Diabetes Risk Factors Analysis With Feature Selections

Bertrand Flanet

Introduction

Health-related telephone **survey** collected annually by the CDC.

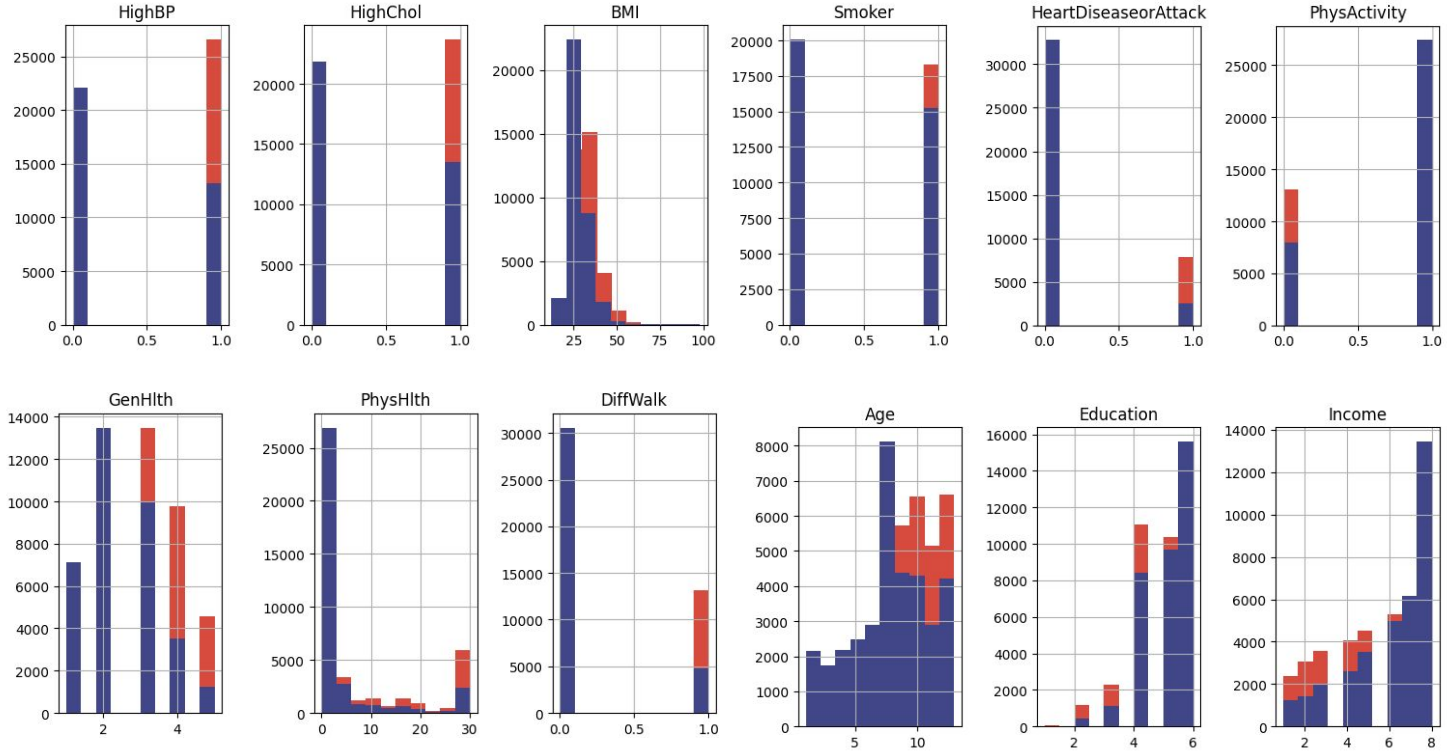
Dataset

- All records are from **2015**
- Original dataset: **441,455** entries
- Given dataset: **70,692** entries
- Individuals split between **two groups**:
  - **Negative** diabetes diagnosis
  - **Positive** diabetes or pre-diabetes diagnosis

