Breast Cancer Data set – Appling Machine Learning Algorithms

# Tala Owais 20198048

Introduction

**In this report For the Decision Tree max width have been set to 2 -in case feature selection has been applied – otherwise 4 when feature selection is not used .Entropy has been used as the division criteria .As for the SVM the best features for gamma and C have been found using the GridSearchCV and the rbf kernel is used since the data is not linearly separable .As For KNN 5 neighbors have used to model the data. Finally , the Sequential ANN , I have used 2 hidden layers as it has proved that it works better than one layer in this situation by trial and error , the first hidden layer has 12 neurons and the second layer has neurons as much as the selected features numbers .the activation function for the first hidden layer is relu as it does not activate all the neurons at the same time.The activation function for the second hidden layer is relu and the activation function for the output layer is sigmoid since this is a binary classification problem .**

**Five Preprocessing Experiments have been conducted in this project .The first experiment does all the previously conducted preprocessing steps in addition to scaling the dataset before modeling the ANN. The second experiment does the same as the first one except that the oversampling step has been omitted. The third experiment is the same as the second but the Missing values has been dropped not imputed.In the fourth experiment , is the same as the third experiment except that the oversampling step has been conducted .Experiment no. five ,All the preprocessing steps done in experiment four but with the feature selection step omitted.Table I. shows the different preprocessing steps done in each experiment. The second section of the paper discusses the algorithms results in each experiment .The third section of the project paper discusses the conclusion.**

**Table I.** PREPROCESSING STEPS IN EACH EXPERIMENT

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Experiment 1** | **Experiment2** | **Experiment3** | **Experiment4** | **Experiment5** |
| 1. Features Encoding  2. Missing Values **-KNN Imputer**  3. train/test split 0.66 -0.34  **4. Feature Selection**  **5.OverSampling –SMOTENC**  6. Scaling xtrain and xtest sets to model the ANN. | 1. Features Encoding  2. Missing Values -**KNN Imputer**  3. train/test split 0.66 -0.34  **4. Feature Selection**  6. Scaling xtrain and xtest sets to model the ANN. | 1. Features Encoding  2. Missing Values –**Dropping Missing values**  3. train/test split 0.66 -0.34  **4. Feature Selection**  6. Scaling xtrain and xtest sets to model the ANN. | 1. Features Encoding  2. Missing Values –**Dropping Missing values**  3. train/test split 0.66 -0.34  4. **Feature Selection**  **5.OverSampling –SMOTENC**  6. Scaling xtrain and xtest sets to model the ANN. | 1. Features Encoding  2. Missing Values –**Dropping Missing values**  3. train/test split 0.66 -0.34  **5.OverSampling –SMOTENC**  6. Scaling xtrain and xtest sets to model the ANN. |

Section2:ALGORITHMS RESULTS:

**In this section we see the different algorithms we applied and their results .The algorithms with the best performance was decided depending on the Sensitivity (Recall) , Specificity , average precision as these measures are the most important measures in the medical field [1].**

2.1 Support vECTOR MACHINE:

**Using the above mentioned five experiments , experiment 5 have shown the sensitivity (recall) ,specificity and average precision results .Table.II shows all experiments results for the SVM models.**

**Table II.** SVM RESULTS IN THE FIVE EXPERIMENTS

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Precision | Recall | f1-score | support | | 'no-recurrence-events' | 0.82716 | 0.858974 | 0.842767 | 78 | | 'recurrence-events' | 0.352941 | 0.3 | 0.324324 | 20 | | Accuracy | 0.744898 | 0.744898 | 0.744898 | 0.744898 | | macro avg | 0.590051 | 0.579487 | 0.583546 | 98 | | weighted avg | 0.730381 | 0.744898 | 0.736963 | 98 |  1. Experiment 1 Results | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.795918 | 1 | 0.886364 | 78 | | 'recurrence-events' | 0 | 0 | 0 | 20 | | Accuracy | 0.795918 | 0.795918 | 0.795918 | 0.795918 | | macro avg | 0.397959 | 0.5 | 0.443182 | 98 | | weighted avg | 0.633486 | 0.795918 | 0.705473 | 98 |   b.Experiment 1 Results |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Precision | Recall | f1-score | support | | 'no-recurrence-events' | 0.705263 | 1 | 0.82716 | 67 | | 'recurrence-events' | 0 | 0 | 0 | 28 | | Accuracy | 0.705263 | 0.705263 | 0.705263 | 0.705263 | | macro avg | 0.352632 | 0.5 | 0.41358 | 95 | | weighted avg | 0.497396 | 0.705263 | 0.583366 | 95 |   c. Experiment 3 RESULTS | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.8 | 0.895522 | 0.84507 | 67 | | 'recurrence-events' | 0.65 | 0.464286 | 0.541667 | 28 | | Accuracy | 0.768421 | 0.768421 | 0.768421 | 0.768421 | | macro avg | 0.725 | 0.679904 | 0.693369 | 95 | | weighted avg | 0.755789 | 0.768421 | 0.755646 | 95 |   d. experiment 4 results |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | Recall | f1-score | support | | 'no-recurrence-events' | 0.859375 | **0.820896**  **(Specificity)** | 0.839695 | 67 | | 'recurrence-events' | 0.612903 | **0.678571**  **(Sensitivity)** | 0.644068 | 28 | | Accuracy | 0.778947 | 0.778947 | 0.778947 | 0.778947 | | macro avg | 0.736139 | 0.749733 | 0.741881 | 95 | | weighted avg | **0.786731** | 0.778947 | 0.782036 | 95 |   E. experiment 5 results | |

