HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

# Defining Corporate Health Classes

Anastasia Makarevich, Min Wu, Ion Barbu, Moritz Müller

# Meet the Team

**Ion Barbu**

**Min Wu**

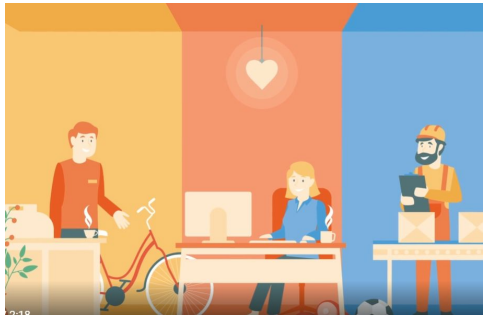
**Anastasia Makarevich**

**Moritz Müller**

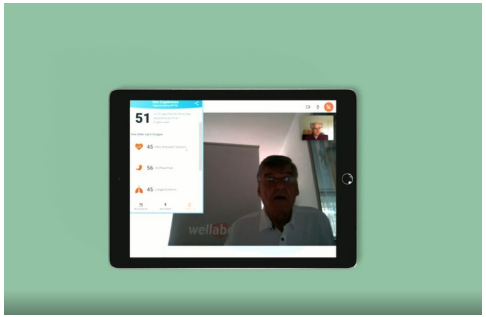
# Project Motivation

# Wellabe GmbH

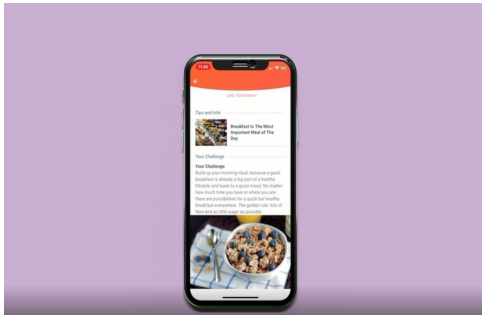
Working to promote healthy habits in the workplace. The emphasis is on prevention by identifying risk factors.



On-Site Health Check-ups




Video Consultations



Digital Prevention Programs

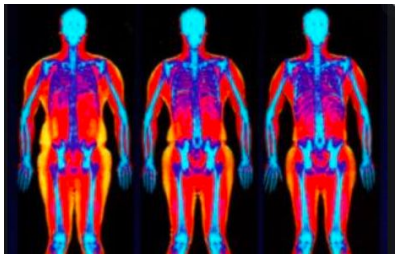
# Wellabe Patient Data

Wellabe collects patient data through examinations, laboratory results and questionnaire data related mainly to the following feature groups:

Respiratory System

Cardiovascular System

Metabolism



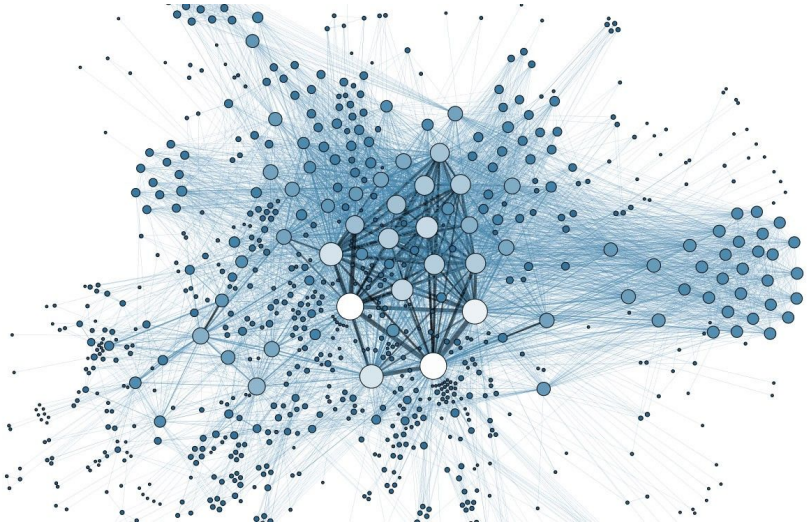
Body Composition

# Wellabe Synthetic Patient Data

Synthetic data is generated through computer simulation

It was generated in order for features to maintain the original distributions of wellabe

The synthetic dataset used has 50,000 samples versus wellabe’s original 8,000



# National Health and Nutrition Examination Survey (NHANES)

Survey and research program built on multiple studies aiming to assess the health and nutritional status of the American population.

