IN-STK5000
Project 1

Fall 2023

Helene Bøsei Olsen, Sander Finnset Ørnes,
Even Tronstad, Andreas Christian Poole

Scenario - Background

The client - Public health authorities

- **Business objective**: “Improve the health and well-being of all people in the population”*
- Diagnosing and treatment handled by Private health service

The Diabetes epidemic

- Undetected and untreated diabetes is a major health issue
- Direct negative impact on the Business Objective

Funding

- Limited funds, need a cheap scalable solution
- Public funded mass testing too expensive

Goals

- Increase awareness of diabetes
- Increase testing of high risk individuals
  - Especially in parts of population with low test rates
- Reduce undetected and untreated diabetes

*: Inspired by US health authority mission statement,
https://www.hhs.gov/about/strategic-plan/2022-2026/introduction/index.html

Scenario – Why undetected?

High marginal cost of a Doctors visit:

- Highly skilled personnel with high hourly rates
- Doctors Business unscalable
- Can only assess / treat one patient at the time

High threshold for Doctors visit:

- Many patients in low income demographics hesitant to spend money on doctors visits
- A negative test is a waste of money for many people

Helping people identify themselves as high risk of having diabetes might push to seek testing:

- Easier to budget a visit if there is a high chance it will be worth it

Scenario – Our proposal

- Machine Learning system for self evaluation of Diabetes risk
  - Based on users own input
  - Data-Driven decision: Instant advice on booking a Doctors visit

- Make accessible on a website
  - Highly scalable ~0 marginal cost, accelerate number of detected cases
  - Input*: Symptoms / risk factors, generic personal data**

- Supported by marketing campaign
  - Focus on Quality of life improvement from treatment to incentivise use of website
  - Possibility of network effects by «word of mouth»

- Target adults
  - Children out of scope, at least for now
  - Min age set to 16

*: Our further analysis will identify what input is needed, **: Gender, Age, height, weight

The Data

Small dataset covering 546 individuals

The target:
    Diabetes

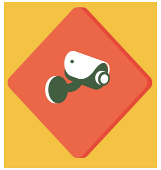
24 features:
- Personal: Age, Gender, Race, Occupation, Weight, Height, General Practitioner
- Medical: Risk factors and symptoms

How can this type of data be collected?
    Routine/ diabetes specific checkups at doctors office.

    Surveys or questionnaires

- Requires informed consent from participants

Data Hazards



Risk to privacy:

- Combination of doctor, age, occupation, height, race, and gender enough to identify individuals.

Reinforces existing biases:

- Everyone has the right to be treated equally
- Who is represented in the data.
- Need to collect more representative data

Ranks or classified people

- Predictions are based on user-input - not related to external information about user.
- Clear explanations of the models recommendation
- Clear information about the models limitations.

Dangers of Misuse:

- Assumptions about correlations
- Include domain experts

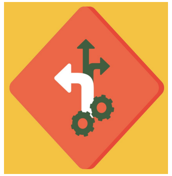
Data Hazards

Difficult to understand

- Interpretable and explainable
- Well documented and open source code

Lacks informed consent

- Assume explicit and informed consent in accordance with GDPR
- Important for future data collection

