Diabetes Risk Prediction

IE 500 – Data Mining Project

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Diabetes: A Growing Disease with Serious Complications

Diabetes

- Blood sugar disease
- Body unable to produce or use insulin effectively

Prevalence in the US

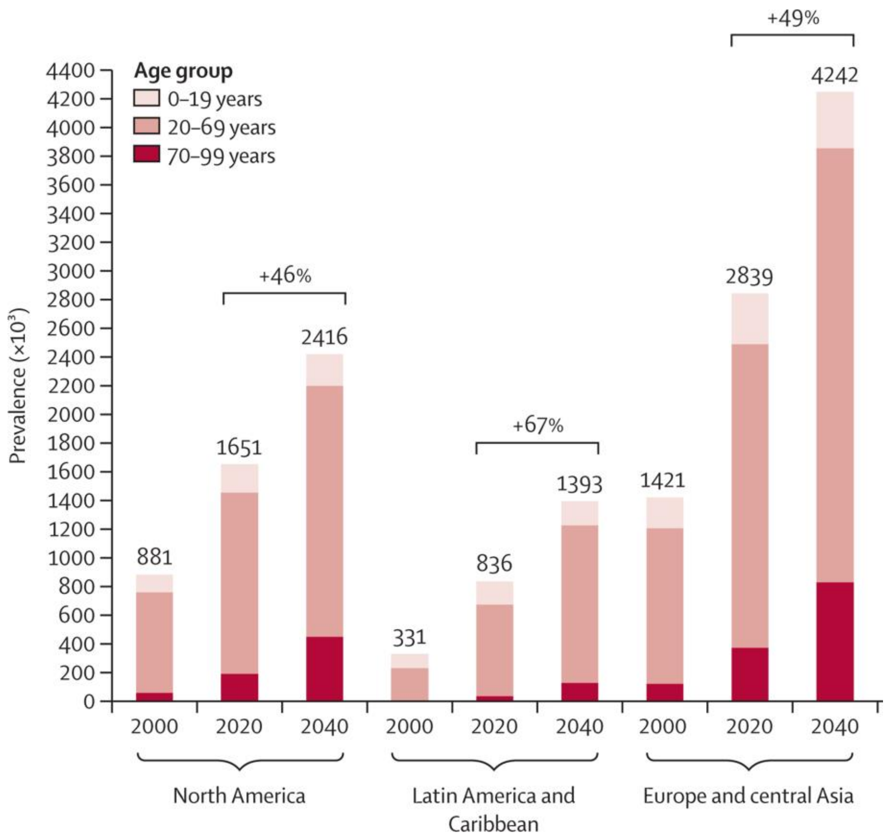
- 34+ million Americans diagnosed (CDC, 2018)
- 88 million at risk due to prediabetes
- $400 billion annual costs

Possible Complications

Heart diseases

Kidney Failure

Vision Loss



Source: *Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: a modelling study*

Diabetes: A Growing Disease with Serious Complications

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Diabetes

- Blood sugar disease
- Body unable to produce or use insulin effectively

Prevalence in the US

- 34+ million Americans diagnosed (CDC, 2018)
- 88 million at risk due to prediabetes
- $400 billion annual costs

Possible Complications

Heart diseases

Kidney Failure

Vision Loss

Project Goal

Develop accurate predictive model to enable early diabetes detection and mitigate disease progression

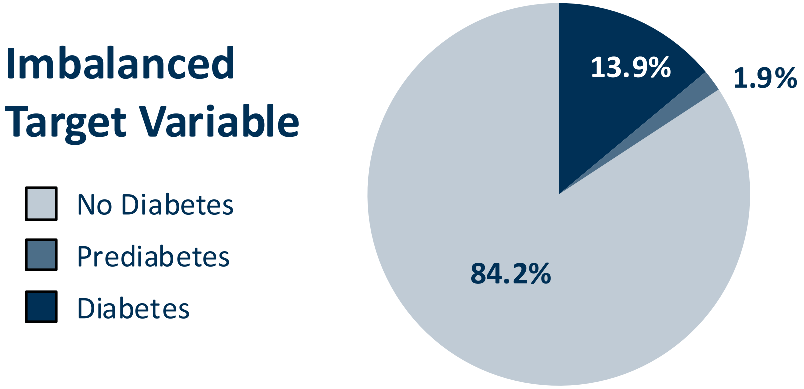
Dataset Overview and Preprocessing Steps