Includes 72,000 patient samples from 1999-2012

**Laboratory**

**Dietary**

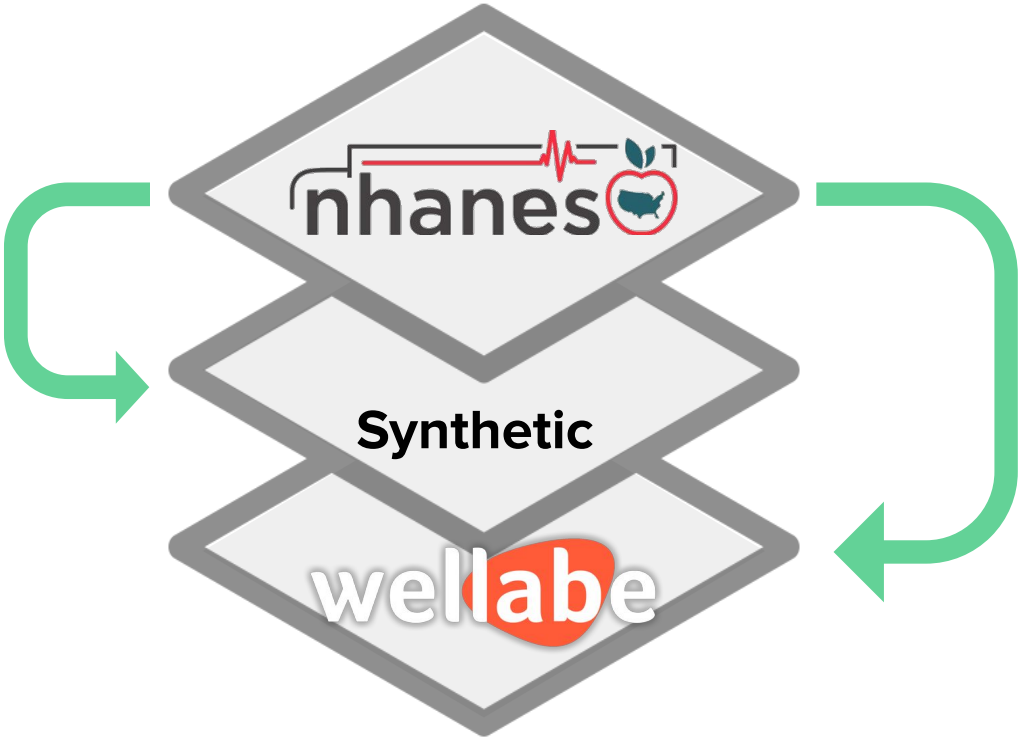
**Demographic**

**Questionnaire**

**Examination**

# Where do we come in?

**Training** on the **NHANES** dataset and **validating** on the **synthetic** wellabe dataset




**Training** on the **NHANES** dataset and **testing** on the **real** wellabe dataset

# Content

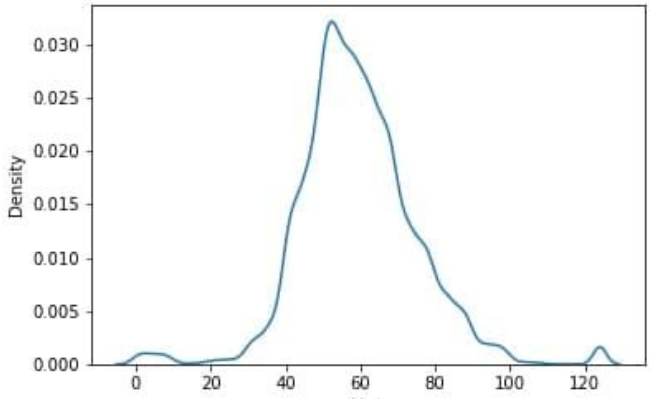
1. Understanding the model transfer problem
2. Defining metrics and procedures on our test regression model
3. Expanding this test model to classification

# When train and test data are different

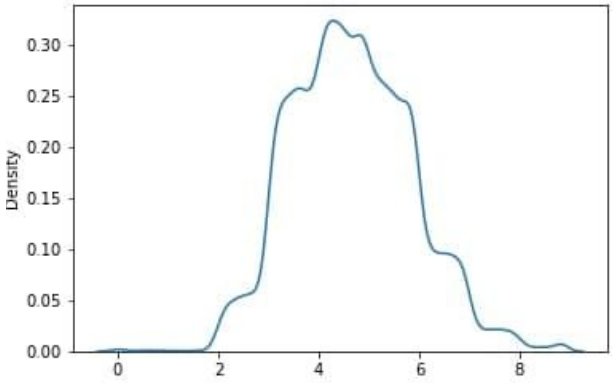
Wellabe is representative of the German corporate space and NHANES of the American population

# Synthetic Dataset Limitations

- Small clusters exist at outlier values
- Non-smooth distributions
- Violation of medical inequalities




High-Density Lipoprotein



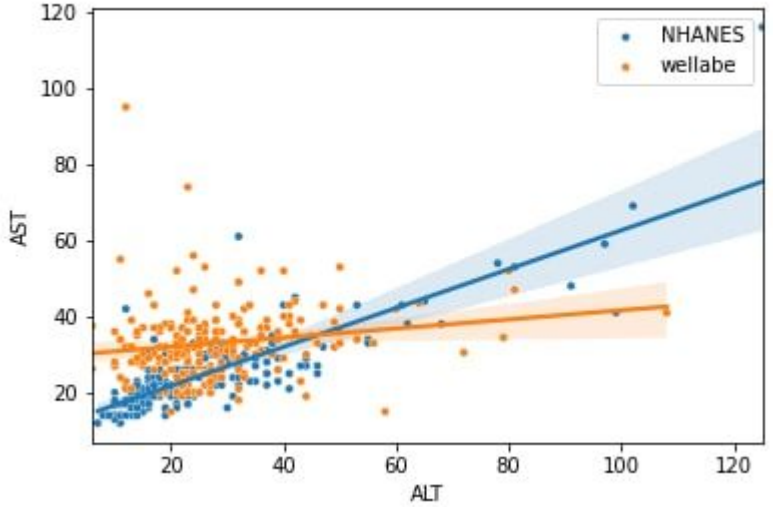

Forced Vital Capacity (Liters)

$$BMI = \frac{WEIGHT}{(HEIGHT/100)^2}$$

# Key Differences

Feature correlations do not hold in certain cases.

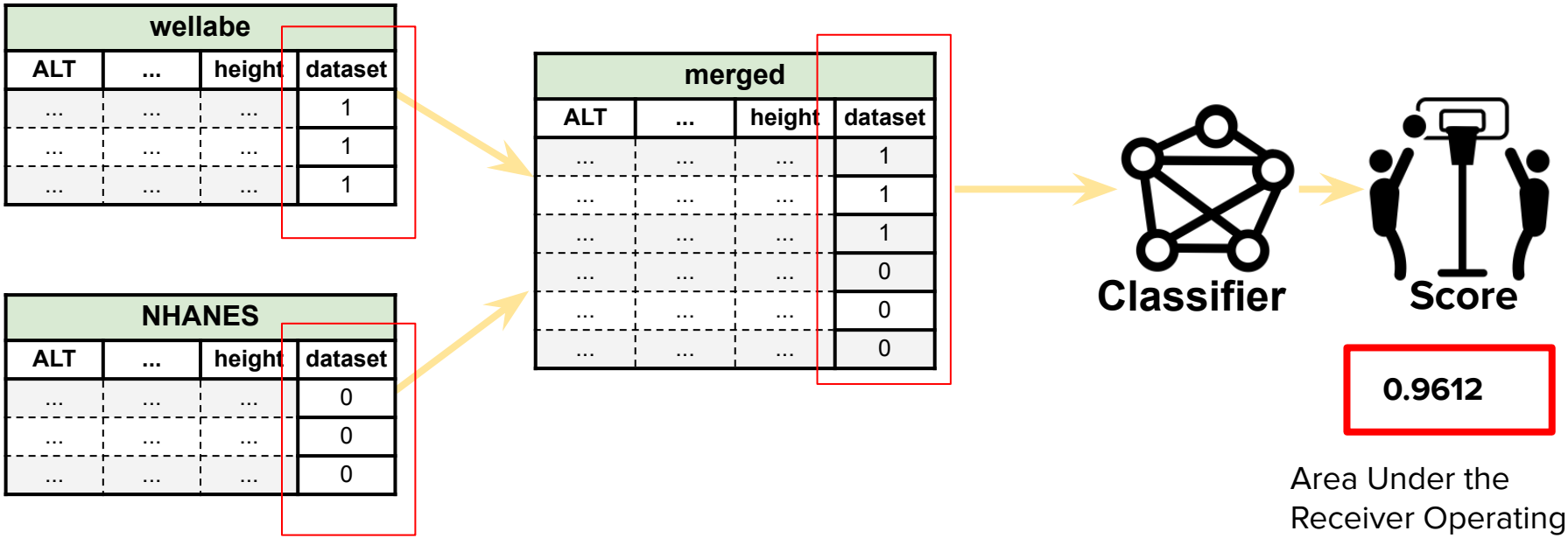
ALT and AST are highly correlated features which is shown in NHANES but not in the synthetic wellabe