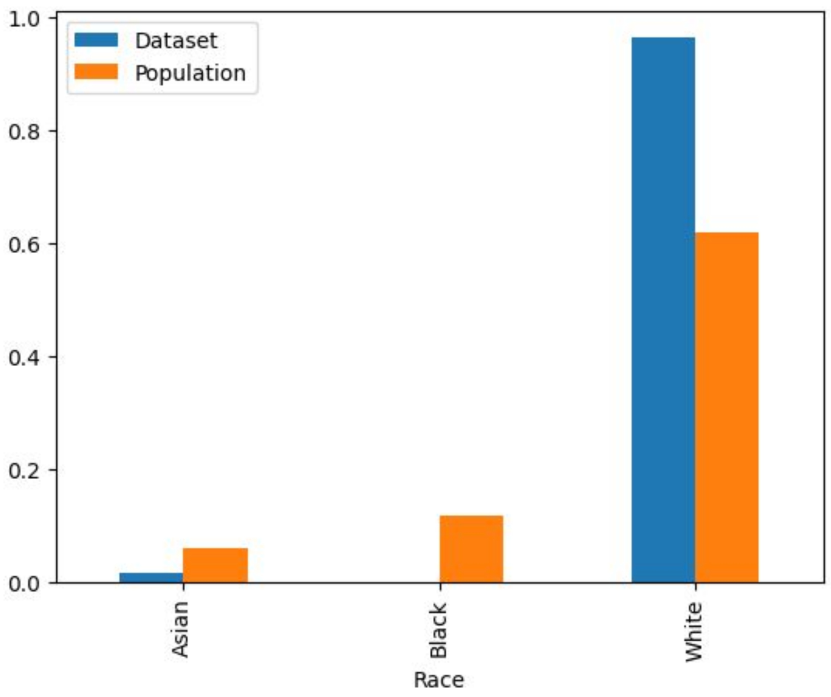


Automate decision making

- Creating decisions instead of replacing decisions
- Final prognosis always determined by a doctor
- Risk of false negative result

Data set - Challenges

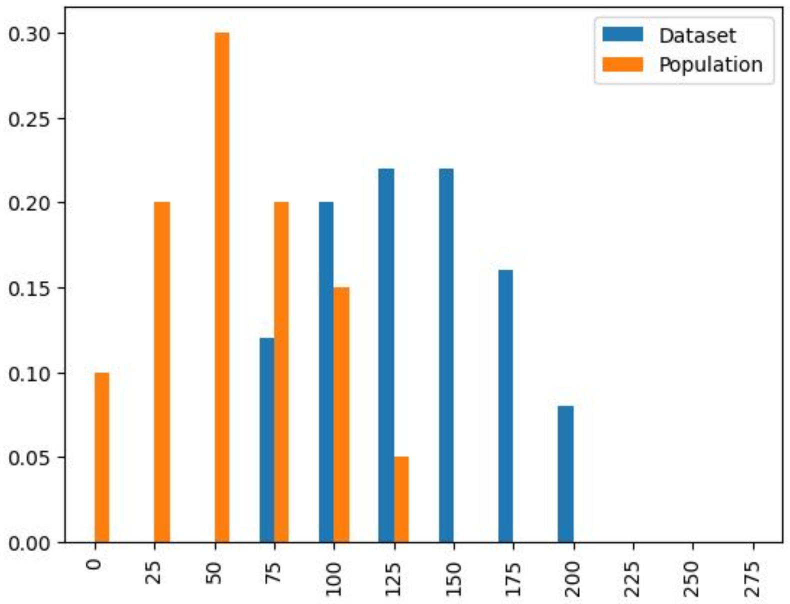
- Skewed heavily towards the **white demographic,** does not accurately represent the broader population



https://en.wikipedia.org/wiki/United_States. Population numbers does not sum to 1, as more labels are used in wiki stats

