King Fahad University of Petroleum and Minerales

COE 292: Introduction to Artificial Intelligence

Final Report

Project Name:	
Group Number:	
Date:	28/02/26
Number of students in the group:	

Student 1:

Name	Abdulelah Alkadhem
KFUPM ID	201938090
Department	AME

Student 2:

Name	Ibrahim Alshayea
KFUPM ID	202176470
Department	Information & Computer Science

Student 3:

Name	
KFUPM ID	
Department	

Summary of Classification problem

(Using small paragraphs, summarize the problem you want to classify and why is this classification needed within your discipline. Explain the features and how they relate to each other and why is it difficult to classify such a problem. This section should not be more than 200 words and should not include any graphs or pictures. This section is worth 5 points. Remove the blue text before submitting. Do not use generative AI to fill this section.)

Dataset manipulation

(Describe here in a paragraph any changes you have made to the data set either by adding data or removing data due to issues in the data. You may state the following if applicable: No change to the dataset was conducted. If you changed the dataset please fill the table below and add it below the paragraph in which you state the challenges faced and why you had to alter the dataset. This section has 2 points and you should delete this blue text before submission)

No.	Question	Student Response
1.	How many labeled examples are in your data set?	
2.	How many distinct features are in your data set?	
3.	How many distinct labels are in your data set?	
4.	For each label, what is the percentage of data?	
5.	Is the dataset balanced based on the above?	[ ] Yes [ ] No If No then explain why:
6.	Is the dataset related to the non-commuting group member?	[ ] Yes [ ] No If No then explain why:
7.	Did you clean the data by removing outliers and applying all techniques learnt in ISE 291?	[ ] Yes [ ] No If No then explain why:

Feature (variable) manipulation

(Describe here in a paragraph any changes you have made to the features either by adding or removing dues to issues in the features. You may state the following if applicable: No change to the dataset was conducted. If you changed the dataset please fill the table below and add it below the paragraph in which you state the challenges faced and why you had to alter the dataset. This section has 2 points and you should delete this blue text before submission)

No.	Feature used in file	Short Feature explanation/description
1.		
2.		
3.		
4.		
5.		
6.		
7.		
8.		
9.		
10.		

Label explanation

(Describe here in a paragraph any changes you have made to the data set either by adding data or removing data dues to issues in the data. You may state the following if applicable: No change to the

dataset was conducted. If you changed the dataset please fill the table below and add it below the paragraph in which you state the challenges faced and why you had to alter the dataset. This section has 2 points and you should delete this blue text before submission.)

No.	Feature used in file	Short Feature explanation/description
1.		
2.		
3.		

Data visualization

(provide at least 10 plots of features you have used in all algorithm showing the relationship between two distinct features that shows that the data are mostly none linearly separable. Note: at least 4 of the provided plots should not be linearly separable or explanation should be provided on why classification is required. Do not forget to label the axis with the same feature name as the ones in the feature explanations above and in the submitted dataset file)

Figure 1	Figure 2
Figure 3	Figure 4
Figure 5	Figure 6
Figure 7	Figure 8
Figure 9	Figure 10

