Weather Prediction

Rain prediction based on:

Principal Component Analysis and

Decision Trees

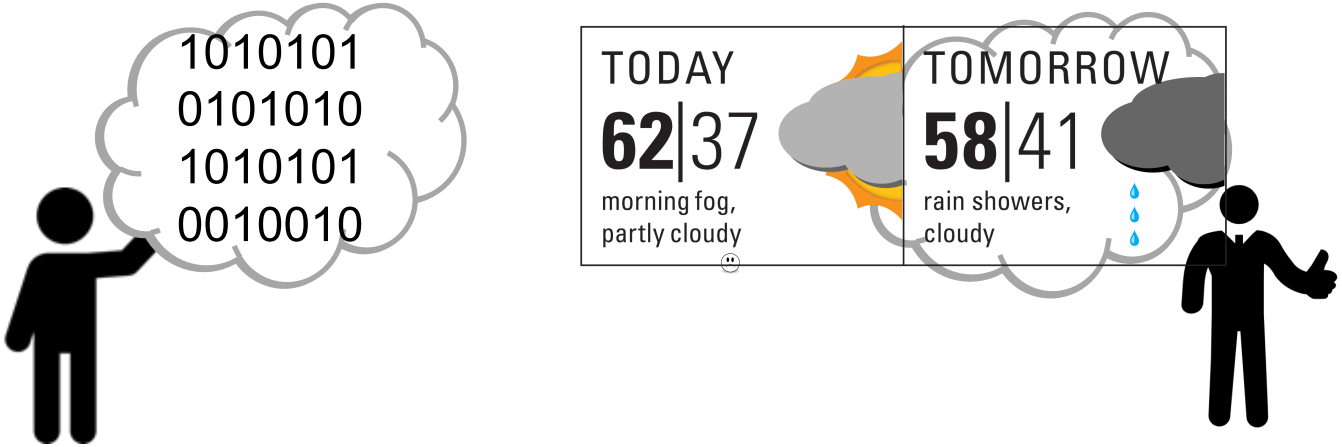
Aditya Raj
Institut für Informatik

Overview

- Problem statement
- Background
- Input data analysis
- Evaluation

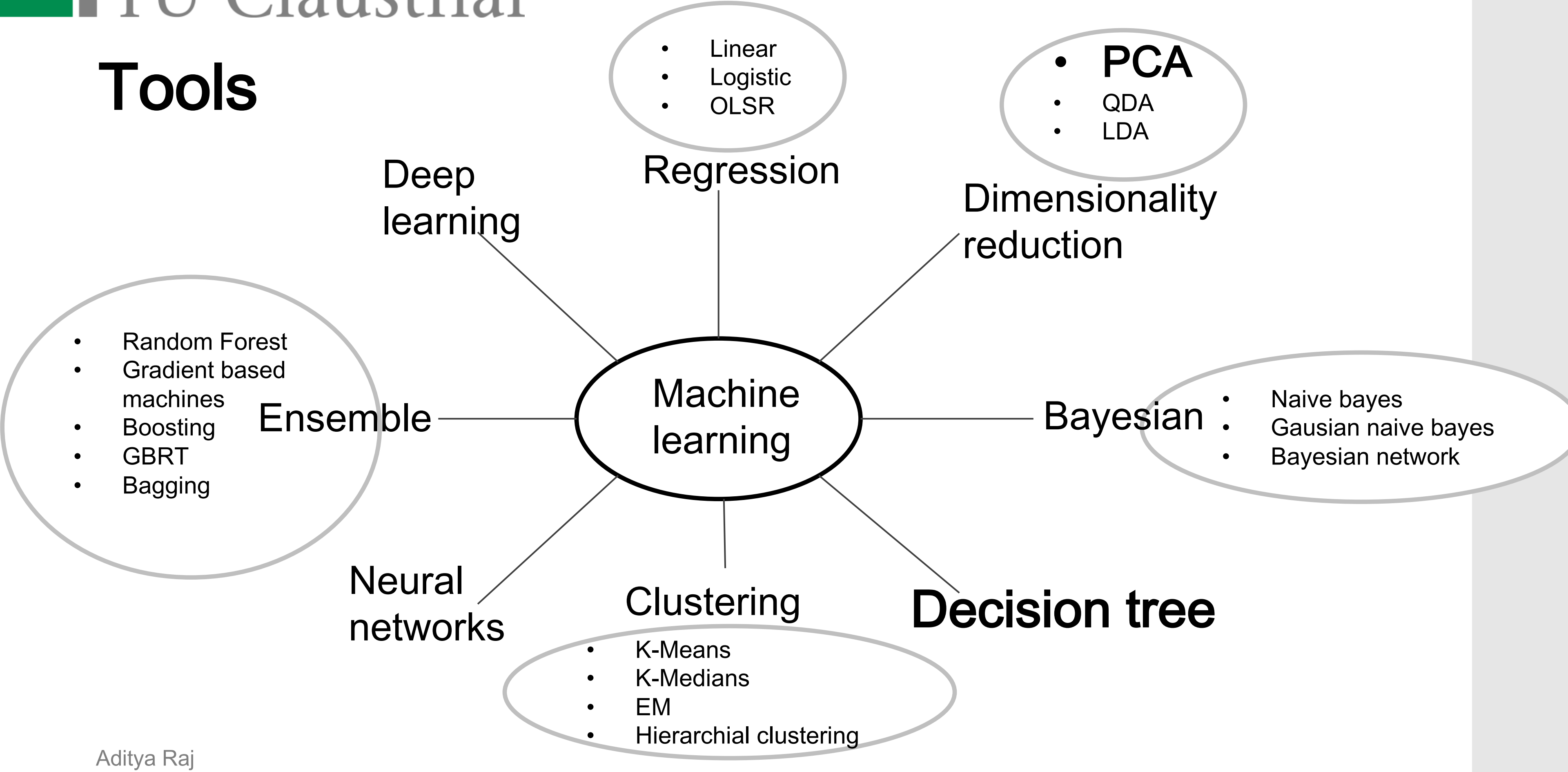
PROBLEM STATEMENT

What’s tomorrow?



