFILE:

BMI: 73

General Health: "Excellent"

Income: > $ 75,000

Age: 12 panda years 48 human years

Physical Health: Active once a week

Education: GED with some college

Mental Health: 29 days of poor mental health

DATASET EXPLAINED

AGE (1 - 13)	GENDER	EDUCATION? (1-6)	INCOME (1-8)	FRUITS?	VEGGIES?	HEALTHCARE?	EXPENSIVE DOCTORS?
1 = 18 to 24yr	0 = Female	1 = No Education	1 = < $10,000	0 = No	0 = No	0 = No	0 = No
13 = > 80yr	1 = Male	6 = +4yr College	8 = > $75,000	1 = Yes	1 = Yes	1 = Yes	1 = Yes

GENERAL HEALTH	MENTAL HEALTH	PHYSICAL ACTIVITY?	PHYSICAL HEALTH	DIFFICULTY WALKING?	ALCOHOL CONSUMPTION	SMOKER?	STROKE?
Excellent to Poor	Days of Poor Mental Health	0 = No	Active in last 30 days?	0 = No	0 = Little to None	0 = No	0 = No
1 to 5	1 to 30	1 = Yes	1 to 30	1 = Yes	1 = Yes, Heavy	1 = Yes	1 = Yes

BMI	HIGH BP	HEART DISEASE	HIGH CHOLESTROL	CHECKED CHOLESTROL?
INTS	0 = No	0 = No	0 = No	0 = No
	1 = Yes	1 = Yes	1 = Yes	1 = Yes

DIABETES OUTCOME
0 = No diabetes
1 = Pre-diabetic
2 = diabetic

EXTRACT, TRANSFORM, LOAD

Source:
The Behavioral Risk Factor Surveillance System Survey collected by the CDC annually.

For this project, the data used to build the machine learning models is from a 2015 survey containing responses from over 440k individuals.

Clean, clean, and CLEAN!

- NULLS
- Correct Data Types
- Duplicates

Import Database:

- Tools:
  - SQL Postgres
  - Amazon Web Services (AWS)
  - Google Colab / PySpark

SQL QUERIES

- Males and Females with good BMI index (18-25 kg/m2) and count of diabetes outcomes.

	diabetes_012 numeric (3,1)	sex numeric (3,1)	Total Count bigint	Percentage double precision
1	0.0	0.0	52407	20.658703878902553
2	0.0	1.0	29094	11.468779564806054
3	1.0	0.0	585	0.23060548722800378
4	1.0	1.0	354	0.13954588457899716
5	2.0	0.0	2853	1.1246452223273415
6	2.0	1.0	2609	1.0284610532954903

- Males and Females with bad BMI index (<18 and >25 kg/m2) and count of diabetes outcomes.

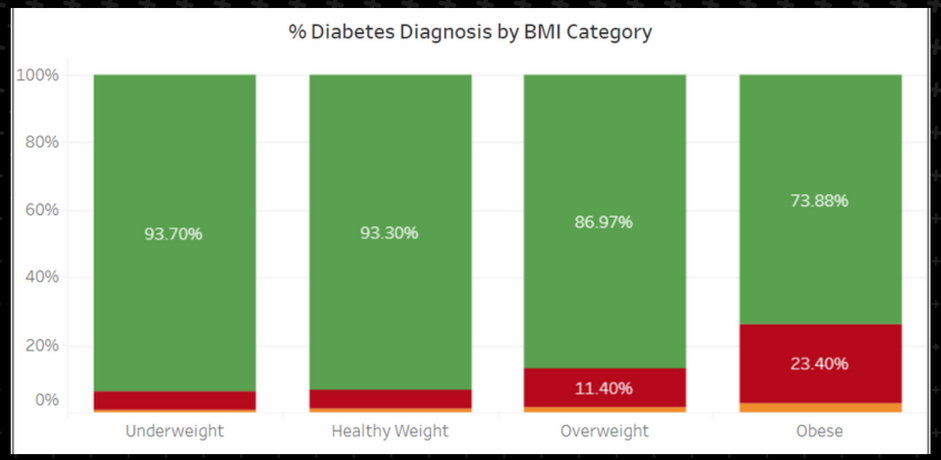
	diabetes_012 numeric (3,1)	sex numeric (3,1)	Total Count bigint	Percentage double precision
1	0.0	0.0	68552	27.023021128981394
2	0.0	1.0	63650	25.09066540523494
3	1.0	0.0	2019	0.7958845789971618
4	1.0	1.0	1673	0.6594922737306843
5	2.0	0.0	15558	6.1329233680227055
6	2.0	1.0	14326	5.64727215389467