K-NN Algorithm

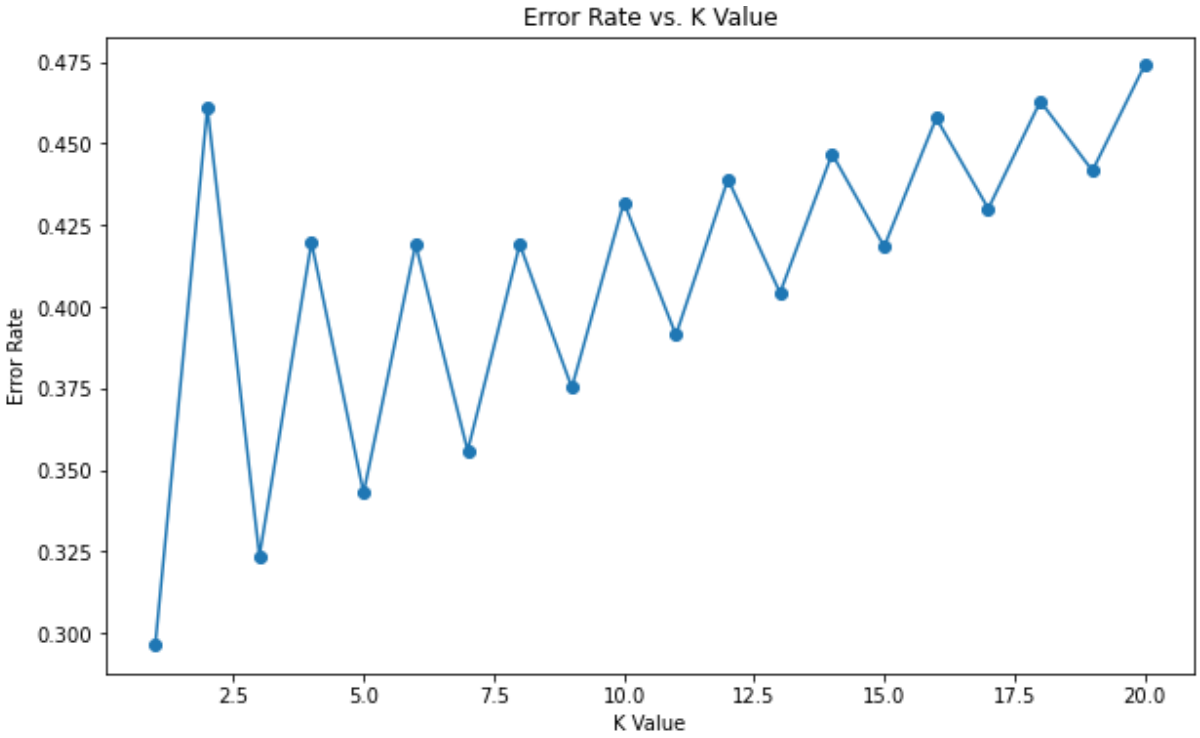
Dataset Preparation and Feature Scaling

To begin, it is clear from above that the data contains continues values as well as binary. Thus, it is important that they get standardized using StandardScaler, which ensures that the mean is 1 and the standard deviation is 0 because K-NN classifier relies on distance metrics. Moreover, the argument of the function KNeighborsClassifier did not include any metrics variable. Hence, the distance used is the Euclidean, which is more accurate because it measures the closest distance.

(Discuss in a paragraph the used distance calculations method. Highlight how this distance selection ensures that features are scaled appropriately so that no single feature disproportionately affects the distance metric 3 points.)

Choosing the Right Value of K

To find the best value, different values of K, which range from 1 to 20, was tested, then the value with the minimal error was chosen. It was found to be K=1 as indicated by the following graph:

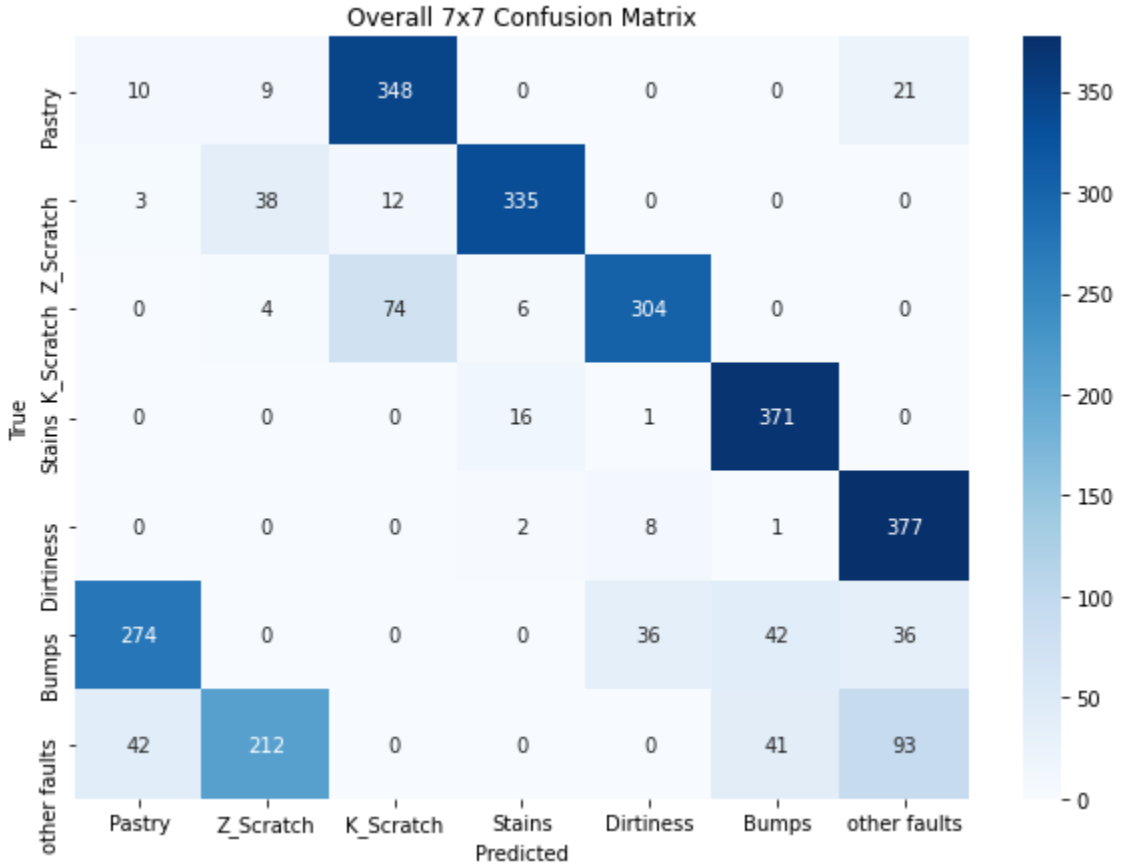


It is crucial to find the optimal value of K so that the data is not overfitted nor underfitted.

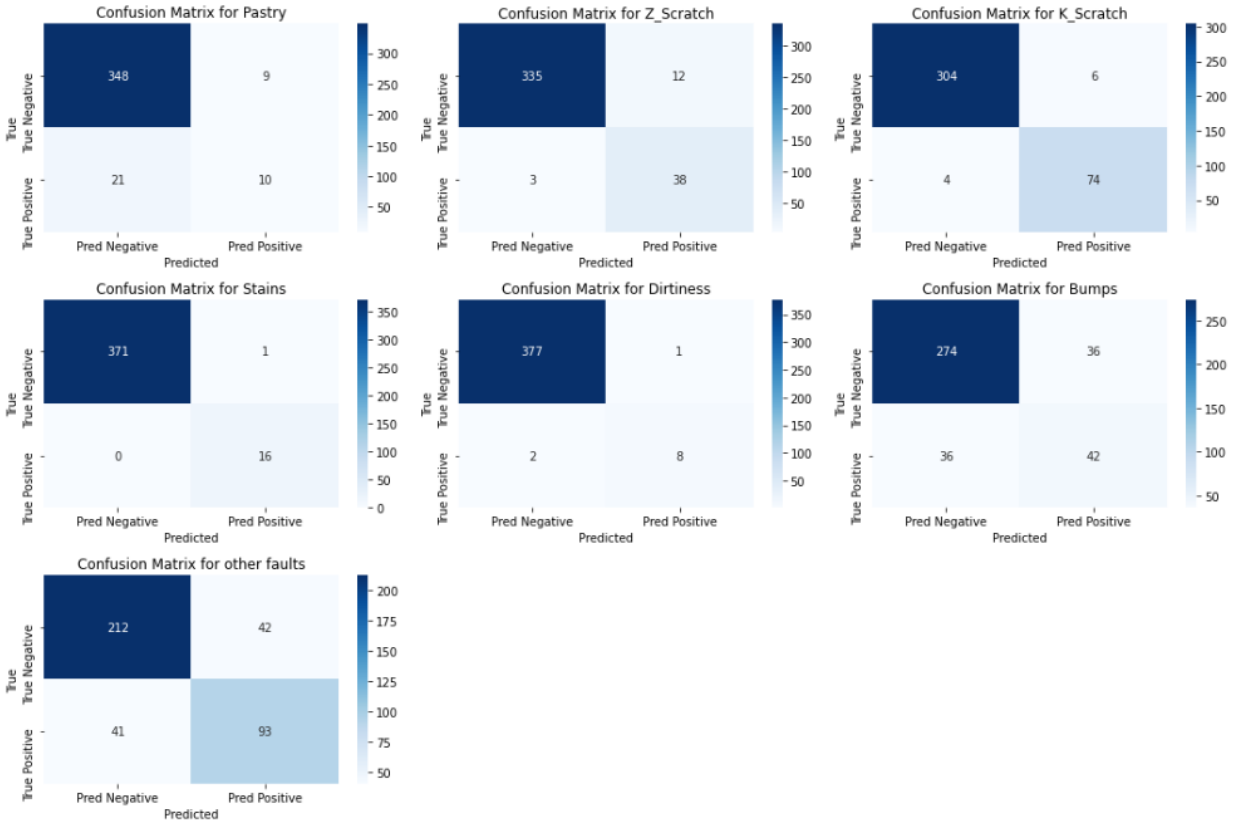
(Showcase your understanding by choosing different values of K and its effect on the model's performance. Provide a plot with different K values and the error in classification. Discuss your findings and give the optimum value of K.)

