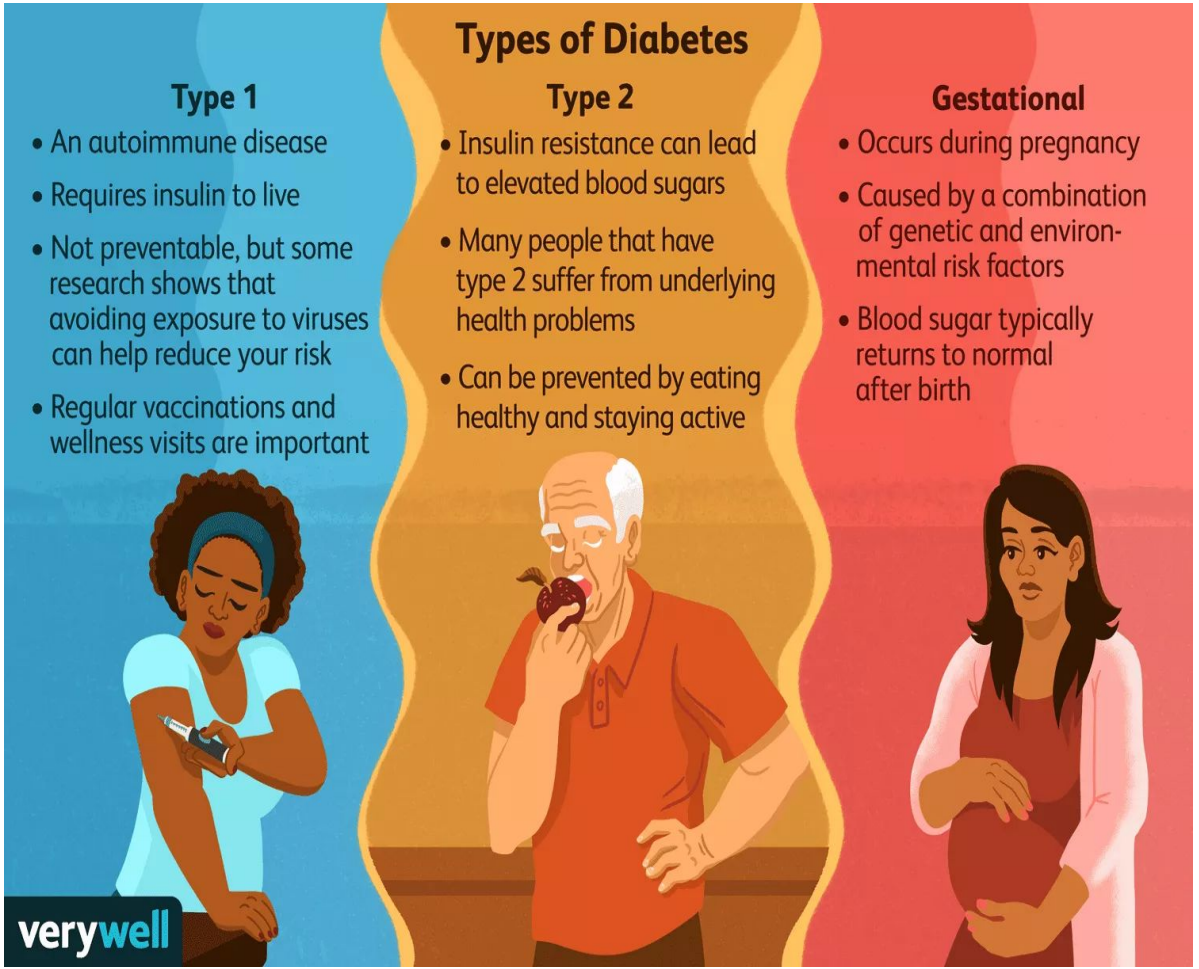
Diabetes Health Indicators

Group 3:
Anas Al-Rasbi
Tyler Avret
Thilak Raj Murugan
Tyler Wilson
ZK Zhao

What is Diabetes?