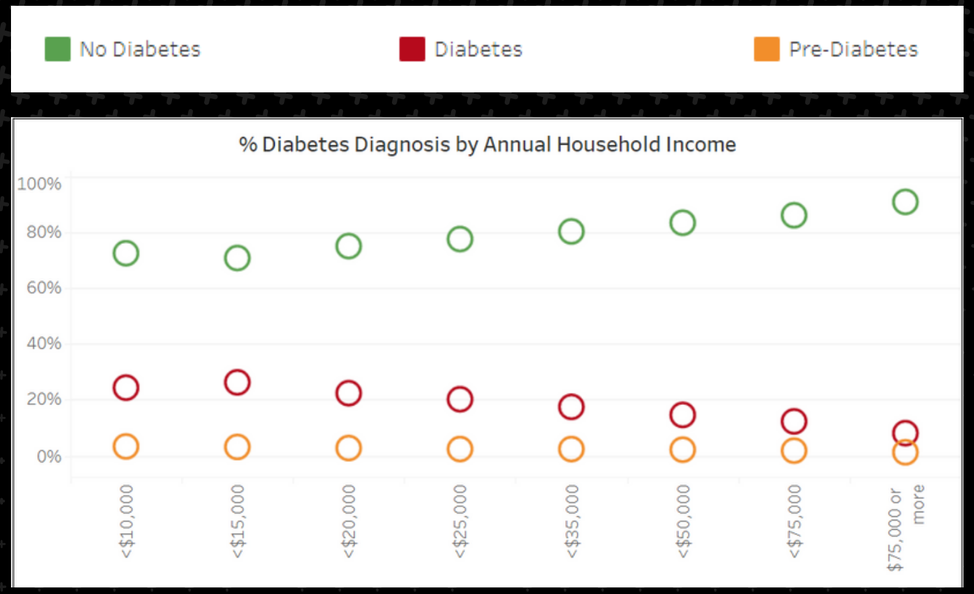
DIABETES OUTCOME
0 = No diabetes
1 = Pre-diabetic
2 = diabetic

PYSPARK & COLAB

diabetes_012	highbp	highchol	cholcheck	bmi	smoker	stroke	heartdiseaseorattack	physactivity	fruits	veggies	hvyalcoholconsump	anyhealthcare	nodocbccost	genhlth	menthlth	physhlth	diffwalk	sex	age	education	income
0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0	1.0
0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0	8.0
0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	0.0	11.0	3.0	6.0
0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	3.0	0.0	0.0	0.0	11.0	5.0	4.0
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	2.0	0.0	1.0	10.0	6.0	8.0
0.0	1.0	0.0	1.0	30.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	3.0	0.0	14.0	0.0	0.0	9.0	6.0	7.0
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	3.0	0.0	0.0	1.0	0.0	11.0	4.0	4.0
2.0	1.0	1.0	1.0	30.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	5.0	30.0	30.0	1.0	0.0	9.0	5.0	1.0
0.0	0.0	0.0	1.0	24.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	1.0	8.0	4.0	3.0