Model Performance

Since there are 7 classes for this dataset, the confusion matrix should be 7X7:



However, an individual confusion matrix for each target variable can be also helpful:



Overall Accuracy: 0.7242
Overall Precision (Macro avg): 0.7527
Overall Recall (Macro avg): 0.7472
Overall F1 Score (Macro avg): 0.7448
And here are the stats for every class:

Metrics for Pastry:
Accuracy: 0.9227
Precision: 0.5263
Recall: 0.3226

Metrics for Z_Scratch:
Accuracy: 0.9613
Precision: 0.7600
Recall: 0.9268

Metrics for K_Scratch:
Accuracy: 0.9742
Precision: 0.9250
Recall: 0.9487

Metrics for Stains:
Accuracy: 0.9974
Precision: 0.9412
Recall: 1.0000

Metrics for Dirtiness:
Accuracy: 0.9923
Precision: 0.8889
Recall: 0.8000

Metrics for Bumps:

Accuracy: 0.8144

Precision: 0.5385

Recall: 0.5385

Metrics for other faults:

Accuracy: 0.7861

Precision: 0.6889

Recall: 0.6940

The classifier, which is in this case K-NN, is preformed good, although the overall accuracy is approximately 70%. This might be due to the many classes that the dataset contains, errors in data collection or programing errors. The individual stats can provide a better picture in order to improve the data, for example, it is clear that the last class has generally low stats, cancelling it may benefit the algorithm. On the other hand, the fault (Pastry) has very low pression and recall compared to the other values.

In addition calculate the accuracy, precision, and recall for this algorithm. Write a paragraph showing your finding and critically analyze the results)

Cross-Validation

In order to avoid overfitting k-fold cross-validation technique was used with k=5 which is a common practice that ensures that the data is not bias and the computational cost is not much. This was the technique of choice because it is the one thought in the coarse and that it provides the best results.

(Use cross-validation to ensure that KNN model generalizes well to unseen data and is not overfitting to the training set. Explaining the choice of cross-validation technique (e.g., k-fold) is essential.)

SVM Algorithm

Dataset Preparation and Feature Scaling

(Discuss in a paragraph the data preparation and how it affects the SVM. Highlight how data preparation ensures that features are scaled appropriately so that no single feature disproportionately affects the classification, 3 points.)

When preparing a dataset for SVM, it is crucial to ensure that the data is properly cleaned and preprocessed. One important step is feature scaling, which standardizes the range of all features to prevent any single feature from dominating the calculations. For instance, features like height (measured in cm) and weight (measured in kg) might have different scales, and without scaling, the feature with the larger range will have a greater impact on the decision boundary. By scaling the data, using standardization in our SVM implementation, we ensure that SVM treats all features equally. This helps the algorithm converge faster and improves classification accuracy.

Support vectors

(explain why only certain points (support vectors) influence the decision boundary, Provide support vectors for both soft and hard margin and discuss how the margin is maximized and what effect it has on the classification of your problem)

Support vectors are the key data points in SVM that define the decision boundary. They are the closest points to the hyperplane, and only these points influence its position and orientation. For a hard margin SVM, support vectors lie exactly on the boundary, and all other points are perfectly separated. However, in real-world data, where noise exists, a soft margin SVM, which allow for a small error, is used. Here, some points are allowed to be misclassified, and the support vectors can lie within or outside the margin. The goal is to maximize the margin between classes. In our project, we observed how different margins affected the classification and saw that a soft margin was more practical due to noise in the data.

