



Predicting the probability of the breast cancer in patients

Part II

A photograph of three surgeons in an operating room, wearing blue scrubs, masks, and surgical caps. The central surgeon is wearing a head-mounted display and is holding a surgical instrument. The background is slightly blurred, showing medical equipment.

Team members

Brendan Lim (A0216513N)

Laura Do (A0280548X)

Wong Cheuk Wah (A0280543H)



Outline of this report

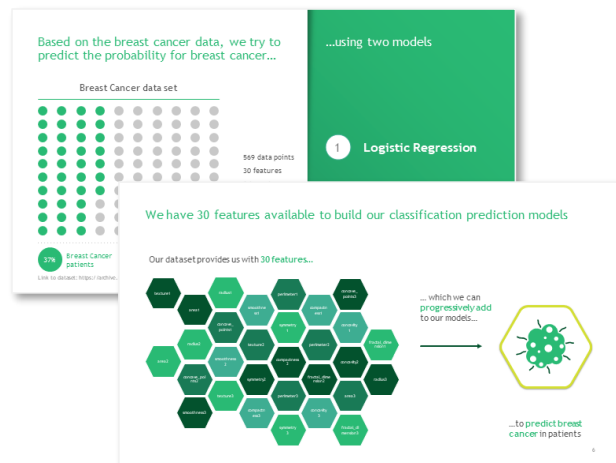
- Business problem and summary of findings in bonus challenge 1
- Breast cancer prediction with random forest (bagging)
- Breast cancer prediction by boosting with stumps
- Breast cancer prediction with support vector machines
- Improvement of prediction with ensemble method?
- Our conclusion

A photograph of surgeons in an operating room, overlaid with a semi-transparent teal filter. The surgeons are wearing blue scrubs, surgical masks, and caps. One surgeon in the center is wearing a head-mounted display (HMD) and is focused on a task. Another surgeon is visible on the right, also in profile. The background shows typical operating room equipment.

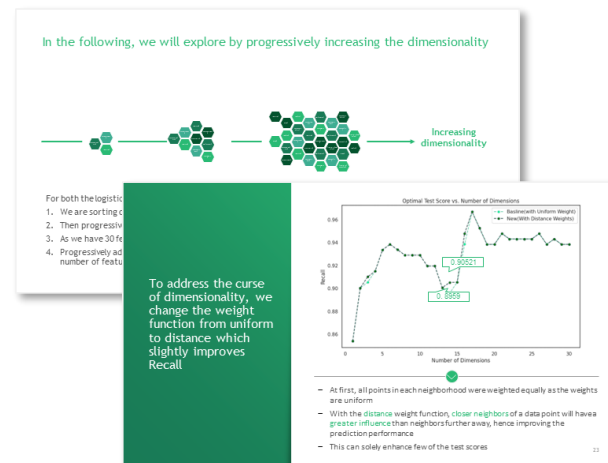
Breast Cancer Prediction and previous findings from challenge 1

After overcoming the curse of dimensionality, we concluded logistic regression performs better than KNN in predicting breast cancer in patients

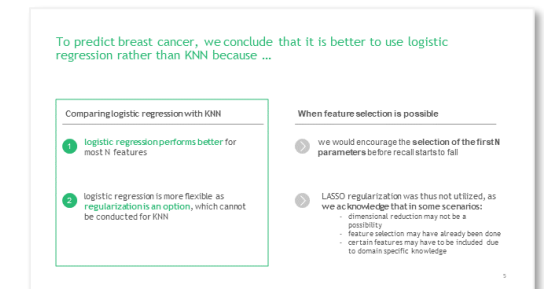
Recall from bonus challenge 1 that we tested two classification models to predict breast cancer in patients



In our analysis, we detected the curse of dimensionality which we addressed trying different methods



Comparing recall score of both models, we concluded logistic regression performs better than KNN



In bonus challenge 2, we would like to see if other classification models can outperform logistic regression in predicting breast cancer in patients

In summary, we are comparing **six classification techniques** to see...

...which classification model performs best in...

...predicting breast cancer in patients?

Logistic Regression

done on bonus 1, see appendix

K-Nearest Neighbors

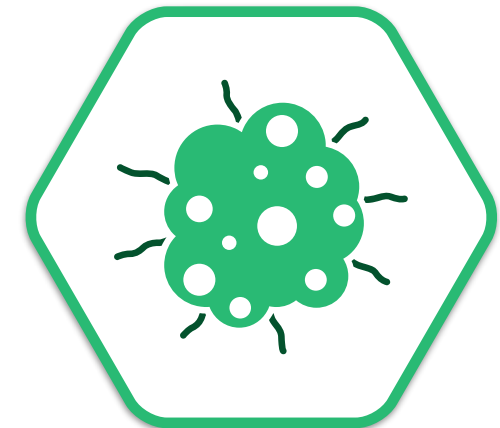
Random forest

Boosting with stumps

Support Vector Machines

Stacking (ensemble method)

We are using Recall to compare performance as we want to reduce the probability of the model predicting the absence of breast cancer if the patient has cancer



A photograph of surgeons in an operating room, overlaid with a semi-transparent teal filter. The text is centered over the image.

Breast cancer prediction with random forest

To prevent overfitting, we average a number of decision trees

This method is called **Random Forest**

There are three parameters in the Random Forest model which we can tweak:

- *number of trees ($n_estimators$)*
- *parameter for each tree (max_depth)*
- *number of features to consider at each split ($max_features$)*



Using the GridSearch method, we try different parameters in the Random Forest model:

- n estimators: 50, 100, 200
- max depth: None, 2, 10, 20
- max features: 2, 'sqrt', 10



Based on best estimator and recall, the best performing Random Forest operated under the following parameters:

- n estimators: 50
- max depth: None
- max features: 'sqrt'



Recall score for

- Class 0 (Benign): 0.95
- Class 1 (Malignant): 0.89

A photograph of surgeons in an operating room, overlaid with a semi-transparent teal filter. The text is centered over the image.

Breast cancer prediction with boosting with stumps

Boosting with stumps

involves the use of multiple decision trees with depth 1, with every tree considering the misclassified data points from previous iterations

This method is used as a means to reduce bias and combine weak learners into a single strong learner



For the purpose of GridSearch, the parameters considered were:

- number of trees: 50, 100, 150
- tree depth: 1, 3, 5



Notably, while depths of more than 1 were considered, the grid search result still returned an optimal set of parameters of:

- depth =1
- number of trees = 150



Recall scores for

- Class 0 (Benign): 0.99
- Class 1 (Malignant): 0.92

A photograph of surgeons in an operating room, overlaid with a semi-transparent teal filter. The text is centered over the image.

