**Assignment - 03**



**Group Members:**

**Name Roll No**

Zuraiz Ahmed 231370227

Abdul Rehman 231370197

Ibrahim 231370216

Ahmad Ali 231370208

**Submitted To:**

Mam Sumbal Fatima

**Submission Date:**

14 - Feb - 2025

**Group Members & Their Tasks**

1. **Zuraiz Ahmed**

* Introduction to Classification and Regression: Explain what classification and regression are, and how they differ.
* Explanation of Confusion Matrix: Describe what a confusion matrix is and explain its four components (True Positives, False Positives, True Negatives, False Negatives).

1. **Abdul Rehman**

* Practical Calculation of Confusion Matrix, Accuracy, and Precision: Construct a confusion matrix and calculate accuracy and precision for the given spam email classification data.
* Explanation of Accuracy and Precision: Define accuracy and precision, and explain how they are different.

1. **Ahmad Ali**

* Explanation of Why Precision is Important: Discuss why precision is more important than accuracy in some cases, especially in spam detection.
* Impact of Misclassification: Explain how false positives and false negatives can affect a user’s experience in a spam email classifier.

1. **Ibrahim**

* Interpretation of Model Performance: Analyze whether the spam email classifier performed well based on the confusion matrix and metrics.
* Final Conclusion and Summary: Summarize the key points of the assignment and provide a final reflection on the importance of precision in classification tasks.

**Part 1: Classification and Performance Metrics**

1. **Classification**

* Classification is when a model predicts categories or labels. For example, deciding if an email is "spam" or "not spam" is a classification task.
* Regression is when a model predicts a continuous value, like predicting the price of a house or the temperature. The main difference is that classification deals with categories, while regression deals with numbers.

1. **Confusion Matrix**

* A confusion matrix is a table that helps us see how well a classification model is performing. It shows the actual vs. predicted results and helps us understand the mistakes the model is making. The four parts of a confusion matrix are:

**Four components**

* **True Positives (TP):** The model correctly predicts the positive class (e.g., correctly identifying an email as spam).
* **False Positives (FP):** The model incorrectly predicts the positive class (e.g., marking a non-spam email as spam).
* **True Negatives (TN):** The model correctly predicts the negative class (e.g., correctly identifying an email as not spam).
* **False Negatives (FN):** The model incorrectly predicts the negative class (e.g., failing to identify a spam email).

**3. Accuracy and Precision**

* Accuracy tells us how often the model is correct overall. It’s calculated as

Accuracy =

For example, if the model is correct 75% of the time, the accuracy is 75%.

* Precision tells us how often the model is correct when it predicts the positive class (e.g., spam). It’s calculated as

For example, if the model predicts 6 emails as spam and 5 of them are actually spam, the precision is 83.3%.

#### **4. Precision more important than Accuracy**

* Precision is more important than accuracy when the cost of false positives is high. For example, in spam email detection, if the model marks a legitimate email as spam (false positive), the user might miss an important email. So, it’s better to have high precision (few false positives) even if some spam emails are missed (false negatives).

**Part 2: Practical Calculation**

We are given data on actual and predicted labels for 12 emails. Our task is to create a confusion matrix and calculate accuracy and precision.

|  |  |  |
| --- | --- | --- |
| Email No. | Actual Label  (1 = Spam, 0 = Not Spam) | Predicted Label  (1 = Spam, 0 = Not Spam) |
| 1 | 1 | 1 |
| 2 | 0 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |
| 5 | 0 | 1 |
| 6 | 1 | 1 |
| 7 | 0 | 0 |
| 8 | 1 | 1 |
| 9 | 0 | 0 |
| 10 | 1 | 0 |
| 11 | 0 | 0 |
| 12 | 1 | 1 |

**Step 1: Confusion Matrix**

From the data, we count:

* **True Positives (TP):** Emails correctly predicted as spam (1).  
  TP = 5 (Emails 1, 4, 6, 8, 12)
* **False Positives (FP):** Emails incorrectly predicted as spam (1).  
  FP = 1 (Email 5)
* **True Negatives (TN):** Emails correctly predicted as not spam (0).  
  TN = 4 (Emails 2, 7, 9, 11)
* **False Negatives (FN):** Emails incorrectly predicted as not spam (0).  
  FN = 2 (Emails 3, 10)

The confusion matrix looks like this:

|  | **Predicted Spam (1)** | **Predicted Not Spam (0)** |
| --- | --- | --- |
| **Actual Spam (1)** | TP = 5 | FN = 2 |
| **Actual Not Spam (0)** | FP = 1 | TN = 4 |

**Step 2: Calculate Accuracy**

Accuracy =  = = = 0.75 OR 75%

**Step 3: Calculate Precision**

Precision= = = ≈0.833 or 83.3%

### Part 3: Interpretation & Reflection

#### **1. Model performance**

* The model did okay, with 75% accuracy and 83.3% precision. However, it missed 2 spam emails (false negatives) and marked 1 legitimate email as spam (false positive). The high precision is good, but the model could be improved to reduce mistakes.

**2. High precision or high accuracy**

* In a spam classifier, high precision is more important than high accuracy. This is because marking a legitimate email as spam (false positive) is worse than missing some spam emails (false negatives). High precision ensures that most emails marked as spam are actually spam.

**3. Misclassified email**

* **False Positive (FP):**A legitimate email marked as spam could cause the user to miss important emails, like work-related messages or personal emails. This can be frustrating.
* **False Negative (FN):** A spam email that is not marked as spam could clutter the inbox and expose the user to phishing or scams.