Anomaly Detection – Multivariate data

**Use case 1:**

Breast cancer dataset from UCI.

Total number of observations – 468 observations

Total number of columns – 11 columns

In Class column, If the value is 2, it is benign, If value is 4, it is malignant.

Data contains 458 cases benign and 10 cases malignant.

**Solution:**

**Approach 1 – Local Outlier Factor**

Package – DMwR

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted (Benign) | Predicted (Malignant) |
| Actual (Benign) | 392 | 66 |
| Actual (Malignant) | 5 | 5 |

Accuracy = (392+5)/468 = 0.85

Misclassification – (66+5)/468 = 0.15

False positive(Type I) - 66/458 = 0.14

False Negative(Type II) – 5/10 = 0.5

**Approach 2 – k-means Clustering**

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted (Benign) | Predicted (Malignant) |
| Actual (Benign) | 441 | 17 |
| Actual (Malignant) | 0 | 10 |

Accuracy = (441+10)/468 = 0.96

Misclassification – (17+0)/468 = 0.03

False positive(Type I) - 17/458 = 0.03

False Negative(Type II) – 0/10 = 0

**Approach 3 – One class SVM**

Package – e1071

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted (Benign) | Predicted (Malignant) |
| Actual (Benign) | 406 | 52 |
| Actual (Malignant) | 0 | 10 |

Accuracy = (406+10)/468 = 0.88

Misclassification – (52+0)/468 = 0.11

False positive(Type I) - 52/456 = 0.11

False Negative(Type II) – 0/10 = 0

**Approach 4 – Mahanalobis distance**

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted (Benign) | Predicted (Malignant) |
| Actual (Benign) | 438 | 20 |
| Actual (Malignant) | 2 | 8 |

Accuracy = (438+8)/468 = 0.95

Misclassification – (20+2)/468 = 0.05

False positive(Type I) - 20/458 = 0.04

False Negative(Type II) – 2/10 = 0.2

**Approach 5 – using Outliers function**

**Package – outliers**

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted (Benign) | Predicted (Malignant) |
| Actual (Benign) | 455 | 3 |
| Actual (Malignant) | 4 | 6 |

Accuracy = (455+6)/468 = 0.985

Misclassification – (3+4)/468 = 0.01

False positive(Type I) - 3/458 = 0.006

False Negative(Type II) – 4/10 = 0.4

**Approach 6 – Using z-score**

|  |  |  |
| --- | --- | --- |
|  | Predicted (Benign) | Predicted (Malignant) |
| Actual (Benign) | 429 | 29 |
| Actual (Malignant) | 1 | 9 |

Accuracy = (429+9)/468 = 0.935

Misclassification – (29+1)/468 = 0.06

False positive(Type I) - 29/458 = 0.06

False Negative(Type II) – 1/10 = 0.1