Question

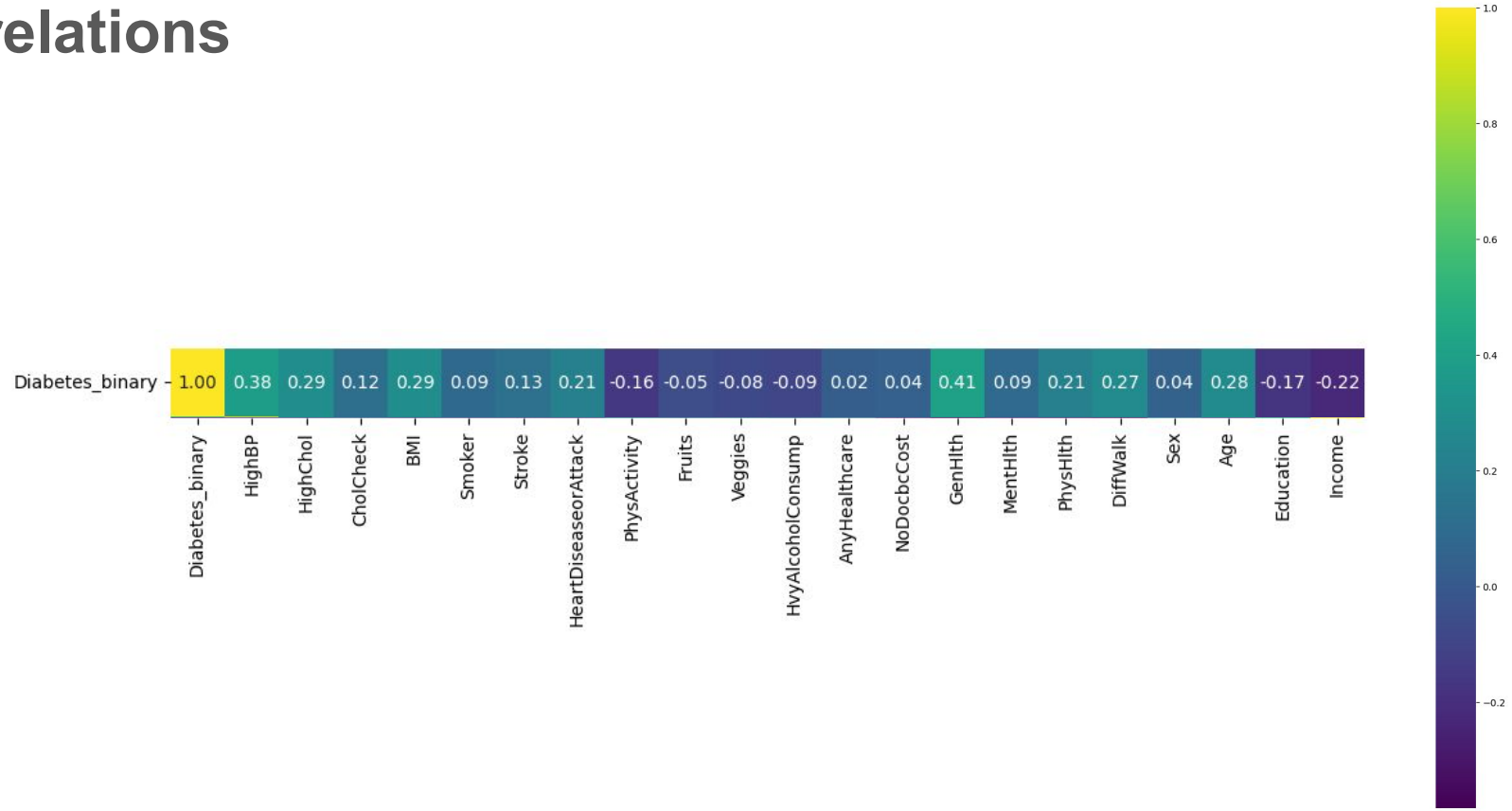
What risk factors are most predictive of diabetes?