BACKGROUND

Tools



Machine learning

Regression

- Linear
- Logistic
- OLSR

Dimensionality reduction

- **PCA**
- QDA
- LDA

Bayesian

- Naive bayes
- Gausian naive bayes
- Bayesian network

Decision tree

Clustering

- K-Means
- K-Medians
- EM
- Hierarchial clustering

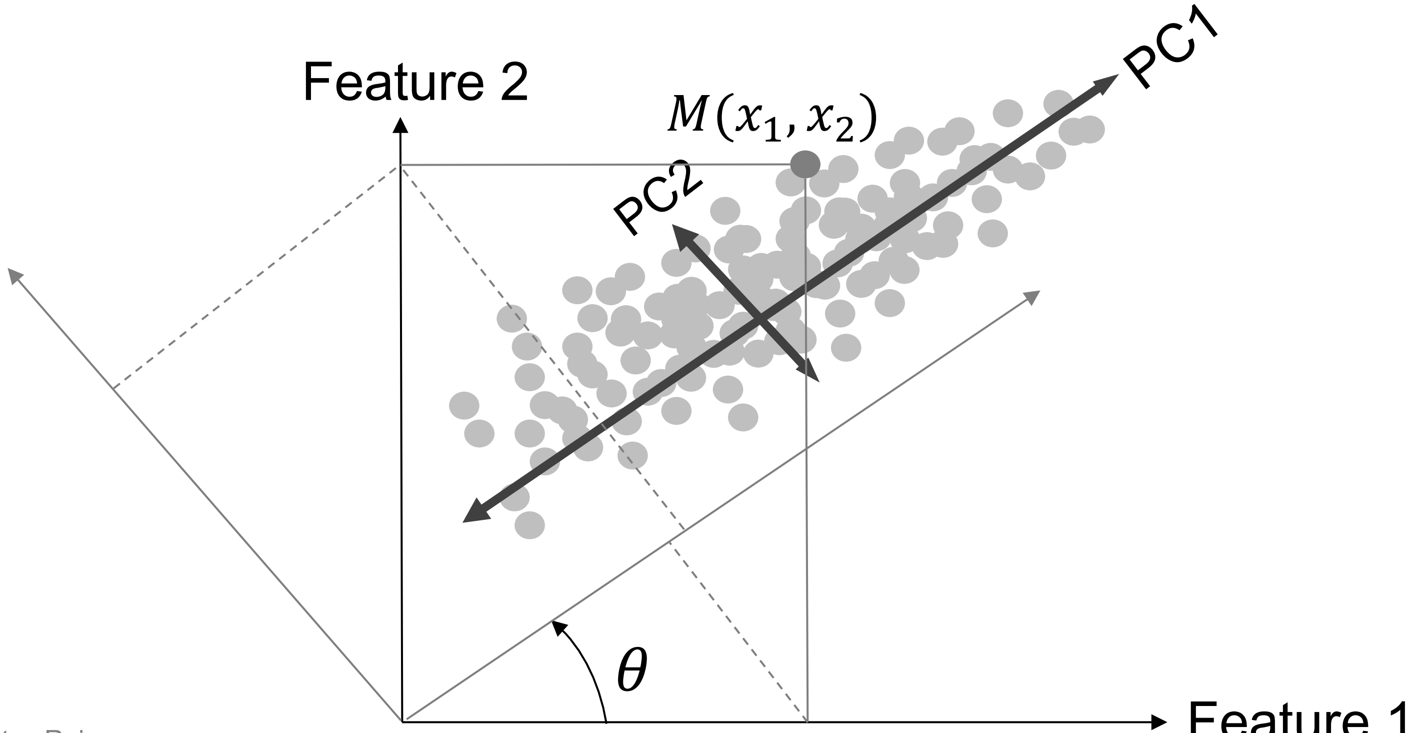
Neural networks

Ensemble

- Random Forest
- Gradient based machines
- Boosting
- GBRT
- Bagging

Deep learning

Principal Component Analysis: Algebraic View



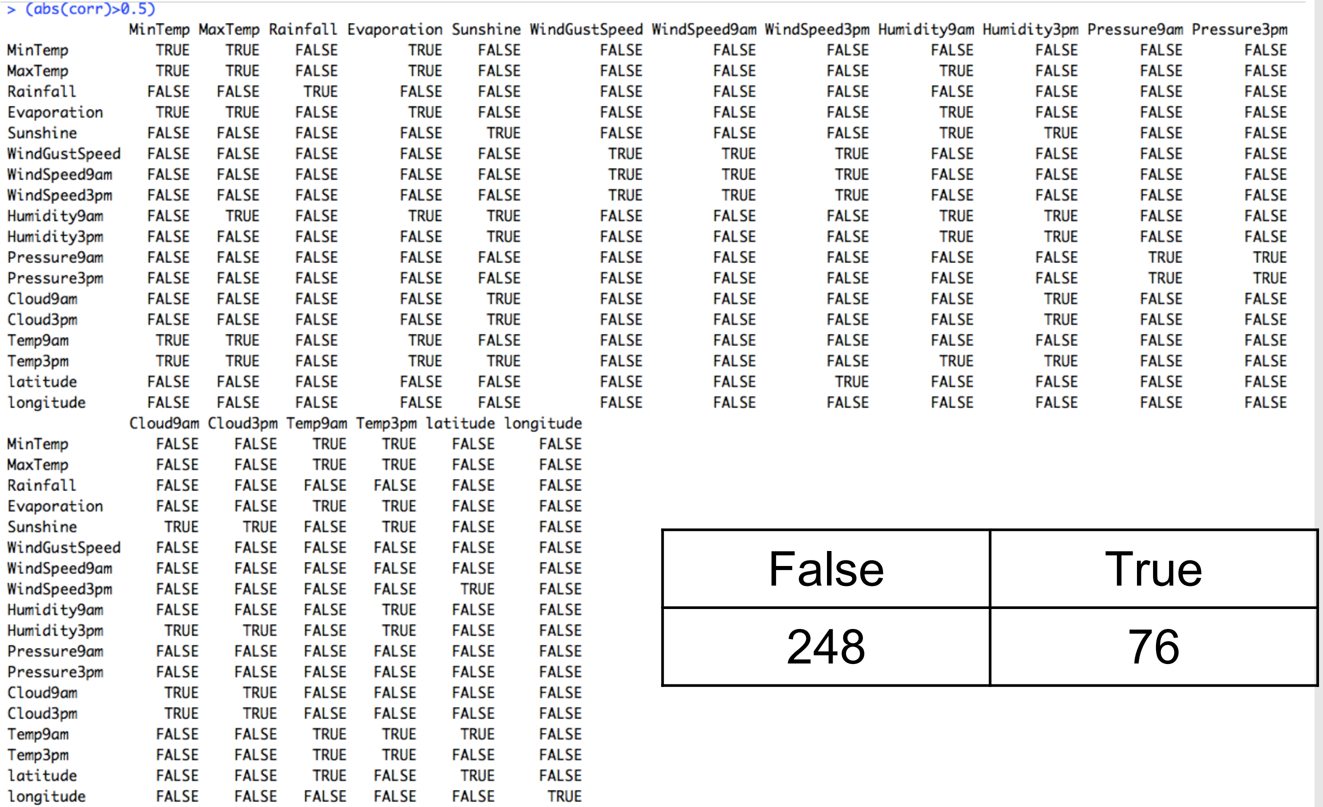
INPUT ANALYSIS

Input structure

- 10948 Observations
- About:
  - Cateogrical:Location, RainToday, RainTomorrow
  - Numerical:temp, rainfall, evaporation, pressure, humidity, windSpeed, cloud, longitude, latitude