Nhanes correlation: 0.77
Wellabe synthetic correlation: 0.18

# Shift happens
# Model Transfer Problem

# Can we distinguish between the datasets?

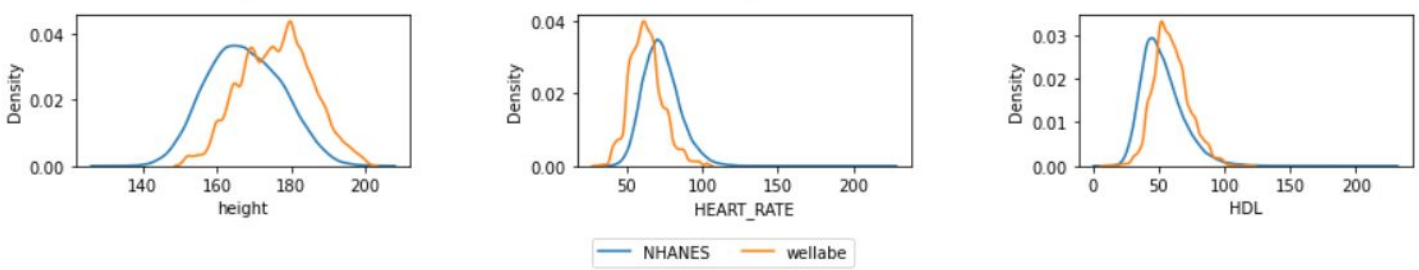



Area Under the Receiver Operating Characteristic Curve

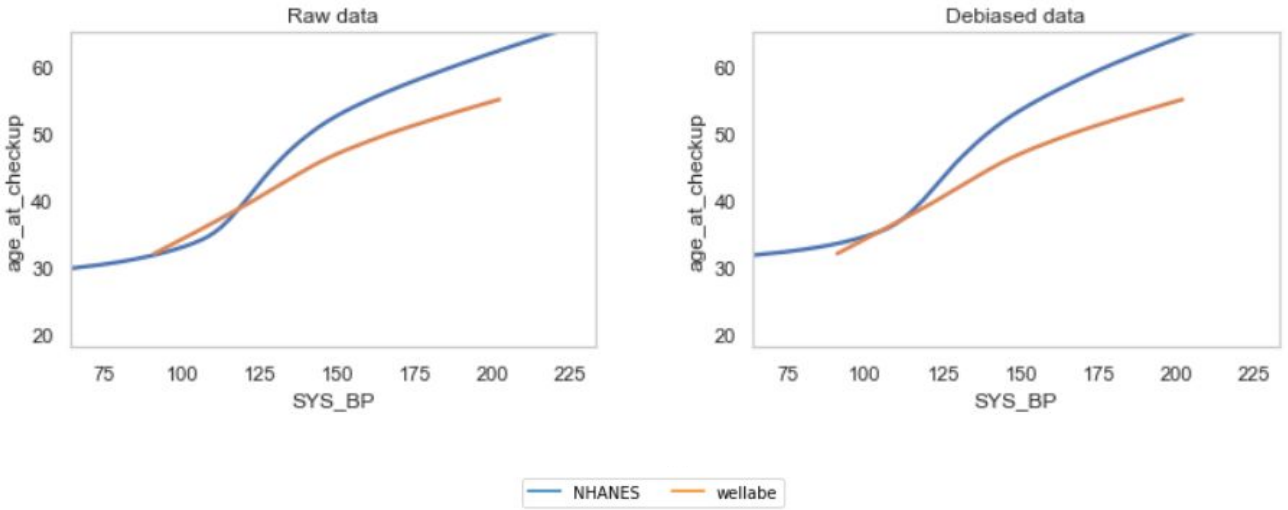
ROC/AUC

# Dataset Shift Types

- Dataset shift: joint distributions are different
- Covariate shift:
  - The covariates have different distributions in train and test
  - The relationship between covariates and target is the same in train and test
  - Not necessarily a problem



# Concept Shift

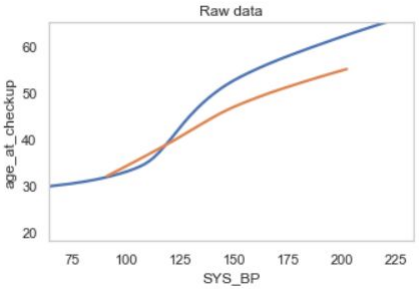



**Debiased:**

- Ethnicities proportions are corrected to be more representative of German population
- People taking medication were removed from the dataset

# Dataset Shift Types

- Dataset shift: joint distributions are different
- Covariate shift:
  - The covariates have different distributions in train and test
  - The relationship between covariates and target is the same in train and test
  - Not necessarily a problem
- Concept shift
  - When the dependency between covariates and targets is different in train and test
  - Systolic Blood Pressure (SYS_BP) and Age (age_at_checkup)



# Methods for Dataset Shift

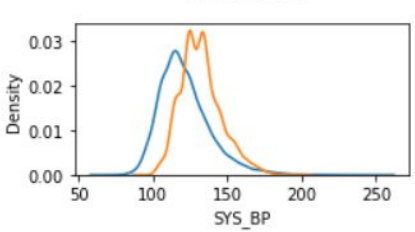
- **General goal:** Mimic the joint distribution of synthetic dataset using NHANES
- **Application:** covariate shift, concept shift
- **Methods we used:**
  - Naive:
    - Proportions matching
    - De-biasing
    - Resampling with nearest neighbours matching

# Methods for Dataset Shift

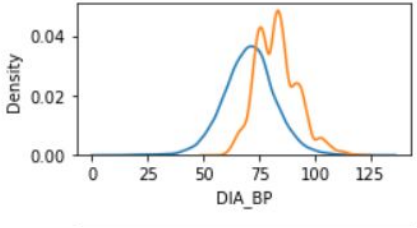
- **General goal:** Mimic the joint distribution of synthetic dataset using NHANES
- **Application:** covariate shift, concept shift
- **Methods we used:**
  - Naive:
    - Proportions matching
    - De-biasing
    - Resampling with nearest neighbours matching
  - Importance weighting
    - Logistic regression
    - Kullback-Leibler Importance Estimation Procedure (KLIEP)
    - Boosted Decision Tree Reweighting

$$w(\boldsymbol{x}) := \frac{p_{\mathrm{te}}(\boldsymbol{x})}{p_{\mathrm{tr}}(\boldsymbol{x})}.$$