Data set - Challenges

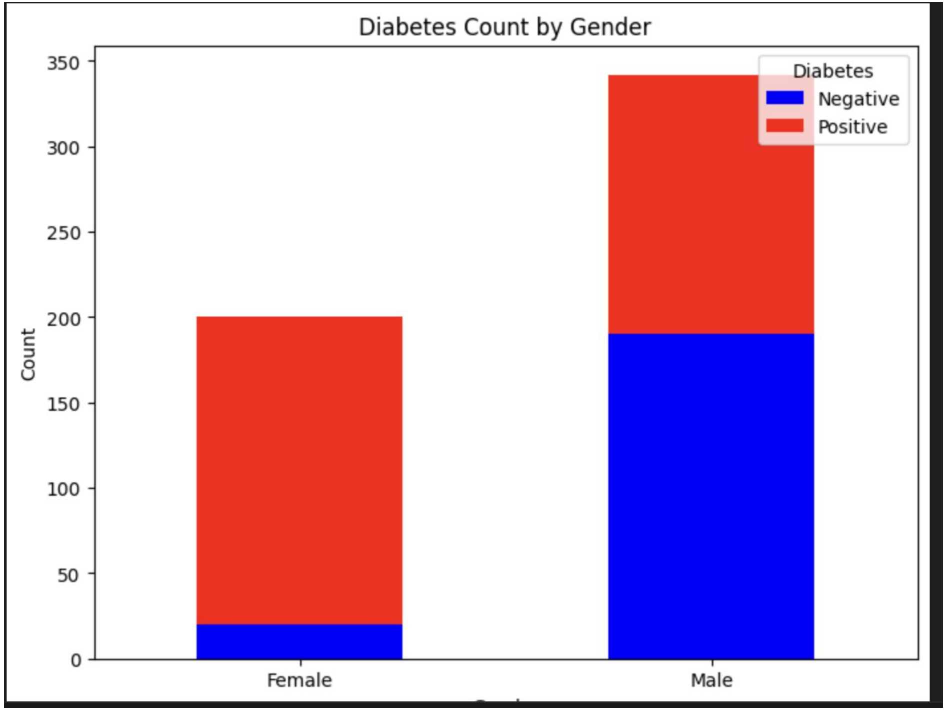
- Skewed towards individuals with **high income**



*: Income, kUSD, estimated from average salaries of occupations.
These numbers are made up and not based on real income stats.

Data set - Challenges

- The dataset is significantly skewed toward **male**
- **Over-representation of diabetes** compared to general public.
- Almost all female in data set have diabetes



50-50 in the population is assumed.

Dealing with bias and data cleaning

- Remove female
  - Due to inaccurate diabetes representation
- Remove non white
  - To few, better to limit scope than reinforce bias

- Normalise input:
  - Casing (Yes/No -> yes/no)
  - Metrics (m -> cm)
- Remove 26 duplicates

Split the data into train and test sets (80/20)

- The following analysis is solely performed on the train set.

Outliers

Univariate

- Establish rules for what is an outlier
    - Upper and lower limits
    - Outlier if < lower or > upper
- By domain knowledge:
    - Sensible min, max values
        - Age, Height, weight
    - Children deemed out of scope, min age set to 16
    - Other real instances outside interval might be possible, but more likely errors
- By statistical tools
    - IQR Score
        - Scale: 1.5
    - Z-score
        - Abs. value exceeding 3
    - We choose conservative values
        - Min. for lower, Max for upper
        - If lower negative, set to zero

Boundaries for univar. outliers:

	Lower	Upper
Age	16.0	120.00
Height	110.0	240.00
Weight	30.0	200.00
Temperature	36.4	37.61
Urination	0.0	4.93

Note! This analysis is done on the training set.

Univariate outliers

Initial look at min / max values:

	Age	Height	Weight	Temperature	Urination
min	-22.0	154.01	46.11	36.46	0.83
max	377.0	194.24	125.95	37.44	12.00

Height, weight and Temperature, within bounds.
Age and Urination needs further investigation

Sources:

- Negative values for numerics: Mistake in collection
- Age >= 150: Mistake in collection
- High natural values for urination might be possible, but > 10l seems very unlikely

How they might occurred:

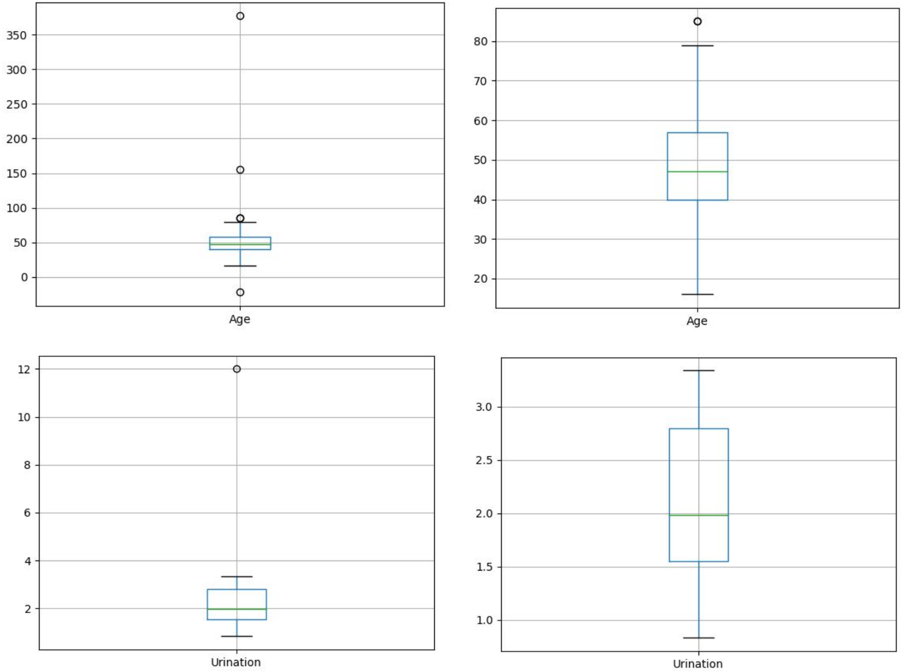
- Typos, digitisation/scanning errors

Handling:

- Age between 0 to 16
    - Delete sample
- Other
    - Assume error
    - Replace with missing value
    - Handled further by process for missing data

Box plots for Age and Urination:

Before handling: After handling:



After handling univariate outliers:

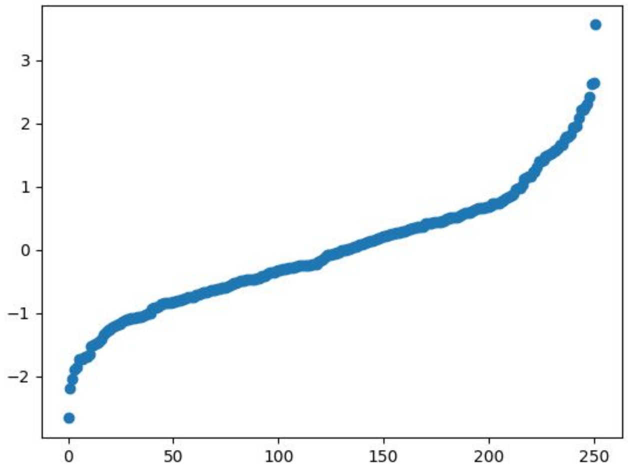
	Age	Height	Weight	Temperature	Urination
min	16.0	154.01	46.11	36.46	0.83
max	85.0	194.24	125.95	37.44	3.34

Multivariate outliers

- Z score for Euclidean distance to mean on standardized values
- Outlier if > 3

Multivariate outliers:

	Age	Gender	Race	Occupation	Height	Weight	Urination	Temperature
242	58.0	Male	White	Retired	192.74	125.95	2.79	37.1



A very tall and heavy person

Would be an combined outlier if height and weight was uncorrelated

But deemed to not be an outlier due to high correlation between height and weight

To account for this correlation we increase limit to 4

Handle:
Delete sample (no cases in our data set)

Missing Data

Extent:

1.3% of the data cells are missing

Unsystematic, randomly distributed.

- Evenly distributed among instances:
  - 23.4% of rows missing one value
  - 4.2% of rows missing two or more values
- Most affected: Sudden Weight Loss, Partial Paresis, Urination, Muscle Stiffness, Age, Occupation.

Handling:

Derive from other features:

- Obesity calculated from height and weight using BMI formula. Threshold: 30
- Polydipsia from Urination. Threshold: 2.5

Impute with best non assuming guess:

- Binary features set to false - assume Missing Not At Random
- Numerical features set to mean of training set

Note! This analysis is done on the complete data set.