Input structure

- cor > 0.5



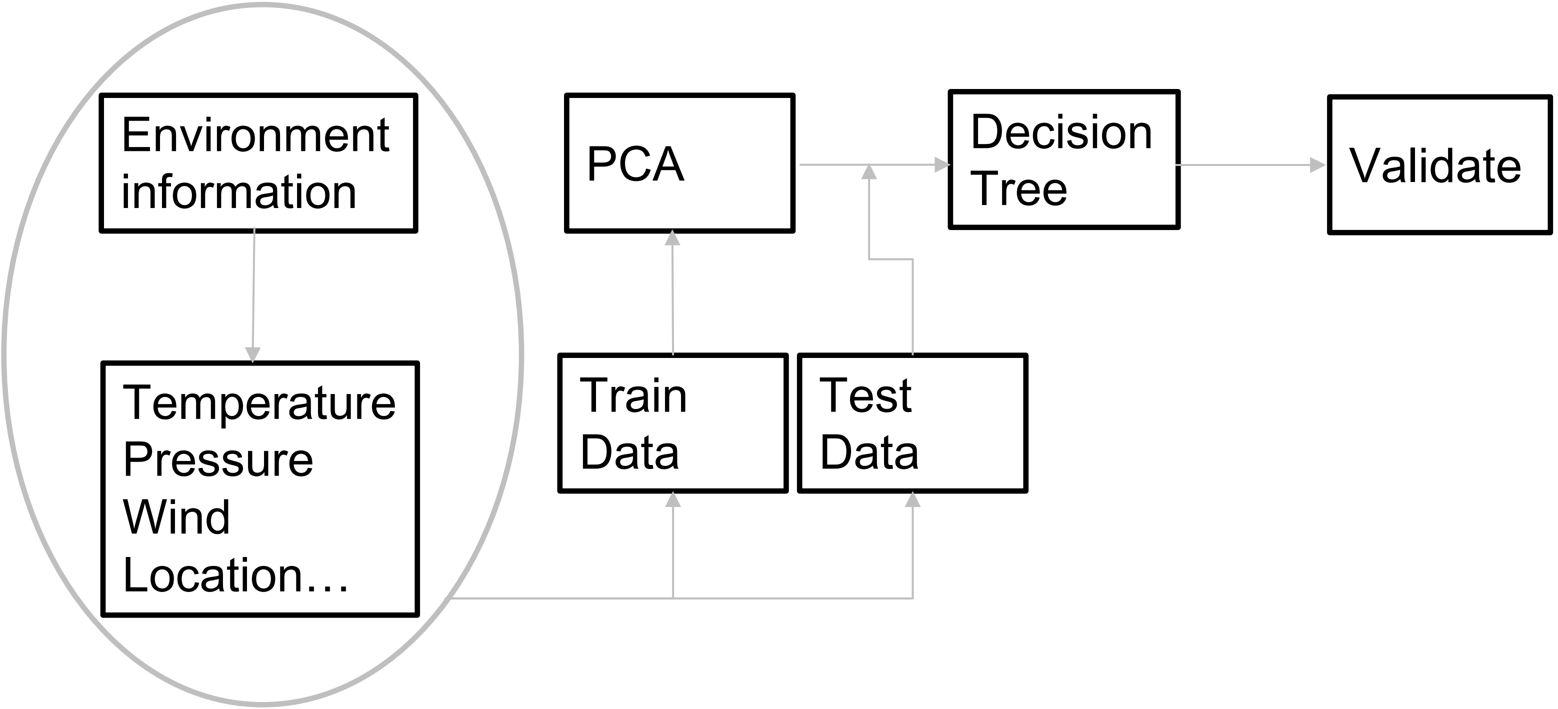
```
> (abs(corr)>0.5)
              MinTemp MaxTemp Rainfall Evaporation Sunshine WindGustSpeed WindSpeed9am WindSpeed3pm Humidity9am Humidity3pm Pressure9am Pressure3pm
MinTemp          TRUE    TRUE    FALSE        TRUE    FALSE         FALSE        FALSE        FALSE       FALSE       FALSE       FALSE       FALSE
MaxTemp          TRUE    TRUE    FALSE        TRUE    FALSE         FALSE        FALSE        FALSE        TRUE       FALSE       FALSE       FALSE
Rainfall        FALSE   FALSE     TRUE       FALSE    FALSE         FALSE        FALSE        FALSE       FALSE       FALSE       FALSE       FALSE
Evaporation      TRUE    TRUE    FALSE        TRUE    FALSE         FALSE        FALSE        FALSE        TRUE       FALSE       FALSE       FALSE
Sunshine        FALSE   FALSE    FALSE       FALSE     TRUE         FALSE        FALSE        FALSE        TRUE        TRUE       FALSE       FALSE
WindGustSpeed   FALSE   FALSE    FALSE       FALSE    FALSE          TRUE         TRUE         TRUE       FALSE       FALSE       FALSE       FALSE
WindSpeed9am    FALSE   FALSE    FALSE       FALSE    FALSE          TRUE         TRUE         TRUE       FALSE       FALSE       FALSE       FALSE
WindSpeed3pm    FALSE   FALSE    FALSE       FALSE    FALSE          TRUE         TRUE         TRUE       FALSE       FALSE       FALSE       FALSE
Humidity9am     FALSE    TRUE    FALSE        TRUE     TRUE         FALSE        FALSE        FALSE        TRUE        TRUE       FALSE       FALSE
Humidity3pm     FALSE   FALSE    FALSE       FALSE     TRUE         FALSE        FALSE        FALSE        TRUE        TRUE       FALSE       FALSE
Pressure9am     FALSE   FALSE    FALSE       FALSE    FALSE         FALSE        FALSE        FALSE       FALSE       FALSE        TRUE        TRUE
Pressure3pm     FALSE   FALSE    FALSE       FALSE    FALSE         FALSE        FALSE        FALSE       FALSE       FALSE        TRUE        TRUE