Comparing Groups with Negative and Positive Diabetes diagnosis



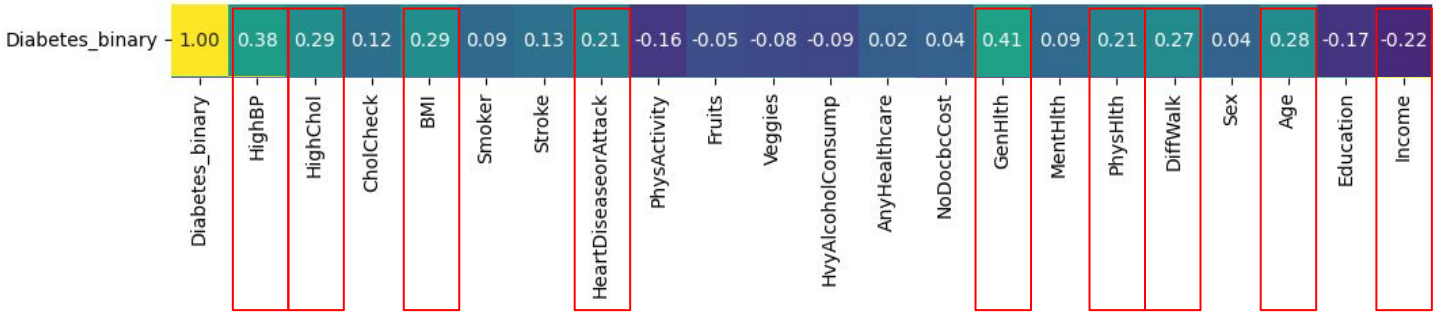
- HighBP
- HighCol
- BMI
- Smoker
- HeartDiseaseorAttack
- PhysActivity
- GenHlth
- PhysHlth
- DiffWalk
- Age
- Education
- Income

Correlations



Correlations

- Income
- Age
- DiffWalk
- PhysHlth
- GenHlth
- HeatDiseaseOrAttack
- BMI
- HighCol
- HighBP