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Dataset

- Preprocessed Behavioral Risk Factor Surveillance System (BRFSS) dataset
- 253,680 observations with 22 features
  - 1 target variable | ***Diabetes_012***
  - 14 binary features | e.g., *Smoker, Stroke, HighBP*
  - 4 ordinal features | e.g., *Education, Age*
  - 3 numerical features | e.g., *BMI or MentHlth*

Imbalanced Target Variable



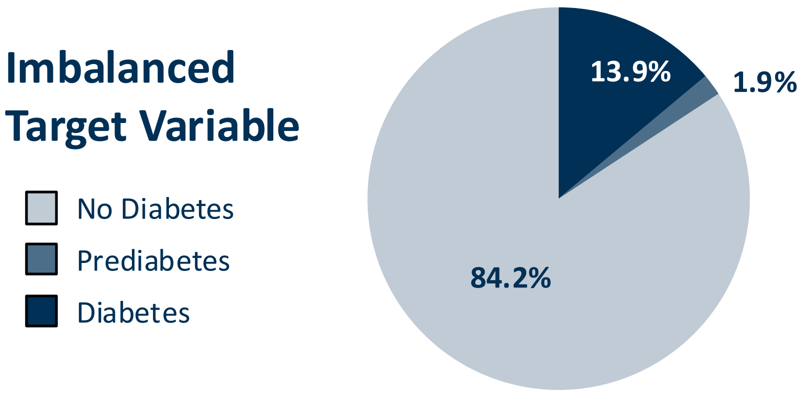
Dataset Overview and Preprocessing Steps

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Dataset

- Preprocessed Behavioral Risk Factor Surveillance System (BRFSS) dataset
- 253,680 observations with 22 features
  - 1 target variable | ***Diabetes_012***
  - 14 binary features | e.g., *Smoker, Stroke, HighBP*
  - 4 ordinal features | e.g., *Education, Age*
  - 3 numerical features | e.g., *BMI or MentHlth*

Imbalanced Target Variable



Preprocessing

- Inconsistency checks (e.g., outlier detection, missing values, etc.)
- Merging *prediabetes* and *diabetes* creating binary target
- Processing of numerical features
  - Normalization of *MentHlth* and *PhysHlth*
  - Binning of *BMI* into medical classes, i.e., *Underweight, Normal Weight, Overweight, Obesity*

Defining Baseline Strategies and Selecting Binary Classification Models

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Baselines

Majority Class

Always predicting the most common class

Accuracy: 0.84
Recall on Diabetes: 0

Stratified

Random predictions based on class distributions

Accuracy: 0.73
Recall on Diabetes: 0.16

Highest Correlation

Prediction based on threshold of highest correlating feature

Accuracy: 0.79
Recall on Diabetes: 0.39

Defining Baseline Strategies and Selecting Binary Classification Models

Baselines

Majority Class

Always predicting the most common class

Accuracy: 0.84
Recall on Diabetes: 0

Stratified

Random predictions based on class distributions

Accuracy: 0.73
Recall on Diabetes: 0.16

Highest Correlation

Prediction based on threshold of highest correlating feature

Accuracy: 0.79
Recall on Diabetes: 0.39

Model Selection

Distance-Based

K-Nearest Neighbors
Nearest Centroids

Tree-Based

Decision Trees
Random Forest
AdaBoost Decision Tree

Kernel-Based

Support Vector Machines

Linear

Logistic Regression

Probabilistic

Naïve Bayes

Deep Learning

Neural Network

Model Training via Cross-Validation including Over- and Undersampling Techniques

Increasing Minority Class | Decreasing Majority Class

Random Oversampling	SMOTE Oversampling	SMOTE Tomek and Links	Random Undersampling
Duplicate random samples of minority class	Generate synthetical new samples of minority class	SMOTE Oversampling **+** removing “noisy” samples of majority class	Removing random samples of majority class

Model Training via Cross-Validation including Over- and Undersampling Techniques

Increasing Minority Class — Decreasing Majority Class

Random Oversampling	SMOTE Oversampling	SMOTE Tomek and Links	Random Undersampling
Duplicate random samples of minority class	Generate synthetical new samples of minority class	SMOTE Oversampling **+** removing "noisy" samples of majority class	Removing random samples of majority class

Cross Validation

Parameter grid includes model-specific hyperparameters as well as sampled datasets.

