Mathematical Underpinnings of ML

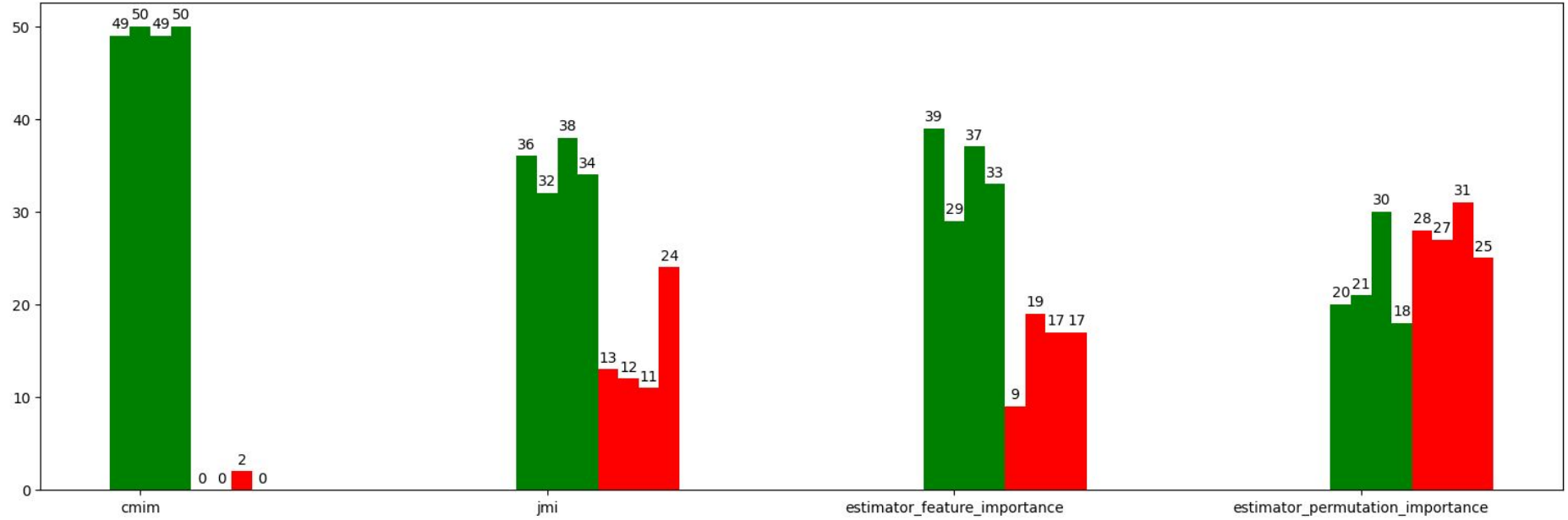
Feature Selection

Ivan Mialeshka

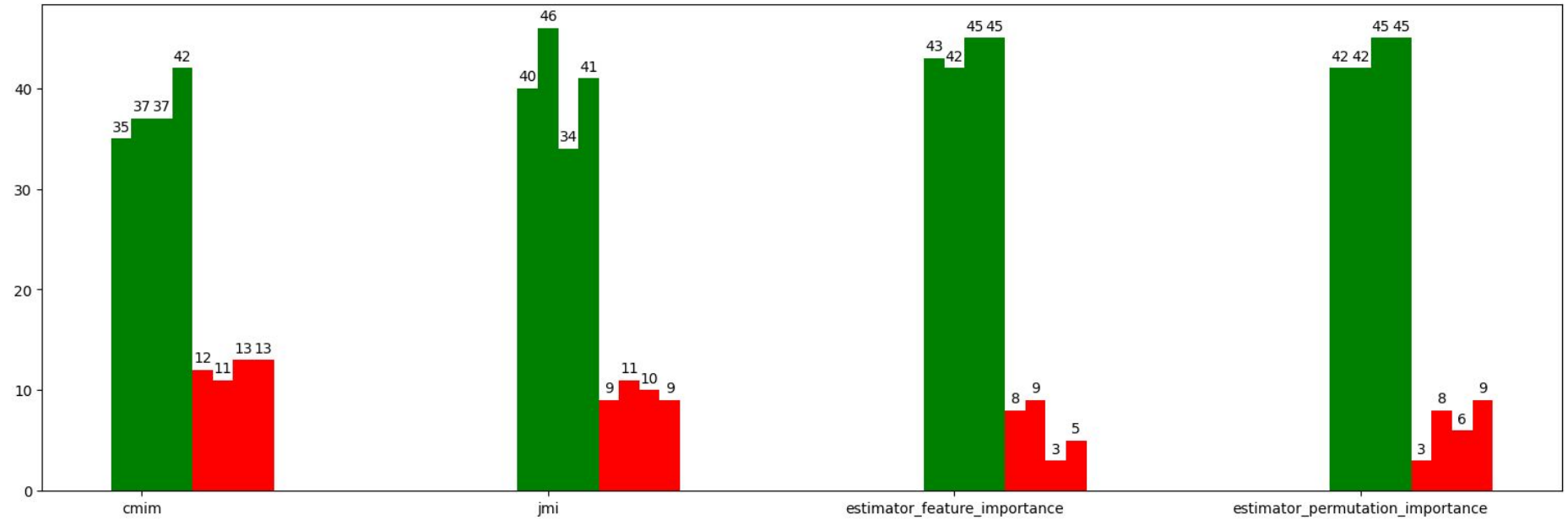
Methods

- JMI
- CMIM
- Random forest feature importance
- Random forest permutation importance

Artificial data (high sparsity)



Artificial data (low sparsity)