2.2 Decision TREE:.

**Using the above mentioned five experiments ,experiment 4 have shown the sensitivity (recall) ,specify and average precision results .Table.III shows all experiments results for the decision tree.**

**Table III.** Decision Tree RESULTS IN THE FIVE EXPERIMENTS

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.819277 | 0.871795 | 0.84472 | 78 | | 'recurrence-events' | 0.333333 | 0.25 | 0.285714 | 20 | | accuracy | 0.744898 | 0.744898 | 0.744898 | 0.744898 | | macro avg | 0.576305 | 0.560897 | 0.565217 | 98 | | weighted avg | 0.720105 | 0.744898 | 0.730638 | 98 |  1. Experiment 1 Results | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Precision | recall | f1-score | support | | 'no-recurrence-events' | 0.824176 | 0.961538 | 0.887574 | 78 | | 'recurrence-events' | 0.571429 | 0.2 | 0.296296 | 20 | | Accuracy | 0.806122 | 0.806122 | 0.806122 | 0.806122 | | macro avg | 0.697802 | 0.580769 | 0.591935 | 98 | | weighted avg | 0.772595 | 0.806122 | 0.766905 | 98 |   b.Experiment 2 Results |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.755814 | 0.970149 | 0.849673 | 67 | | 'recurrence-events' | 0.777778 | 0.25 | 0.378378 | 28 | | accuracy | 0.757895 | 0.757895 | 0.757895 | 0.757895 | | macro avg | 0.766796 | 0.610075 | 0.614026 | 95 | | weighted avg | 0.762288 | 0.757895 | 0.710765 | 95 |   c. Experiment 3 RESULTS | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Precision | recall | f1-score | support | | 'no-recurrence-events' | 0.820896 | **0.820896**  **(Specificity)** | 0.820896 | 67 | | 'recurrence-events' | 0.571429 | **0.571429**  **(Sensitivity)** | 0.571429 | 28 | | Accuracy | 0.747368 | 0.747368 | 0.747368 | 0.747368 | | macro avg | 0.696162 | 0.696162 | 0.696162 | 95 | | weighted avg | **0.747368** | 0.747368 | 0.747368 | 95 |   d. experiment 4 results |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | Support | | 'no-recurrence-events' | 0.776316 | 0.880597 | 0.825175 | 67 | | 'recurrence-events' | 0.578947 | 0.392857 | 0.468085 | 28 | | accuracy | 0.736842 | 0.736842 | 0.736842 | 0.736842 | | macro avg | 0.677632 | 0.636727 | 0.64663 | 95 | | weighted avg | 0.718144 | 0.736842 | 0.719927 | 95 |   E. experiment 5 results | |

2.3 K-NEAREST NEIGHBORS:

**Using the above mentioned five experiments ,experiment 5 have shown the best sensitivity (recall) ,specificity and average precision results . Table.IV shows all experiments results for K nearest neighbors.**