- Ensure robust model performance
- Select best parameter set
- Exploitation of training data

Comparison and Evaluation of the Best Models of Each Classifier

	Baseline Stratified	Logistic Regression	Decision Tree	Random Forest	AdaBoost Tree	SVM	KNN	Nearest Centroid	Naive Bayes
Accuracy									
Precision 0									
Precision 1									
Recall 0 (specificity)									
Recall 1 (sensitivity)									
F1-Score 0									
F1-Score 1									

Classes: 0 (no-diabetes), 1 (diabetes)

Comparison and Evaluation of the Best Models of Each Classifier

	Baseline Stratified	Logistic Regression	Decision Tree	Random Forest	AdaBoost Tree	SVM	KNN	Nearest Centroid	Naive Bayes
Accuracy	*0.7343*	0.728	0.6988	0.7535	0.7467	0.4792	0.7433	0.6902	0.7356
Precision 0	*0.8419*	0.9431	0.9428	0.9341	0.9335	0.8809	0.9303	0.937	0.9354
Precision 1	*0.1547*	0.3395	0.3153	0.3583	0.3507	0.1857	0.3448	0.3051	0.3411
Recall 0 (specificity)	***0.8429***	0.7205	0.684	0.7611	0.753	0.4415	0.7516	0.6779	0.737
Recall 1 (sensitivity)	*0.1538*	0.7678	0.7781	0.7131	0.7132	0.6809	0.6988	0.7562	0.7279
F1-Score 0	***0.8424***	0.817	0.7928	0.8388	0.8336	0.5882	0.8314	0.7866	0.8244
F1-Score 1	*0.1542*	0.4708	0.4488	0.477	0.4702	0.2918	0.4617	0.4348	0.4645

Classes: 0 (no-diabetes), 1 (diabetes)

Comparison and Evaluation of the Best Models of Each Classifier

	Baseline Stratified	Logistic Regression	Decision Tree	Random Forest	AdaBoost Tree	SVM	KNN	Nearest Centroid	Naive Bayes	Neural Network
Accuracy	*0.7343*	0.728	0.6988	0.7535	0.7467	0.4792	0.7433	0.6902	0.7356	0.7775
Precision 0	*0.8419*	0.9431	0.9428	0.9341	0.9335	0.8809	0.9303	0.937	0.9354	0.9191
Precision 1	*0.1547*	0.3395	0.3153	0.3583	0.3507	0.1857	0.3448	0.3051	0.3411	0.3753
Recall 0 (specificity)	***0.8429***	0.7205	0.684	0.7611	0.753	0.4415	0.7516	0.6779	0.737	0.8068
Recall 1 (sensitivity)	*0.1538*	0.7678	0.7781	0.7131	0.7132	0.6809	0.6988	0.7562	0.7279	0.6205
F1-Score 0	***0.8424***	0.817	0.7928	0.8388	0.8336	0.5882	0.8314	0.7866	0.8244	0.8593
F1-Score 1	*0.1542*	0.4708	0.4488	0.477	0.4702	0.2918	0.4617	0.4348	0.4645	0.4677

Classes: 0 (no-diabetes), 1 (diabetes)