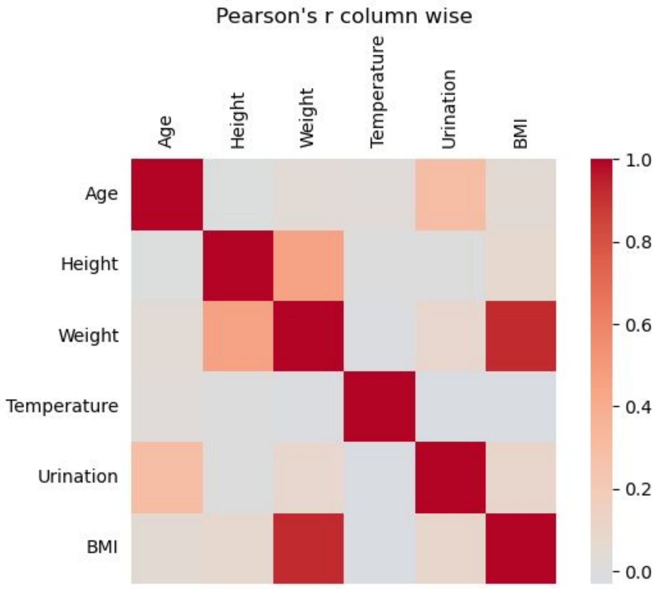
	Count	Percentage
Sudden Weight Loss	16	2.93
Partial Paresis	15	2.75
Urination	14	2.56
Muscle Stiffness	12	2.20
Age	12	2.20
Occupation	12	2.20
Alopecia	10	1.83
Race	9	1.65
Height	8	1.47
Obesity	8	1.47
Genital Thrush	8	1.47
Delayed Healing	8	1.47
GP	7	1.28
Polydipsia	6	1.10
Itching	6	1.10
Weakness	5	0.92
Weight	5	0.92
Irritability	5	0.92
Gender	4	0.73
Visual Blurring	4	0.73
Polyphagia	2	0.37
TCep	0	0.00
Temperature	0	0.00
Diabetes	0	0.00

Correlations - Numeric vs. Numeric

Pearson’s r

Interesting correlation:

- Urination - age
- Temperature and nothing



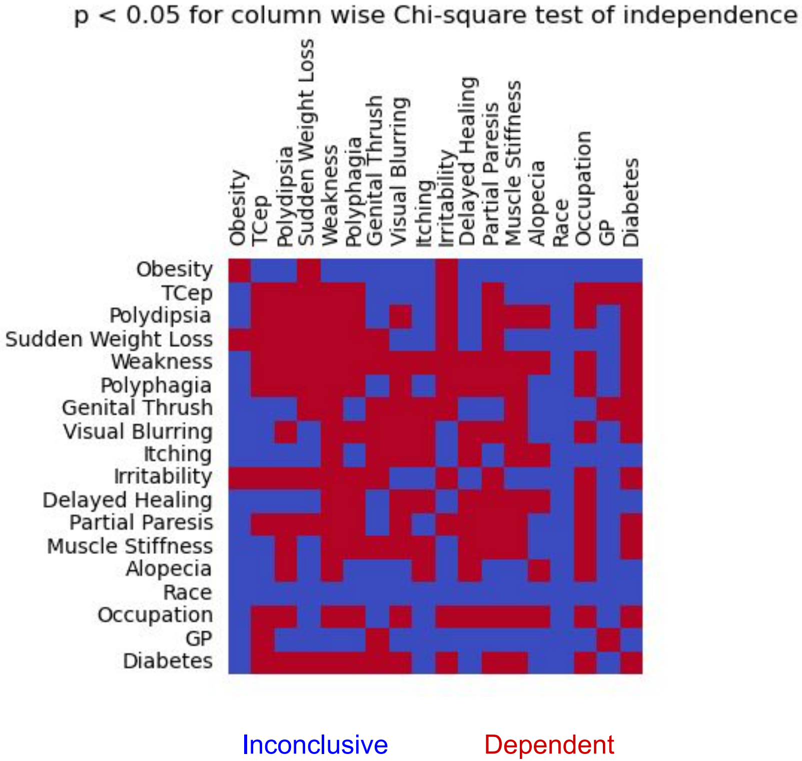
Correlation measures are performed after handling outliers and missing values, as correlation measures can be influenced by both of these factors.