Breast cancer prediction with support vector machine

Support vector machine (SVM) enables the model to learn non-linear decision boundaries, which helps to deal with nonlinearities that can occur in the features

There are three parameters in the SVM model we wanted to find the best parameters for our data set for



The GridSearch parameters used includes:

- C: 0.01, 0.1, 0.5, 1, 5, 10, 100
- kernel: 'linear', 'poly', 'rbf', 'sigmoid'
- gamma: 1, 0.1, 0.01, 0.001



The results from the GridSearch are found to be:

- C= 5
- gamma= 0.01
- kernel= 'rbf'



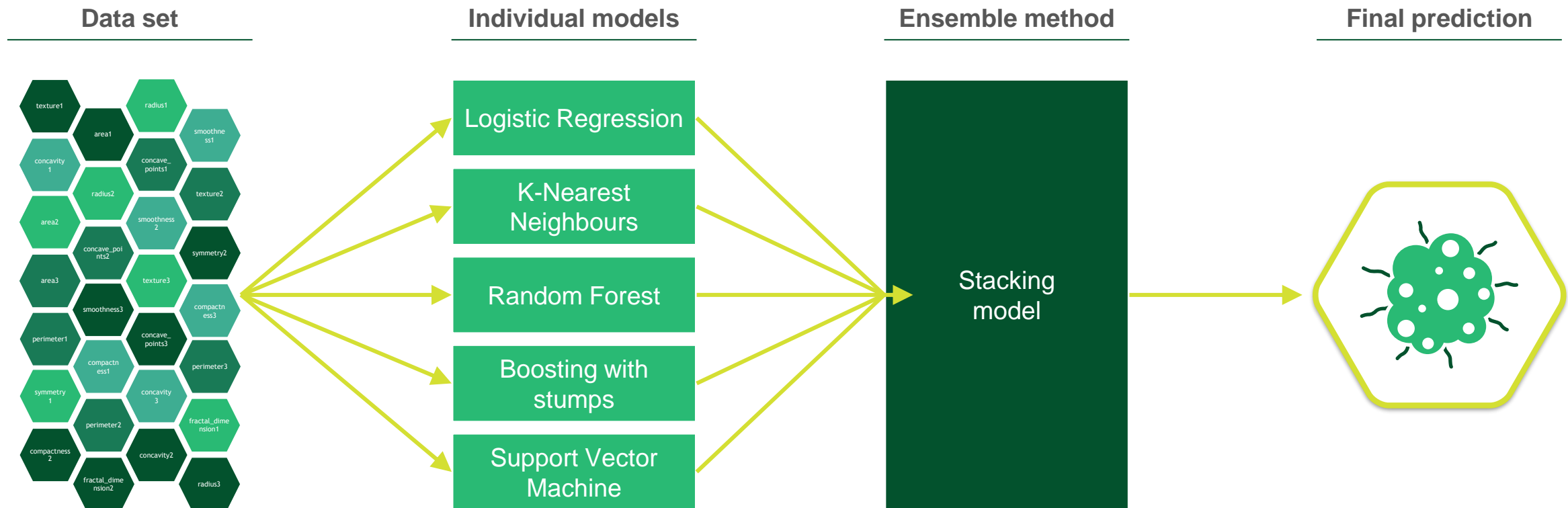
Recall scores for

- Class 0 (Benign): 0.99
- Class 1 (Malignant): 0.94

A photograph of surgeons in an operating room, overlaid with a semi-transparent teal filter. The text is centered over the image.

Breast cancer prediction with ensemble method

Stacking is a method to combine several models to refine the final prediction



We found the best hyperparameters for breast cancer prediction in each individual method before including them in the stacking model

Comparing the Recall scores across all models and methods, Stacking does the best job in prediction breast cancer in patients

Recall scores	Logistic Regression	K-Nearest Neighbours	Random Forest	Boosting with stumps	Support Vector Machine	Stacking model
Class 0 (no breast cancer)	0.971963	0.953271	0.953271	0.990654	0.990654	0.981308
Class 1 (breast cancer)	0.953125	0.859375	0.906250	0.921875	0.937500	0.953125
Weights in Stacked model	1.78	0.80	1.04	1.08	1.48	

> For the application of cancer prediction, we prioritise the Recall of class 1 (malignant) over the Recall of class 0 (Benign)

The stacking model was re-trained with only the best 3 individual models (excl. 2 worse performing models) and again benchmarked


Recall scores	Logistic Regression	Boosting with stumps	Support Vector Machine	Stacking model	Stacking model 2
Class 0 (no breast cancer)	0.971963	0.990654	0.990654	0.981308	0.990654
Class 1 (breast cancer)	0.953125	0.921875	0.937500	0.953125	0.953125
Weights in Stacked model	2.08	1.54	1.63		

➤ Recall of class 0 (Benign) of the new stacking model slightly improved

- The stacking model computes the final prediction by combining the output of the underlying models. The strength of each individual estimator is used
- By removing the ones with lower recall scores in Class 1 (i.e., Random Forest and KNN), we examined whether the stacking model would be more robust

A photograph of surgeons in an operating room, overlaid with a semi-transparent teal filter. The central surgeon is wearing a surgical cap, mask, and large magnifying glasses, holding a surgical instrument. Other surgeons are visible in the background, also in scrubs and caps. The word "Conclusion" is centered in white text.

Conclusion



After having tested **different classification and ensemble methods** across bonus challenges 1 and 2...

...we can confidently say that we are able to **predict breast cancer in patients with a recall of 0.99** (class 0, no cancer) **and 0.95** (class 1, cancer) using an ensemble model

The background of the slide is a photograph of surgeons in an operating room, wearing blue scrubs, surgical masks, and caps. The image is dimmed and serves as a backdrop for the text.

Appendix

Bonus Challenge 1 report



Outline of this report (bonus challenge 1)

- The curse of dimensionality
- Business problem: Prediction breast cancer in patients
- Logistic regression
- K nearest neighbors
- Our conclusion

A photograph of surgeons in an operating room, overlaid with a semi-transparent teal filter. The central surgeon is wearing a surgical cap, mask, and glasses with an attached light. They are holding a surgical instrument. Other surgeons are visible in the background, also in scrubs and caps. The text "Curse of dimensionality" is centered over the image.

Curse of dimensionality

What is the curse of dimensionality and how we can show it with our dataset

The curse of dimensionality...



...refers to working with high-dimensional data which is due to large number of features



...occurs during analyzing the data to identify patterns and training the model



...is the model's decreasing performance with more features leaving all else constant as it becomes harder to generalize

We are aiming to show the curse in our dataset by...



... reducing and adding features into our models and observing model performances in each steps



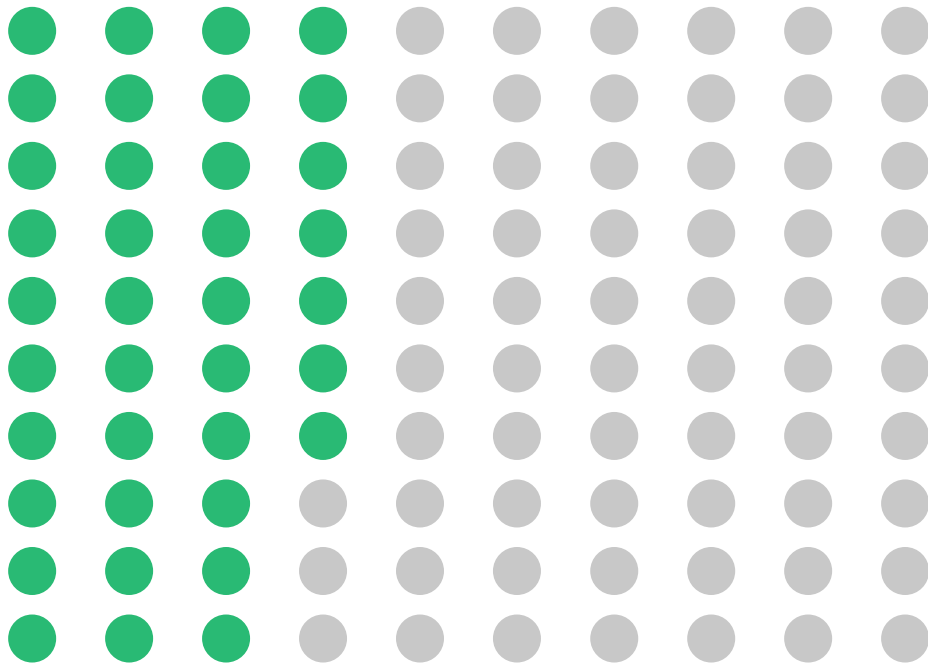
... showing that the model's ability to accurately predict the target is better with fewer than more features

A photograph of surgeons in an operating room, overlaid with a semi-transparent teal filter. The central surgeon is wearing a surgical cap, mask, and large magnifying glasses, holding a surgical instrument. Other surgeons are visible in the background, also in scrubs and caps. The text "Predicting breast cancer" is centered over the image.

