IN-STK5000 PROJECT 1

Diabetes classification for automatic decision-making

Amir, Cornelius, Espen & Torstein

Structure

01
Building a scenario

02
Data analysis & processing

03
Feature selection

04
Classification & evaluation

01

Building a scenario

Building a scenario

Overview dataset:

- Small dataset (520 rows)
- Each row represents an individual
  - Medical record
  - Collected by hospitals
- Variables are information regarding the individual
  - General information: Age, Gender, Height, Weight, etc.
  - Medical information: Medical diagnoses, etc.
- Target variable: Diabetes

Data-driven decision making

- More complex relationship to a disease
- Analyze multiple features at once
- Find best models

Scenario

Self screening for school children on diabetes related features to detect risk of diabetes, and encourage individuals with a high risk to engage in a clinical check with their doctor

Recommend at risk students to see a doctor

- X features you can self report
- Score > y ⟶ go check in with your doctor

Overview scenario

- Target:
  - Self screening → Efficient & easy
  - Detect diabetes & risk of diabetes
- Target group:
  - School children → detect diabetes early
- Benefit:
  - Diabetes is costly for society
  - Our scenario
    - School system
      - Detect Diabetes early
      - Prevention and awareness
      - Less cost of collection

Overview scenario

- Goal
  - Send as many of the at risk people to a clinical check
  - High recall:
    - Want small proportion of false negative
      - Rather alert more people with low risk of diabetes than not alerting people with high risk of diabetes

Concerns

- Privacy
  - Sensitive information - medical records
  - Identity - Gender, Doctor, Occupation, etc.
- Biases
  - Skewed dataset
    - Adults
    - White
    - With a complex medical record
    - Large proportion with diabetes

Mitigate concerns

- Should not be able to identify people from dataset
  - Remove sensitive information/features
    - Doctor
    - Occupation
    - Race
- More data
  - Important
  - Better prediction
  - Anonymize data further

02

Data analysis & processing

Data analysis & processing

Initial data analysis

- 520 entries (patients)
- 24 variables
  - Multi-valued categorical variables
  - Binary categorical variables
  - Continuous variables
- Target value: Diabetes
  - binary: “yes”/”no”
- 0.87% of data is missing
  - 0 to 7 missing values for each variable
  - 95 individuals have missing values
    - 1 to 3 per individual
    - 13 individuals with 2 or more missing values

Interesting findings

- Interesting findings:
  - Differences in the data for each variable
    - Reasoning: data from different hospitals
    - Examples:
      - “Yes”/”No” or “yes”/”no”
      - Height measured in cm or m
  - TCep – whether the individual has had tattoos or cosmetic enhancing procedures
    - Initially seems like an irrelevant variable
  - Close to homogenous in terms of race (99% white)

Initial data cleanup

- Initial data cleanup:
  - Convert all reported string entries to lower case letters:
    - “Yes”/”No” ⟶ “yes”/”no”: Only lowercase letters entries
  - Remove individuals with 2 or more missing values
    - 13 rows/individuals
    - Dataset: 520 entries ⟶ 507 entries

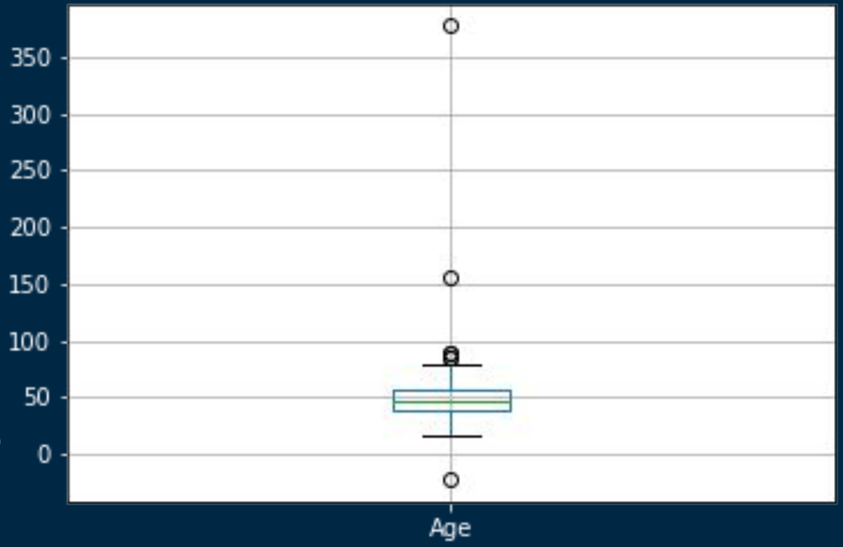
Outliers

- Categorical variables – hard to find outliers
- Numerical variables
  - Temperature low variance: mean, max, min, 50% all close to each other

	count	mean	std	min	50%	max
Age	504.0	48.623016	19.844944	-22.0000	47.000	377.00
Height	501.0	161.056607	37.945772	1.5239	169.490	195.82
Weight	503.0	67.825905	18.188628	21.8800	66.670	128.11
Temperature	507.0	37.000533	0.208257	36.4700	37.000	37.57
Urination	500.0	2.329340	1.061961	0.9600	2.295	15.00