Evaluation

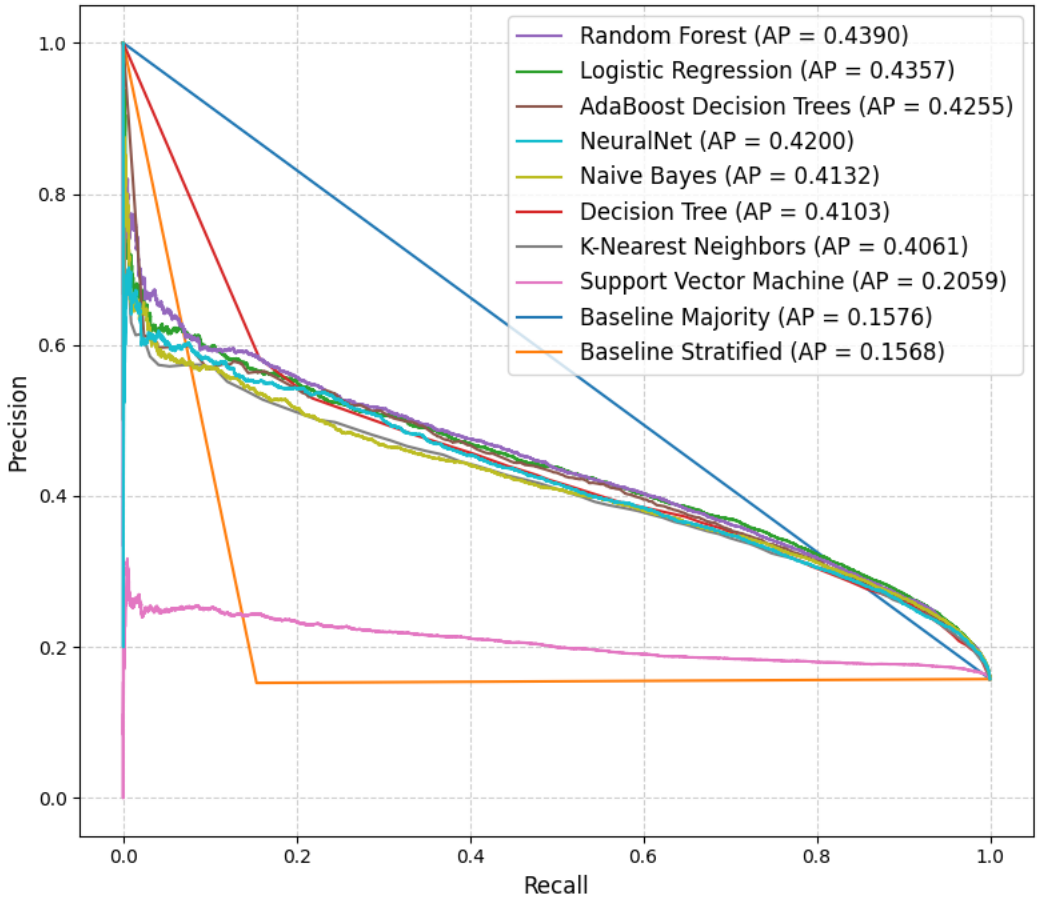


Fig 1: Precision-Recall Curve with Average Precision (AP)