Cloud9am        FALSE   FALSE    FALSE       FALSE     TRUE         FALSE        FALSE        FALSE       FALSE        TRUE       FALSE       FALSE
Cloud3pm        FALSE   FALSE    FALSE       FALSE     TRUE         FALSE        FALSE        FALSE       FALSE        TRUE       FALSE       FALSE
Temp9am          TRUE    TRUE    FALSE        TRUE    FALSE         FALSE        FALSE        FALSE       FALSE       FALSE       FALSE       FALSE
Temp3pm          TRUE    TRUE    FALSE        TRUE     TRUE         FALSE        FALSE        FALSE        TRUE        TRUE       FALSE       FALSE
latitude        FALSE   FALSE    FALSE       FALSE    FALSE         FALSE        FALSE         TRUE       FALSE       FALSE       FALSE       FALSE
longitude       FALSE   FALSE    FALSE       FALSE    FALSE         FALSE        FALSE        FALSE       FALSE       FALSE       FALSE       FALSE
              Cloud9am Cloud3pm Temp9am Temp3pm latitude longitude
MinTemp          FALSE    FALSE    TRUE    TRUE    FALSE     FALSE
MaxTemp          FALSE    FALSE    TRUE    TRUE    FALSE     FALSE
Rainfall         FALSE    FALSE   FALSE   FALSE    FALSE     FALSE
Evaporation      FALSE    FALSE    TRUE    TRUE    FALSE     FALSE
Sunshine          TRUE     TRUE   FALSE    TRUE    FALSE     FALSE
WindGustSpeed    FALSE    FALSE   FALSE   FALSE    FALSE     FALSE
WindSpeed9am     FALSE    FALSE   FALSE   FALSE    FALSE     FALSE
WindSpeed3pm     FALSE    FALSE   FALSE   FALSE     TRUE     FALSE
Humidity9am      FALSE    FALSE   FALSE    TRUE    FALSE     FALSE
Humidity3pm       TRUE     TRUE   FALSE    TRUE    FALSE     FALSE
Pressure9am      FALSE    FALSE   FALSE   FALSE    FALSE     FALSE
Pressure3pm      FALSE    FALSE   FALSE   FALSE    FALSE     FALSE
Cloud9am          TRUE     TRUE   FALSE   FALSE    FALSE     FALSE
Cloud3pm          TRUE     TRUE   FALSE   FALSE    FALSE     FALSE
Temp9am          FALSE    FALSE    TRUE    TRUE     TRUE     FALSE
Temp3pm          FALSE    FALSE    TRUE    TRUE    FALSE     FALSE
latitude         FALSE    FALSE    TRUE   FALSE     TRUE     FALSE
longitude        FALSE    FALSE   FALSE   FALSE    FALSE      TRUE
```

