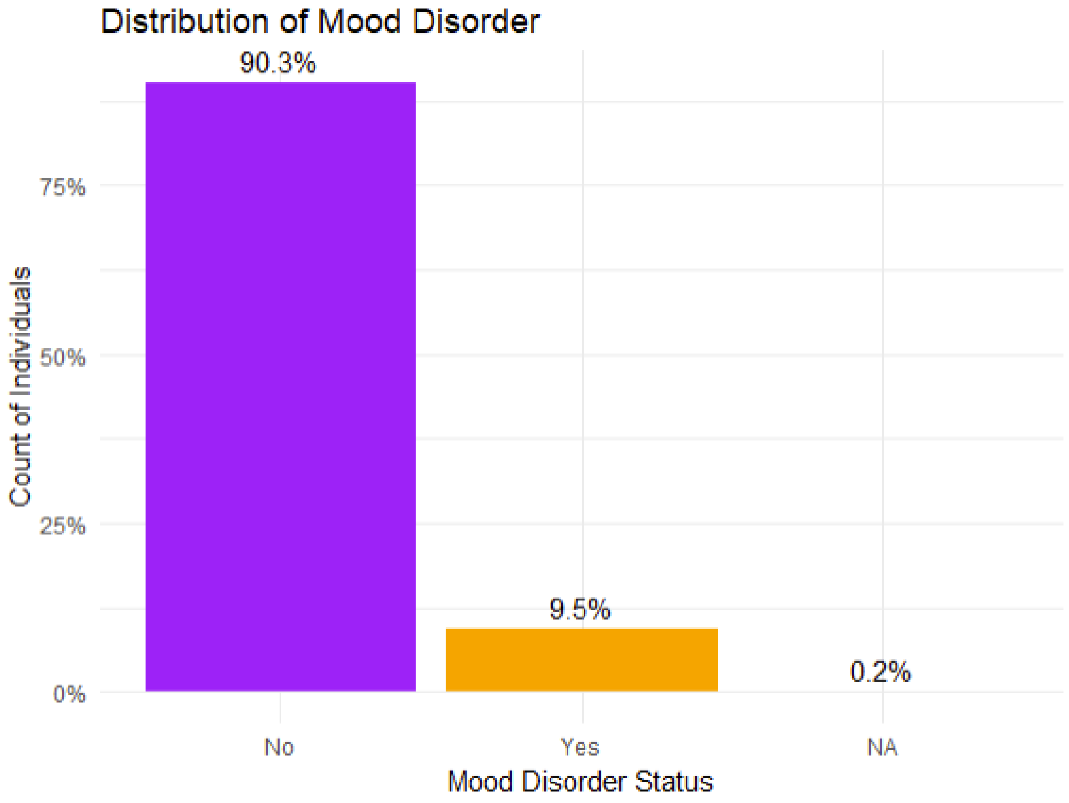
Mood Disorder Prediction Using Statistical Learning Techniques

Statistical Inference and Learning

Federica Andreazza

Goal: Predict mood disorder risk in the general population using survey data.

- Dataset:
  - Canadian Health Survey 2019–2020.
  - 108,252 observations and 50 variables.
  - 55,150 records, 37 predictors after cleaning.
- Target variable: **Mood_disorder**
  - It indicates whether a person has a mood disorder such as depression, bipolar disorder, mania or dysthymia.
  - **Acknowledged Class Imbalance**
  - **(90% 'No', 9.5% 'Yes')**
- Evaluation:
  - Standard metrics (Accuracy, Recall, Specificity, Sensitivity, AUC-ROC) on an 80/20 train-test split.



Dataset Overview

- Most previous studies use clinical data; this project uses large-scale public survey data for broader applicability.
- Sample Insights:
  - Wide range of demographic, socioeconomic, health, and lifestyle variables
  - Most respondents: married, post-secondary education, good/very good mental health
  - Mostly non-smokers, non-drug users, sufficiently active; alcohol use common
  - Majority are food secure and insured
- Data Preparation:
  - Ensured all variables were recoded to match documented categories (e.g., fruit/vegetable consumption, smoking status)
  - Removed variables with >30% missing data
  - Removed incomplete cases

What are the main risk factors for mood disorder in the general population?

Which statistical learning techniques best predict mood disorder risk?

Model Approaches

- Models Used:
  - Logistic Regression (full, stepwise, top 10/20 predictors)
  - Lasso Regression (cross-validation)
  - Quadratic Discriminant Analysis (QDA)
  - Naive Bayes (full, top 20 predictors)
  - K-Nearest Neighbors (cross-validation)
- Training and evaluation:
  - 80% training, 20% test split

Model Selection

- Stepwise Regression (forward, backward, bidirectional):
  - Best balance at 20 variables (lowest Cp)
  - 10 variables also a good parsimonious model
- Key Predictors:
  - Mental health status
  - Absence of anxiety disorder
  - Food insecurity
  - Gender, pain, stress, income

Model Evaluation Metrics

- Metrics Used:
  - **Accuracy**: closeness of a measured value to a standard or known value.
    - N.B.: alone can be misleading with imbalanced data!
  - **Precision**: closeness of two or more measurements to each other.
  - **Recall**: the proportion of all actual positives that were classified correctly as positives.
  - **AUC-ROC curve:** Measures the model's ability to distinguish between classes across all thresholds.
    - **Sensitivity (recall)**: true positive rate, quantifies how well a test identifies true positives.
    - **Specificity**: true negative rate, quantifies how well a test identifies true negatives.
  - **Confusion matrices.**

