SVM

Statistical Prediction Modeling *DATA 2204*

Jumana Obeid

100832624

Assignment #5
6th of April 2022

DURHAM COLLEGE
SUCCESS MATTERS

b. Rational Statement (summary of the problem or problems to be addressed by the PPT) – 2%

We would like to classify -based on the features provided if this kind of cancer is benign or malignant. For that, we are using the Support Vector Machines Algorithm.SVM performs classification by finding the hyperplane that maximizes the margin between the two classes.

The independent variables(features):

1. Clump Thickness – 1-10
2. UofCSize - Uniformity of Cell Size 1-10
3. UofShape - Uniformity of Cell Shape 1-10
4. Marginal Adhesion - 1-10
5. SECSize - Single Epithelial Cell Size 1-10
6. Bare Nuclei - 1-10
7. Bland Chromatin - 1-10
8. Normal Nucleoli - 1-10
9. Mitoses - 1-10

Dependent Variable Class –

1. Benign (i.e. No Cancer) - 2,
2. Malignant (i.e. Cancer) - 4

c. Present the Learning Curve for the Original SVM Model and explain three (3) insights – 3%

```
Estimator: SVM
[[87  2]
 [ 0 48]]
              precision    recall  f1-score   support

           2       1.00      0.98      0.99        89
           4       0.96      1.00      0.98        48

    accuracy                           0.99       137
   macro avg       0.98      0.99      0.98       137
weighted avg       0.99      0.99      0.99       137
```

1. The accuracy of the module .99 which is almost perfect, based on F1 which is the harmonic mean between precision and recall.

2. Precision, is .99 , almost perfect , precision is out of all predictions we got 99% right.

3. Recall is .99 almost perfect , recall is out of all truths we got 99% right.

d. Present and explain three (3) key insights from the Optimized SVM model classification report, but first use SMOTE to balance the Classes. – 7%

```
Optimized Model

Model Name: SVC()

Best Parameters: {'clf__C': 100, 'clf__gamma': 0.01, 'clf__kernel': 'poly'}

 [[87  2]
 [ 0 48]]

              precision    recall  f1-score   support

           2       1.00      0.98      0.99        89
           4       0.96      1.00      0.98        48

    accuracy                           0.99       137
   macro avg       0.98      0.99      0.98       137
weighted avg       0.99      0.99      0.99       137
```

1. The best kernel method is polynomial Kernel .Which creates non-linear kernel, it generates new features by applying the polynomial combination of all the existing features.
2. The best c is 100.The penalty to miss classifying the data is moderate ,in other words it is ok to miss classify some points.
3. The best gamma is .01, low gamma is preferred that the decision boundary is not so curvy by that it is reducing variance.(reducing overfitting)

e. State and explain three (3) recommendations for Mr. John Hughes for next steps. – 3%

We recommend the following:

1. Add more data to reduce the problem of overfitting.
2. Remove less important features.
3. Increase regularization SVM. Higher the C parameter