Machine Learning

Coursework 2: Practical Assignment Report

Q1

Model Implementation

To fit the breast cancer detection model, a Logistic Regression trained with Stochastic Gradient Descent (SGD) is used. The model predicts the type of tumours, classifying them as being either benign or malignant, based on features extracted from the Breast Cancer Wisconsin (Diagnostic) data set, which includes 30 features describing cell nuclei. The code used a sigmoid activation function to map predictions to the range [0, 1] and binary cross-entropy as the loss function to evaluate the accuracy of the prediction.

The model was trained at a learning rate of 0.0001 and iterations of 1000 since these hyperparameters were found through experimentation to give a good balance between convergence speed and stability without making the model diverge or overfit. Initialisation of weights and biases to zero, followed by iterative updating using gradients of the loss function, is done to improve the accuracy of the predictions.

Model Evaluation

The model was evaluated using accuracy, precision, recall, and F1 score metrics on the test set. The results showed:
Accuracy: 50.00% - The model correctly classified half of the test instances.
Precision: 100.00% - Every instance classified as malignant was indeed malignant, but the number of predicted positives was very low.
Recall: 20.00% - Only 20% of the actual malignant cases were correctly identified, which is concerning in medical contexts where missing malignant tumours can have serious consequences.
F1 Score: 33.33% - The low F1 score indicates an imbalance between precision and recall, suggesting insufficient effectiveness in identifying malignant cases.

Q2

Evaluation Metrics

The logistic regression model for breast cancer detection was evaluated to determine the most suitable metrics for assessing its performance, focusing on whether precision or recall should be prioritised. The risk of oversight for serious conditions such as malignant tumours must be very small in medical contexts; hence, the most relevant measure will have to be recall.

Model Evaluation

The model was evaluated using accuracy, precision, recall, and F1 score:

Accuracy: This is the ratio of correctly predicted observations to total observations. While accuracy can sometimes be informative, it is not in the case of imbalanced datasets, like this, where there are roughly 500+ benign cases versus about 100 malignant cases. The high accuracy could mean simply that it was correctly classifying the benign cases and missed many of the malignant ones. The result of the accuracy evaluation is 50.00%, which means the model correctly predicted half of the cases.

Precision refers to the number of true positives among the positive predictions that the model made, which, in this case of breast cancer detection, helps in ascertaining the model's reliability upon its prediction of a tumour being malignant. A high precision value implies there will be hardly any false positives; this will save baseless follow-up tests and reduce anxiety on the part of the patient. At the same time, with high precision alone, recall cannot be low because it is far worse to miss a malignant case. The result of the precision evaluation is 100.00%, which means all predicted malignant cases were indeed malignant, indicating no false positives.

Recall is the measure of the model's ability to correctly identify all the positive instances. It is defined as the ratio of True Positives to True Positives + False Negatives. High recall is important in medical contexts since a model needs to catch as many malignant cases as possible to make sure treatment is timely. A model with low recall will fail to detect malignant tumours, which might be disastrous for the patient. The result of the recall evaluation is 20.00%, which means the model correctly identified only 20% of the actual malignant cases, suggesting a high risk of missing malignant tumours.

F1-score: F1-score is the harmonic mean of Precision and Recall. It, therefore, gives a proper balance between the two measures. Especially, when there is an imbalance in class distribution, or the problem needs a metric that accounts for both, then F1-score becomes of most use. A low F1-score says that either Precision or Recall is way higher in magnitude as compared to the other; this says that probably a model isn't picking malignant with so high a precision value. The result of the F1 Score evaluation is 33.33%, the low F1 score reflects an imbalance between precision and recall, with the model being highly specific but not sensitive enough.

Importance of Recall in Medical Context

The primary objective, in the case of medical diagnosis, would be to minimize the possibility of overlooking critical conditions. The most dangerous situation related to this is the so-called False Negatives (Type II Errors), i.e., when the model predicts a tumour as benign while it is malignant, which is dangerous as such cases may lead to delayed or missed treatment. Recall will, therefore, be very important in this problem because the metric will tell how well a model performs in finding all malignant cases.

This is particularly disturbing because, in this case, a recall of 20% implies that 80% of malignant cases are missed. This will result in a huge number of patients with undiagnosed malignant tumours, which is unacceptable in healthcare. While precision is also important to avoid false positives and unnecessary medical procedures, the potential harm caused by false negatives far outweighs the inconvenience of false positives.

Precision vs. Recall Trade-off

The high precision at 100% indicates no misclassifications of benign tumours being malignant, so there would be no false positives. At the same time, low recall means that many positive cases are not being labelled as such, which is much worse in this context.