Diabetes is a chronic disease that occurs when the body doesn't produce enough insulin or can't use it properly, resulting in high blood sugar levels



Data

- Behavioral Risk Factor Surveillance System (BRFSS) survey responses conducted by the CDC (telephone-based survey)
- A clean dataset of over 250,000 survey responses from BRFSS 2015
- 21 Features

```
Index(['Diabetes_binary', 'HighBP', 'HighChol', 'CholCheck', 'BMI', 'Smoker',
       'Stroke', 'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'Veggies',
       'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost', 'GenHlth',
       'MentHlth', 'PhysHlth', 'DiffWalk', 'Sex', 'Age', 'Education',
       'Income'],
      dtype='object')
```

Task

- Find the feature variables that best predict diabetes
  - Are certain standalone features that are strong indicators of diabetes?

Challenges

- 250,000 is a lot of data
  - SVM could not handle that much
  - C = 1 takes ~18 secs, C = 50 takes ~ 680 secs, C = 100 takes ~ 850 secs

- Random Forest
  - Quick for KNN, Very slow for Decision Tree

- Unbalanced vs Balanced Dataset
  - Unbalanced accuracy capped out at ~86%
  - Balanced accuracy ~75%

Logistic Regression

- 80 - 20 (train-test) split
- GridSearchCV to find best hyperparameters

Note: The best hyperparameters were identified based on accuracy.

```
param_grid = [
    {'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000],
     'penalty': ['l2'],  # l2 for lbfgs
     'solver': ['lbfgs', 'newton-cg', 'sag']},

    {'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000],
     'penalty': ['l1'],  # l1 for liblinear and saga
     'solver': ['liblinear', 'saga']},
]
```

Results

Imbalanced Data-Set (87-13)

Best Parameters: { C: 0.001, Penalty: l2, Solver: lbfgs}

Accuracy: Train: **86.4%** Test: **86.3%**

F1 Score:

- **Class 0 (Non-Diabetic):** High F1-score (**0.92**)
- **Class 1 (Diabetic):** Low F1-score (**0.22**)

There is significant bias in the model. It performs well for the majority class (non-diabetic) but fails to generalize to the minority class (diabetic).

Balanced Data-Set (50-50)

Best Parameters: { C: 0.1, Penalty: l1, Solver: liblinear}

Accuracy: Train: **74.8%** Test: **74.5%**

F1 Score:

- **Class 0 (Non-Diabetic):** F1-score (**0.74**)
- **Class 1 (Diabetic):** F1-score (**0.75**)

While accuracy is slightly lower than with the imbalanced dataset, this approach ensures good performance for both classes, reducing bias and enhancing real-world applicability.

KNN Model

- Feature Selection
  - Trained a random forest with 200 decision trees to select input features.
    - **SelectFromModel:** A meta transformer which does feature selection.
    - Decreased time to train and fine-tune hyper-parameters.
- A kNN model was then trained on the transformed features.
- Tuning Hyper-parameters
  - No. of neighbors (Unbalanced/Balanced data): **70 / 100**
  - **Uniform** vs weighted distance between neighbors
    - Weighted distance performs fits training data well but has lower validation and test accuracy compared to uniform distance.

KNN Results

- Testing accuracy of 86% and 73% for unbalanced and balanced data.
- F1 score
  - 0.93 / 0.1 in predicting a person as non-diabetic / diabetic
  - 0.74 / 0.71 in predicting a person as diabetic / non-diabetic
- Selected features:
  - High Blood pressure, BMI,
  - General Health, Mental Health, Physical Health,
  - Age, Education, Income

