Breast Cancer Data Analysis Project

Anna Källén and Mason Simmons

Used Data

- University of Wisconsin
- Not large by modern standards
- Comprises of physical tumor measurements
- Tumor type as defining feature

Preprocessing

- Data already largely processed
  - Complete,
- Noise managed by bin-mean smoothing
  - 10 bins to remove outliers fully

In [183]:
```
# This was a nightmare to program

def set_to_mean(row, bin_name, col, frame):
    meaned_values = []
    bin_interval = row[bin_name]
    for val in frame[col]:
        if val in bin_interval:
            meaned_values.append(val)
    row[col] = np.mean(meaned_values)
    return row

data2 = data.drop(columns = ['diagnosis'])
columns = data2.columns
drop_columns = []
for col in columns:
    bin_name = 'bin_' + col
    data2[bin_name] = pd.qcut(data2[col], q=10, precision=5)
    drop_columns.append(bin_name)
for col in columns:
    bin_name = 'bin_' + col
    data2 = data2.apply(lambda f: set_to_mean(f, bin_name, col, data2), axis=1)
data2 = data2.drop(columns = drop_columns)
data2
```

Out[183]:

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_
0	18.324483	12.804035	120.581034	1042.763158	0.122521	0.219140	0.265454	0.132232	0.235960	
1	21.453750	17.354821	142.741071	1438.157895	0.086125	0.075380	0.099248	0.073251	0.182196	
2	21.453750	20.553860	142.741071	1438.157895	0.110893	0.157588	0.172896	0.132232	0.207674	
3	11.691754	20.553860	79.494386	367.069643	0.122521	0.219140	0.265454	0.132232	0.235960	
4	21.453750	14.920175	142.741071	1438.157895	0.101166	0.130484	0.172896	0.132232	0.182196	
...	...	...	...	...	...	...	...	...	...	...
564	21.453750	21.816034	142.741071	1438.157895	0.110893	0.114774	0.265454	0.132232	0.175559	
565	21.453750	27.803158	142.741071	1438.157895	0.097740	0.101148	0.129684	0.091150	0.175559	
566	15.914561	27.803158	104.774912	787.471930	0.086125	0.101148	0.099248	0.056049	0.161972	
567	21.453750	27.803158	142.741071	1438.157895	0.122521	0.219140	0.265454	0.132232	0.235960	
568	9.342052	23.718214	59.495439	267.081034	0.074103	0.040022	0.005425	0.005294	0.154554	

Classification: K-Neighbors

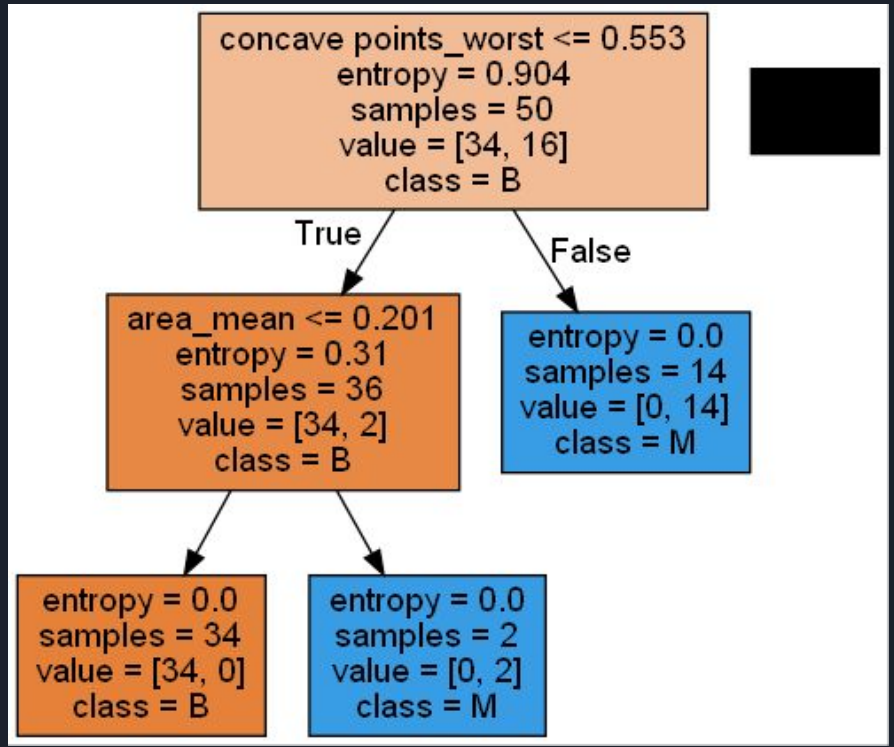
- Testing data 1/10 of total
- Neighbor count of 5
- Similar to clustering
- Recall: TP/(TP + FN)
- Precision: TP/(TP+FP)

```
[[41  1]
 [ 0 15]]
```

	precision	recall	f1-score	support
B	1.00	0.98	0.99	42
M	0.94	1.00	0.97	15
accuracy			0.98	57
macro avg	0.97	0.99	0.98	57
weighted avg	0.98	0.98	0.98	57

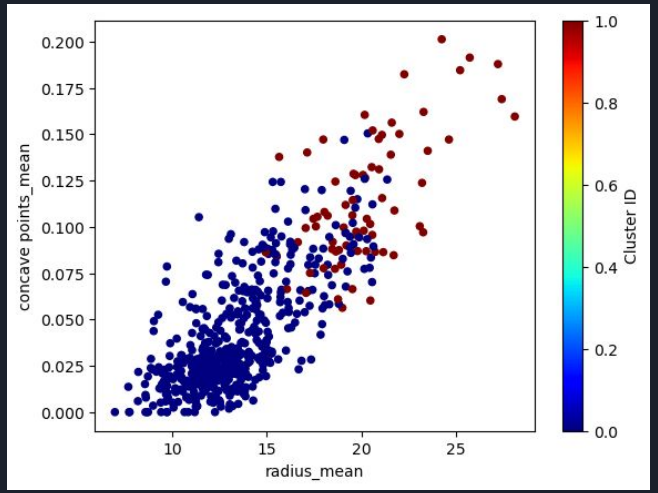
Classification: Decision Tree

- Max depth 7 required initially
- Not many attributes actually needed



Clustering: K-Means

- Preprocessed differently
- Some attributes deemed redundant
- Not always evident in slices




```
Recall: 0.719758064516129
Precision: 1.0
F-score: 0.8370457209847596
```

Conclusion

- Clear patterns observable
  - Not complex, either
- Improvements to methodology are possible