	count	mean	std	min	50%	max
Age	504.0	48.623016	19.844944	-22.0000	47.000	377.00

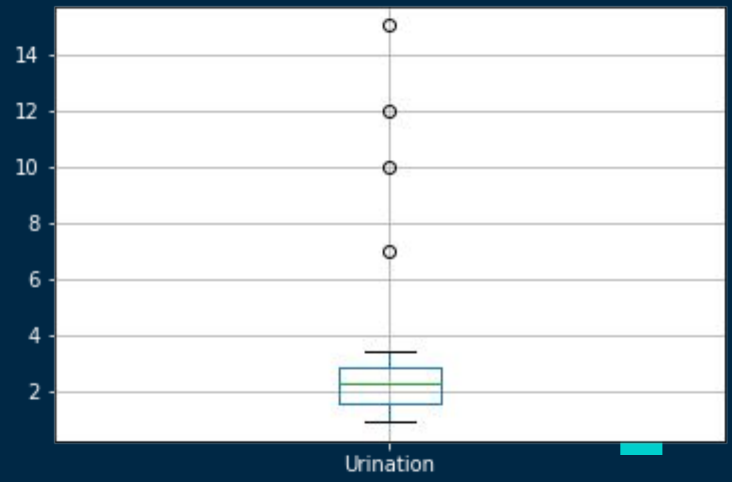
- **Age:**
  - min value = - 22
  - max value = 377
  - 3 Outliers: - 22, 155, 377
  - No reliable explanation
  - Remove outliers
    - dataset:
      - 507 entries ⟶ 504 entries



	count	mean	std	min	50%	max
Urination	500.0	2.329340	1.061961	0.96	2.295	15.0

- **Urination:**
  - max value 15.0
    - 15 L seems like a lot per day
  - 4 outliers:
    - 7.0, 10.0, 12.0, 15.0
    - See no clear reason → remove
    - Dataset:
      - 504 entries → 500 entries

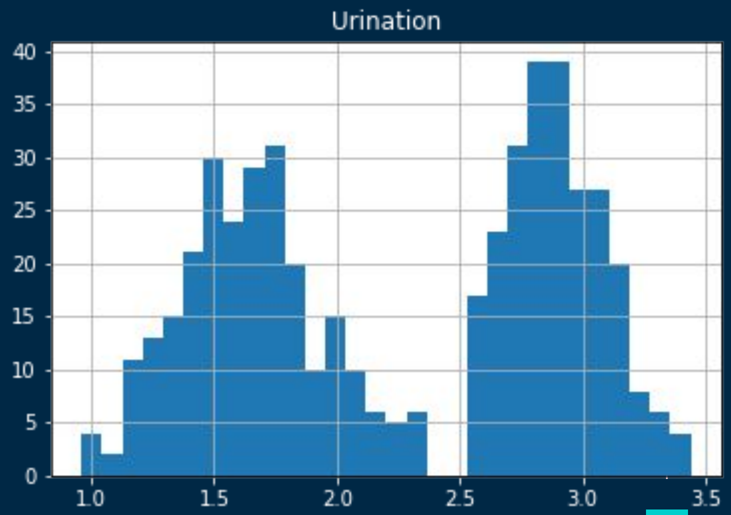
Urination



Urination

- After removing outliers
  - Bi distribution
- No values between 2.37 & 2.55
  - clear cut off between distributions
- Diabetes ratio
  - low - urination = 25.8% diabetes
  - high - urination = 80.9% diabetes
    - Clear difference
- **Manipulation**
  - **Urination → Urination_high**
    - **Cutoff at urination = 2.4**

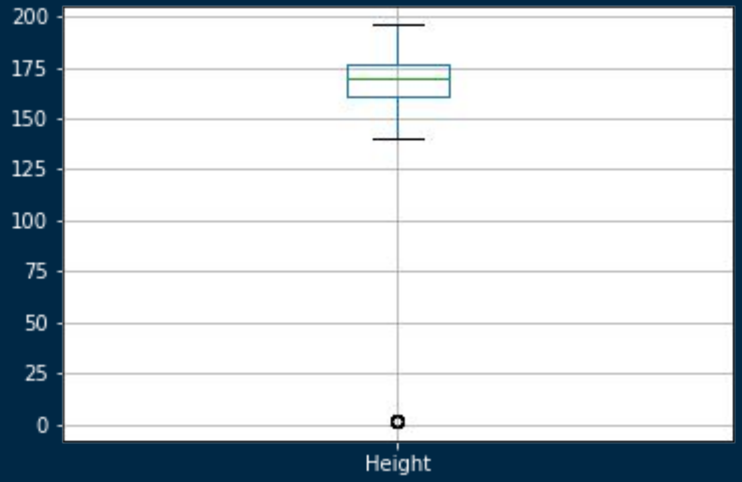
Urination Bi distribution



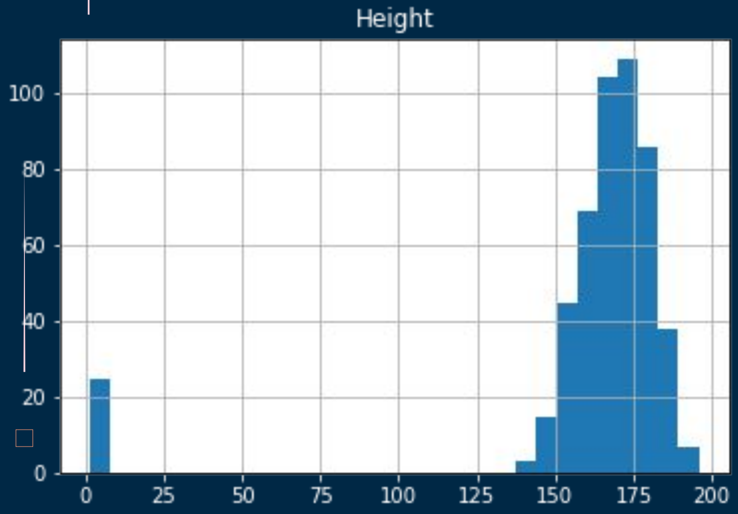
	count	mean	std	min	50%	max
Height	501.0	161.056607	37.945772	1.5239	169.49	195.82