False	True
248	76

Process Model

- Dividing observations into
  - 75% training data and
  - 25% test data
- Principal Component Analysis on the train data
- Rpart Model using PC for test data
- Evaluation of test data

Numbers to patterns

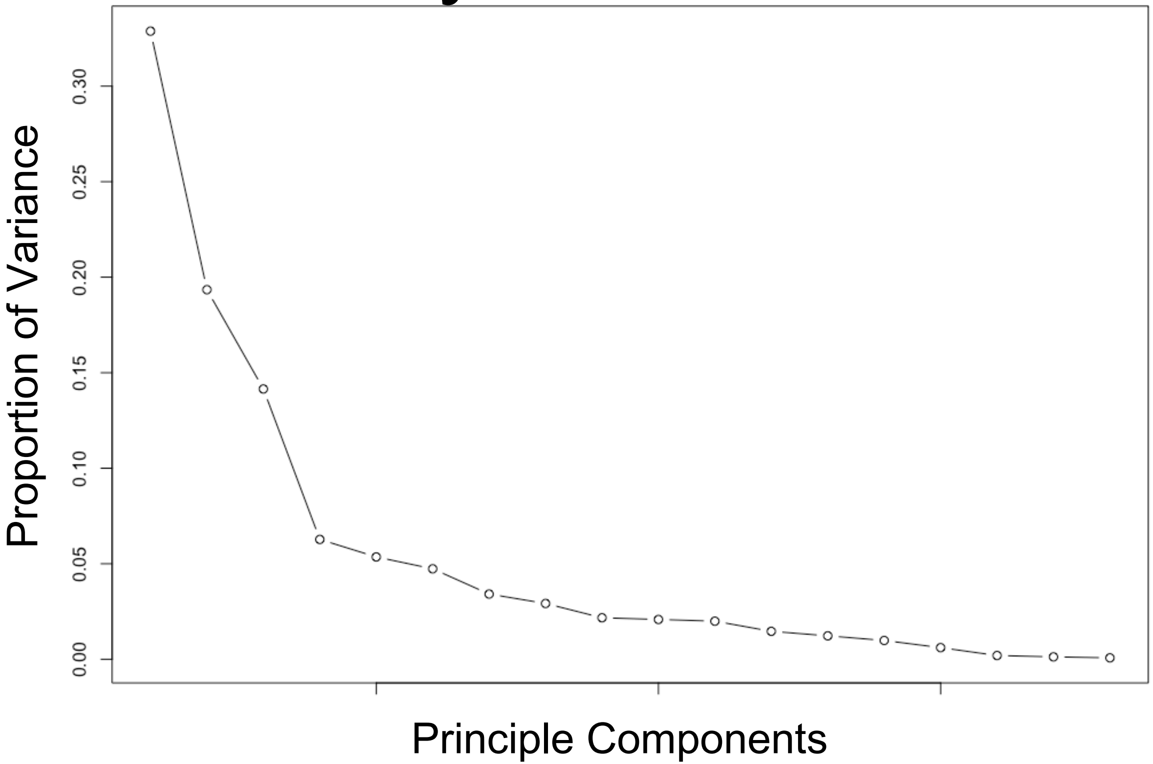


Principal Component Analysis

- "sdev”: Standard deviation on principal components
- "rotation”: Rotation axes
- "center" : Center at (0,0)
- "scale" : Normalise data
- "x”: Transformed training points