Note! BMI was added as a new feature

Correlations - Independence between categorical features

Interesting correlations:

- Diabetes - TCep (must assume spurious)
- Diabetes dependent on most features
- In general, lots of interdependence of all features

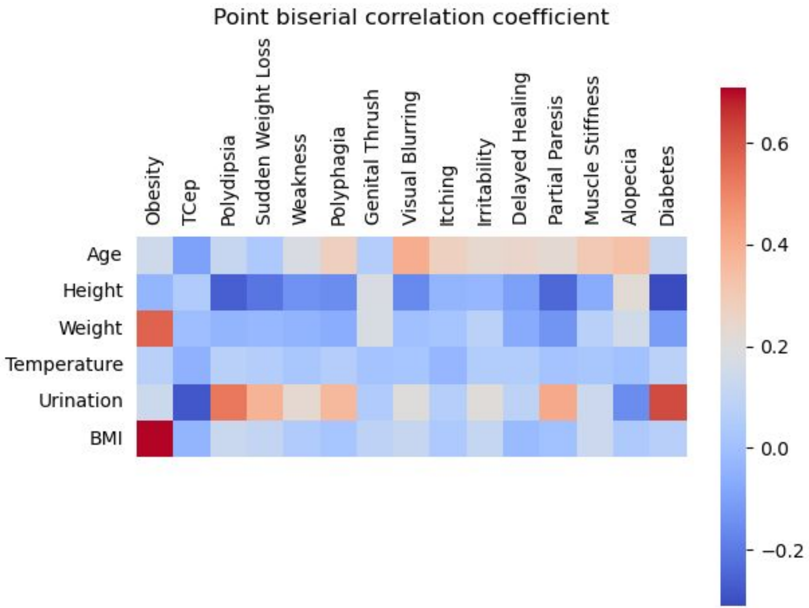


Note: Blue entries do not imply independence, but that the hypothesis test failed to detect dependence.

Correlations - numeric and categorical

Interesting correlations:

- Urination and diabetes
- Age is weakly correlated with many symptoms
- BMI and diabetes are not correlated



Point biserial coefficient is a special case of Pearson's r

Correlations - Summary

Medical correlations:

- Polydipsia (excessive thirst), Urination and Diabetes

Confounders:

- E.g. sudden weight loss, weakness and polyphagia are probably correlated since they might have the same cause (big caloric deficit), without one causing the other.

Difficult to determine causality:

- Age and weakness/visual blurring/alopecia/muscle stiffness/delayed healing/partial paresis
  - Possible explanation for why age is not included in the final features i.e. a confounder

Spurious correlation:

- TCep and Diabetes

Feature Selection

Low Variance:
- Temperature has low variance and therefore low predictive value

Correlated features:
- Urination - Polydipsia
  - We select one on the basis of ease of measuring for the end user.
  - Urination would require measurement over 24 hours, Polydipsia should be readily known
- Obesity - BMI - Weight
  - Information on Obesity and BMI contained in Height and Weight, but we test explicitly for correlation

Should not impact whether someone has diabetes:
- TCep (tattoos or cosmetic enhancing procedures)
- Occupation
- GP (General Practitioner)

Feature Selection

Correlated with the target:

- Some features have low correlation with diabetes
    - E.g. Obesity
- But we have chosen not to remove any features only based on correlation with diabetes
    - They may have non-linear predictive power in combination with other features

In summary we **drop** the following features:
Occupation, Temperature, Obesity, BMI, Urination, TCep, GP (Gender, Race were removed earlier)

Decision trees have automatic feature selection, further reducing the number of features used for prediction.

Summary of cleaned dataset

Who remains?

- White males who are
    - retired or have high-income jobs
    - and are predominantly above the age of 40

What remains?