- **Height:**
  - min value = 1.5239
  - max value = 195.82
    - Different scales – cm & m
    - No values between 2.00 and 130.00
      - Below 2.00 = meters
      - Above 130.00 = cm
  - **Converting m to cm**
    - Multiply all values below 2.00 by 100

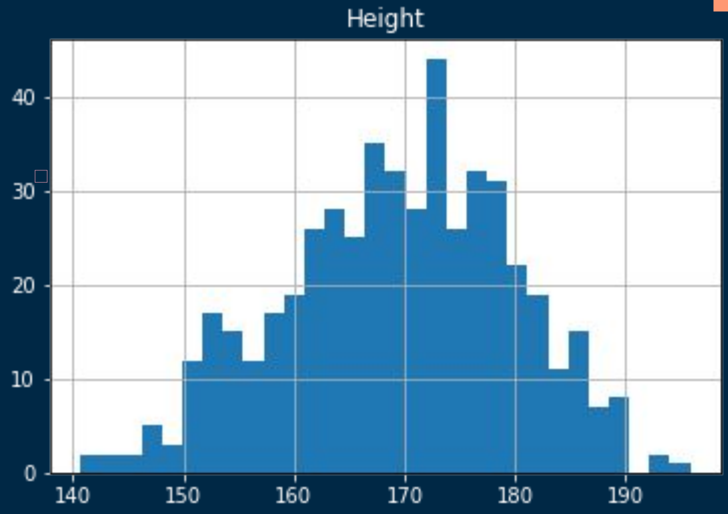
Height



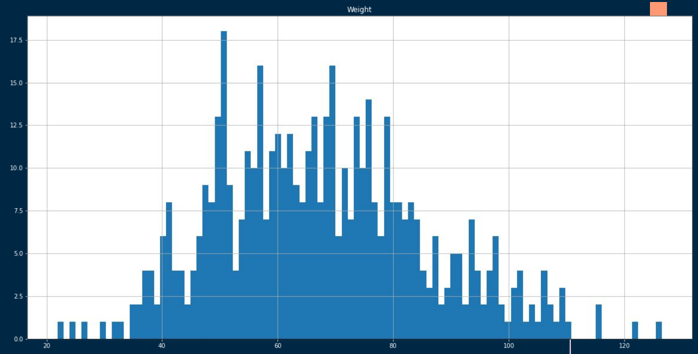
Height before transformation



Height after transformation



	count	mean	std	min	50%	max
Weight	496.0	67.694173	18.089261	21.88	66.58	126.53



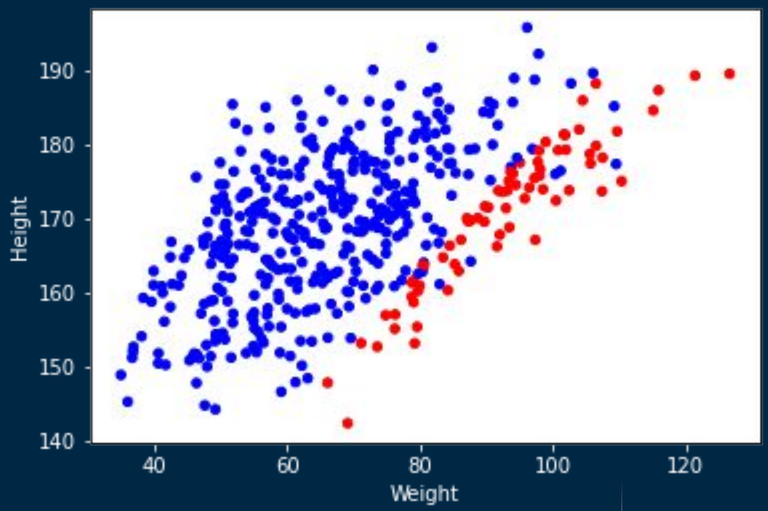
- **Weight:**
  - Complicated case:
    - min value = 21.88
      - Unrealistic
      - Possible reasons:
        - Children: no individuals below the age of 16
        - Amputations: consequence of diabetes
        - Incorrect reporting
        - Different scales