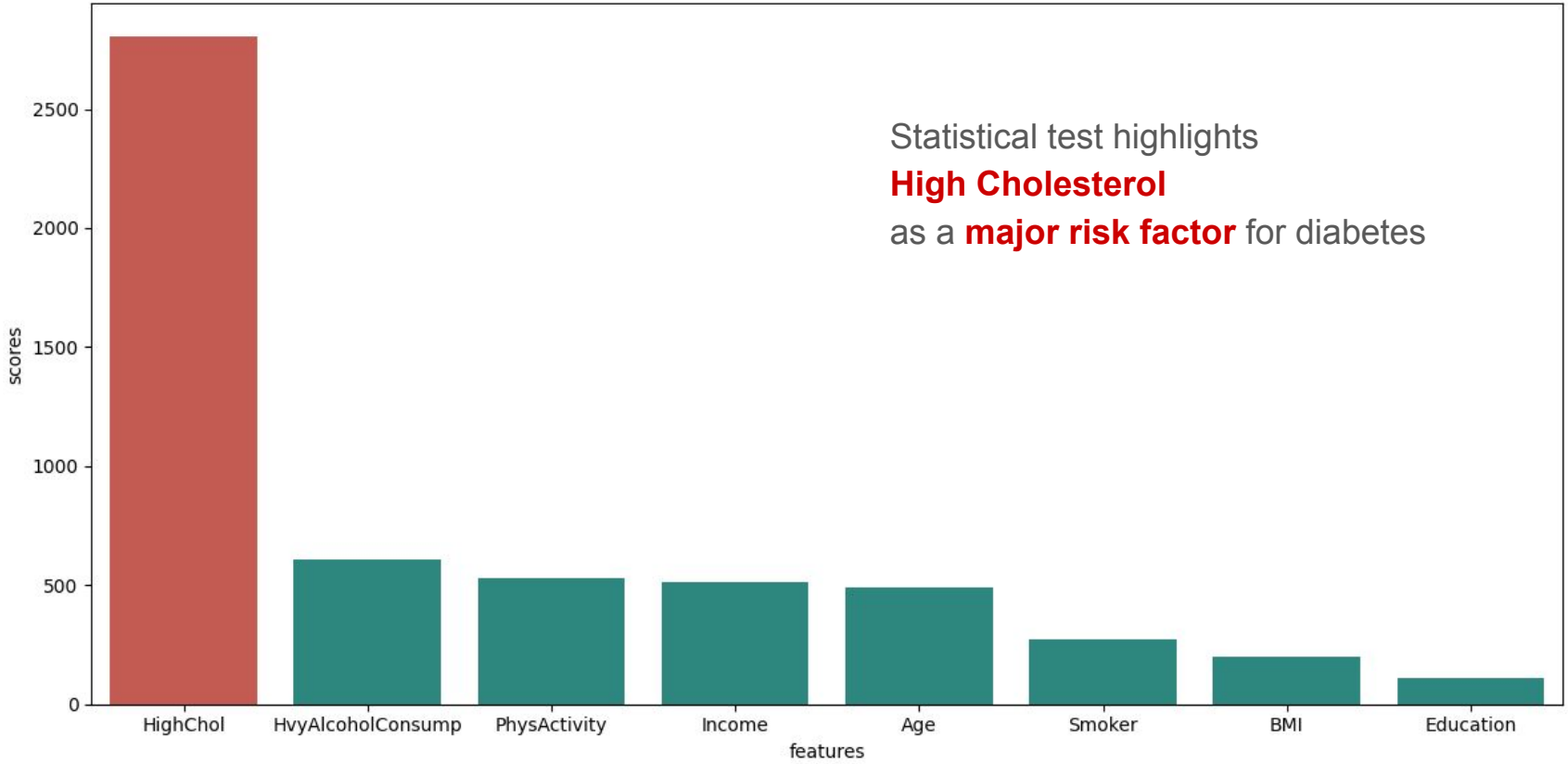


Limits and first observations

- **No strong correlation**, hence no feature that ‘wins it all’
- **Correlation** may **not indicate** a direct **cause**