Predicting breast cancer

Based on the breast cancer data, we try to predict the probability for breast cancer...

Breast Cancer data set



569 data points

30 features

1 target

37% Breast Cancer patients

No Breast Cancer patients

63%

Link to dataset: <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>

...using two models

1

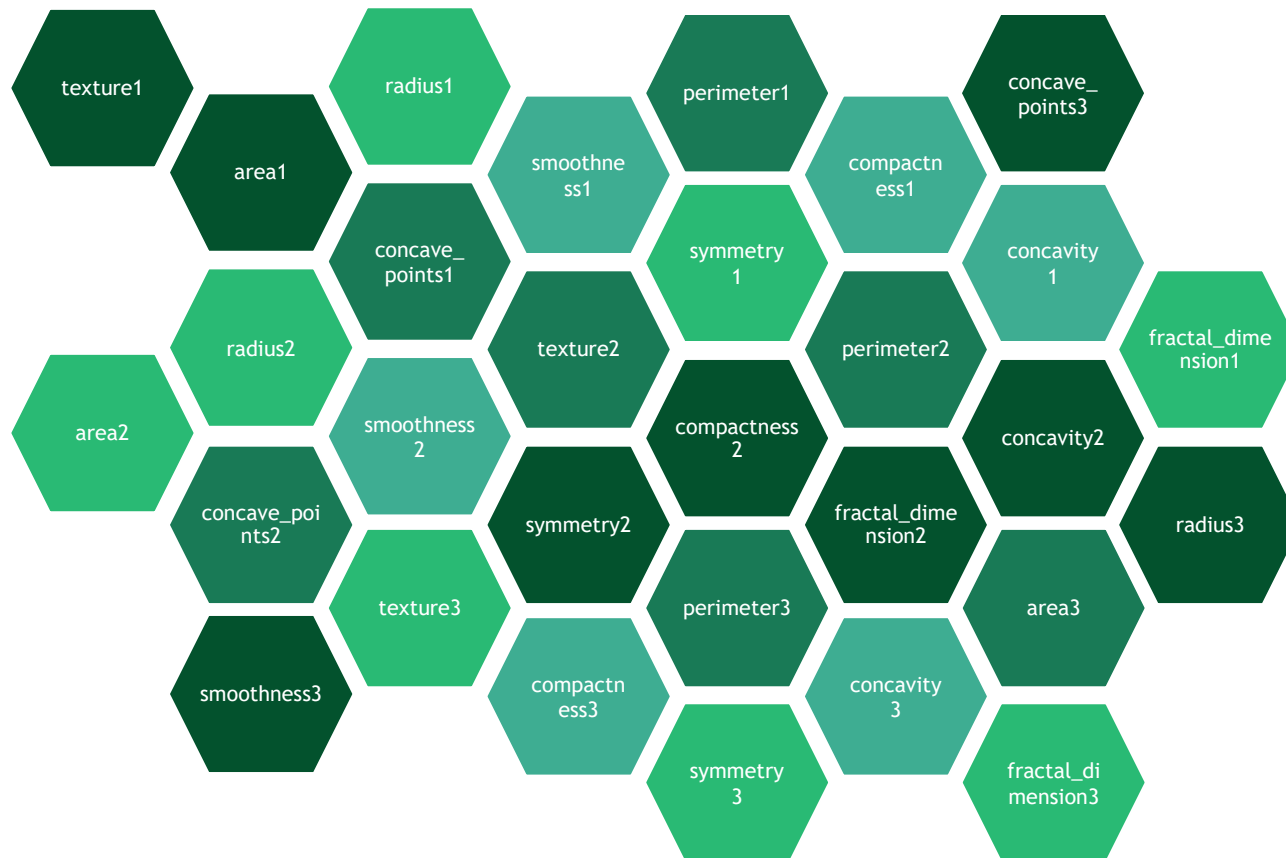
Logistic Regression

2

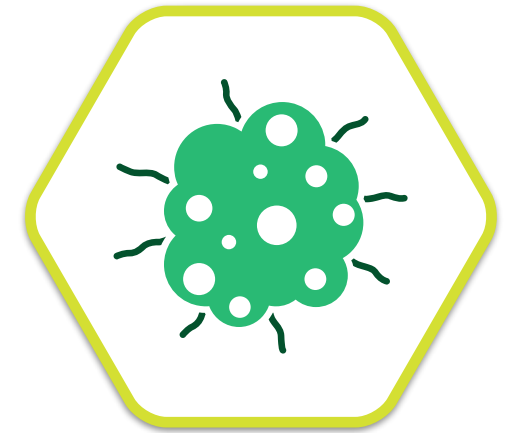
K nearest neighbors

We have 30 features available to build our classification prediction models

Our dataset provides us with **30 features...**

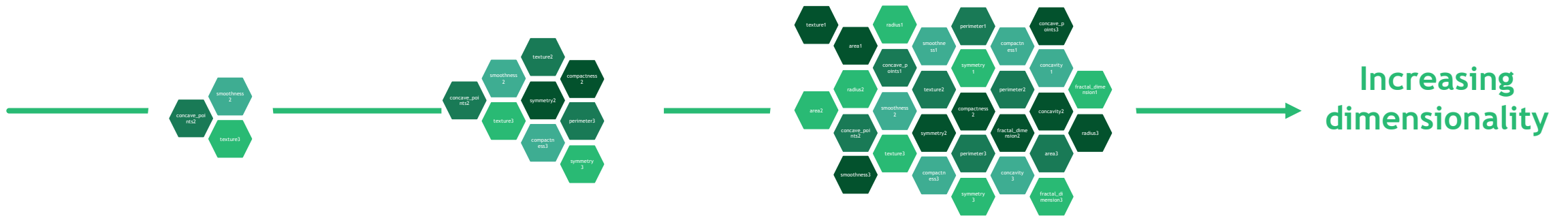


... which we can **progressively add** to our models...



...to **predict breast cancer** in patients

In the following, we will explore by progressively increasing the dimensionality



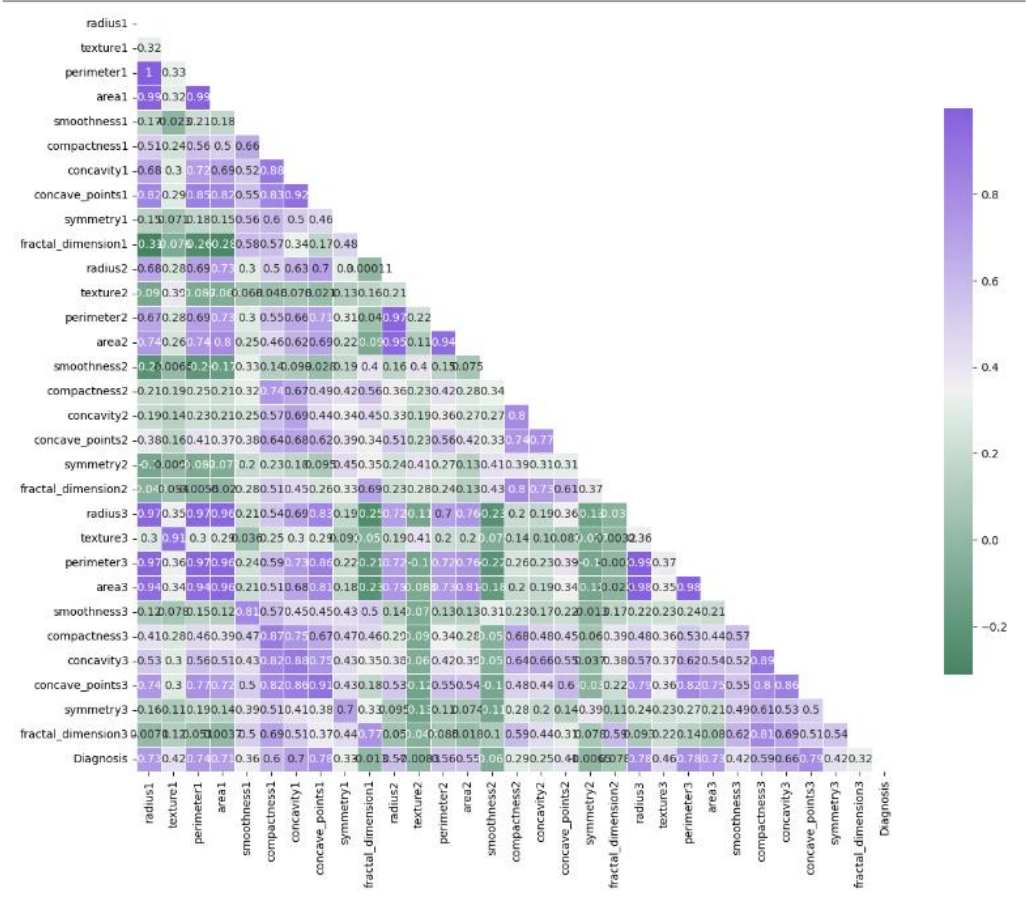
For both the logistic regression and KNN models