1 TOP FACTORS RELATED TO DIABETES

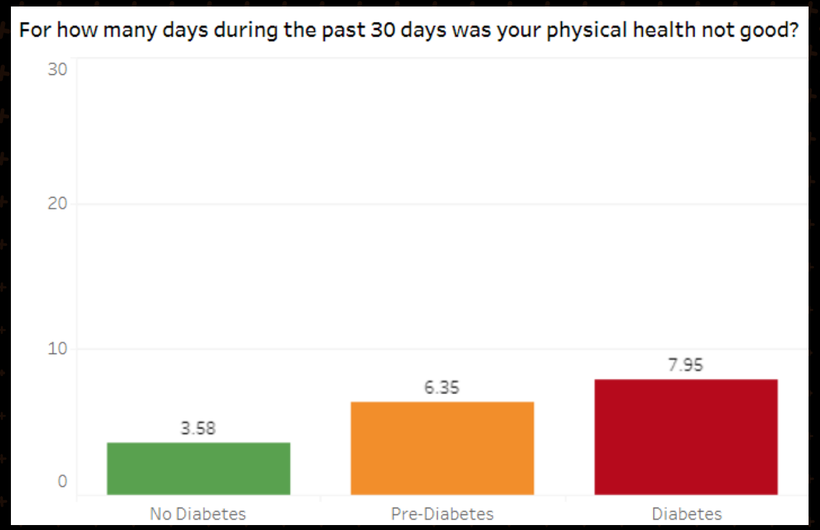


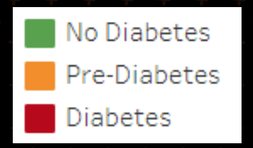


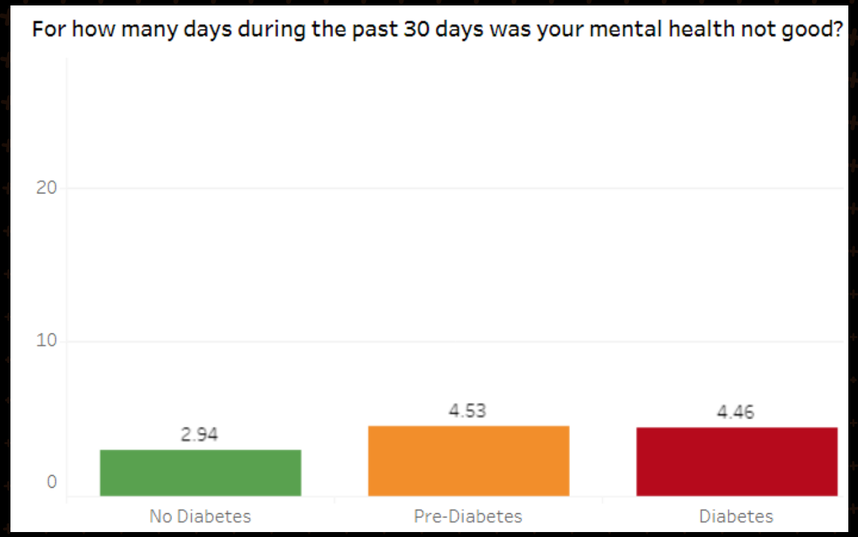
% Diabetes Diagnosis by Age - Heat Map

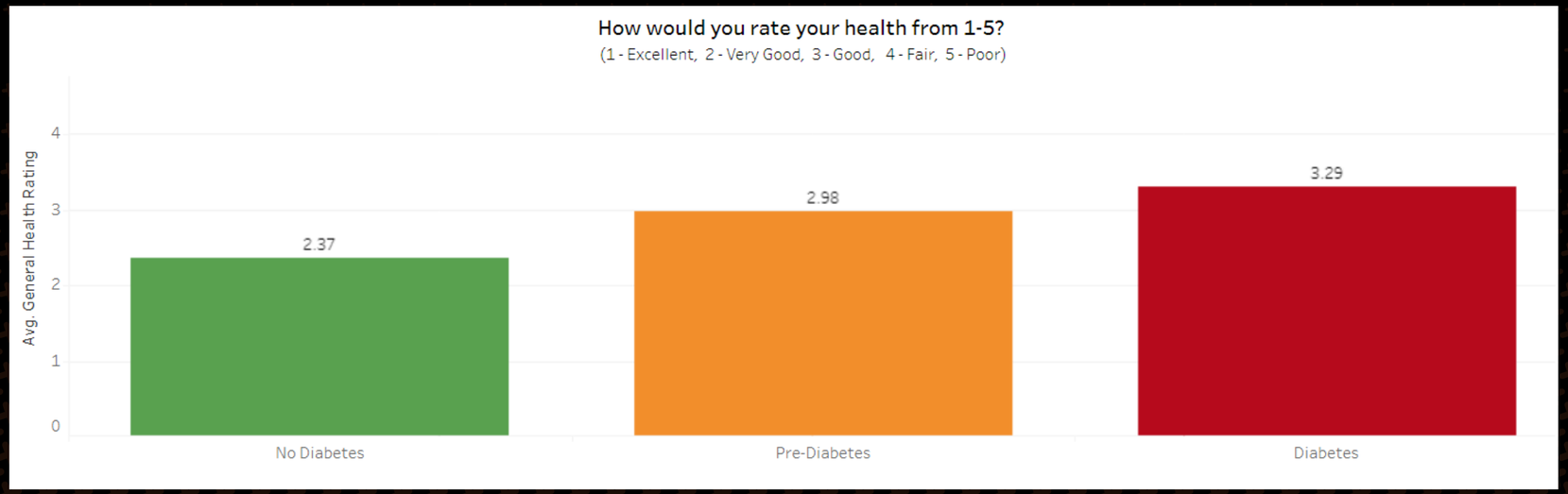
Age Group	No Diabetes	Pre-Diabetes	Diabetes
18-29	98%	1%	2%
30-39	95%	1%	4%
40-49	91%	1%	8%
50-59	85%	2%	13%
60-69	79%	2%	19%
70-79	76%	3%	22%
80+	79%	3%	18%