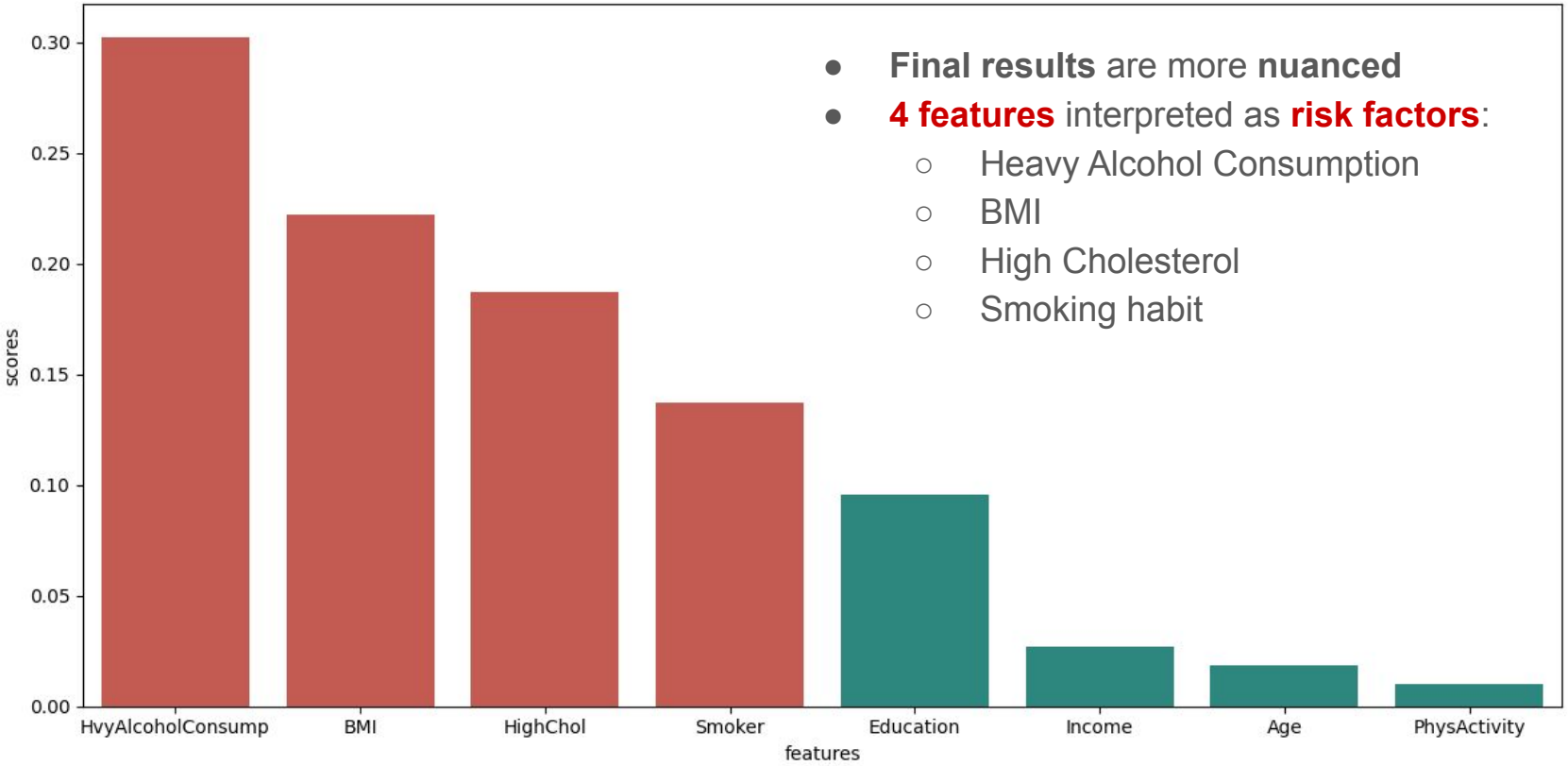
- **Categorizing features** in groups for better interpretability:
  - **Physiological Conditions**: **BMI, High Cholesterol**
  - **Socio-economic Conditions**: **Income, Age, DiffWalk**
  - **Symptoms:** **PhysHlth, GenHlth, HeatDiseaseOrAttack, HighBP**

Chi2 features selection



Statistical test highlights
High Cholesterol
as a **major risk factor** for diabetes

Random Forest Features Selection



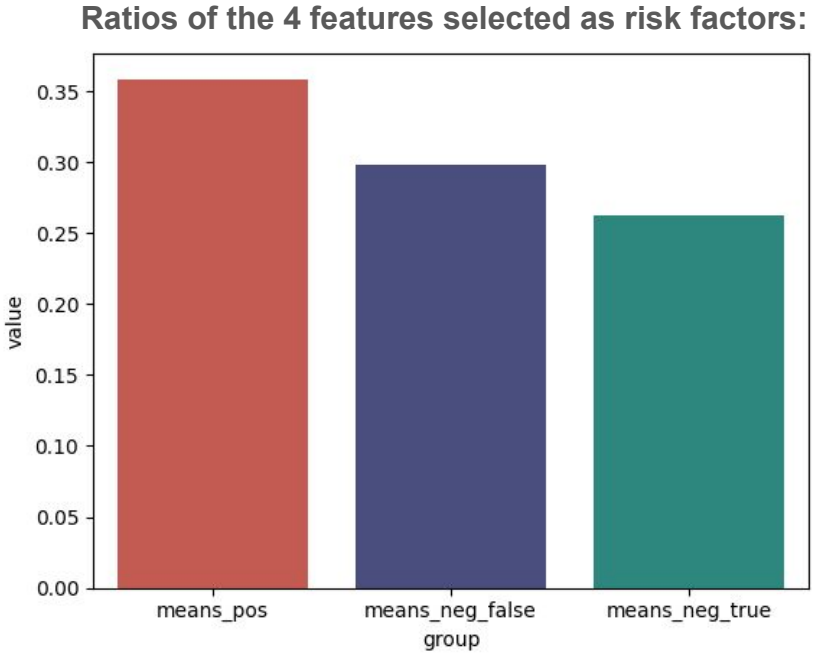
- **Final results** are more **nuanced**
- **4 features** interpreted as **risk factors**:
  - Heavy Alcohol Consumption
  - BMI
  - High Cholesterol
  - Smoking habit