**Table IV.** K NEAREST NEIGHBORS RESULTS IN THE FIVE EXPERIMENTS

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.835821 | 0.717949 | 0.772414 | 78 | | 'recurrence-events' | 0.290323 | 0.45 | 0.352941 | 20 | | accuracy | 0.663265 | 0.663265 | 0.663265 | 0.663265 | | macro avg | 0.563072 | 0.583974 | 0.562677 | 98 | | weighted avg | 0.724495 | 0.663265 | 0.686807 | 98 |  1. Experiment 1 Results | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.807229 | 0.858974 | 0.832298 | 78 | | 'recurrence-events' | 0.266667 | 0.2 | 0.228571 | 20 | | Accuracy | 0.72449 | 0.72449 | 0.72449 | 0.72449 | | macro avg | 0.536948 | 0.529487 | 0.530435 | 98 | | weighted avg | 0.69691 | 0.72449 | 0.709089 | 98 |   b.Experiment 2 Results |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.8 | 0.955224 | 0.870748 | 67 | | 'recurrence-events' | 0.8 | 0.428571 | 0.55814 | 28 | | accuracy | 0.8 | 0.8 | 0.8 | 0.8 | | macro avg | 0.8 | 0.691898 | 0.714444 | 95 | | weighted avg | 0.8 | 0.8 | 0.778611 | 95 |   c. Experiment 3 RESULTS | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.815385 | 0.791045 | 0.80303 | 67 | | 'recurrence-events' | 0.533333 | 0.571429 | 0.551724 | 28 | | Accuracy | 0.726316 | 0.726316 | 0.726316 | 0.726316 | | macro avg | 0.674359 | 0.681237 | 0.677377 | 95 | | weighted avg | 0.732254 | 0.726316 | 0.728961 | 95 |   d. experiment 4 results |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | Recall | f1-score | support | | 'no-recurrence-events' | 0.854545 | **0.701493**  **(Specificity)** | 0.770492 | 67 | | 'recurrence-events' | 0.5 | **0.714286**  **(Sensitivity)** | 0.588235 | 28 | | accuracy | 0.705263 | 0.705263 | 0.705263 | 0.705263 | | macro avg | 0.677273 | 0.707889 | 0.679364 | 95 | | weighted avg | **0.750048** | 0.705263 | 0.716774 | 95 |   E. experiment 5 results | |

2.4 Artificial Neural NETWORk:

**Using the above mentioned five experiments ,experiment 4 have shown the sensitivity (recall) ,specify and average precision results .Table.V shows all experiments results for the decision tree.**

**Table V.** ANN RESULTS IN THE FIVE EXPERIMENTS

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.887097 | **0.705128**  **(Specificity)** | 0.785714 | 78 | | 'recurrence-events' | 0.361111 | **0.65**  **(Sensitivity)** | 0.464286 | 20 | | accuracy | 0.693878 | 0.693878 | 0.693878 | 0.693878 | | macro avg | 0.624104 | 0.677564 | 0.625 | 98 | | weighted avg | **0.779753** | 0.693878 | 0.720117 | 98 |  1. Experiment 1 Results | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.841463 | 0.884615 | 0.8625 | 78 | | 'recurrence-events' | 0.4375 | 0.35 | 0.388889 | 20 | | Accuracy | 0.77551 | 0.77551 | 0.77551 | 0.77551 | | macro avg | 0.639482 | 0.617308 | 0.625694 | 98 | | weighted avg | 0.759022 | 0.77551 | 0.765845 | 98 |   b.Experiment 2 Results |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.747126 | 0.970149 | 0.844156 | 67 | | 'recurrence-events' | 0.75 | 0.214286 | 0.333333 | 28 | | accuracy | 0.747368 | 0.747368 | 0.747368 | 0.747368 | | macro avg | 0.748563 | 0.592217 | 0.588745 | 95 | | weighted avg | 0.747973 | 0.747368 | 0.693598 | 95 |   c. Experiment 3 RESULTS | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.782609 | 0.80597 | 0.794118 | 67 | | 'recurrence-events' | 0.5 | 0.464286 | 0.481481 | 28 | | Accuracy | 0.705263 | 0.705263 | 0.705263 | 0.705263 | | macro avg | 0.641304 | 0.635128 | 0.6378 | 95 | | weighted avg | 0.699314 | 0.705263 | 0.701972 | 95 |   d. experiment 4 results |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | f1-score | support | | 'no-recurrence-events' | 0.777778 | 0.626866 | 0.694215 | 67 | | 'recurrence-events' | 0.390244 | 0.571429 | 0.463768 | 28 | | accuracy | 0.610526 | 0.610526 | 0.610526 | 0.610526 | | macro avg | 0.584011 | 0.599147 | 0.578991 | 95 | | weighted avg | 0.663557 | 0.610526 | 0.626294 | 95 |   E. experiment 5 results | |

SECTION3: Conclusion:

**I have conducted various experiments , accordingly have found that the SVM, KNN performances in the fifth experiment with this binary classification problem was the best .**

Appendix

**Attached with this document you will find the code and the applying algorithms phase related figures and tables.**

Referrences

[1] <https://medium.com/@alon.lek/should-i-look-at-precision-recall-or-specificity-sensitivity-3946158aace1>