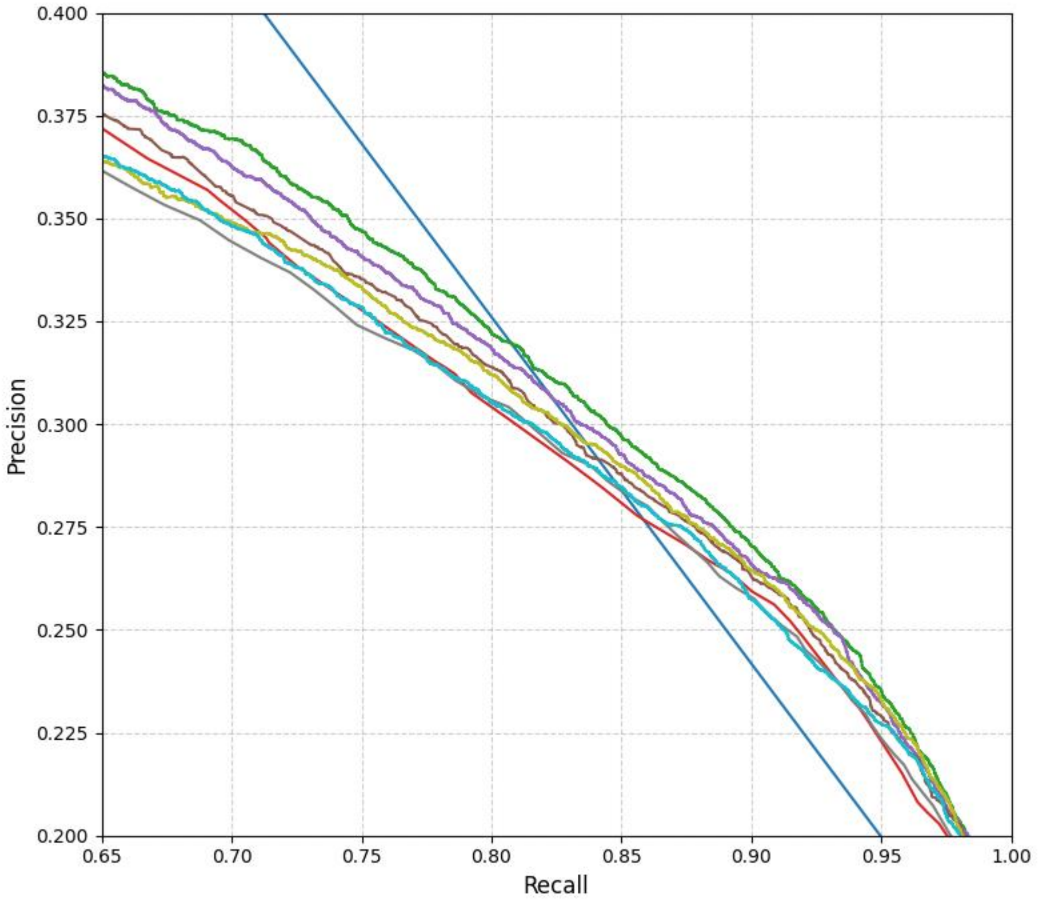


Fig 2: Zoomed in Precision-Recall Curve

Key Take-Aways

Data Exploration → Sampling → Models → Metrics → Outlook

Key Take-Aways

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Data Exploration → Sampling → Models → Metrics → Outlook

Imbalanced Dataset

- No Diabetes (0): 86.07%
- Diabetes (1): 13.93%

Key Take-Aways

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

(Random) Oversampling works best for most models

- Balancing underrepresentation of minority class (Diabetes 1)

Data Exploration → Sampling → Models → Metrics → Outlook

Imbalanced Dataset

- No Diabetes (0): 86.07%
- Diabetes (1): 13.93%

Key Take-Aways

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

(Random) Oversampling works best for most models

- Balancing underrepresentation of minority class (Diabetes 1)

Data Exploration → Sampling → Models → Metrics → Outlook

Imbalanced Dataset

- No Diabetes (0): 86.07%
- Diabetes (1): 13.93%

Random Forest is best performing "traditional" model

- Averages predictions of multiple decision trees

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Key Take-Aways

Data Exploration → Sampling → Models → Metrics → Outlook

Imbalanced Dataset

- No Diabetes (0): 86.07%
- Diabetes (1): 13.93%

(Random) Oversampling works best for most models

- Balancing underrepresentation of minority class (Diabetes 1)

Random Forest is best performing "traditional" model

- Averages predictions of multiple decision trees

Recall on positive class is important for our use case

- False Positives more bearable than False Negatives

Key Take-Aways

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Data Exploration → Sampling → Models → Metrics → Outlook

Imbalanced Dataset

- No Diabetes (0): 86.07%
- Diabetes (1): 13.93%

(Random) Oversampling works best for most models

- Balancing underrepresentation of minority class (Diabetes 1)

Random Forest is best performing "traditional" model

- Averages predictions of multiple decision trees

Recall on positive class is important for our use case

- False Positives more bearable than False Negatives

Neural Networks very promising

- Good results with simple network and little training

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Thank you!

Any questions? Let’s discuss!

Team Information and Contact Details

11 - Support Vector Superstars

Matthias Fast, 2111111 – matthias.fast@students.uni-mannheim.de

Philipp Gänz, 1736316 – philipp.robert.gaenz@students.uni-mannheim.de

Salome Heckenthaler, 1742998 – salome.heckenthaler@students.uni-mannheim.de

Patricia Paskuda, 2119717 – patricia.paskuda@students.uni-mannheim.de

Benedikt Prisett, 2119134 – benedikt.prisett@students.uni-mannheim.de