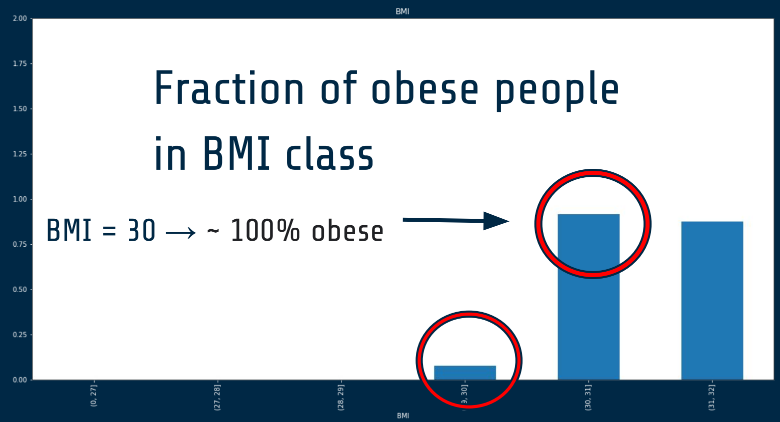
- Convert to BMI:
  - **BMI = kg/m2**
  - Create a cut off for underweight people
- BMI does not take gender into account, or age above 18 years old
  - NHI lists BMI <= 16 as underweight degree 3 (most extreme case)
    - Our cut off = BMI < 15
    - Eliminate 24 individuals
      - Dataset:
        - 500 entries → 476 entries
- What about overweight people?
  - More likely to weight 120kg than 20kg
  - Heavy people can have a lot of muscles
- Did not use Z-score or other outlier detection methods because that would cut off more heavy than underweight people

- **Obesity:**
  - 1 missing value
  - BMI have a clear relationship with Obesity
  - Easy to fill using a BMI/Obesity cut off
    - BMI >= 30 ⟶ Obese
    - For individual with missing Obesity value:
      - Obesity = “yes” if BMI >= 30
      - Obesity = “no” if BMI < 30

Height vs weight

Red = Diabetes positive





Dataset overview after outlier removal

	count	mean	std	min	50%	max
Age	465.0	48.030108	12.153435	16.000000	47.000000	90.000000
Height	462.0	169.302381	10.184845	142.300000	169.905000	195.820000
Weight	464.0	69.230927	17.111345	34.990000	68.135000	126.530000
Temperature	468.0	36.997778	0.208792	36.470000	36.995000	37.570000
Urination	461.0	2.269544	0.671928	0.960000	2.360000	3.440000
BMI	458.0	24.004288	4.953578	15.023426	23.467955	35.999868

- Continuous data look good
- 476 individuals
- Still some missing values

Missing values

- **Race:**
  - 99% white individuals
  - Assumption: the last 1% has no significant impact
  - Too few entries: consider non-white people outliers and remove
  - Dataset:
    - 476 entries ⟶ 468 entries

Deleting rows with missing values:

- Delete rows with missing values in the following variables (numerical and multi-categorical):
        - Age (3 missing values)
        - Gender (2 missing values)
        - Height (6 missing values)
        - Weight (4 missing values)
        - GP (7 missing values)
        - Occupation (2 missing values)
    - Exists method to fill:
            - prediction on age, gender, weight → predict height
                - combinations of these
            - occupation based on age (retired)
        - Cost greater than reward to implement such methods
    - Entries still affected by missing data : 24
    - Dataset:
        - 468 entries → 444 entries

Correlation – Methods

- Three primary correlation measurements used as a guideline
  - Chi-square test - Nominal variables vs Nominal variables
  - ANOVA f-test - Numeric variables vs Nominal variables (without rank)
  - Pearson's correlation - Numeric variables vs Numeric variables
- Chi-square hypothesis test:
  - H_0 – that the nominal variable is independent of the nominal variable
  - H_A – that the nominal variable is dependent on the nominal variable
- ANOVA hypothesis test:
  - H_0 – that the numeric variable is independent of the nominal variable
  - H_A – that the numeric variable is dependent on the nominal variable
- Pearson correlation coefficient
  - Coefficient ranging from -1 to 1

Correlation – Features vs target

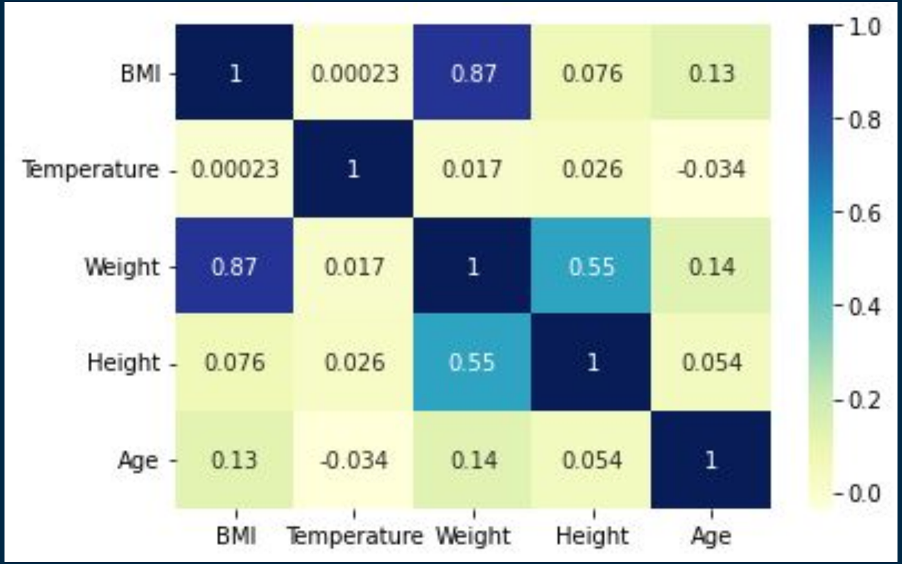
- Variables that cannot be excluded from being independent from the target variable with – Significance level: 0.05:
  - The chi-square test: four nominal variables
  - The ANOVA f-test: Temperature and BMI

	chi2-score	p-value
Genital Thrush_yes	3.0	0.0830
Delayed Healing_yes	1.8	0.1773
Obesity_yes	1.2	0.2676
Itching_yes	0.1	0.7814

	f_classif-score	p-value
Height	42.9	0.0000
Age	7.7	0.0058
Weight	6.0	0.0143
Temperature	1.8	0.1783
BMI	0.4	0.5290