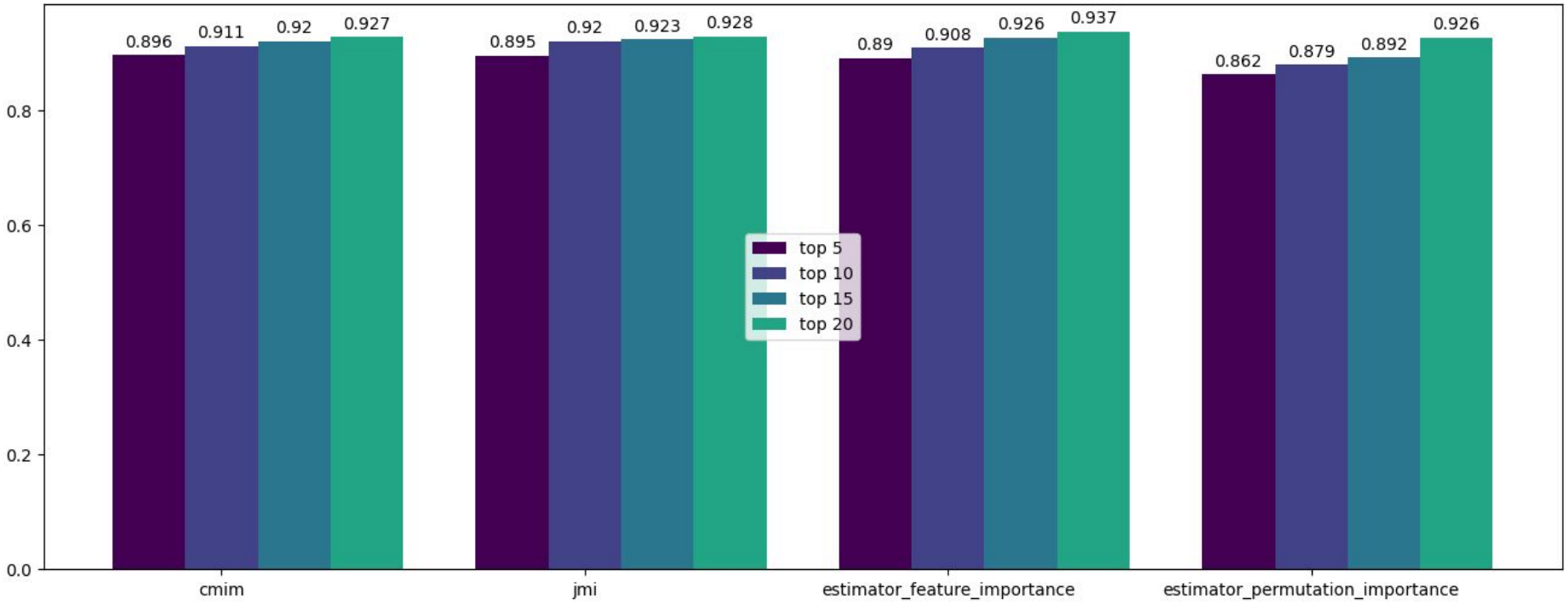
Real-world datasets

- **Musk (Version 2)**
  - features=166
  - samples=6598
- **Breast Cancer Wisconsin (Diagnostic)**
  - features=30
  - samples=569
- **Optical Recognition of Handwritten Digits**
  - features=64
  - samples=5620