Results Interpretation…

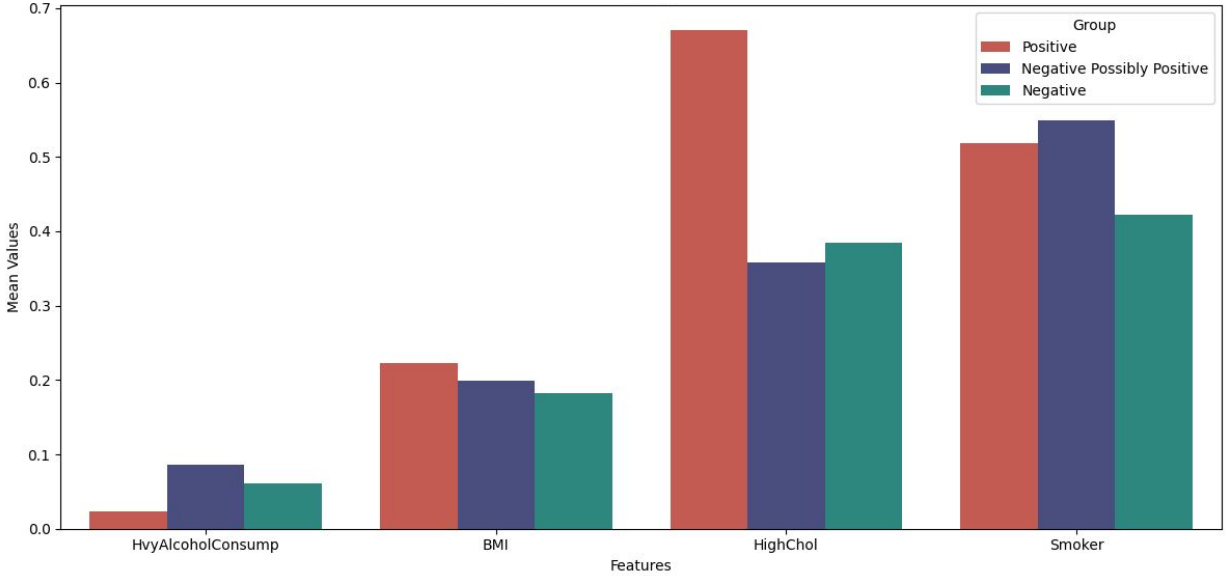
- No single features clearly standing out: could there be a **multifactorial aspects** to pathogenesis?
- Indication of **correlation** **may not mean** **causation**
- **3 out of 4** selected features fall into the **"Physiological Conditions" category**:
  - Possible connexion between individual health habits and diabetes diagnosis
    - Are these features **influenced by other factors**?
    - **Role** of these potential **factors** to prevent **poor** “**Physiological Conditions**”?

…and potential Group at Risk

- **Group** with **no** **health insurance**
- **No access** to **health care** due to **cost**



…and potential Group at Risk



- **Heavy Alcohol** and **Smoking** may indicate individual's **change of habit** once diagnosed with diabetes.
- **Cholesterol** may not have been thoroughly tested on the group potentialy at risk; or **sole high cholesterol isn't a risk factor**?

Conclusion and Refining Results

- Expand the analysis by **integrating more historical data**
- **Differences over time** might indicate **more accurate connexions** between **conditions** and **pathogenesis**
- Possible **multifactorial** aspect and considering **other factors** such as:
  - **Environmental conditions**
  - **Pollution**
  - **Genetic contribution**