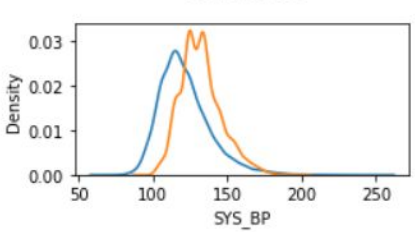
# Marginals: Systolic and Diastolic Blood Pressure



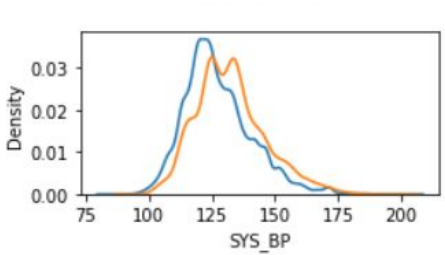

**Original**



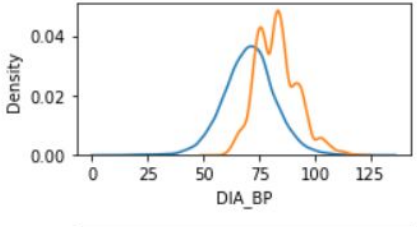

# Marginals: Systolic and Diastolic Blood Pressure

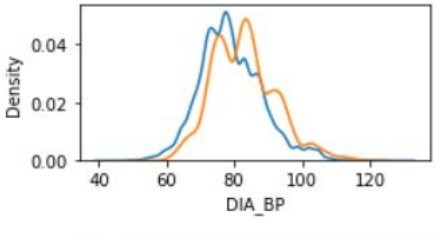