1. We are sorting our features in a meaningful order,
2. Then progressively adding the next best feature into our model.
3. As we have 30 features, we train and fit our models 30 times, each time with one more feature.
4. Progressively adding one features after another allows us to see the models' performance while increasing the number of features (= dimensions).

Please see on the following slide for the order

To order the features in a meaningful order to later progressively add them to the models, we look at their correlation to the target 'Diagnosis'

Correlation between 30 features and target 'Diagnosis'



Features ordered according to their correlation

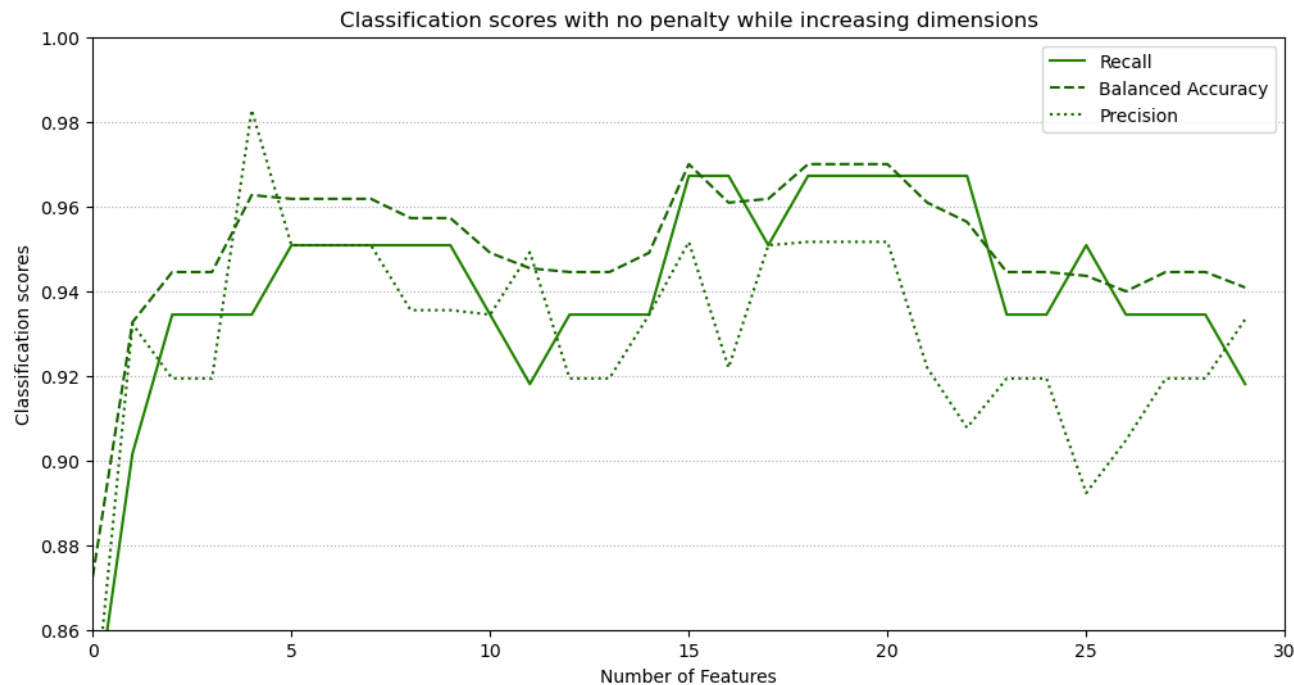
Diagnosis	1.000000	texture3	0.456903
concave_points3	0.793566	smoothness3	0.421465
perimeter3	0.782914	symmetry3	0.416294
concave_points1	0.776614	texture1	0.415185
radius3	0.776454	concave_points2	0.408042
perimeter1	0.742636	smoothness1	0.358560
area3	0.733825	symmetry1	0.330499
radius1	0.730029	fractal_dimension3	0.323872
area1	0.708984	compactness2	0.292999
concavity1	0.696360	concavity2	0.253730
concavity3	0.659610	fractal_dimension2	0.077972
compactness1	0.596534	smoothness2	0.067016
compactness3	0.590998	fractal_dimension1	0.012838
radius2	0.567134	texture2	0.008303
perimeter2	0.556141	symmetry2	0.006522
area2	0.548236	Name: Diagnosis, dtype: float64	

We will pass the features one by one according to this order into our logistic regression and KNN models and observe their performance in prediction breast cancer in patients

A photograph of surgeons in an operating room, overlaid with a semi-transparent teal filter. The text 'Logistic Regression' is centered in white. The image shows surgeons in green scrubs and blue surgical caps, with one surgeon in the foreground wearing a head-mounted display and holding a surgical instrument.

Logistic Regression

The baseline logistic regression suffers the curse of dimensionality



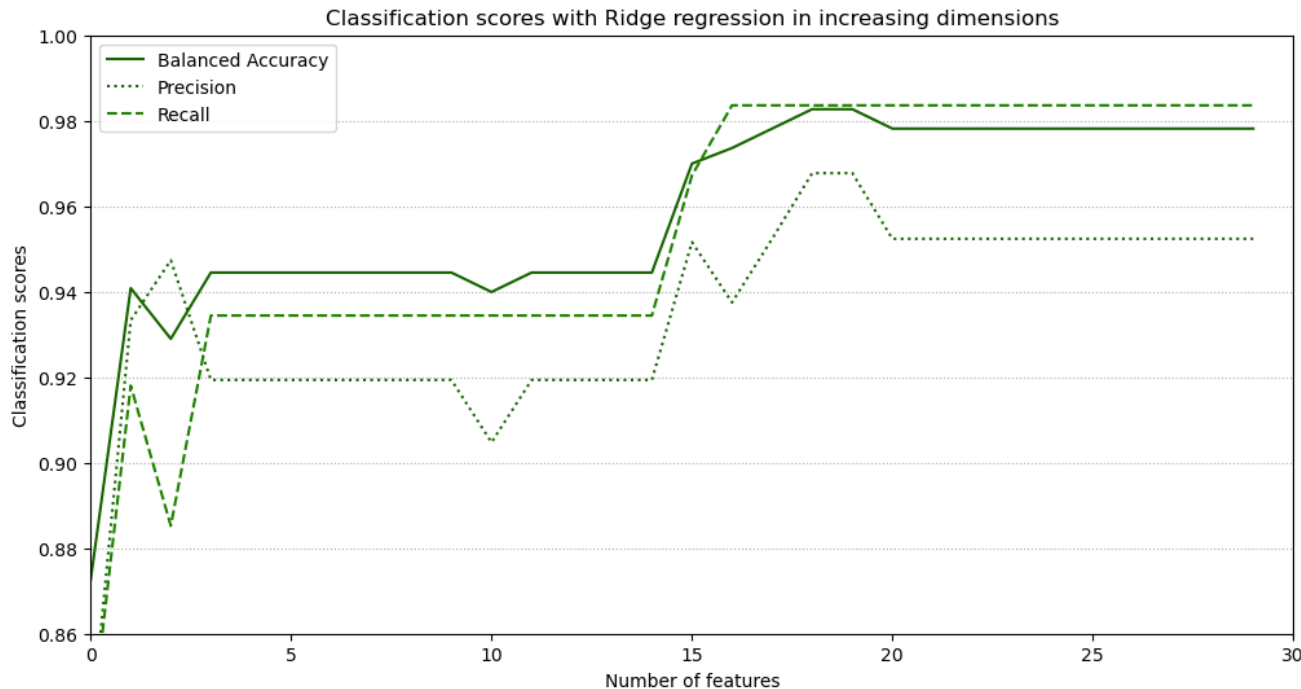
The baseline logistic regression model does not penalize the coefficients