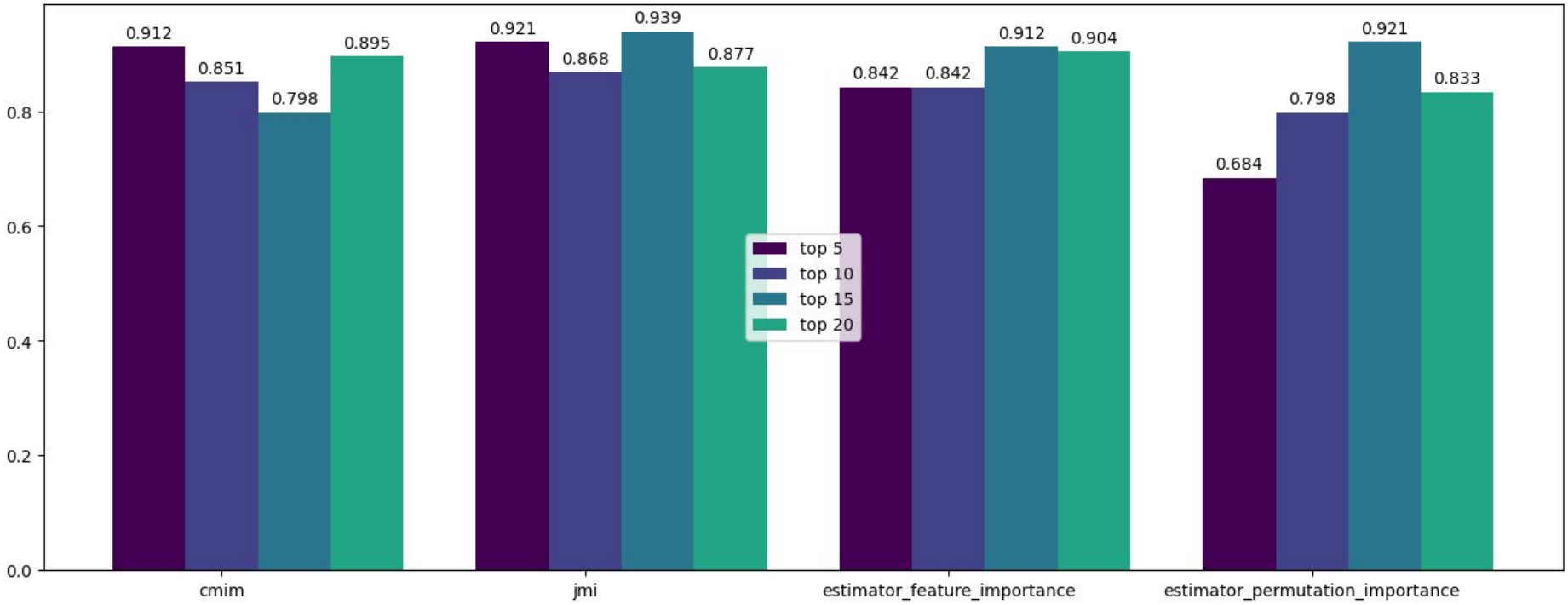
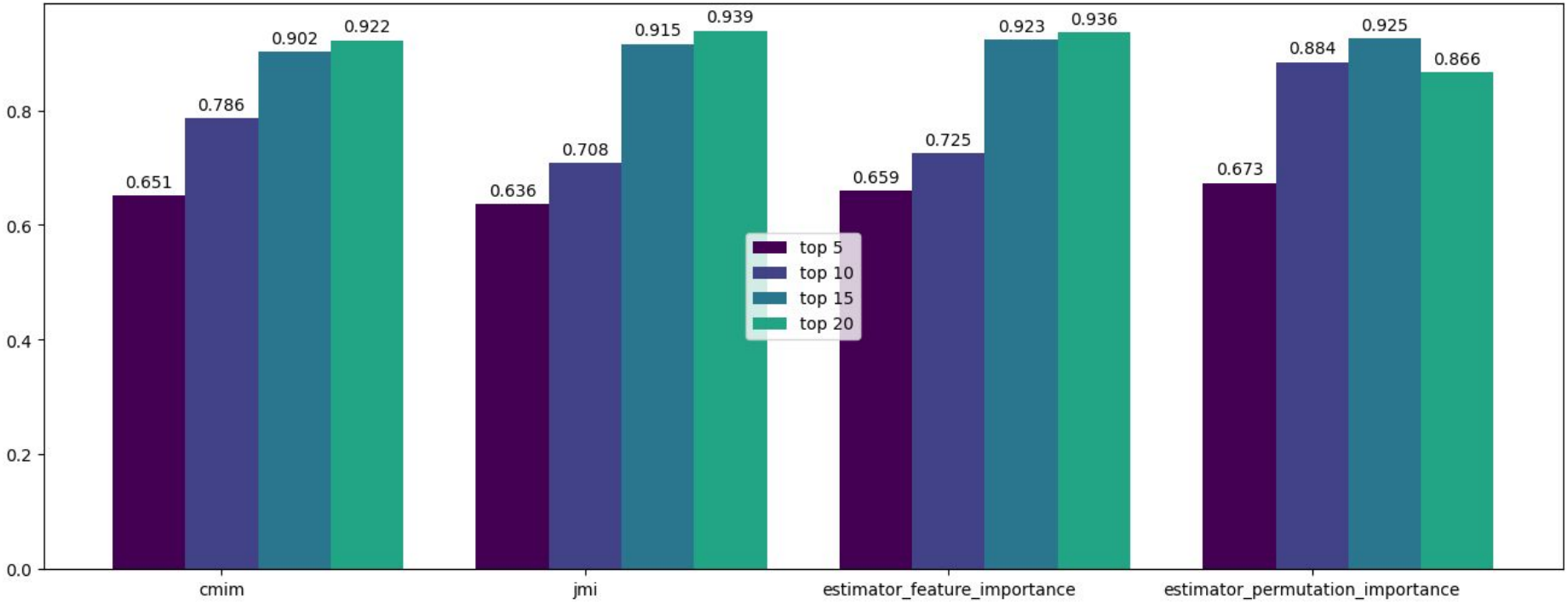
Classifier

```
    return nn.Sequential(
        nn.Linear(in_features, hidden_state),
        nn.ReLU(),
        nn.Linear(hidden_state, hidden_state),
        nn.ReLU(),
        nn.Linear(hidden_state, classes),
        nn.Softmax() if classes > 1 else nn.Sigmoid()
    )
```

Molecules Musk classification



Breast Cancer Wisconsin (Diagnostic)



Optical Recognition of Handwritten Digits