**Original**

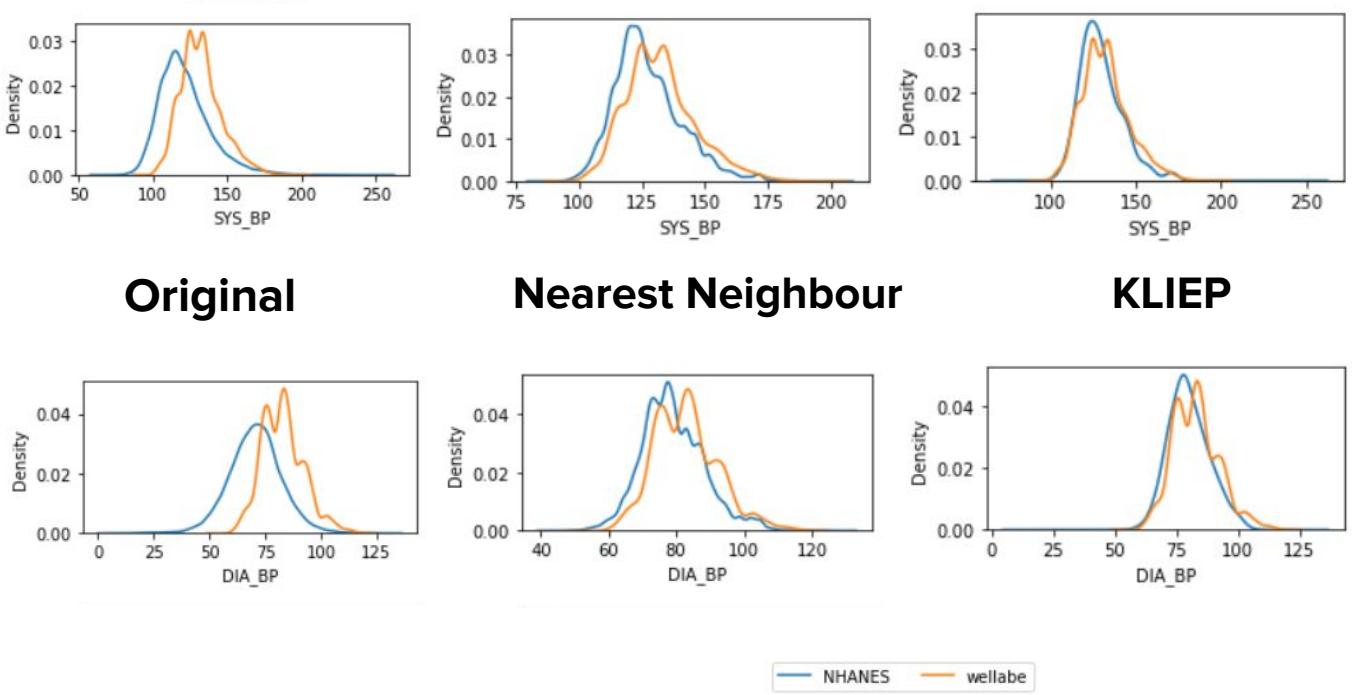



**Nearest Neighbour**

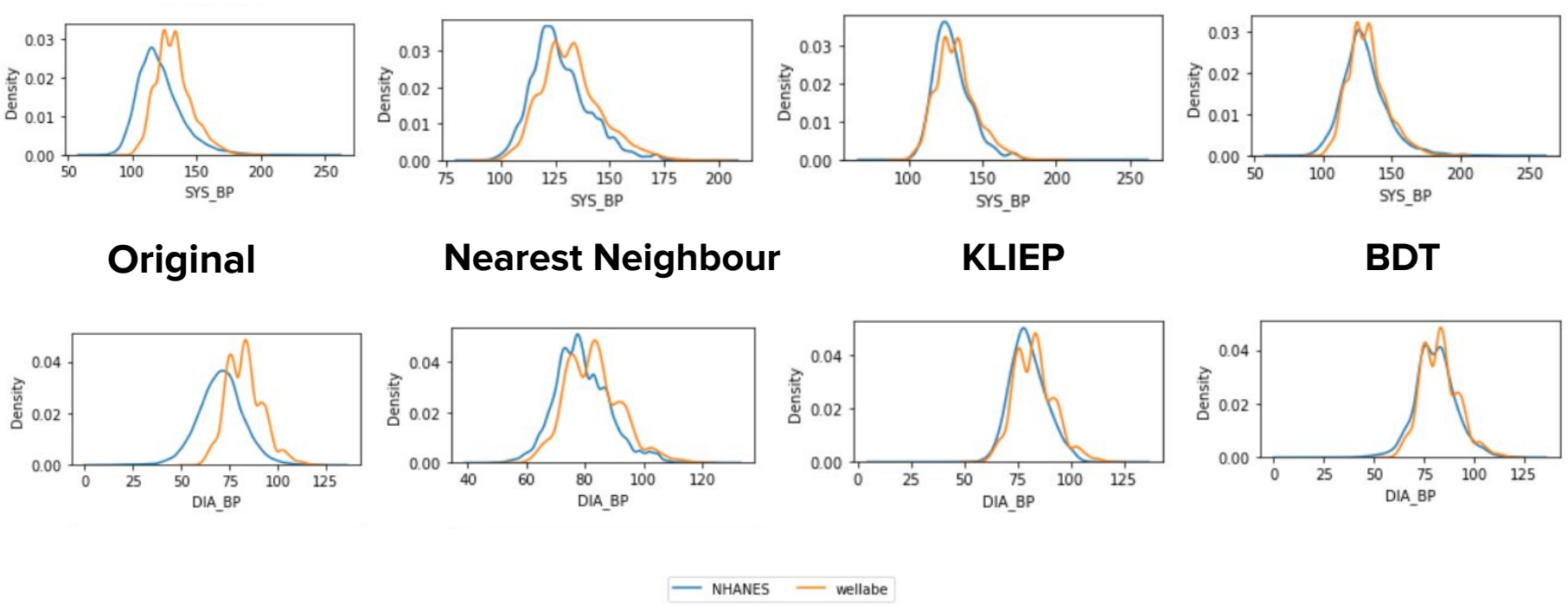


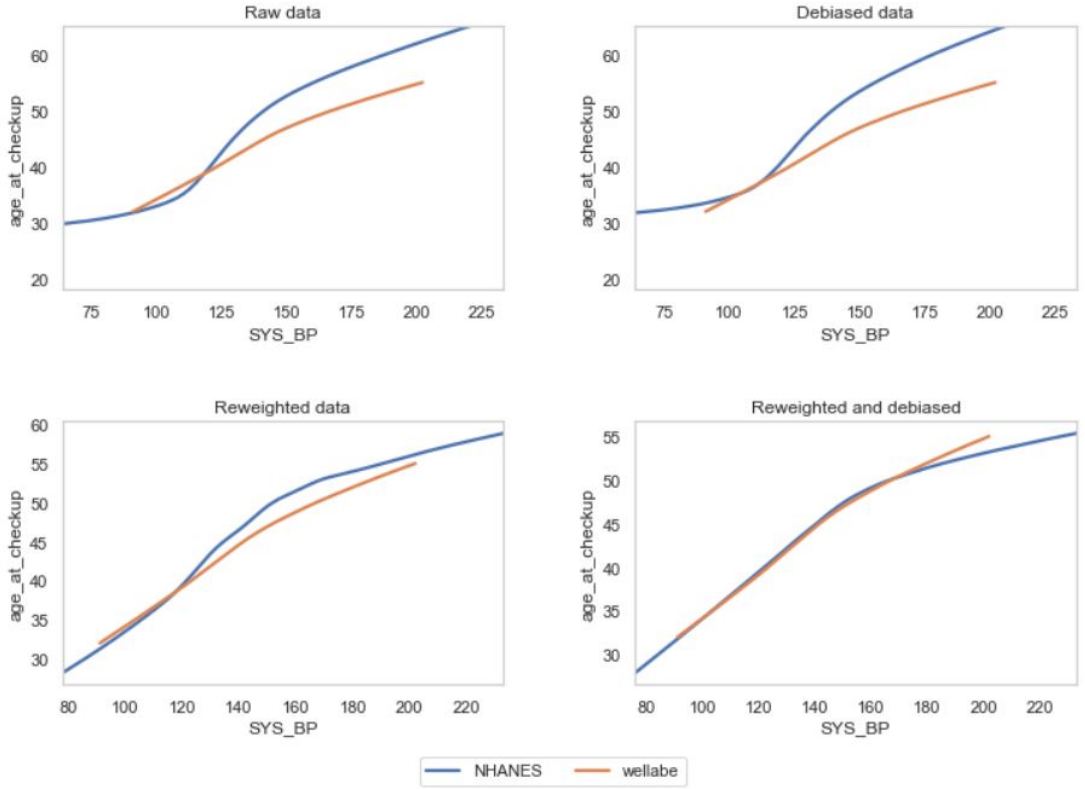

# Marginals: Systolic and Diastolic Blood Pressure



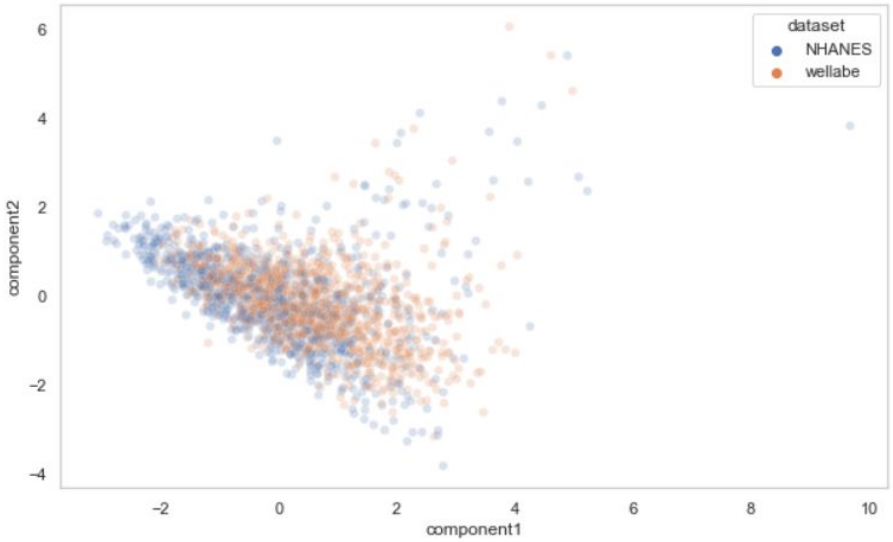
# Marginals: Systolic and Diastolic Blood Pressure