While we pass more and more features into the model, the **classification scores**

- improve up until approx. 5 features
- decrease starting from approx. 20 features

Hence, after the 20th dimension we observe the **model suffering from the curse of dimensionality**

Regularizing with Ridge regression does reduce effect of the curse of dimensionality



Introducing Ridge regression into our baseline model **penalizes coefficients**

While we pass more and more features into the model, the **classification scores**

- keep improving until the 20th dimension
- and manages to retain classification scores beyond

Hence, when regularizing with Ridge regression, we do not observe the model suffering from the curse of dimensionality anymore

We are choosing to look at Recall as we think this metric is most important for doctors

Definition of recall

"Out of all the positive examples, how many are predicted as positive?"

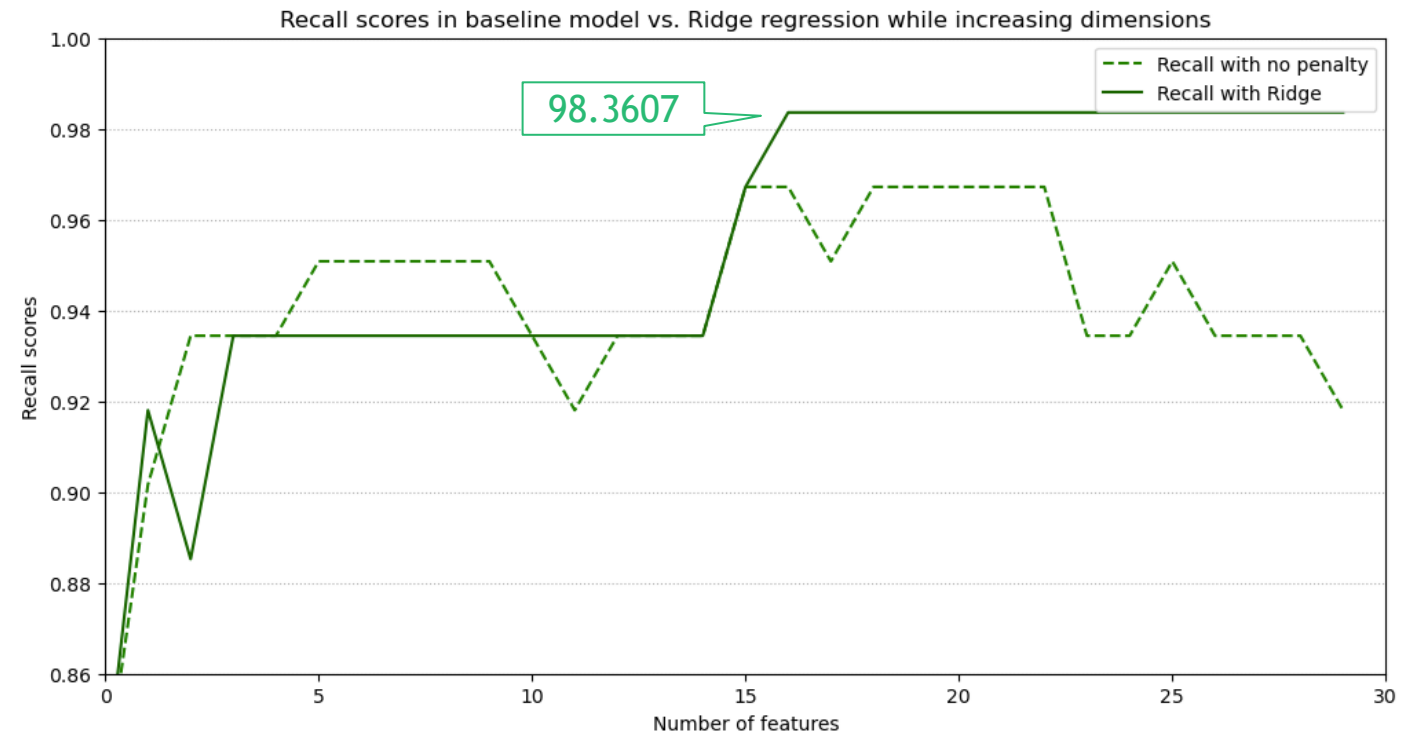
$$\text{recall} = \frac{tp}{tp + fn}$$



For this business problem, we choose to primarily compare recall to minimize the number of patients who do have breast cancer, but it goes undetected

Hence, we prefer false positives over false negatives

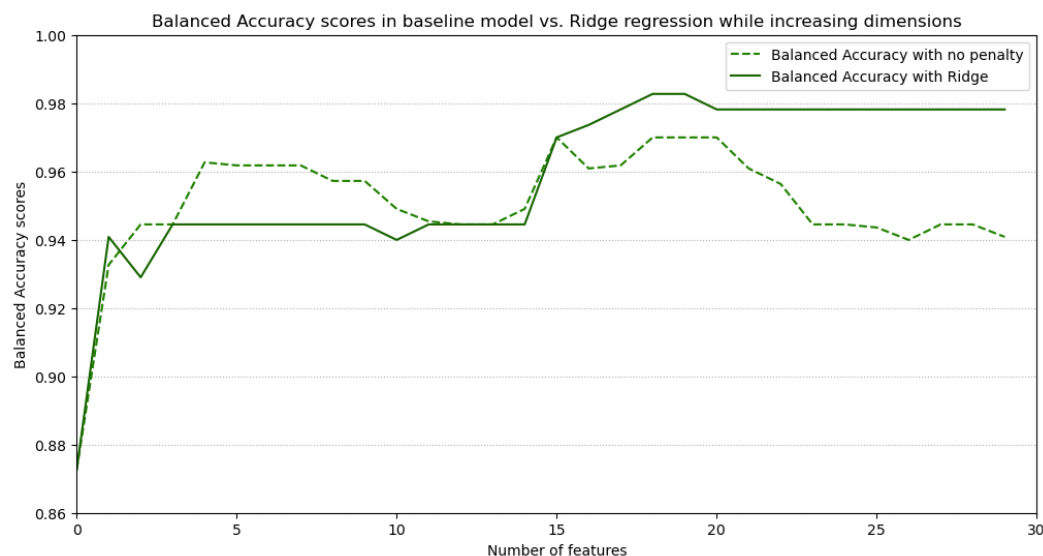
Comparing Recall from both models, we can more clearly see the success in addressing the curse of dimensionality