Precision and recall are a classic trade-off in medical diagnostics. The model that maximises recall will have more malignant cases identified, but the side effect will be more false positives, leading to more follow-up tests and procedures. On the other hand, a model tuned for maximum precision will ensure that the only types of tumours being classified as malignant are ones it is quite sure indeed contain malignancy, whereas in doing so, this system risks completely missing several cases of actual malignancies. In that case, one should look to optimise the recall, hence catching cases even at the cost of a higher percentage of false alarms.

Conclusion

Given the medical context, recall is the most critical metric for ensuring malignant cases are not missed. The current model's recall of 20.00% is inadequate for healthcare purposes despite achieving perfect precision. Lowering the decision threshold and further hyperparameter tuning will probably improve the recall. This would help strike a better balance between recall and precision, enhancing the model's reliability in medical diagnosis.

Q3

This analysis investigates the effect of hyperparameter tuning on the logistic regression model in detecting breast cancer. Learning rate and a number of iterations are two most important hyper-parameters that are varied to see their effects on precision and recall. The goal is to find optimal hyperparameters that balance model performance.

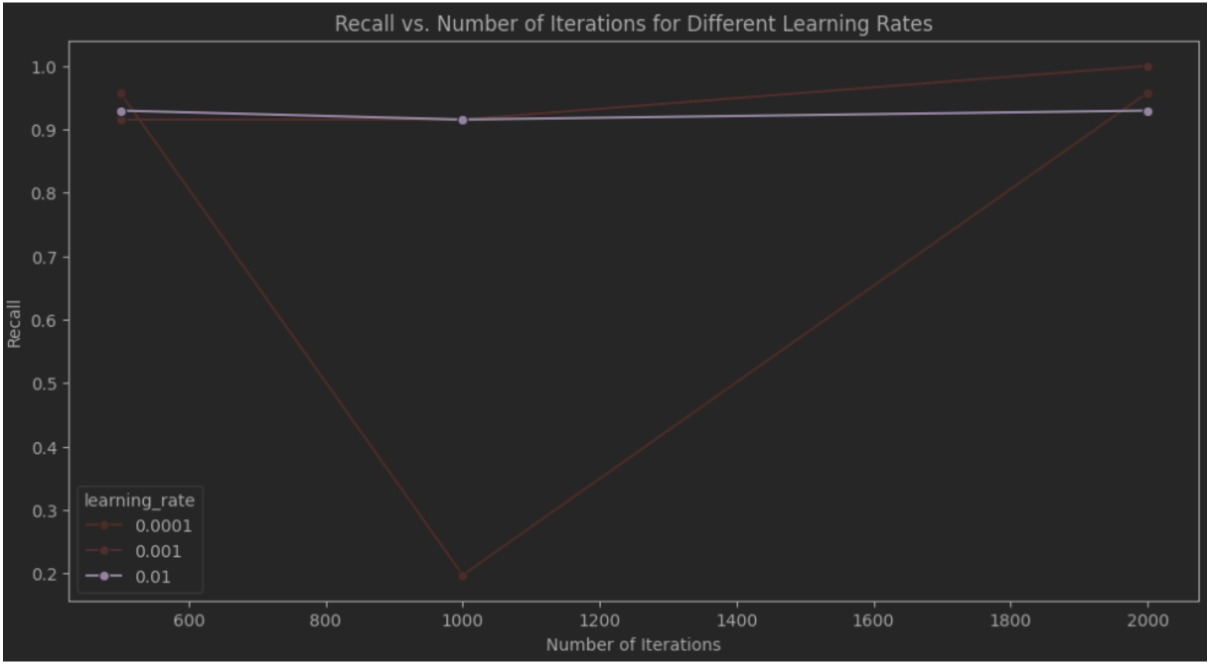
Experimental Setup and Results

The model was trained with different combinations of learning rates (0.01, 0.001, 0.0001) and iterations (500, 1000, 2000). The following table shows the precision and recall values for each combination:

Learning Rate	Iterations	Precision	Recall
0.01	500	0.985075	0.929577
0.01	1000	1.000000	0.915493
0.01	2000	0.985075	0.929577
0.001	500	0.985348	0.915493
0.001	1000	1.000000	0.915493
0.001	2000	1.000000	0.915493
0.0001	500	0.757368	1.000000
0.0001	1000	0.800000	0.197183
0.0001	2000	0.957746	0.957746