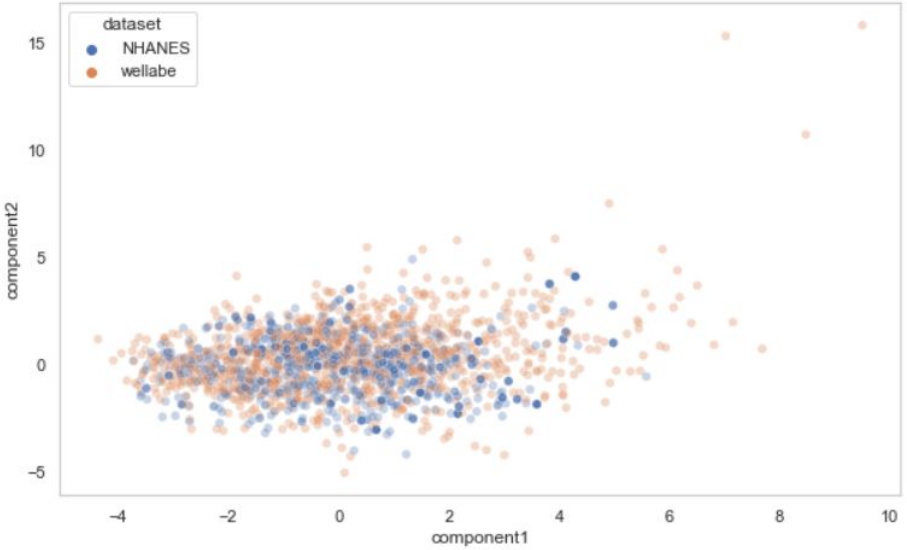


# Concept Shift



# Distribution Transformation



# Methods Comparison

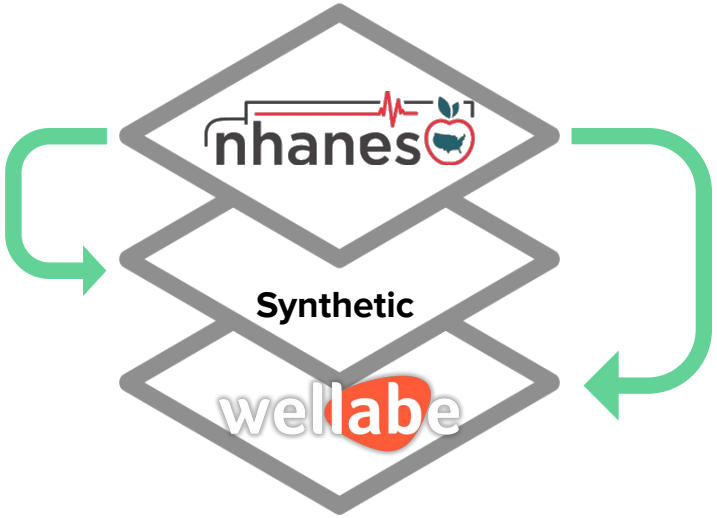
**Lower** ROC-AUC is **better** - means it's harder to distinct between NHANES and wellabe

| Method | ROC-AUC |
|---|---|
| **Original (baseline)** | 0.9612 |
| **NN** | 0.8933 |
| **LogReg** | 0.9360 |
| **KLIEP** | 0.8561 |
| **BDT** | 0.8496 |

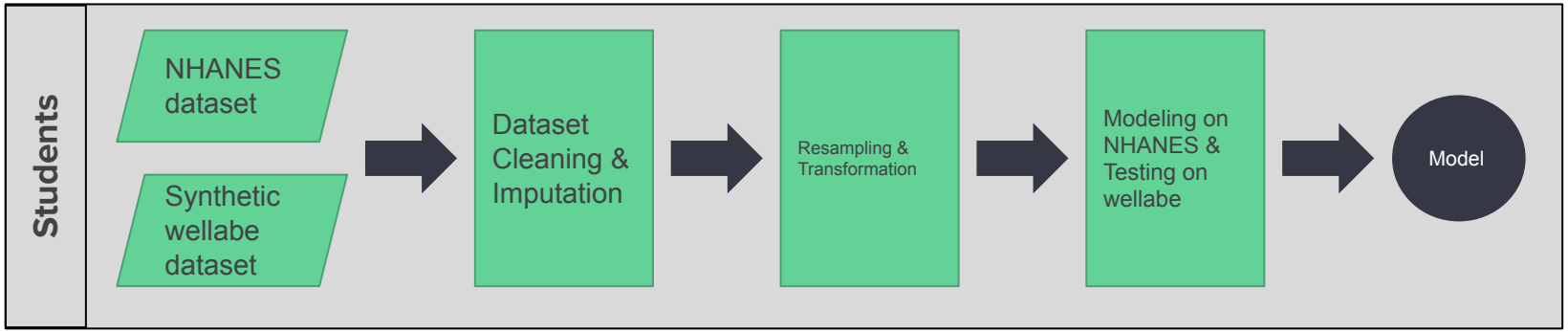
# Age Prediction

# Why age prediction?

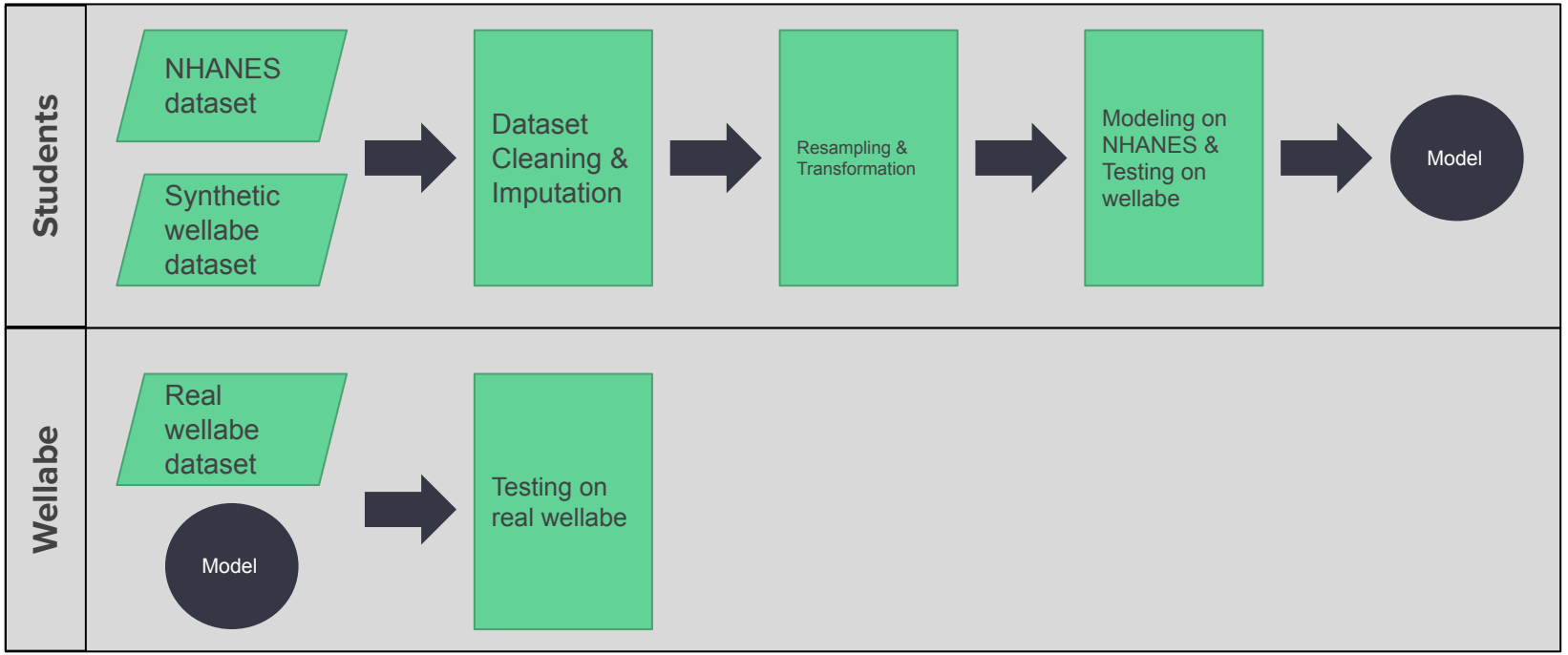
- Validate initial model transfer from NHANES to wellabe
- Simplicity & good interpretation
- Continuous label for range 18 to 65 is contained in both datasets
- Apply KLIEP reweighting to tackle both covariate and concept shift