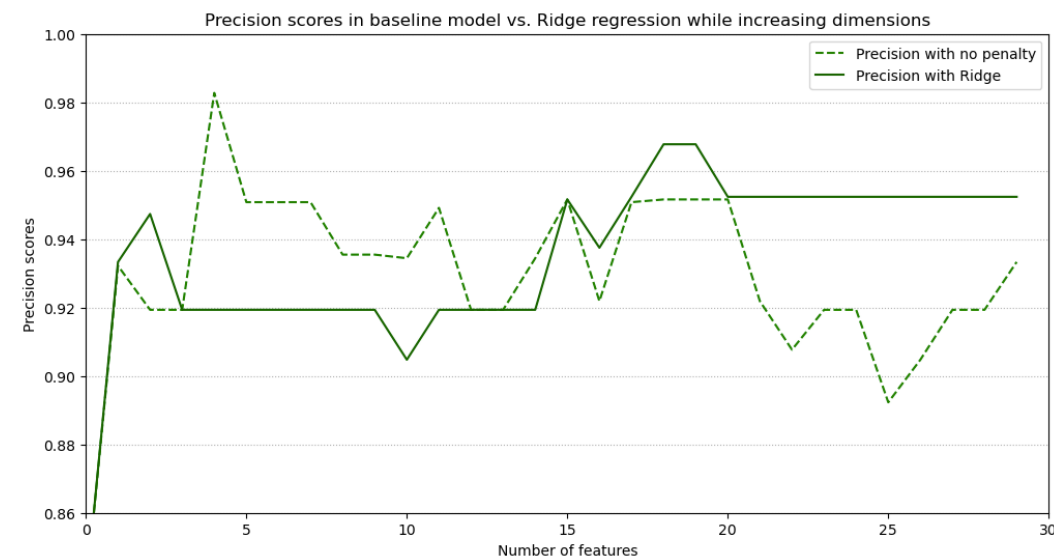
- Looking at Recall, the logistic regression which regularizes with **Ridge does perform better than the logistic regression with no penalty** on coefficients
- Starting from and **beyond the 15th dimension**, regularization does help the model to quite accurately predict whether a patient has breast cancer or not

We yield the same results from looking at Balanced Accuracy and Precision

Observing the curse of dimensionality with Balanced Accuracy starting from the 15th dimension



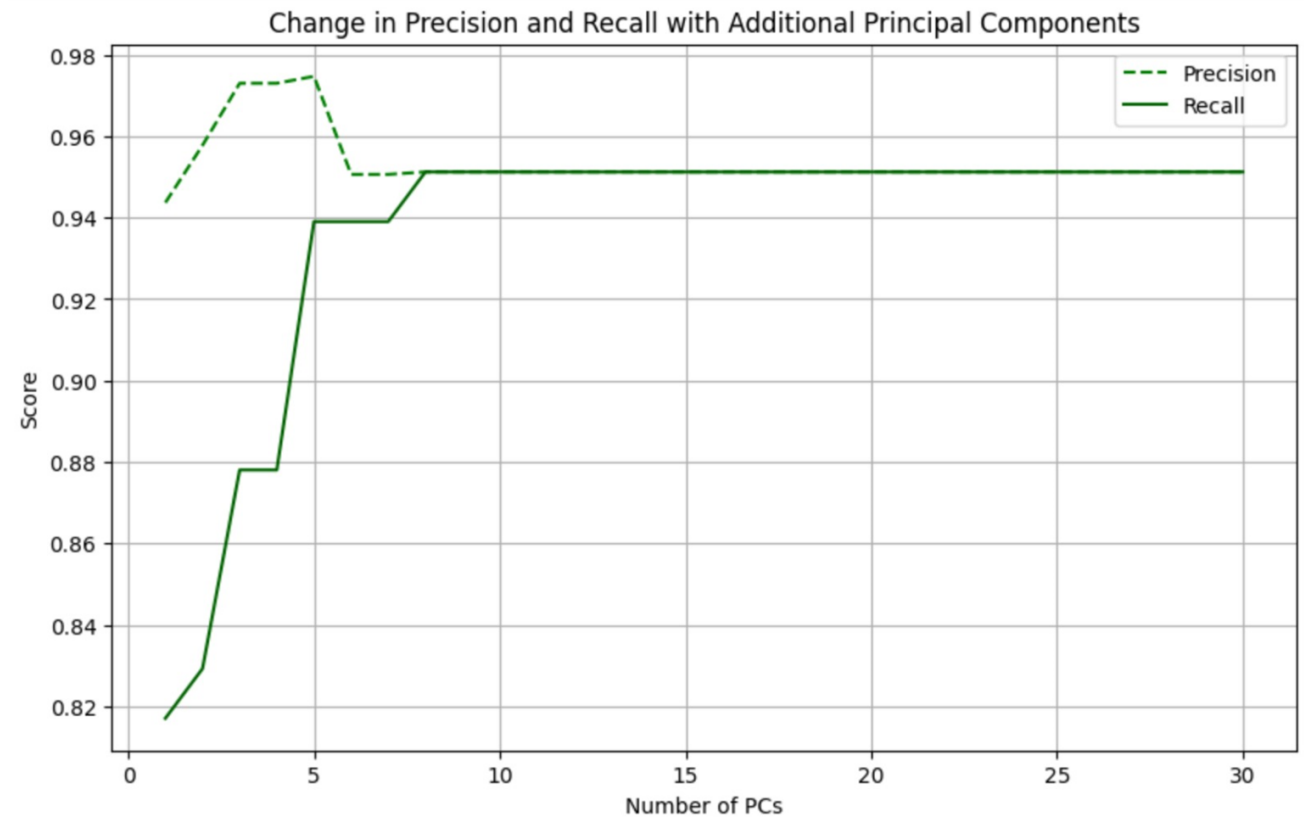
Looking at Precision, the curse of dimensionality starts around the 20th dimension



The PCA method is also successful in addressing the curse of dimensionality

Conducting the PCA method to address the curse of dimensionality,

- PCA contains information from all 30 columns
- we see that Precision and Recall scores stay constant at a high score of approx. 95 after the 8th principal component
- we interpret that it sufficient to extract the first 8 principal components to sufficiently capture the variation of the data



A photograph of surgeons in an operating room, overlaid with a semi-transparent teal filter. The central surgeon is wearing a surgical cap, mask, and glasses, and is holding a surgical instrument. Other surgeons are visible in the background, also in surgical attire. The text "K-nearest neighbors" is centered over the image.