2 HEALTH RATINGS BY DIABETES DIAGNOSIS

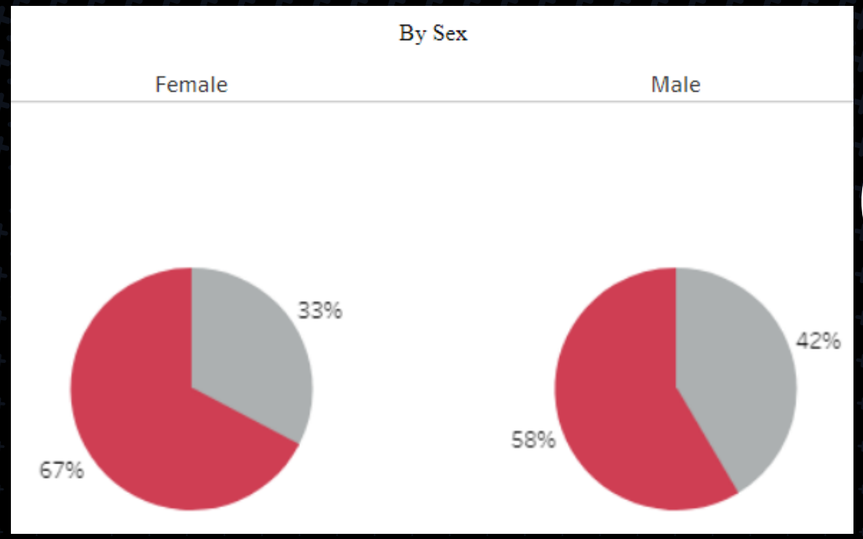


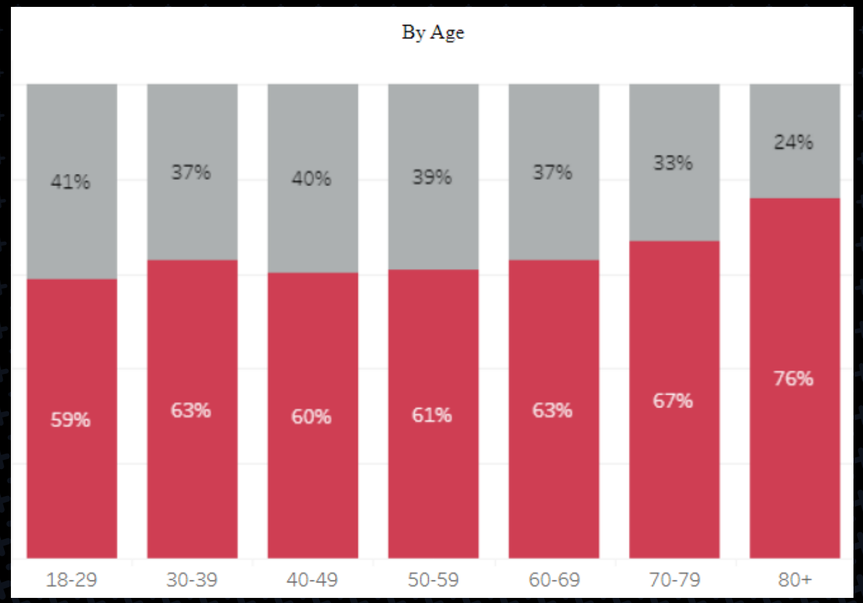


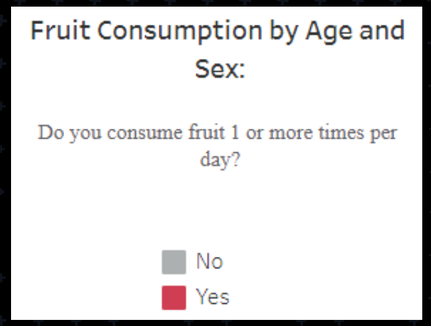


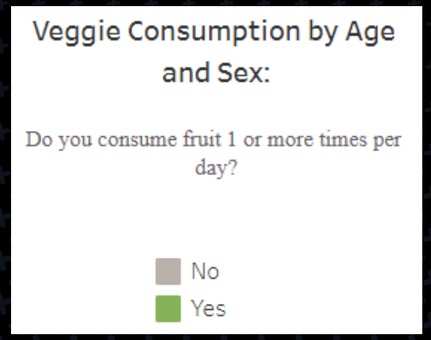


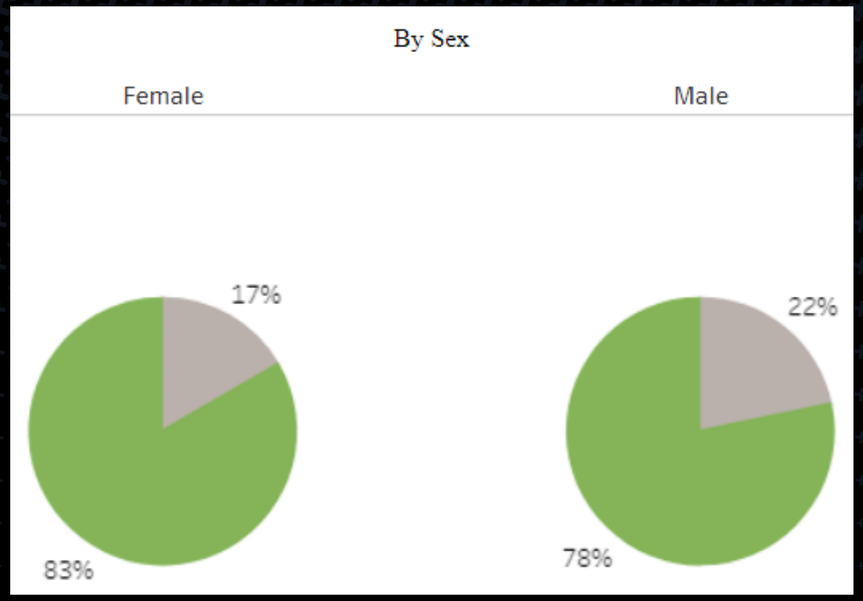
3 FRUITS AND VEGGIES FACTS

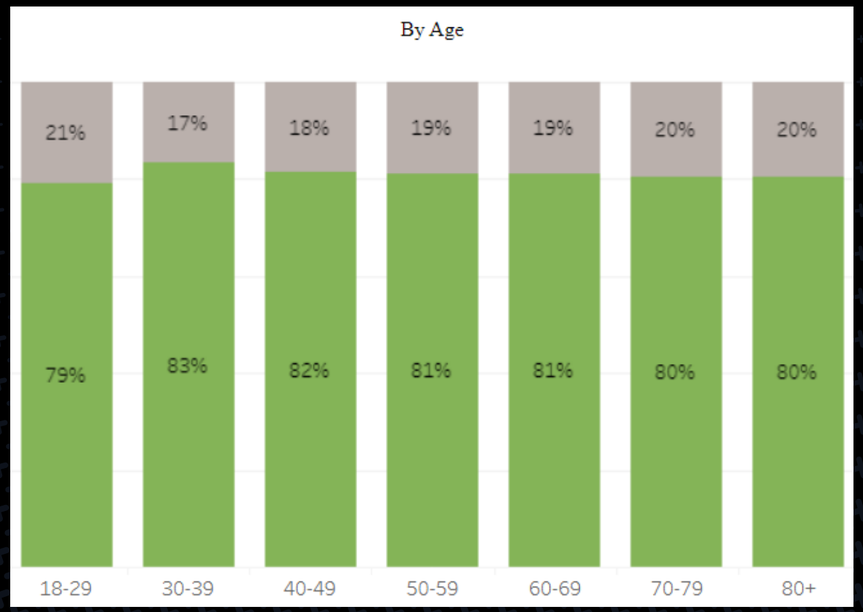












SUPERVISED LEARNING

Random Forest Classifier

Training Score: 99%
Testing Score: 84%

Logistic Regression Model

0 = No diabetes
2 = Diabetic

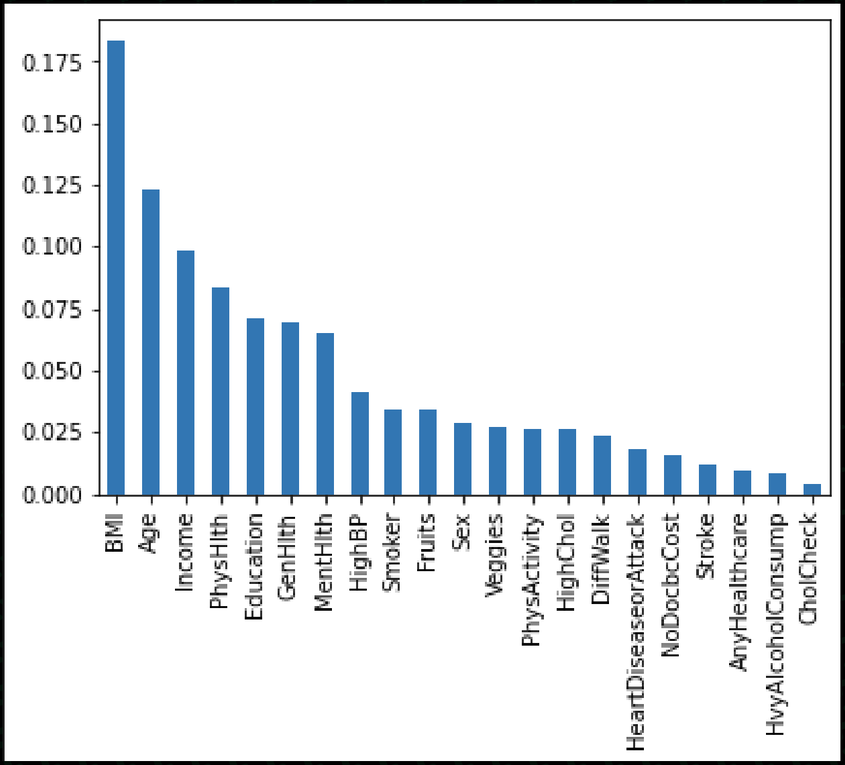
Training Score: 86%
Testing Score: 86%

Confusion Matrix Classification Report

	precision	recall	f1-score	support
0.0	0.87	0.98	0.92	53379
2.0	0.52	0.16	0.24	8884
accuracy			0.86	62263
macro avg	0.70	0.57	0.58	62263
weighted avg	0.82	0.86	0.82	62263

- **Precision:** What percent of your predictions were correct?
- **Recall:** What percent of the positive cases did you catch?
- **F1 Score:** What percent of positive predictions were correct?
  - The model is much more accurate at predicting non-diabetic cases than diabetic cases

Feature Selection



UNSUPERVISED LEARNING

STANDARD SCALER

To uniform the values in dataset.

PREDICT AND APPLY PCA

95% of data is tested then reduced to 19 principal components.

APPLY T-SNE TO PCA DATASET

Applying T-SNE to dataset will form clusters for plotting.

FIND K-MEANS

Using the Elbow Method to find the optimal value of k in K-Means.