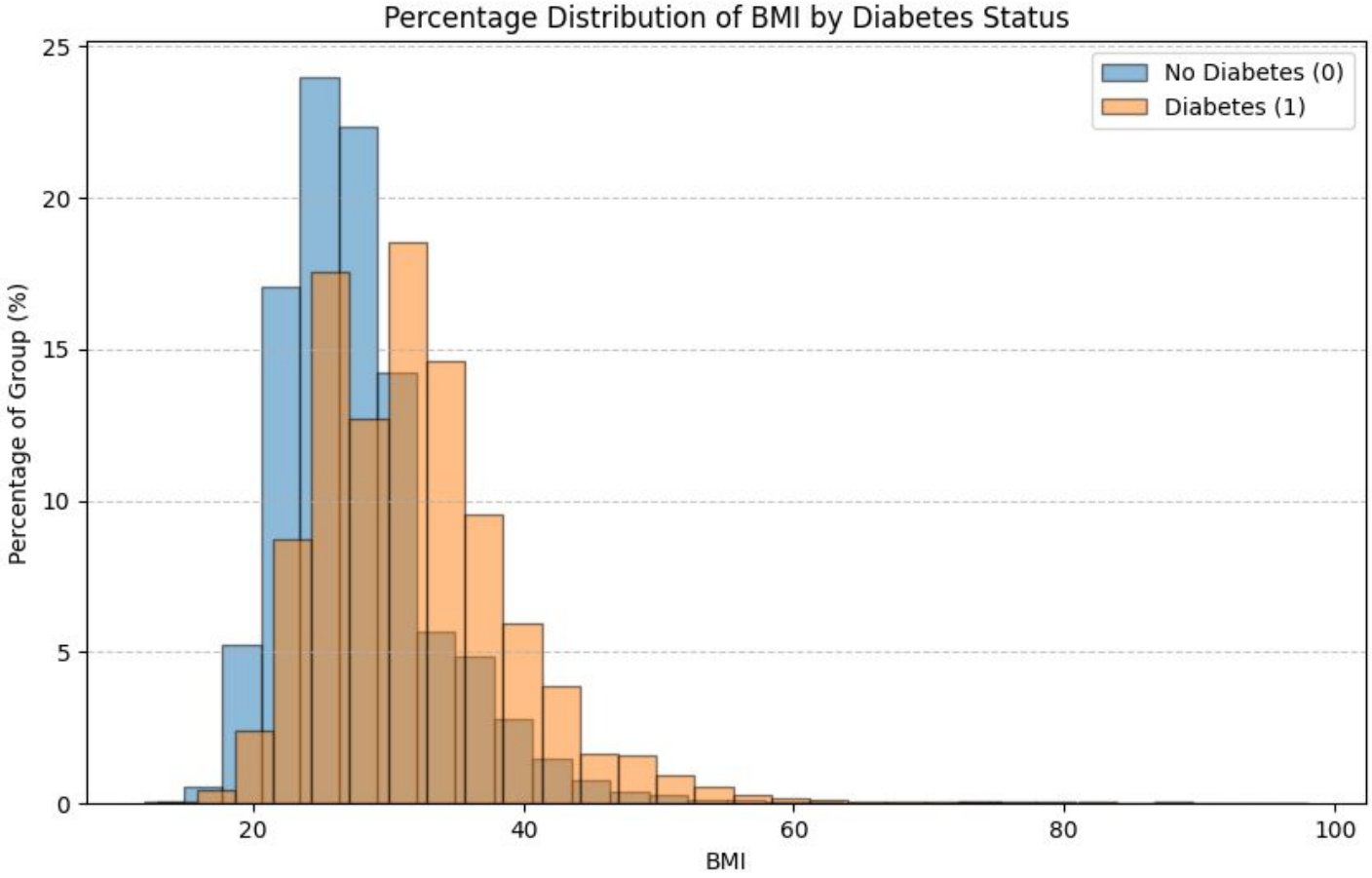
Conclusions

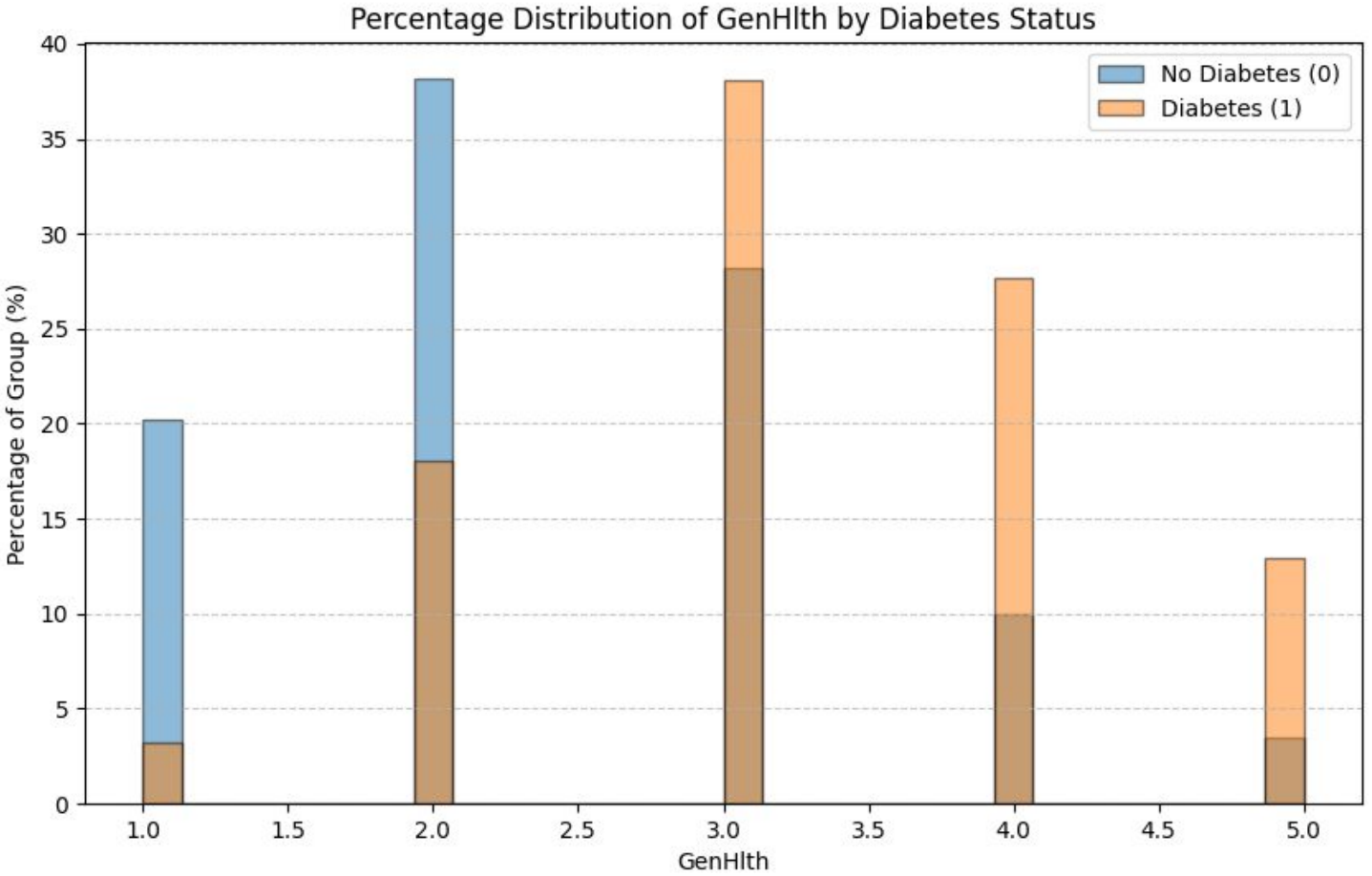
74% accuracy is acceptable, but not as high as we expected