K-nearest neighbors

The baseline KNN model also suffers the curse of dimensionality

Best K values when dimensions are higher

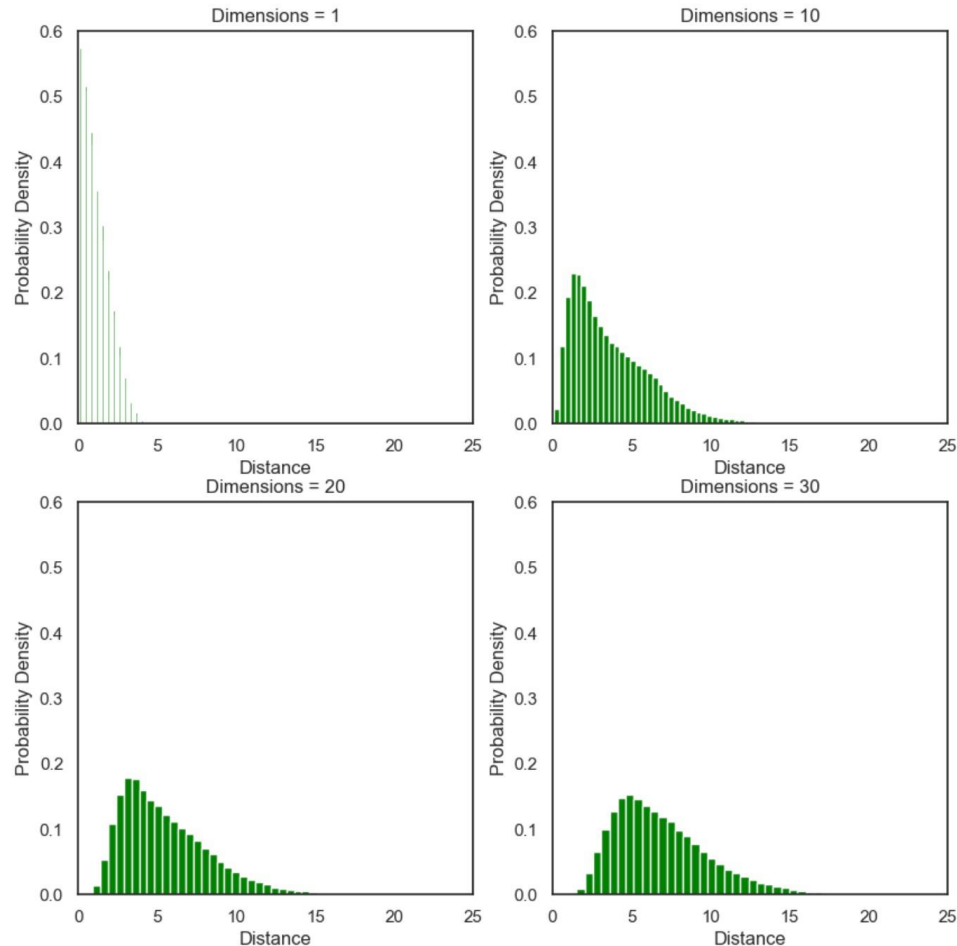
	Number of Dimensions	Best K	Recall Score
0	1	1	0.854042
1	2	1	0.900332
2	3	5	0.905316
3	4	13	0.914950
4	5	3	0.933666
5	6	3	0.938427
6	7	3	0.933776
7	8	3	0.929014
8	9	3	0.929014
9	10	5	0.929014
10	11	3	0.919491
11	12	3	0.919601
12	13	3	0.900664
13	14	3	0.895903
14	15	3	0.905316
15	16	1	0.938538
16	17	1	0.966777
17	18	1	0.952602
18	19	3	0.938317
19	20	1	0.938427
20	21	1	0.947841
21	22	1	0.943079
22	23	1	0.943079
23	24	1	0.943079
24	25	1	0.943079
25	26	1	0.947841
26	27	1	0.938427
27	28	1	0.943079
28	29	3	0.938317
29	30	3	0.938427

Model performance under different dimensions



- Adding dimensions induces noisy and less informative data, leading to a temporary decrease in performance
- Correlation strength may not guarantee all features are equally relevant

In higher dimensions, the data points are located further apart from each other

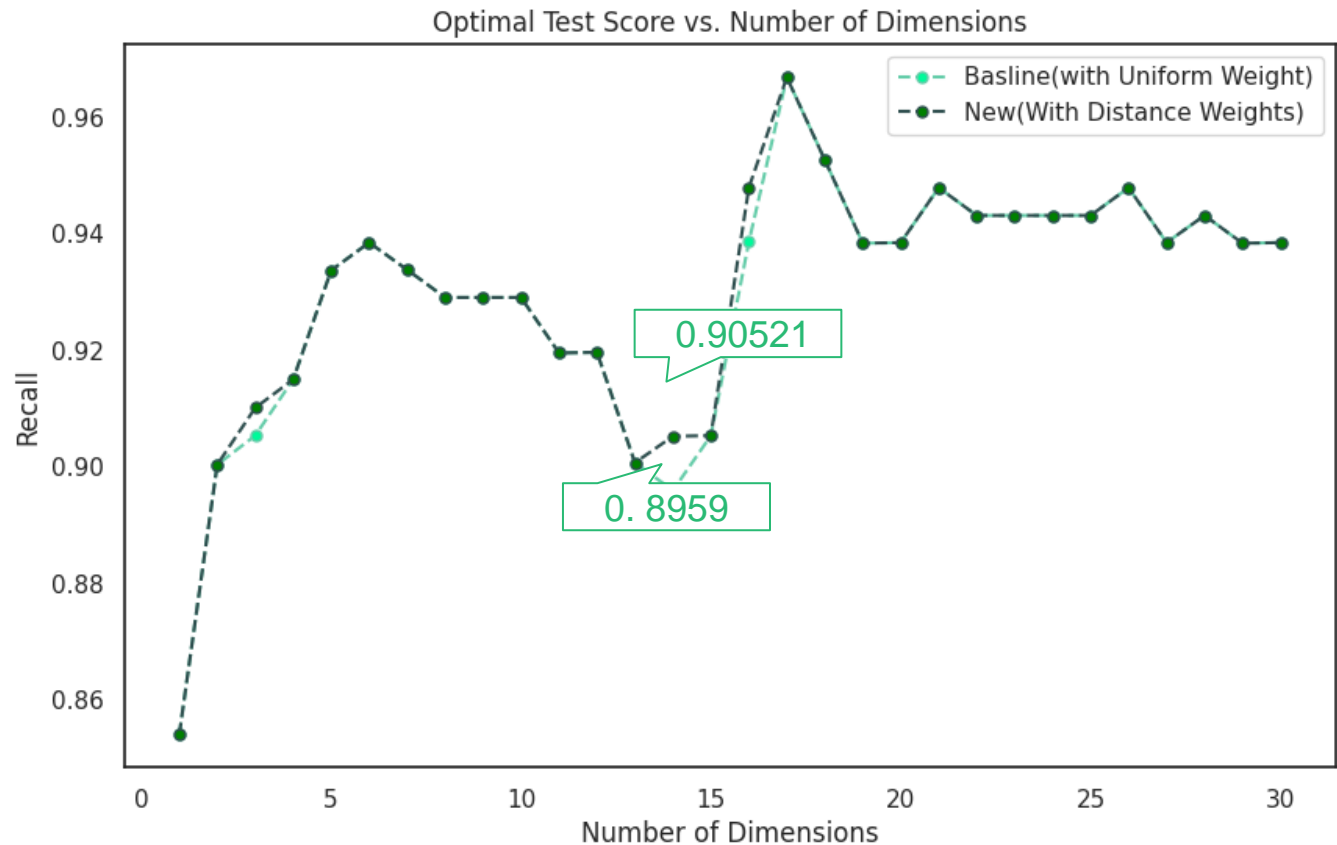


When calculating the **distance between the data points to their neighbors**

- we see that the **distance between them increases** as we move into higher dimensions.
- In one dimension, we see that most distances are very small (approx. 0)
- In 30th dimension, the mean of distances is around 5 and 6