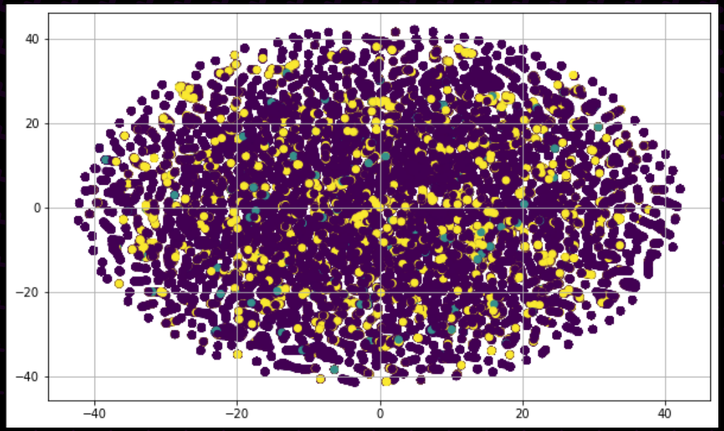
PLOT CLUSTERS

Create 2D and 3D scatter plots to visualize defined clusters.

APPLY RANDOM FOREST CLASSIFIER

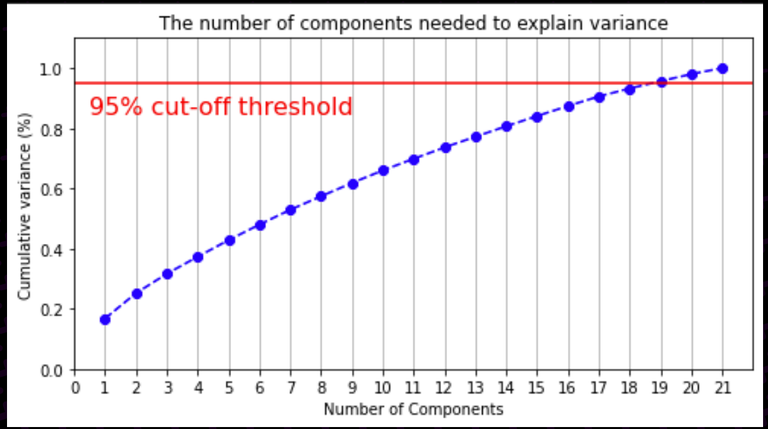
Testing Score: 100%
Training Score: 97%

UNSUPERVISED LEARNING PLOTS 1

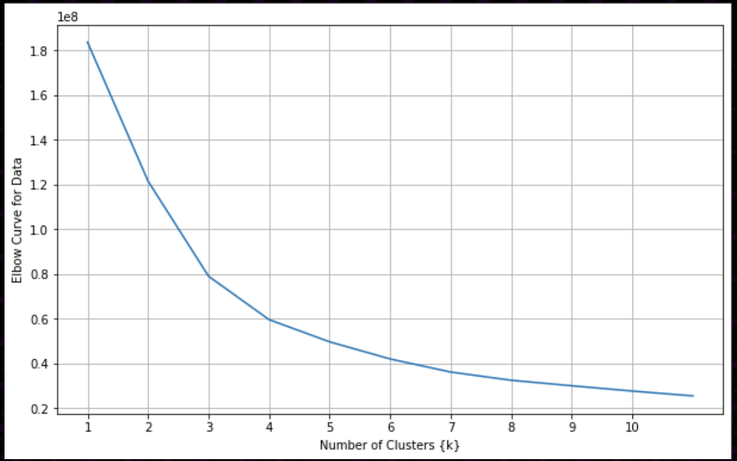


T-SNE TO PCA

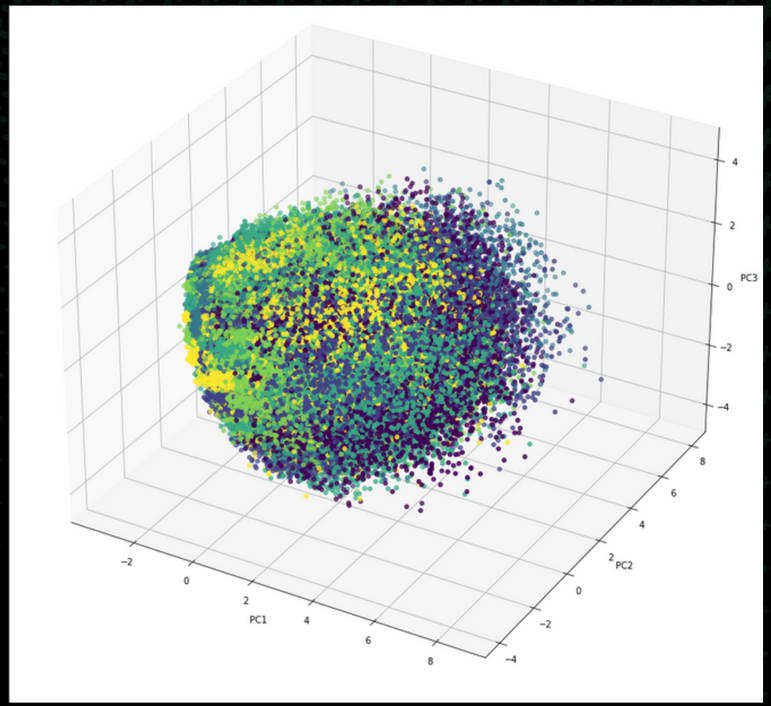
PREDICT AND APPLY PCA



FIND K-MEANS

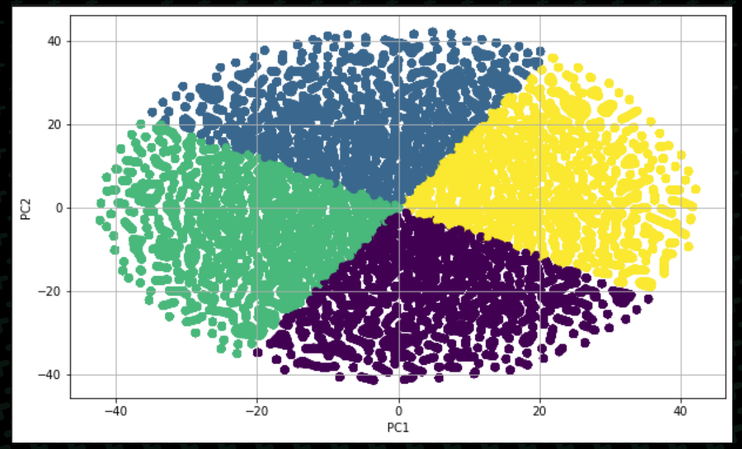


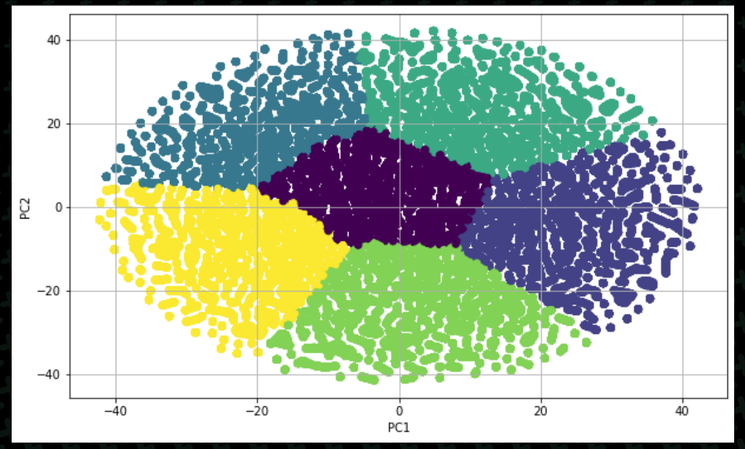
UNSUPERVISED LEARNING PLOTS 2



3D CLUSTERS

2D CLUSTERS





TENSORFLOW

Trial 2

- 3 hidden layers
- RELU for activations, Sigmoid as output
- Utilize only columns considered important
- Increase batch sizes
- No binning technique
- No dropout layers
- Increase learning rate

SCORE: 70%

Trial 24

- 4 hidden layers
- RELU and Sigmoid
- All columns
- Batch size = 32
- Decrease nodes
- Huber loss
- Two dropout layers
- Early Stopping and monitor val_loss
- Learning rate=0.00001