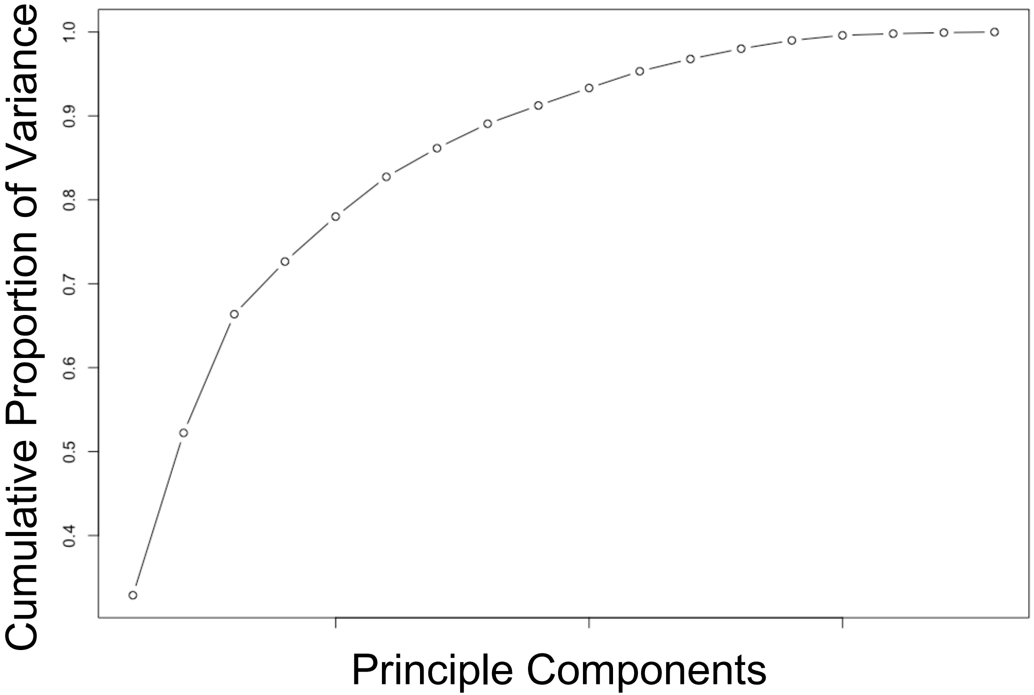
Principal Component Analysis: sdev

- Standard deviation on principal components
  - $variance = \dfrac{pca\$sdev^2}{\sum pca\$sdev^2}$



Principal Component Analysis: sdev

- Standard deviation on principal components
    - $plot(cumsum(variance))$
- First 14 PC
    - Capture 98.99% information

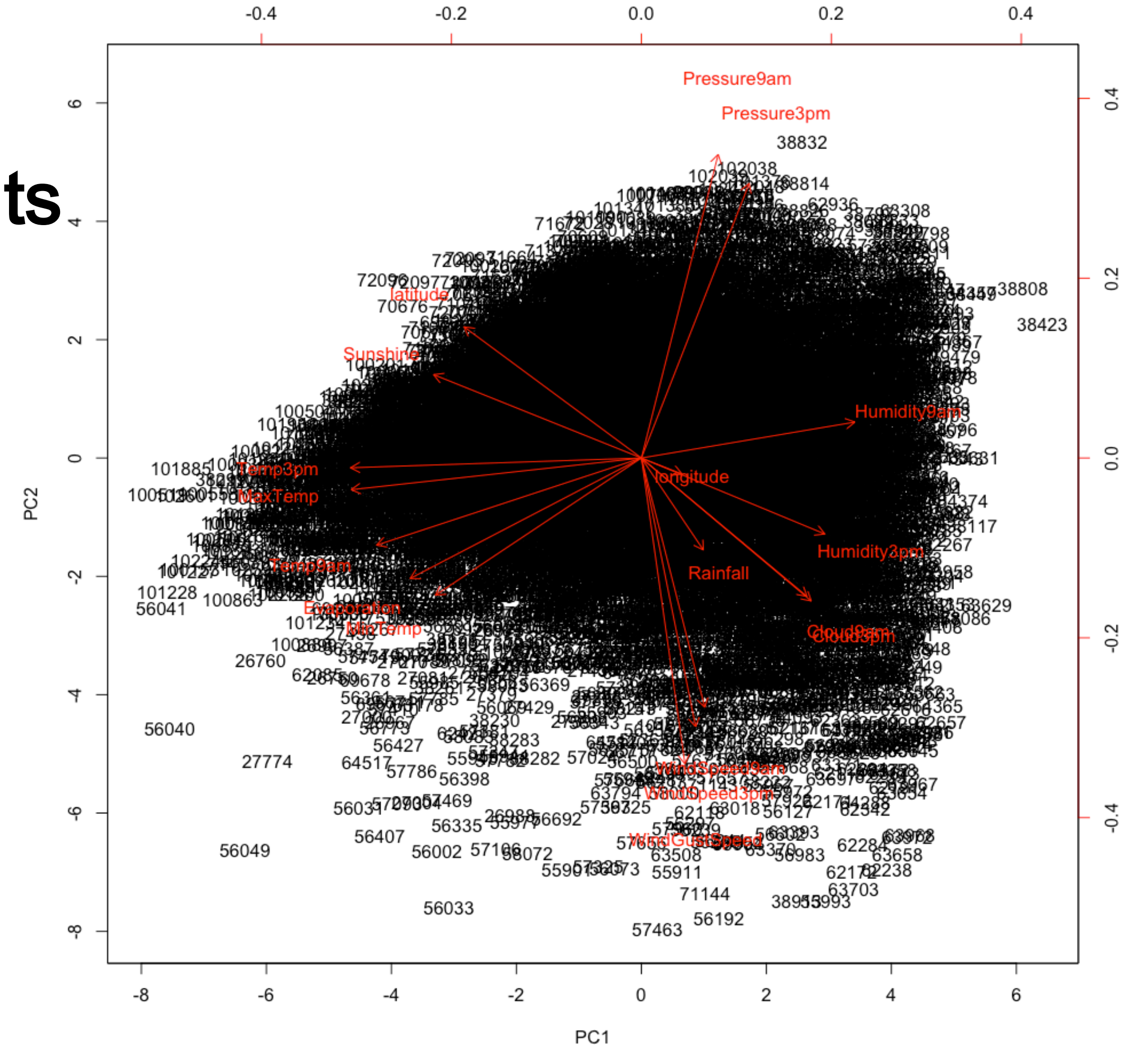


Rpart Prediction

- Run a decision tree
- Transform test data into PCA
  - Using $predict$ function
- Select the first 14 PC
- Make prediction on test data
  - Using $predict$ function