# Model Transfer Pipeline



# Model Transfer Pipeline



# Setup

- Feature Choice:


- Evaluation metric:
  - Mean Average Error (MAE) score
  - Compare difference between model performance on the NHANES and real wellabe dataset (MAE diff)

# Regression Models

- Linear Regression
- Lasso Regression
- Ridge Regression
- ElasticNet
- Support Vector Regression
- Multivariate Adaptive Regression Splines (MARS)
- Generalized Additive Models (GAM)
- Extreme Gradient Boosting (XGBoost)

# Regression Models

- **Linear Regression**
- **Lasso Regression**
- **Ridge Regression**
- **ElasticNet**

**“Linear Models”**

- Support Vector Regression
- Multivariate Adaptive Regression Splines (MARS)
- Generalized Additive Models (GAM)
- Extreme Gradient Boosting (XGBoost)

# Results for Linear Model

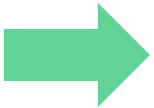
| Model | MAE (NHANES) | MAE (wellabe) | MAE (diff) |
|---|---|---|---|
| Linear | 11.052 | 8.304 | 2.748 |

**... without Reweighting**

# Results for Linear Model

| Model | MAE (NHANES) | MAE (wellabe) | MAE (diff) |
|---|---|---|---|
| Linear | 11.052 | 8.304 | 2.748 |

**... without Reweighting**



| Model | MAE (NHANES) | MAE (wellabe) | MAE (diff) |
|---|---|---|---|
| Linear | 8.388 | 8.346 | 0.042 |

**... with KLIEP Reweighting**

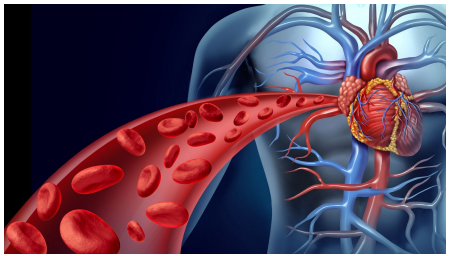
# Classification Tasks

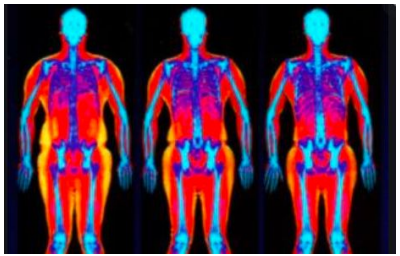
# Classification Task: Diabetes

- Objective: test if the models to identify potential patients at risk
- **Diabetes**: the only label contained in both wellabe and NHANES
- Diabetic population: 8% in NHANES, 0.8% in wellabe
- Glucose level is not available due to too many missing values (51.5%)

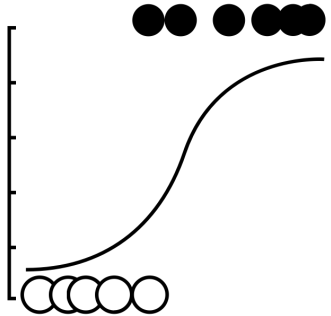
# Classification Task: Discretized Features

- Objective: more varieties & evaluations for model transfer
- **Discretized features**: below, within or above the healthy range
  - Alanine transaminase (ALT)
  - Cholesterol
  - Body mass index (BMI)

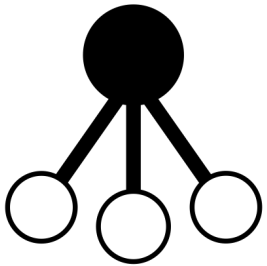




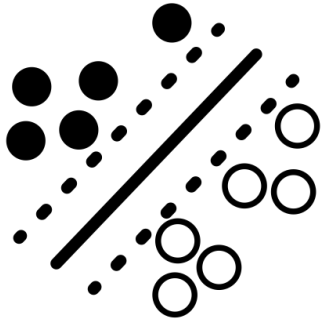
# Models



Logistic regression



Naive Bayes



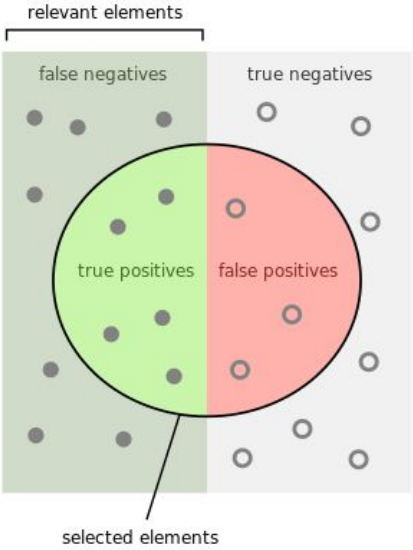
Support Vector Machine

# Evaluation Metric

- Primary metric**: recall**

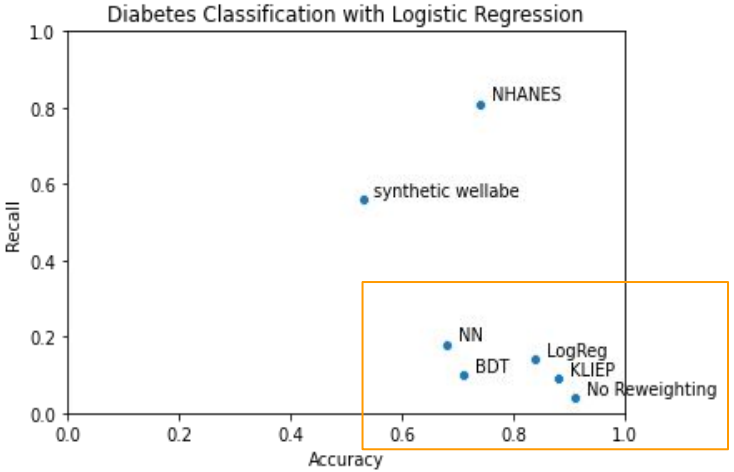
$$\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