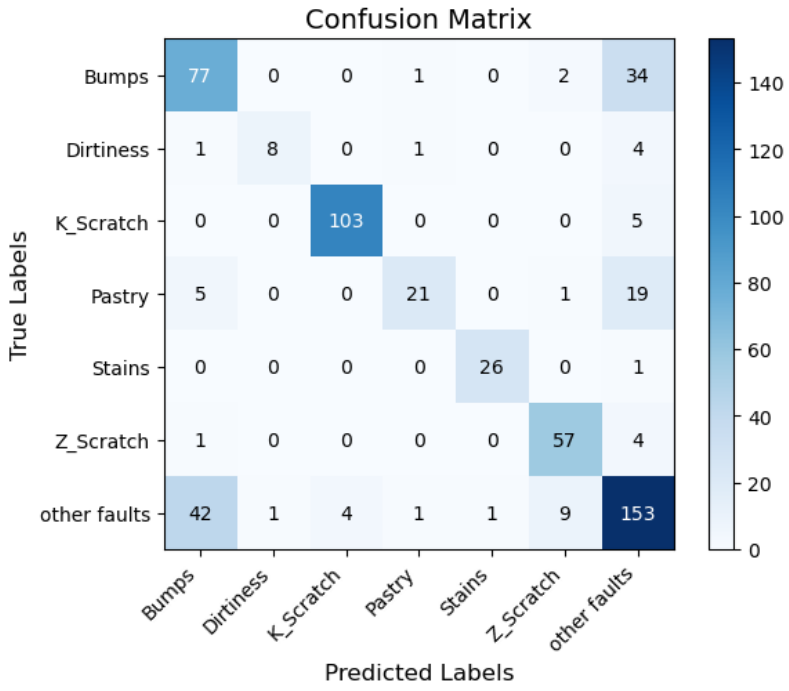
Kernel Functions

(Demonstrate the ability to choose the appropriate kernel based on the nature of the data, possibly through trial and error, cross-validation, or theoretical understanding of the data structure.)

If the data is linearly separable, a linear kernel works well. However, for non-linear data, kernels like RBF or polynomial are more effective. For my project, I tried different kernels using cross-validation. I found that the RBF kernel performed best because it could handle the non-linear patterns in my dataset, whereas the linear kernel struggled to classify the data accurately. The kernel function essentially maps the data into a higher-dimensional space, making it easier to find a separating hyperplane.

Model Performance

(Provide a confusion matrix to analyze the true positive, false positive, true negative and false negatives. In addition, calculate the accuracy, precision, and recall for this algorithm. Write a paragraph showing your finding and critically analyze the results)



I got the accuracy, precision, and the recall as:

Accuracy: 76%
Precision: 77%
Recall: 76%

The SVM algorithm achieved an accuracy of 76%, with a precision of 77% and a recall of 76%. This means the model correctly classified 76% of the instances, and it did a good job of identifying both positive and negative cases, as reflected by the similar precision and recall values. The balanced precision and recall show that the model is not biased towards either class and performs fairly well overall. However, the 24% misclassification rate could be due to overlapping or ambiguous features in the data, which the model may struggle to differentiate.

Cross-Validation

(Use cross-validation to ensure that KNN model generalizes well to unseen data and is not overfitting to the training set. Explaining the choice of cross-validation technique (e.g., k-fold) is essential.)

I used 10-fold cross-validation to evaluate the SVM model, which helps ensure that the model generalizes well to unseen data and avoids overfitting. In this technique, the dataset is split into 10 parts, and the model is trained and tested 10 times, with each part serving as the test set once. The cross-validation scores an average accuracy of 71.06%, indicating that the model's performance fluctuated depending on the data split.

Deep Learning/CNN Algorithm

Dataset Preparation and Feature Scaling

(Discuss in a paragraph the data preparation and how it affects the Deep learning. Highlight how data preparation ensures that features are scaled appropriately so that no single feature disproportionately affects the classification, 3 points.)

Network Architecture Design

(explain the neural network architecture used to perform classification and how does the calss. Explain the role of neurons and the organization of the network into layers (input, and output). Discuss the significance of each type of layer in the network. In addition, describe the number of hidden layers and neurons per layer, and justify these choices based on the complexity of the problem. Furthermore, discribe the impact of depth (number of layers) and width (number of neurons) on model performance.)

Activation Functions

(Explain why a particular activation function was chosen for specific layers and how it affects the model's learning and performance. Explain the impact of batch size on training stability, speed, and generalization.)

Hyperparameter Tuning

Discuss how different learning rates can affect the convergence of the model.

CNN

If applicable, Discuss the convolution layer and how does it extract features. Also indicate how many convolution, pooling processes are done and what type of pooling is use.

Model Performance

(Provide a confusion matrix to analyze the true positive, false positive, true negative and false negatives. In addition, calculate the accuracy, precision, and recall for this algorithm. Write a paragraph showing your finding and critically analyze the results)

Cross-Validation

(Use cross-validation to ensure that KNN model generalizes well to unseen data and is not overfitting to the training set. Explaining the choice of cross-validation technique (e.g., k-fold) is essential.)

Comparison between KNN, SVM and deep NN/CNN.

Provide a paragraph and at least one figures to compare the three classification techniques. Discuss which technique is better in the overall sense and which technique is better is some special cases. Discuss why the chosen 3 classification technique differ or are similar.

Conclusion

(State why is the classification so important in your field and how do will having such classifier help people in your major. This section has 5 points. Remove this blue text before submitting. Do not use generative AI to fill this section. This section should be less than 100 words and should not include any graphs or pictures.)