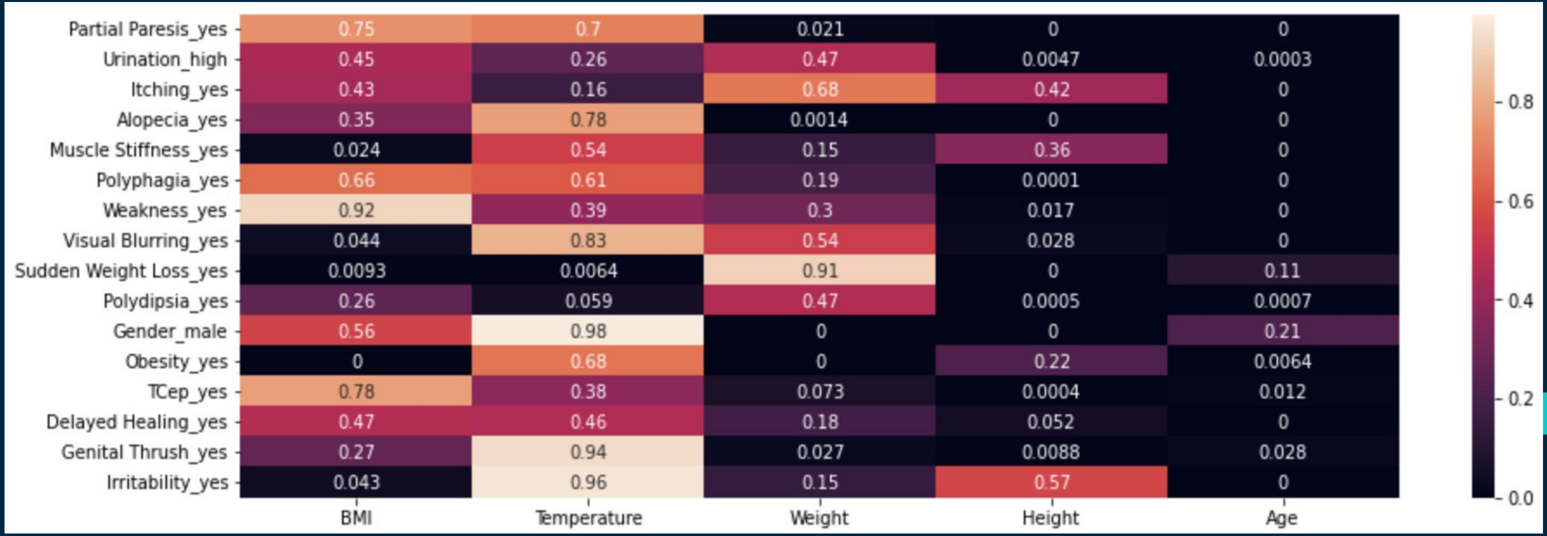
Correlation – Features vs features

- We use Pearson's correlation to check between numeric features
- Matrix shows high correlation between:
  - Weight and BMI
  - Height and Weight