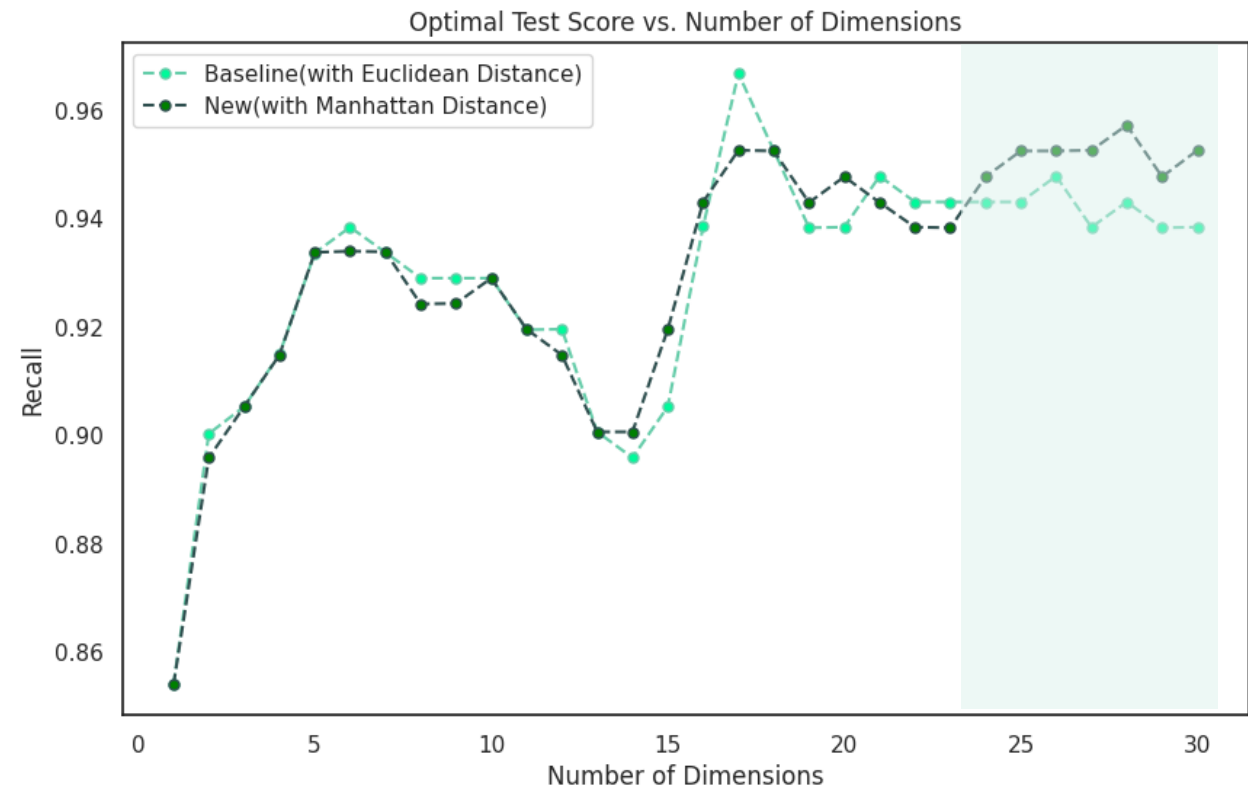
As the distance between data points increases in higher dimensions, it will become **more difficult for the KNN model to find the nearest neighbor**

To address the curse of dimensionality, we change the weight function from uniform to distance which slightly improves Recall

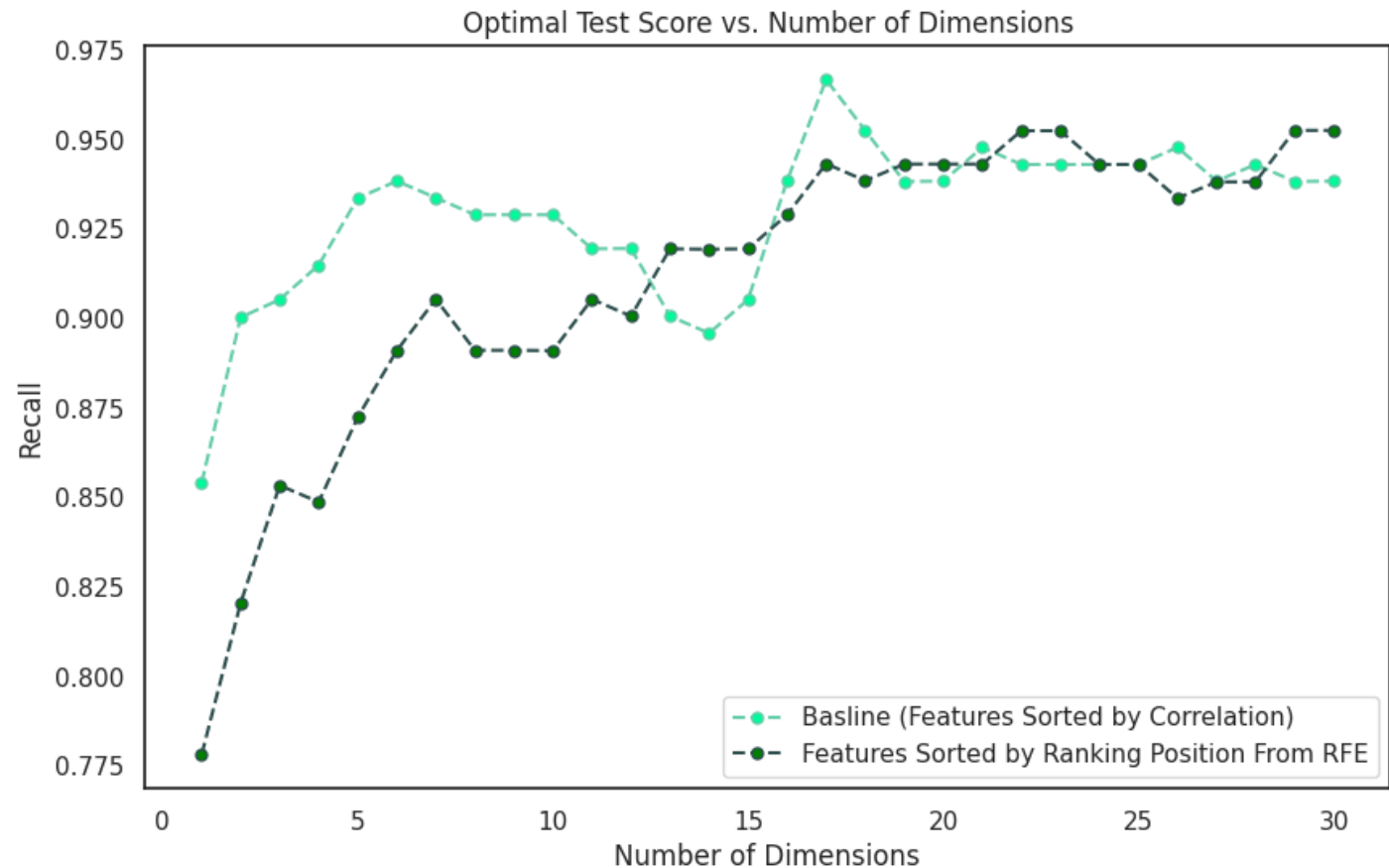


- At first, all points in each neighborhood were weighted equally as the weights are uniform
- With the **distance** weight function, **closer neighbors** of a data point will have a **greater influence** than neighbors further away, hence improving the prediction performance
- This can solely enhance few of the test scores

Manhattan distance
yields higher Recall
than Euclidean
distance when the
dimension increases



Sorting features by ranking with Recursive Feature Elimination (RFE) effectively mitigates the curse of dimensionality



- All the features are ranked with recursive feature elimination using Logistic Regression as the estimator
- This method identifies the most important features effectively, which enables the recall score to keep increasing even in higher dimensions
- On the other hand, ordering features by their correlation strength to the target introduces noise probably because only linearity is assumed

A photograph of surgeons in an operating room, overlaid with a semi-transparent teal filter. The central surgeon is wearing a surgical cap, mask, and large magnifying glasses, holding a surgical instrument. Other surgeons are visible in the background, also in scrubs and caps. The word "Conclusion" is centered in white text.

Conclusion

To predict breast cancer, we conclude that it is better to use logistic regression rather than KNN because ...

Comparing logistic regression with KNN

- 1 **logistic regression performs better** for most N features
- 2 logistic regression is more flexible as **regularization is an option**, which cannot be conducted for KNN

When feature selection is possible

- we would encourage the **selection of the first N parameters** before recall starts to fall
- LASSO regularization was thus not utilized, as **we acknowledge that in some scenarios:**
 - dimensional reduction may not be a possibility
 - feature selection may have already been done
 - certain features may have to be included due to domain specific knowledge