- 252 individuals (of the original 546)
- 15 features (of the original 24)
    - 3 numerical
    - 12 categorical

Remaining features:

- Age
- Height
- Weight
- Polydipsia
- Sudden Weight Loss
- Weakness
- Polyphagia
- Genital Thrush
- Visual Blurring
- Itching
- Irritability
- Delayed Healing
- Partial Paresis
- Muscle Stiffness
- Alopecia

Classification - Pruned Tree

Explainable:

Supports Categorical, Binary and Numerical features (although we use one-hot encoding)

Unbiased and non-linear

Automatic feature Selection is a bonus, relieving workload earlier in the pipeline

Portable - can be deployed without access to original data

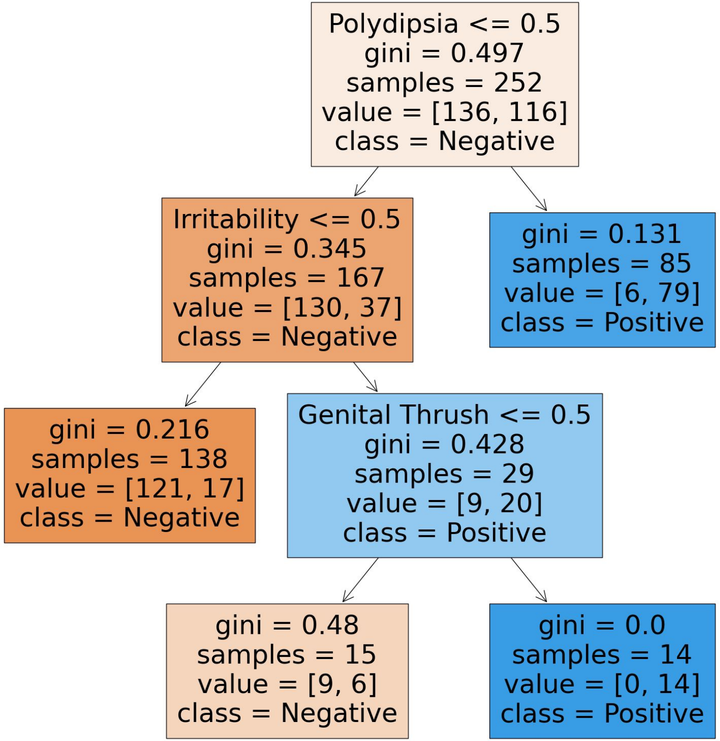
Computationally cheap to train and predict

Commonly used in medical settings

Final model:

- Pruned decision tree with 3 features
- Explainable:
- All features are symptoms of diabetes
- Matches business case, public health agencies need explainability
- Probably generalizes well since all used features are symptoms

Accuracy on test set: 89%



Complexity parameter: alpha = 0.02

Adaptivity and Online Learning

Practical challenges:

We need to:

- Handle outliers and missing data on the fly
- Possibly redefine outlier criterions, e.g. means and standard deviations
- Perform preemptive filtering to avoid input errors.

Outlier detection made robust:

- In general, our model needs to tackle all kinds of unseen outliers

Practical advantages:

We don't need to adaptively update the model:

- Cheap training and small dataset - rerun the whole pipeline!
- Concept drift is likely not a big issue given a good dataset
    - Symptoms remain unchanged
    - However, for our dataset there could be concept drift

Is online learning a good idea?

Challenges for our context and current situation:

- **True labels are only available after proper testing**
    - Will only get access to true positive and false positive labels.
- **Limited adaptability:**
    - The current training data lacks diversity, the system may not effectively adapt to users that are underrepresented in the training data.
    - Great risk of reinforcing biases
- **Difficult to tune the model**
    - Avoiding false positives might yield a high precision, but low recall.
    - E.g., a model which never predicts positives.

Initially:

- When campaign is rolled out we acquire lots of new data
- Can adapt to a wider demographic than the originally skewed data set

Summing up: Applying online learning seems difficult, and proper measures for alleviating the above challenges would need to be in place.