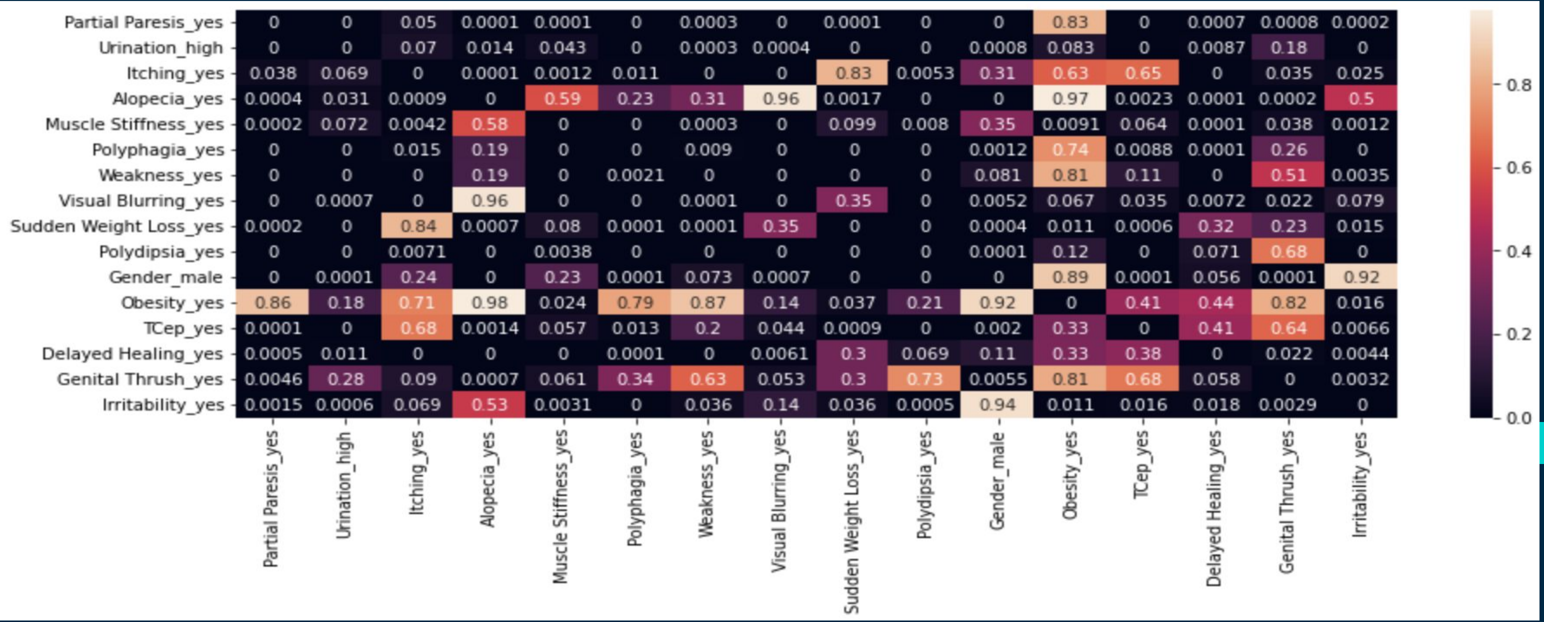
Correlation – Features vs features

- ANOVA f-test to check significant correlation between numeric and nominal features: Displaying p-values
- Black squares indicate possible dependence



Correlation – Features vs features

- Chi-square test to check significant correlation between nominal features: Displaying p-values
- Black squares indicate possible dependence



03

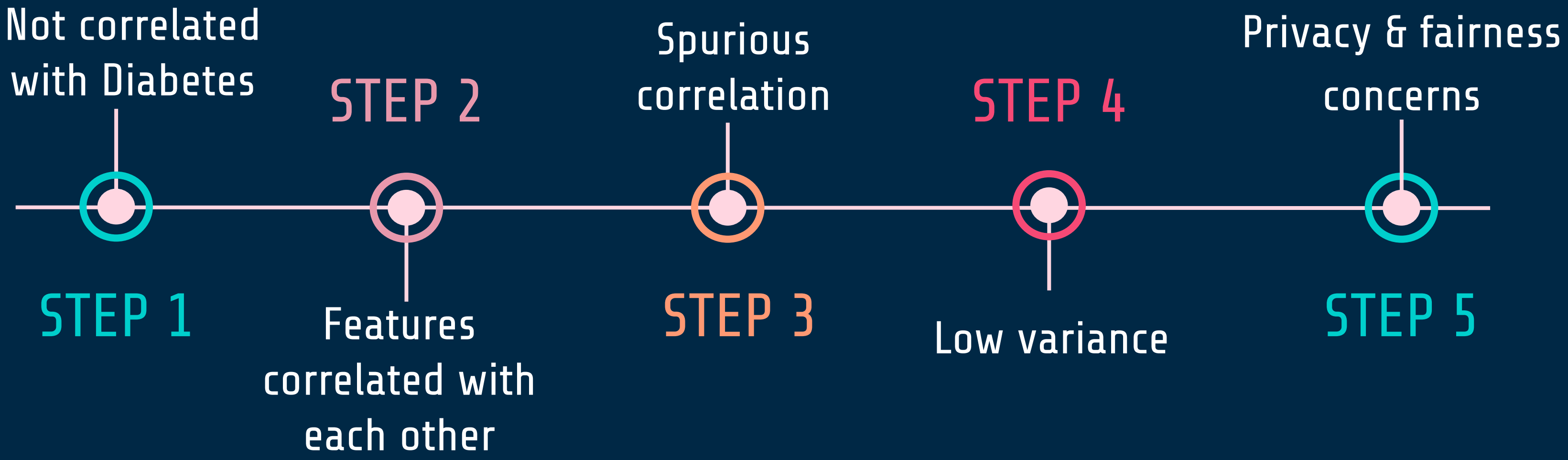
Feature selection

Feature selection

Selecting features – procedure

- Initial dataset – 23 predictive features for Diabetes
    - Removed race due to homogeneous dataset
    - Add BMI as alternative indicator for Obesity
- 23 potential predictive features for Diabetes
- 5 steps to remove features

5 step feature removal procedure



Selecting features – Variables removed

The following variables were removed in each step:

- **Step 1 – No correlation between feature and target:**
  - Muscle Stiffness
  - Genital Thrush
  - Delayed Healing
  - Obesity, Itching
  - Temperature
  - BMI

Selecting features – Variables removed

The following variables were removed in each step:

- **Step 2 – Correlation between features:**
  - Polydipsia - High correlation with Urination_high
    - Backed by high chi-square value relative to other features
  - Age - Correlation with multiple other features
    - Additionally, age is not productive for our use case, since we are targeting our product at school children, that is a homogenous group age wise
  - Height - Correlation with multiple other features
    - Especially weight

Selecting features – Variables removed

The following variables were removed in each step:

- **Step 3 – Spurious correlation**
  - TCEP - Has high correlation with diabetes, we can look for causation by doing a hypothesis test.
  - The hypothesis test yielded a p-value of 4.8e-27 indicating causality.
    - Reverse causality: having diabetes would lead to less likelihood of having a tattoo/cosmetic procedure
      - Patients with diabetes have a higher risk of NOT healing properly after having a tattoo or cosmetic surgery
    - Simply spurious correlation

Selecting features – Variables removed

The following variables were removed in each step:

- **Step 4 – Low variance**
  - Temperature (Already removed in step 1)

Selecting features – Variables removed

The following variables were removed in each step:

- **Step 5 – Privacy & fairness concerns**
  - GP – Privacy
    - No strong reason why your GP (Doctor) should have an impact
    - Supported by findings that most GP's have similar diabetes/patient rate
    - Wouldn't be relevant for our business case anyway
  - Occupation – Privacy
    - Target demographic for the product is children in school ages – no occupation

Selecting features – Variables selected

- After applying our 5 step procedure we end up with the following subset of features:
  - Alopecia_yes
  - Gender_male
  - Irritability_yes
  - Partial Paresis_yes
  - Polyphagia_yes
  - Sudden Weight Loss_yes
  - Urination_high
  - Visual Blurring_yes
  - Weakness_yes
  - Weight

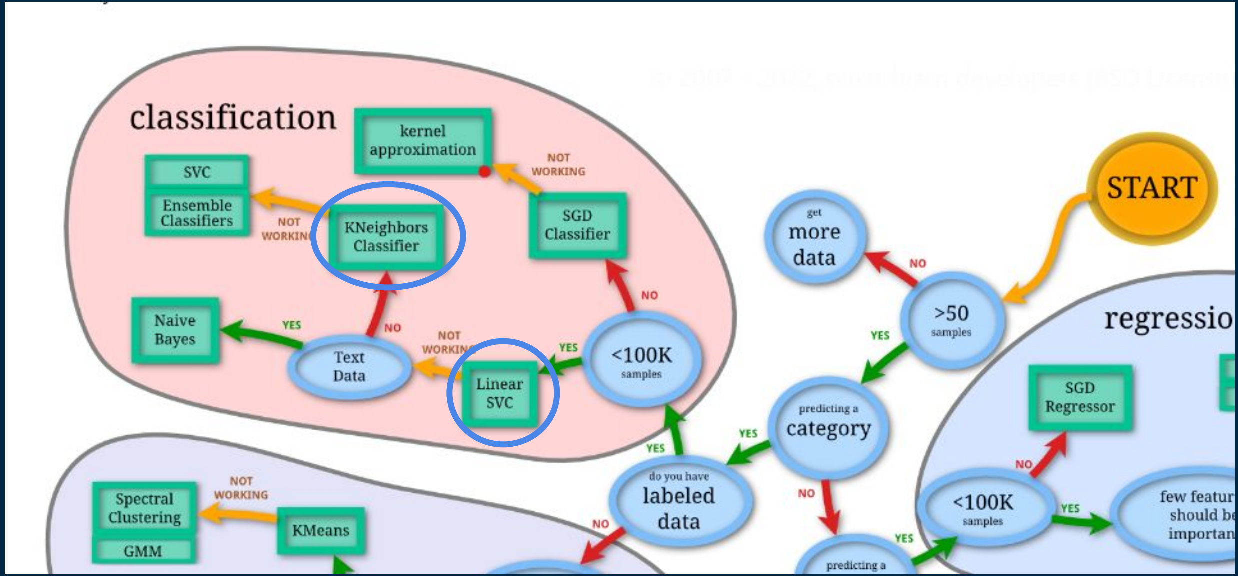
- 10 features
  - 9 categorical – binary
  - 1 numeric

04

Classification & evaluation

Choice of classifier

- Few features and limited data
  - Performance and explainability are two factors that contribute to the decision.
  - Suitable classifiers: support-vector-machine, a classification tree or k-nearest-neighbors.
  - A neural network might give higher performance, but is difficult to interpret, and needs a lot of data / many features.



Source:
https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Choice of classifier

We chose the following 5 classifiers:

1. Linear Support Vector Machine
2. Multi-Layer Perceptron (2 layers, small)
3. Multi-Layer Perceptron (4 layers, large)
4. k-Nearest Neighbors, k=30
5. Decision Tree Classifier

Evaluating our classifier - Performance

- Performance
  - Metric for performance: F_β-score
    - Weighting precision & recall
    - F_2-score suitable as recall is more important than precision
      - Specifically, 2x
    - Punish False Negatives harder
      - We prefer sending to many people to the doctor, to not detecting real cases
- Train-test split
  - We use 80% of data for training, and reserve 20% for testing
- Baseline: predict randomly based on target: 63% accurate

Initial performance

Classifier	Accuracy	Precision	Recall	F1-score	**F2-score**
Linear SVM	0.876	0.877	0.950	0.912	**0.934**

Initial performance

Classifier	Accuracy	Precision	Recall	F1-score	**F2-score**
Linear SVM	0.876	0.877	0.950	0.912	**0.934**
MLP, small	0.933	0.922	0.983	0.952	**0.970**
MLP, large	0.989	1.000	0.983	0.992	**0.987**
kNN, k=30	0.876	0.877	0.950	0.912	**0.934**
Tree	0.910	0.894	0.983	0.937	**0.964**