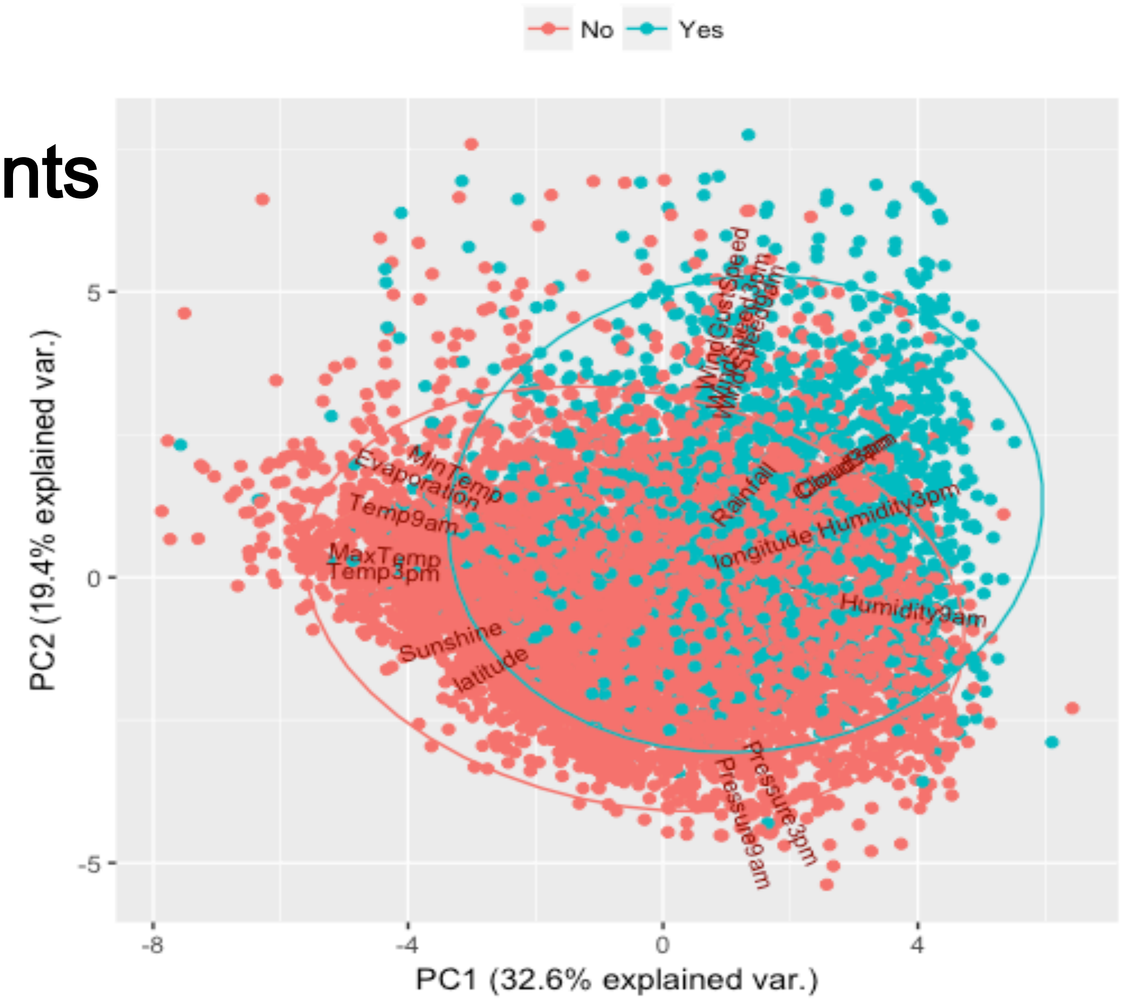
Principal Components

- biplot(pca)
  - PC1 vs PC2

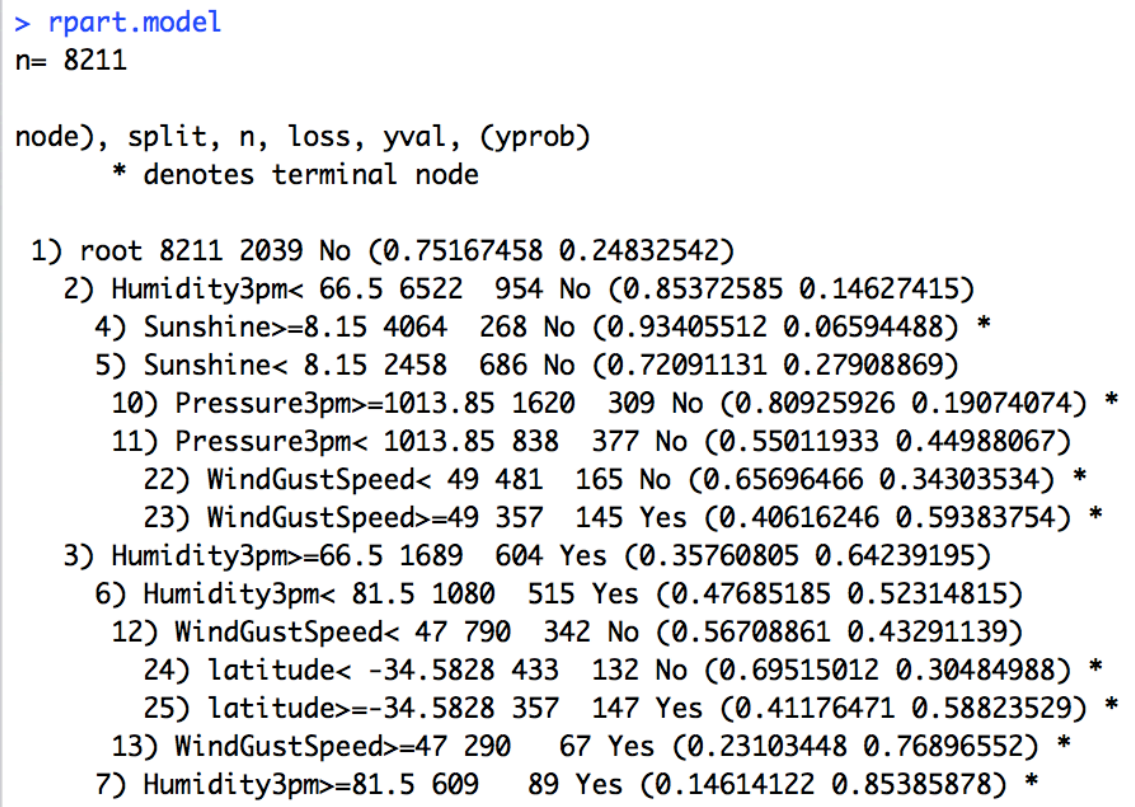


Principal Components

- ggbiplot(pca)
  - PC1 vs PC2



Evaluation results: Decision Tree

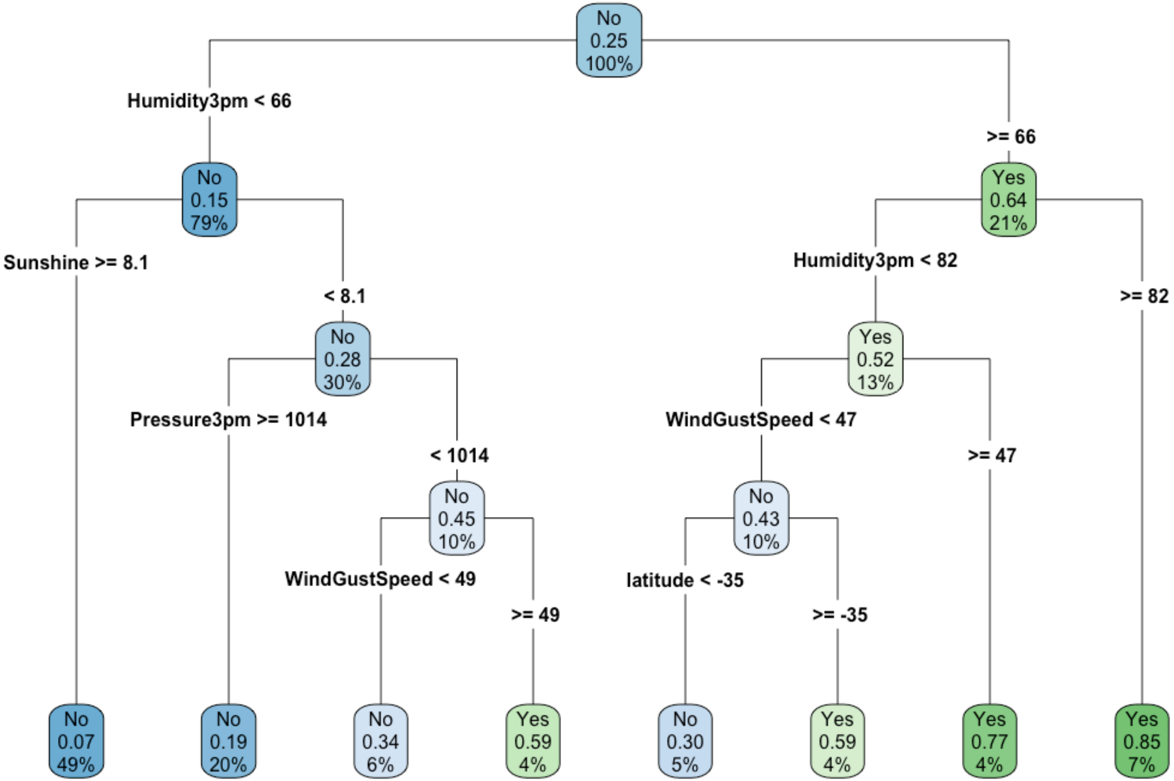


```
> rpart.model
n= 8211

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 8211 2039 No (0.75167458 0.24832542)
   2) Humidity3pm< 66.5 6522  954 No (0.85372585 0.14627415)
     4) Sunshine>=8.15 4064  268 No (0.93405512 0.06594488) *
     5) Sunshine< 8.15 2458  686 No (0.72091131 0.27908869)
      10) Pressure3pm>=1013.85 1620  309 No (0.80925926 0.19074074) *
      11) Pressure3pm< 1013.85 838  377 No (0.55011933 0.44988067)
        22) WindGustSpeed< 49 481  165 No (0.65696466 0.34303534) *
        23) WindGustSpeed>=49 357  145 Yes (0.40616246 0.59383754) *
   3) Humidity3pm>=66.5 1689  604 Yes (0.35760805 0.64239195)
     6) Humidity3pm< 81.5 1080  515 Yes (0.47685185 0.52314815)
      12) WindGustSpeed< 47 790  342 No (0.56708861 0.43291139)
        24) latitude< -34.5828 433  132 No (0.69515012 0.30484988) *
        25) latitude>=-34.5828 357  147 Yes (0.41176471 0.58823529) *
      13) WindGustSpeed>=47 290   67 Yes (0.23103448 0.76896552) *
     7) Humidity3pm>=81.5 609   89 Yes (0.14614122 0.85385878) *
```

Evaluation results: Decision Tree



Evaluation results: Test Data

- Accuracy
  - $\frac{1878+372}{1878+155+332+372} = 82\%$

	No	Yes
No	1878	155
Yes	332	372