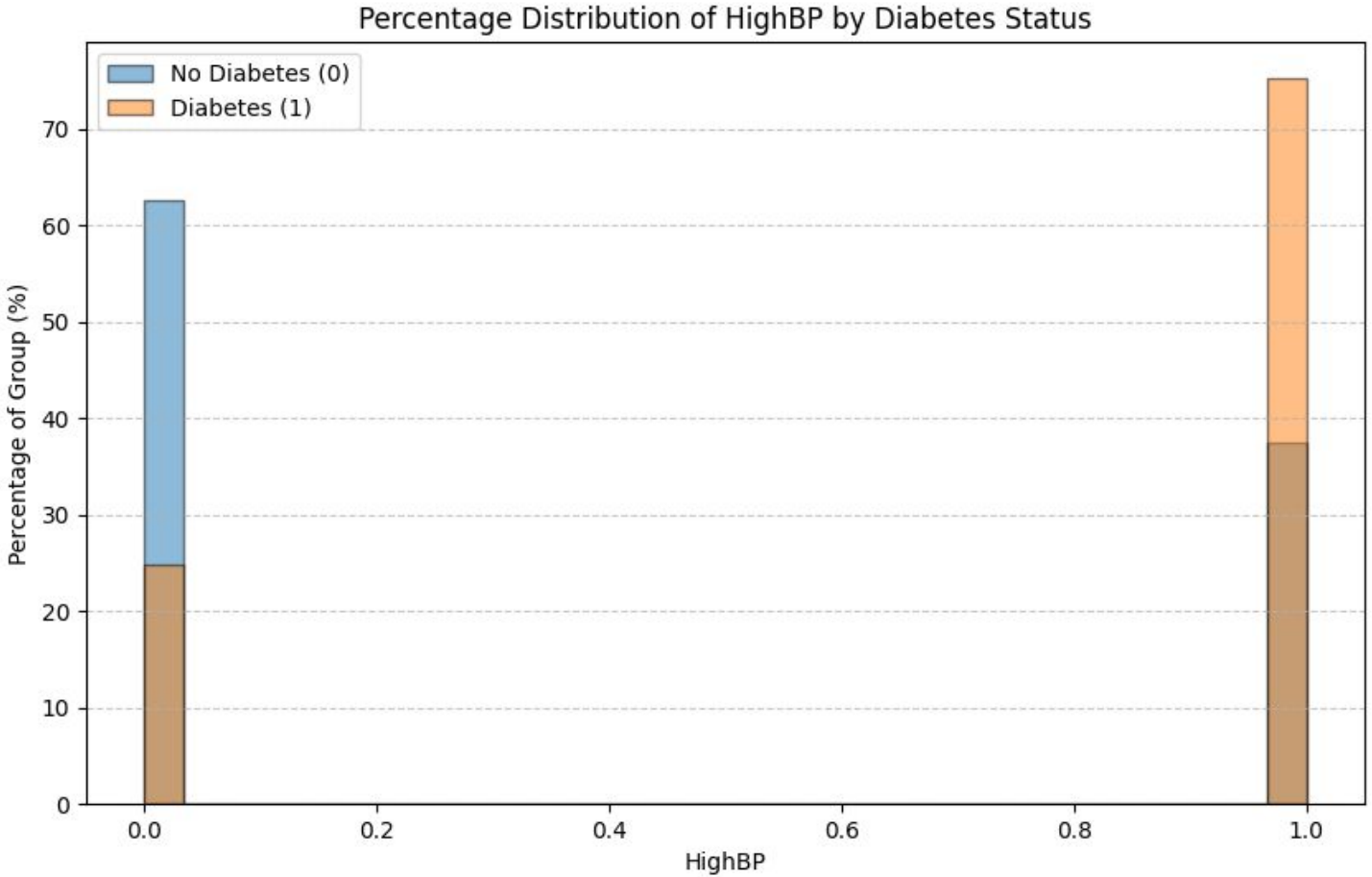
Predicting diabetes was more challenging than anticipated, why?

- Class imbalance (87%/13%) in the initial dataset may have led to overtraining
- Dataset contents may include Type 1 (hereditary) Diabetic data points
- We did not know how long a survey participant had Diabetes before taking the survey
- There were not features that consistently indicated whether or not someone would have diabetes

For example...


Percentage Distribution of BMI by Diabetes Status
No Diabetes (0)
Diabetes (1)
Percentage of Group (%)
0
5
10
15
20
25
20
40
60
80
100
BMI


Percentage Distribution of GenHlth by Diabetes Status
No Diabetes (0)
Diabetes (1)
Percentage of Group (%)
0
5
10
15
20
25
30
35
40
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
GenHlth


Percentage Distribution of HighBP by Diabetes Status
No Diabetes (0)
Diabetes (1)
Percentage of Group (%)
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
HighBP

As a Final Interesting Note

Features that have a negative correlation with Diabetes:

- Income
- Education
- Physical Activity
- Heavy Alcohol Consumption
- Veggies
- Fruits

Questions?