Initial performance

Classifier	Accuracy	Precision	Recall	F1-score	**F2-score**
Linear SVM	0.876	0.877	0.950	0.912	**0.934**
MLP, small	0.933	0.922	0.983	0.952	**0.970**
MLP, large	0.989	1.000	0.983	0.992	**0.987**
kNN, k=30	0.876	0.877	0.950	0.912	**0.934**
Tree	0.910	0.894	0.983	0.937	**0.964**

Initial performance – features chosen

Classifier	F2-score	# feat	Features
Linear SVM	0.934	3	Gender_male, Irritability, Urination_high
MLP, small	0.970	5	Gender_male, Irritability, Partial Paresis, Sudden Weight Loss, Urination_high
MLP, large	0.987	7	Alopecia, Gender_male, Irritability, Partial Paresis, Polyphagia, Sudden Weight Loss, Visual Blurring
kNN, k=30	0.934	3	Gender_male, Irritability, Urination_high
Tree	0.964	10*	Alopecia, Gender_male, Irritability, Partial Paresis, Polyphagia, Sudden Weight Loss, Visual Blurring, Weakness, Weight, Urination_high

*) all features were used

Initial performance – features chosen

Classifier	F2-score	# feat	Features
Linear SVM	0.934	3	Gender_male, Irritability, Urination_high
MLP, small	0.970	5	Gender_male, Irritability, Partial Paresis, Sudden Weight Loss, Urination_high
MLP, large	0.987	7	Alopecia, Gender_male, Irritability, Partial Paresis, Polyphagia, Sudden Weight Loss, Visual Blurring
kNN, k=30	0.934	3	Gender_male, Irritability, Urination_high
Tree	0.964	10*	Alopecia, Gender_male, Irritability, Partial Paresis, Polyphagia, Sudden Weight Loss, Visual Blurring, Weakness, Weight, Urination_high

*) all features were used

Initial performance – coefficients (SVM)

```
SVC(kernel='linear')
[[-1.99902344  1.99902344  2.        ]]
['Gender_male', 'Irritability_yes', 'Urination_high']
```

- This means:
  - Women: if you are either irritable or have high urination levels: see a doctor
  - Men: if you have high urination: see a doctor
    - Else, no need

Evaluating our classifier – Fairness

- We have four variables that can give fairness issues
  - Race
  - Gender
  - GP
  - Occupation
- We will only look at gender fairness
- Rate of FN and FP should be similar for both genders
  - In our case we're more sensitive to FN
- This is evaluated on the test set

Evaluating our classifier – Fairness

Classifier	FN Female	FN Male	Rate of FN (female)
Linear SVM	0	3	0%
MLP, small	0	1	0%
MLP, large	2	1	66.6%
kNN, k=30	2	1	66.6%
Tree	0	1	0%

Anonymization

- We previously removed some features for privacy purposes
- We now further anonymize the dataset using differential privacy
  - Benefits: robust to linkage attacks
  - Toss a coin, 50-50: keep response or generate random entry
  - This gives ln(3)-differentially private data
- Centralized model of differential privacy: we handle the anonymization
  - In a future study, you could do it decentralized (data is anonymized on reporting)
  - This is more sensitive to the amount of data, however
- We use the best feature set from each model, and re-run the experiment after anonymizing the data

Evaluating our classifier – Anonymization

Classifier	Original F2-score	New F2-score	New recall
Linear SVM	0.934	0.912	
MLP, small	0.970	0.788	
MLP, large	0.987	0.724	
kNN, k=30	0.934	0.702	
Tree	0.964	0.707	

Evaluating our classifier – Anonymization

Classifier	Original F2-score	New F2-score	New recall
Linear SVM	0.934	0.912	
MLP, small	0.970	0.788	
MLP, large	0.987	0.724	
kNN, k=30	0.934	0.702	
Tree	0.964	0.707	

Evaluating our classifier – Anonymization

Classifier	Original F2-score	New F2-score	New recall
Linear SVM	0.934	0.912	1.000
MLP, small	0.970	0.788	0.767
MLP, large	0.987	0.724	0.783
kNN, k=30	0.934	0.702	0.883
Tree	0.964	0.707	0.633

Evaluating our classifier - Conclusion

- Considering anonymization, the best classifier seems to be the Linear SVM, which was durable to losing data via the anonymization process
  - Linear SVM does predict all diabetes cases, however
  - The features set it uses is just 3: Gender_male, Irritability, Urination_high
    - This does seem low, but we did see in data analysis that urination had a massive impact
  - The others are known to perform well with more data, so future data collection should still consider a few alternatives
- Our recommendation for future data collection is to collect the 10 features we had before the start of classification (slide 41)
- Even after anonymization our best classifier is ~0.9 F2-score
  - Again, note the recall issues

Example of use - web form

Diabetes screening

This form will be used to perform a simple screening for diabetes, and recommend at-risk persons to seek out a proper diagnostic test.

[redacted]

*Må fylles ut

What is your gender? *

- Male
- Female

In the last week, have you been feeling signs of increased irritability or rage? *

- Yes
- No

In the past week, have you been drinking more than 2.5 liters of fluids per day? *

- Yes
- No

Send

Tøm skjemaet