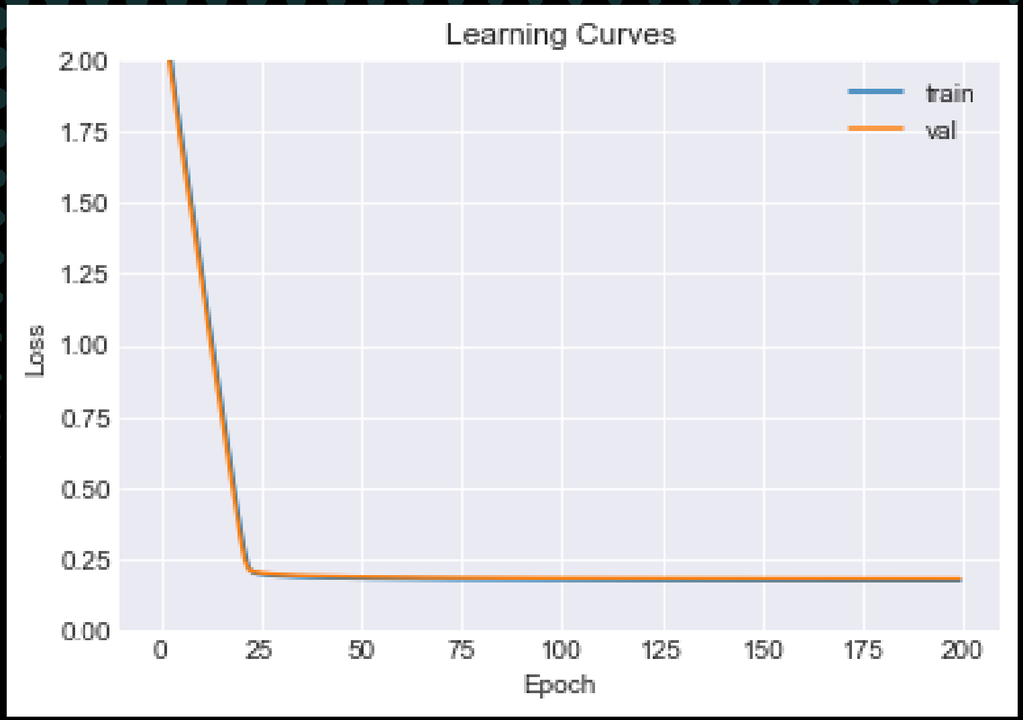
SCORE: 77%

Trial 11

- 5 hidden layers
- Leaky RELU for activations, Sigmoid as output
- Utilize all columns in dataset
- Decrease batch sizes
- Increased nodes
- Binning technique
- Add dropout layers
- Decrease learning rate

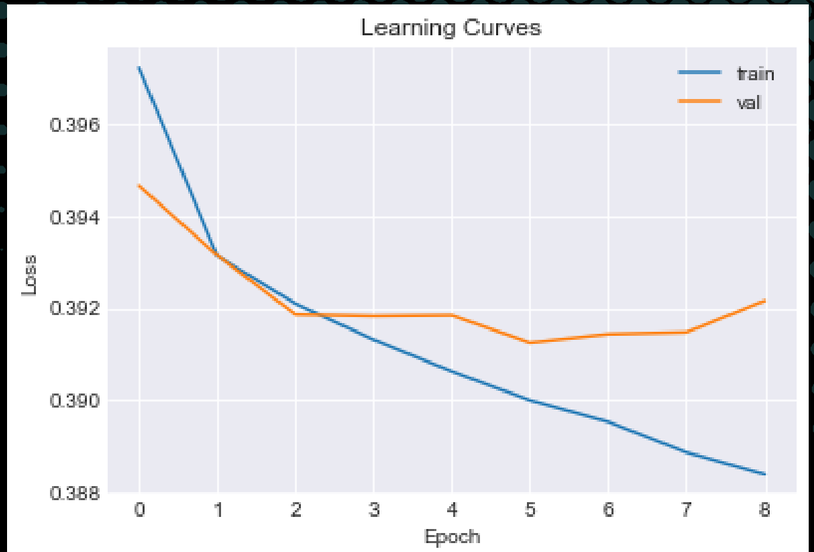
SCORE: 74%

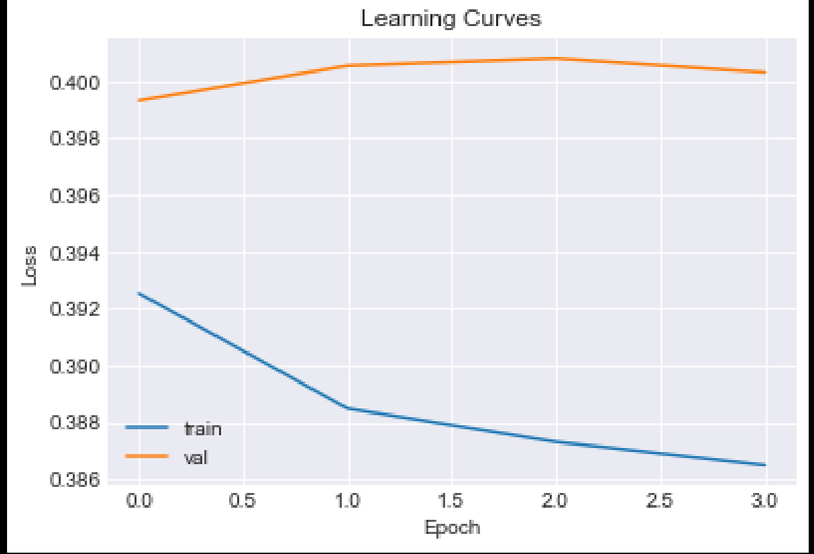
EXCEPTIONAL



VS.

UNEXCEPTIONAL





MARSHMALLOW'S PROFILE:

BMI: 73

General Health: "Excellent"

Income: > $ 75,000

Age: 12 panda years 48 human years

Physical Health: Active once a week

Education: GED with some college

Mental Health: 29 days of poor mental health

DIAGNOSTIC

```
# Do a prediction
marshmallow_data = [[73, 6, 8, 6, 5, 1, 29]]
predicted_class = clf.predict(marshmallow_data)
print(predicted_class)
```

MARSHMALLOW'S PROFILE:

BMI: 73

General Health: "Excellent"

Income: > $ 75,000

Age: 12 panda years 48 human years

Physical Health: Active once a week

Education: GED with some college

Mental Health: 29 days of poor mental health

DIAGNOSTIC

```
# Do a prediction
marshmallow_data = [[73, 6, 8, 6, 5, 1, 29]]
predicted_class = clf.predict(marshmallow_data)
print(predicted_class)
```

NO DIABETES!

THANK YOU FOR LISTENING!

CLICK ME FOR MORE FUN!