Graphical Analysis of Precision and Recall

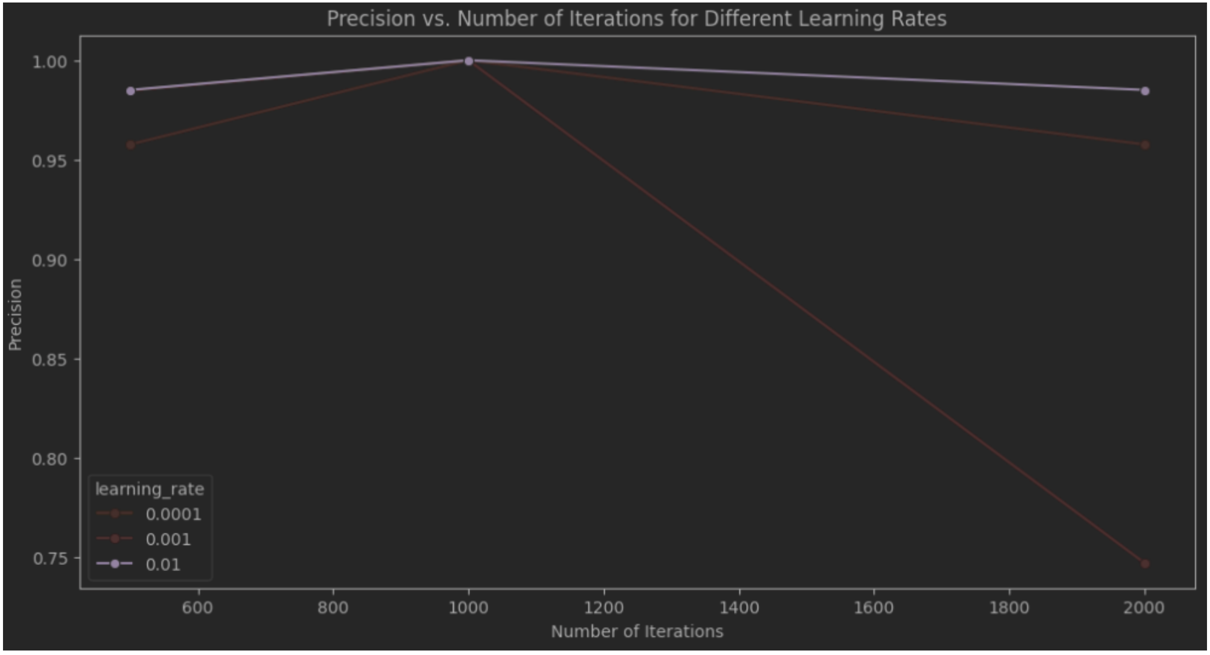
The graphs below illustrate the changes in recall and precision as a function of the number of iterations for different learning rates:



Learning Rate 0.01: Recall remains relatively stable across all iterations, with values hovering around 0.92. This indicates a consistent level of model performance, but precision reached 1.0 for 1000 iterations, indicating potential overfitting.

Learning Rate 0.001: Like in 0.01, the recall in each iteration was kept stable while the value has remained the same at 0.915. Precision values have always been high to the maximum or close to 1.0 throughout and this can be taken as another overfitting point due to which the model tried to create a balance between stable recall and high precisions.

Learning Rate 0.0001: The model performed erratically. At 500 iterations, recall was perfect, at 1.0, but the precision was low at 0.757, which shows that many false positives occurred. With 1000 iterations, precision and recall dropped dramatically, showing divergence or instability of the model. This resulted in recall and precision for 2000 iterations coming back to acceptable values of 0.957 each to show that more iterations were required to stabilise the performance for this lower learning rate.



For learning rate 0.01 and 0.001, precision reached 1.0 for 1000 iterations, indicating a potential risk of overfitting.

For learning rate 0.0001, precision was initially low at 500 iterations but improved at 2000 iterations to approximately 0.957.

Analysis of Results

Learning Rate: The higher the learning rate, 0.01, the better the recall and precision, though with a risk of overfitting at 1000 iterations. A moderate learning rate of 0.001 also yielded high performance, while a low learning rate of 0.0001 resulted in unstable recall, especially at 1000 iterations.

Number of Iterations: Increasing the number of iterations generally increased model stability, at least for the lowest learning rate, but came with overfitting risks in case of a higher learning rate.

The best performance was achieved with a learning rate of 0.01 and iterations of 500, balancing high recall at 0.929 and high precision at 0.985. Then, with a low learning rate of 0.0001, inconsistent performance could be seen, as more iterations are needed for the values to stabilize.

