**Cairo University**

**Faculty of Computers and Artificial Intelligence**

**Machine Learning**

**Assignment 2 Report**

|  |  |
| --- | --- |
| **Team Members** | **IDS** |
| **Ahmed Saad Ahmed** | **20210020** |
| **Maryam Osama Mohamed** | **20211090** |
| **Shahd Mostafa Mahmoud** | **20211054** |
| **Shahd Osama Hamed** | **20211052** |
| **Seif Ibrahim Ramadan** | **20210164** |

Task 2: Implement Decision Tree, k-Nearest Neighbors (kNN) and naïve Bayes

1. Using scikit-learn implement Decision Tree, kNN and Naïve Bayes
2. Compare the performance of your implementations by evaluating accuracy, precision, and recall metrics.

|  |  |  |  |
| --- | --- | --- | --- |
|  | kNN(k=5) | Decision Tree | Naïve Bayes |
| accuracy | 97.20% |  |  |
| precision | 92.00% |  |  |
| recall | 82.14% |  |  |

1. Implement k-Nearest Neighbors (kNN) algorithm from scratch.
2. Report the results and compare the performance of your custom kNearest Neighbors (kNN) implementation with the pre-built kNN algorithms in scikit-learn, using the evaluation metrics mentioned in point 2. Using any missing handling techniques, you chose from task 1.2.

|  |  |  |
| --- | --- | --- |
|  | kNN(k=5) scikit-learn | kNN(k=5) from scratch |
| accuracy | 97.20% | 97.20% |
| precision | 92.00% | 92.00% |
| recall | 82.14% | 82.14% |

Task 3: Interpreting the Decision Tree and Evaluation Metrics Report

1. The effect of different data handling

* Provide a detailed report evaluating the performance of scikitlearn implementations of the Decision Tree, k-Nearest Neighbors (kNN) and naïve Bayes with respect to the different handling missing data technique.
* dropping missing values

|  |  |  |  |
| --- | --- | --- | --- |
|  | kNN(k=5) | Decision Tree | Naïve Bayes |
| accuracy | 95.53% |  |  |
| precision | 87.30% |  |  |
| recall | 80.88% |  |  |

* replacing them with the average

|  |  |  |  |
| --- | --- | --- | --- |
|  | kNN(k=5) | Decision Tree | Naïve Bayes |
| accuracy | 97.20% |  |  |
| precision | 92.00% |  |  |
| recall | 82.14% |  |  |

1. Decision Tree Explanation Report

* Create a well-formatted report that includes a plot of the decision tree and a detailed explanation of how the tree makes predictions.
* Discuss the criteria and splitting logic used at each node of the tree.

1. Performance Metrics Report

* Provide a detailed report evaluating the performance of your implementations of the k-Nearest Neighbors (kNN) from scratch with different k values at least 5 values.
* Include the accuracy, precision, and recall metrics for models.
* Compare these results with the performance of the corresponding algorithms implemented using scikit-learn.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | accuracy | precision | | recall |
| KNN(3) from scratch | 96.60% | | 88.24% | 80.36% |
| KNN(5) from scratch | 97.20% | | 92.00% | 82.14% |
| KNN(7) from scratch | 96.60% | | 88.24% | 80.36% |
| KNN(9) from scratch | 97.40% | | 93.88% | 82.14% |
| KNN(11) from scratch | 97.00% | | 90.20% | 82.14% |
| KNN(5) scikit-learn | 97.20% | | 92.00% | 82.14% |
| Decision Tree |  | |  |  |
| Naïve Bayes |  | |  |  |