Model Performance

Model	Accuracy	Precision	Recall	AUC
Logistic (full)	0.9346	0.2729068	0.7861206	0.879
Logistic (20 coeff)	0.9328	0.2598513	0.7952218	0.874
Logistic (stepAIC)	0.9341	0.2680492	0.7940842	0.879
QDA (20 coeff)	0.8681	0.2082747	0.8418658	0.881
Nave Bayes (full)	0.8818	0.2401978	0.7736064	0.856
Naïve Bayes (20 coeff)	0.9137	0.2780136	0.745165	0.853
Lasso Regression	0.9348	0.314682	0.7485779	0.864
KNN (k=19)	0.9162	0.2338439	0.7121729	0.812

Key Predictors of Mood Disorder Risk

Top Risk Factors:

- Poor, fair, or good self-reported mental health (strongest predictors)
- Severe food insecurity
- Female gender
- Higher education level
- Heavy smoking

Protective Factors:

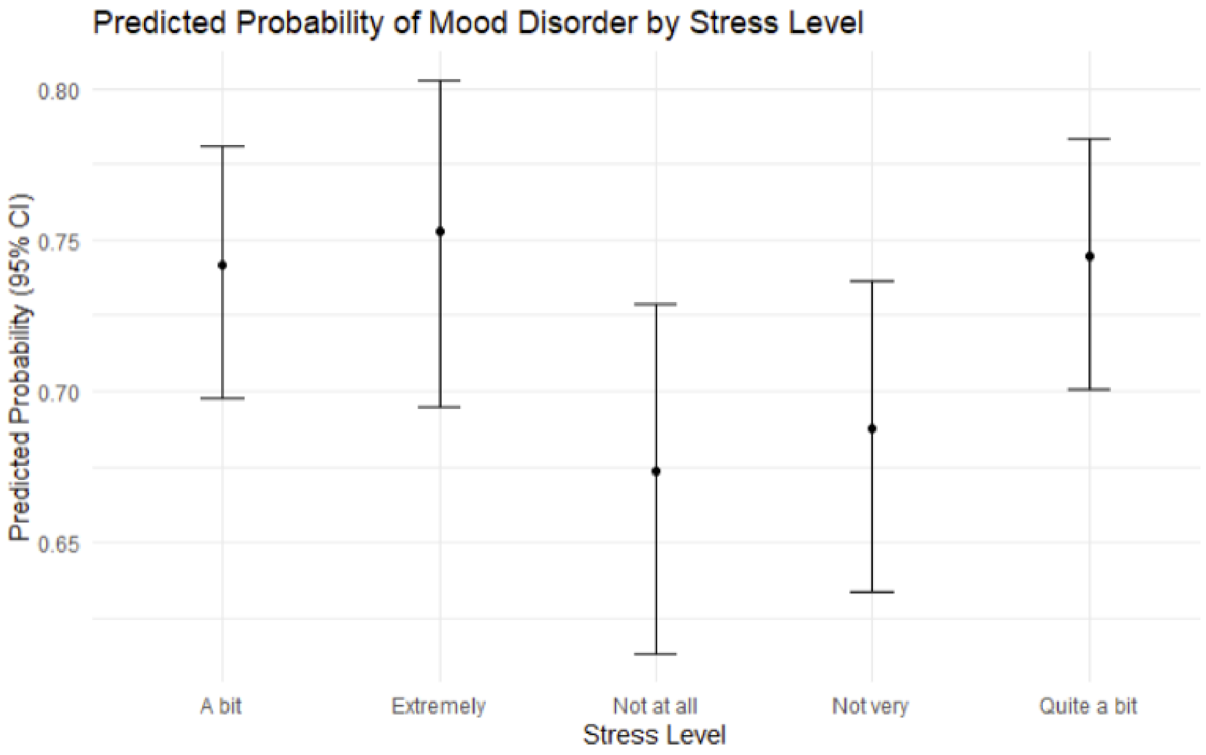
- Absence of anxiety disorder, sleep apnea, or fatigue syndrome
- Higher health utility index (HUI ≥ 0.8)
- Medical cannabis use
- Being foreign-born
- No chronic respiratory condition

Applying Our Model: Example Scenarios.

- **Scenario 1: Gender Differences**
  - Females with high stress & fair mental health: 69–80% predicted risk.
  - Males with similar characteristics: 59–72% predicted risk.
  - Insight: Gender significantly influences risk, even with similar stressors.

- **Scenario 2: Varying Risk Profiles**
  - Young, asthma & moderate stress: 39–54% predicted risk.

- **Scenario 3: Stress Level Comparison**
  - Higher stress increases predicted probability of mood disorder.



Conclusions

- Why This Analysis Matters:
  - Mood disorders have significant impacts on individuals' well-being and societal health.
  - Identifying key predictors is crucial for developing targeted prevention strategies and early intervention programs.
- Key Findings:
  - For the unbalanced dataset, **Logistic Regression** (full, top 20 predictors) emerged as the most suitable model.
  - **Strongest Risk Increase:** Self-reported "Poor," "Fair," and "Good" mental health states.
  - **Key Protective Factors:** Absence of anxiety disorder, sleep apnea, and fatigue syndrome.

Thank you for your attention!