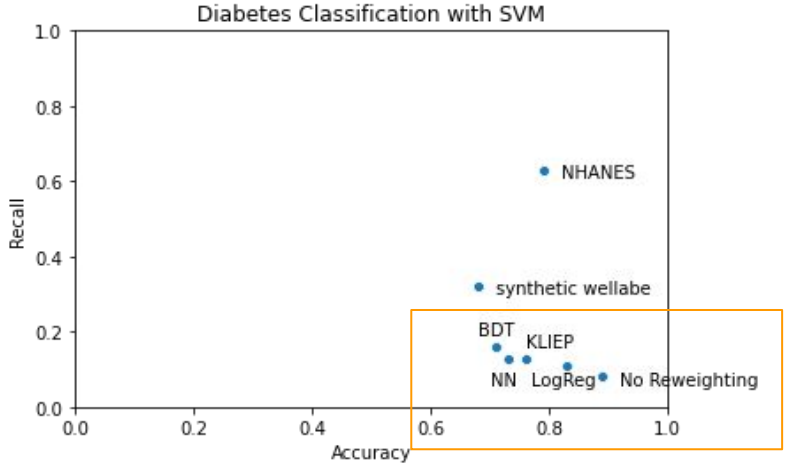
- Identify as many people with sickness as possible at the cost of FP
- **Accuracy** is also considered
- In multi-class classifications, micro-averaging is used to better deal with class imbalance




# Results: Diabetes

- Reweighting methods increase recall at the cost of lowering accuracy
- The synthetic dataset does not preserve the true interdependencies
- Learning from synthetic does not guarantee good results on real wellabe

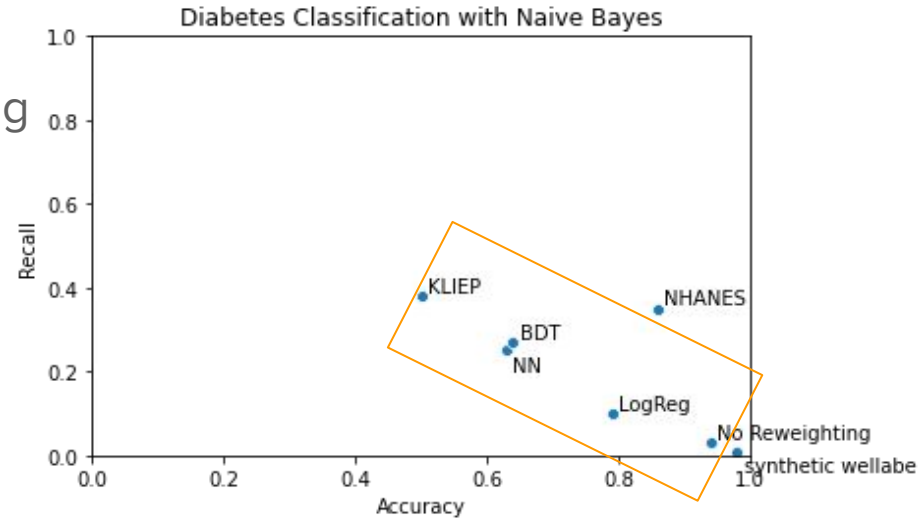


# Results: Diabetes

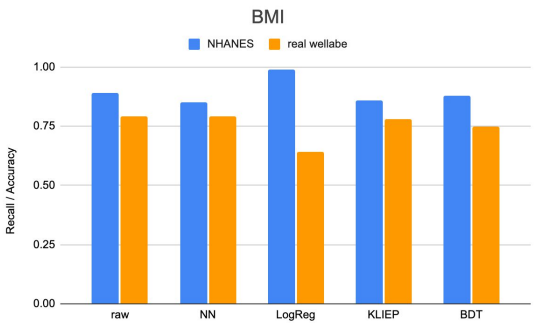
Difference in diabetic populations making modeling Naive Bayes difficult

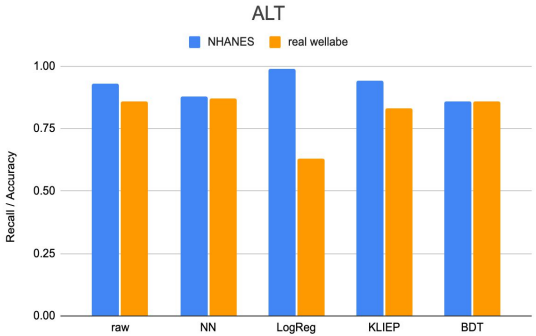
Self-reported diabetes labels, undiagnosed diabetes patients might exist

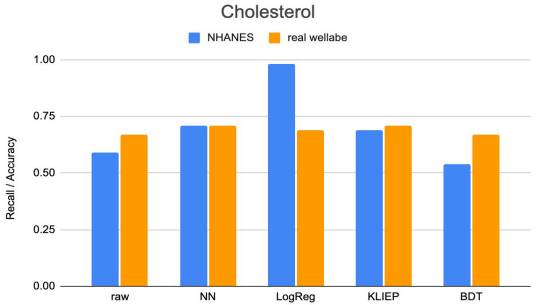


# Results: Discretized Features

- Results from logistic regression
- Accuracy & recall have same values in case of multi-class with micro-averaging
- Differences in score between two datasets are small in most cases
- Rare diseases transfer worse, more balanced labels can transfer better



# Final Remarks

# Lessons Learned

## Dataset Shift

Use dataset prediction method to validate difference between joint distributions

No single best resampling method suited for all prediction cases, but NN performs quite well

Reweighting methods can reduce the difference between synthetic wellabe and NHANES joint distributions

# Lessons Learned

## Dataset Shift

Use dataset prediction method to validate difference between joint distributions

No single best resampling method suited for all prediction cases, but NN performs quite well

Reweighting methods can reduce the difference between synthetic wellabe and NHANES joint distributions

## Synthetic Dataset

Model performance is restricted to the limitations of synthetic wellabe joint distribution

We expect better results when applying reweighting methods directly on real wellabe dataset

# What have we accomplished?

**Model Transfer Evaluation**

- Limitations of synthetic data
- Dataset Shift
- Resampling methods
- Evaluation metric

# What have we accomplished?

**Model Transfer Evaluation**

- Limitations of synthetic data
- Dataset Shift
- Resampling methods
- Evaluation metric

**Model Transfer Pipeline**

- Data Cleaning
- Resampling & Transformation
- Modeling on synthetic dataset
- Testing on real dataset

# What have we accomplished?

**Model Transfer Evaluation**

- Limitations of synthetic data
- Dataset Shift
- Resampling methods
- Evaluation metric

**Model Transfer Pipeline**

- Data Cleaning
- Resampling & Transformation
- Modeling on synthetic dataset
- Testing on real dataset

**First Models for Future Production**

- Asthma prediction
- Medication prediction

# Thank You