Q4

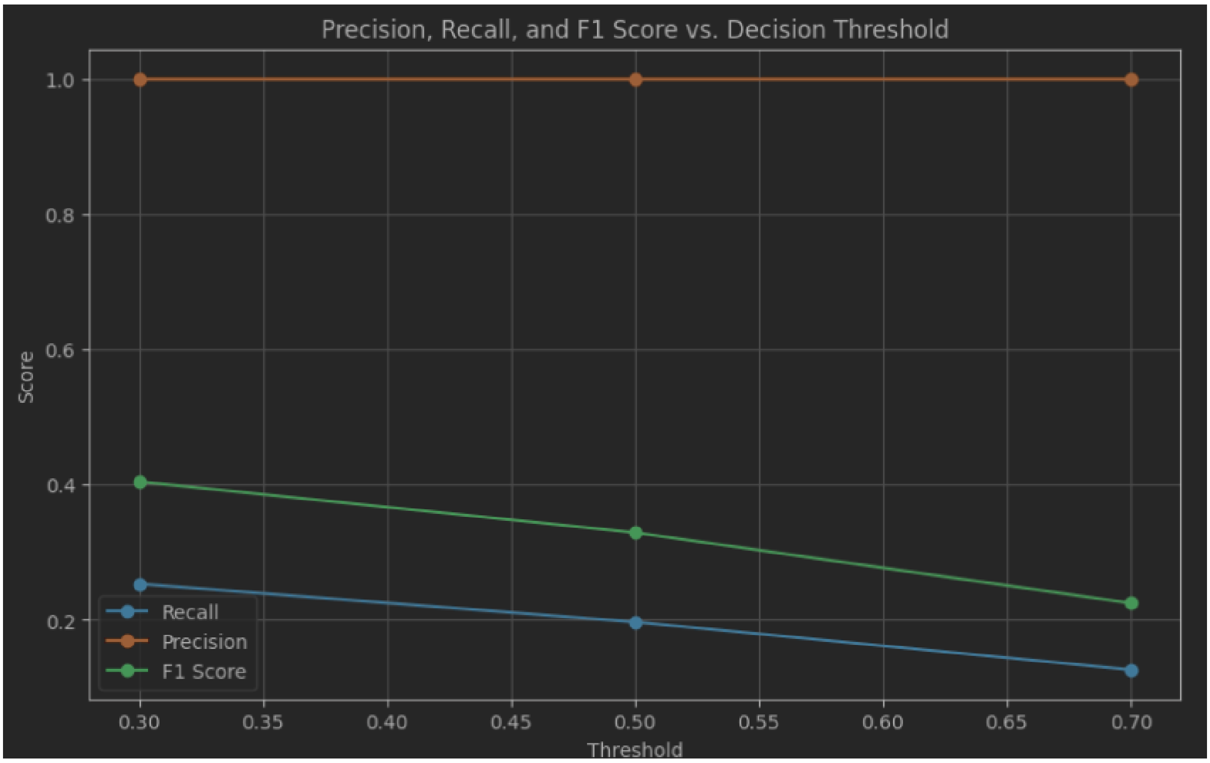
In this question, we consider the effect of changing the decision threshold of the logistic regression model on its various performance metrics, in particular on precision, recall, and F1 score. The goal would be to determine how different thresholds impact the identification of malignant cases, even in the medical context where this is necessary to minimise the number of false negatives.

Experimental Setup and Results

Three decision thresholds were tested: 0.3, 0.5, and 0.7. The table below summarises the metrics obtained for each threshold:

Threshold	Accuracy	Precision	Recall	F1 Score
0.3	0.535088	1.000	0.253521	0.404494
0.5	0.500000	1.000	0.197183	0.329412
0.7	0.456140	1.000	0.126761	0.225000

The graph below illustrates the trends in precision, recall, and F1 score as a function of the decision threshold:



The graph shows how increasing the threshold leads to a decrease in both recall and F1 score, while precision remains constant at 1.0. This graphical representation really depicts the importance of picking the right threshold for balance between recall and precision, especially in critical applications such as cancer detection.

Analysis of Results

Precision: Precision remained at 1.0 for all thresholds, indicating that all positive predictions were correct. However, this does not mean it correctly predicted all cases of malignancy.

Recall: Recall decreased with increasing threshold, ranging from 0.2535 at threshold 0.3 to 0.1268 at threshold 0.7. It means that the lower the threshold, the more cases of malignancy are identified, which is very important in medical diagnostics.

F1 Score: The F1 score also decreased with increasing thresholds, reflecting the trade-off between precision and recall. At threshold 0.3, the F1 score was highest at 0.4045.

Lowering the decision threshold from 0.7 to 0.3 resulted in a notable enhancement of recall, increasing from 0.1268 to 0.2535, thereby rendering it more effective for the detection of malignant cases. Considering the critical role of recall in medical diagnostics, a threshold of 0.3 is recommended to reduce the risk of false negatives.