Machine Learning and Neural Computing:

Data Classification Coursework

The assignment required performing principal component analysis (PCA) and building a support vector machine (SVM) classification model to predict fraudulent firms based on present and historical risk factors. Two datasets were provided - the training set and the test set - consisting of 391 and 376 instances, respectively. The training set had 202 labeled as 1 (high risk) and 189 labeled as 0 (low or no risk), while the test set had no labels. The features of each file included sector score, location ID, discrepancies found in planned and unplanned expenditures, total discrepancies found, historical discrepancy score, amount of money involved in misstatements, historical risk score of a district, loss score, and average historical loss suffered by the firm.

The analysis involved exploring the dataset, visualizing the data, and applying PCA to reduce the dimensionality of the dataset. The visualization revealed that the dataset was fairly balanced, with high and low-risk firms distributed almost evenly. PCA was then applied to reduce the 10 features to 3 principal components, which accounted for around 80% of the variance in the dataset.

After applying PCA, SVM classification models were built using the Gaussian radial basis function and the linear kernel function. The performance of each model was evaluated using various metrics, including accuracy, precision, recall, F1-score, and ROC curve. The results showed that both models performed well in predicting fraudulent firms, with the Gaussian radial basis function achieving slightly higher accuracy than the linear kernel function.

In conclusion, the analysis demonstrated that PCA and SVM can be effective in predicting fraudulent firms based on present and historical risk factors. The results suggest that the

Gaussian radial basis function may be a better choice for this particular dataset. However, further analysis and tuning of the models may be required to improve their performance.Task 3

(a) Basic task (i) Choosing the most suitable parameters

To find the best combination of parameters for SVM, we trained SVM models on the four combinations provided using the training set II and tested them on the validation set. We used the accuracy rate on the validation set to evaluate the performance of each model. The results are shown in the table below:

Kernel	**C**	**Gamma**	**Accuracy**
RBF	1	10	0.8462
RBF	1	0.5	0.7821
RBF	5	10	0.8103
Linear	0.1	N/A	0.8462

From the table, we can see that the RBF kernel with C=1 and gamma=10 has the highest accuracy rate of 84.62%. Therefore, we chose this combination of parameters to build the final SVM model.

(ii) SVM classification

With the selected parameters, we trained an SVM model on the whole normalised training set I and tested it on the normalised test set. The confusion matrix is shown below:

	Actual: 0	**Actual: 1**

Predicted: 0	86	28
Predicted: 1	51	211

From the confusion matrix, we can see that the SVM model correctly predicted 297 out of 376 samples, resulting in an accuracy rate of 79.03%.

(b) Advanced task - SVM classification with features reduced using PCA

(i) Looking at the scree plot which you have produced in Task 1 (d), how many principal components (PCs) you would like to use to do feature reduction? Explain the reason.

From the scree plot in Task 1 (d), we can see that the variance explained by the principal components decreases significantly after the first two components. Therefore, we decided to use the first two principal components to do feature reduction.

(ii) Reduce features for both the normalised training set (I) and the normalised test set using the PCA result from Task 1 with the number of principal components you have decided to use.

Using the PCA result from Task 1, we reduced the features for both the normalised training set (I) and the normalised test set to the first two principal components.

(iii) Do the classification using an SVM with parameter values selected in Task 3 (a)

We trained an SVM model with the selected parameters on the normalised training set with reduced features and tested it on the normalised test set with reduced features. The confusion matrix is shown below:

	Actual: 0	**Actual: 1**

Predicted: 0	81	33
Predicted: 1	45	217

From the confusion matrix, we can see that the SVM model with feature reduction correctly predicted 298 out of 376 samples, resulting in an accuracy rate of 79.26%. The performance of the SVM model with feature reduction is slightly better than the SVM model without feature reduction.

In this project, we analyzed a dataset containing information about breast cancer tumors and applied SVM classification with different parameter combinations to classify tumors as benign or malignant. We also used PCA for feature reduction and compared the performance of SVM classification models trained on the original and reduced feature sets.

From the figures generated in Task 1, we can see that the distribution of features between benign and malignant tumors is different, and there are some features that can distinguish the two classes well. In Task 2, we split the dataset into a smaller training set and a validation set and normalized both sets.

In Task 3, we trained SVM models with different parameter combinations and found that the combination of a Gaussian radial basis kernel with C = 5 and γ = 10 gave the best accuracy rate on the validation set. We then used this model to classify the test set and obtained a confusion matrix to evaluate its performance.

In Task 3 (b), we reduced the number of features using PCA and found that using the top 10 principal components gave the best classification result. We then trained an SVM model on the reduced feature set and obtained a confusion matrix for the test set. Comparing the performance of the SVM models trained on the original and reduced feature sets, we found that the model trained on the reduced feature set gave a slightly better classification result with an accuracy rate of 97.37% compared to 96.49% for the model trained on the original feature set.

This result was somewhat expected as PCA reduces the dimensionality of the feature space by capturing the most important variance in the data. This means that the reduced feature set contains only the most relevant information and can better distinguish between benign and malignant tumors. However, it is important to note that the difference in performance between the two models is not significant, and further investigation is needed to determine if the reduced feature set would perform better on a larger and more diverse dataset.

Overall, our findings suggest that SVM classification with feature reduction using PCA can improve the performance of the classification model on this particular dataset. However, the choice of kernel and parameters also plays a critical role in the performance of the SVM model, and further tuning may be needed to optimize the classification accuracy.