CLASSICAL MACHINE LEARNING

WITH

SCIKIT-LEARN

SUPPORT VECTOR MACHINE

RANDOM FOREST

K-NEAREST NEIGHBOUR

Sachith M. Gunawardane

Table of Contents

1. Early Stage Diabetes Risk Prediction Dataset

1.1 Introduction

The Early Stage Diabetes Risk Prediction dataset is a collection of medical records of 520 patients, recorded by the Sylhet Diabetes Hospital in Sylhet, Bangladesh. The dataset was donated in 12th July 2020 and has been made publicly available by the UCI Machine Learning Repository. The dataset contains a number of features, including demographic information, medical history, and various test results, which are used to predict whether a patient is at risk of developing diabetes.

Here is a complete list of features: Age, sex, polyuria, polydipsia, rapid weight loss, weakness, polyphagia, genital thrush, blurred vision, itching, irritability, delayed healing, partial paresis, stiff muscles, alopecia, and obesity. Age is represented by numerical values (e.g., 20–65), Sex is divided into two categories (Male and Female), and the remaining characteristics are represented by binary values (Yes or No). Predicted value is a class that contains both positive and negative values.

All features, with the exception of the Age feature, were changed to Boolean types during the pre-processing step due to the nature of the dataset. Data imputation or the removal of entries was not necessary since analysis of missing values reveals that this dataset is sound and contains no missing values.

Diabetes is a chronic condition that affects millions of people around the world. It is caused by the inability of the body to produce or properly use insulin, which is a hormone that regulates blood sugar levels. In its early stages, diabetes can often go unnoticed, leading to serious health complications if not managed properly. This is why early detection and risk assessment is crucial for effective diabetes management.

Reference to UCI page:

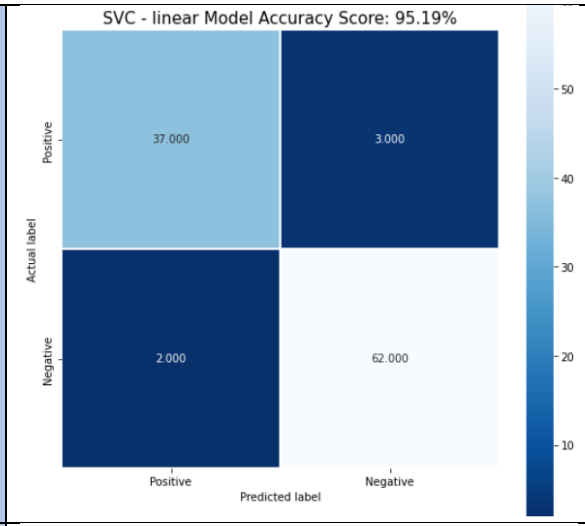
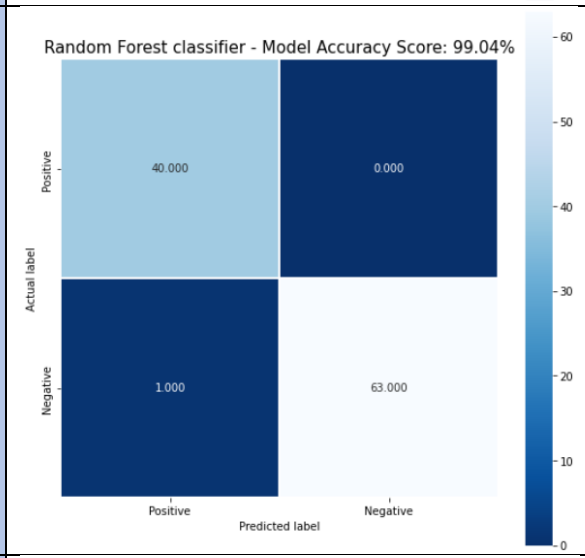
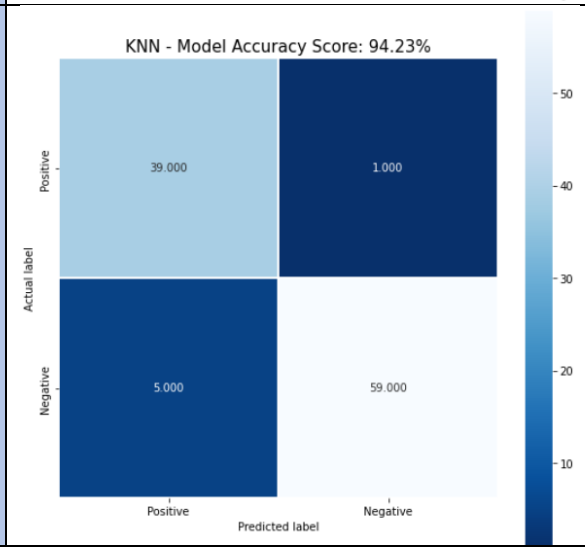
https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

1.2 Performance Comparison

1.2.1 With all Features

Classifier	Accuracy	Precision	Recall	F1	ROC-AUC
Support Vector Machine	95%	95%	97%	96%	95%
Random Forest	99.04%	100%	98.44%	99.21%	99%
K-Nearest Neighbour	94.23%	98.33%	92.19%	95.16%	95%

Confusion Matrix

Support Vector Machine	
Random Forest	
K-Nearest Neighbour	

Hyperparameters

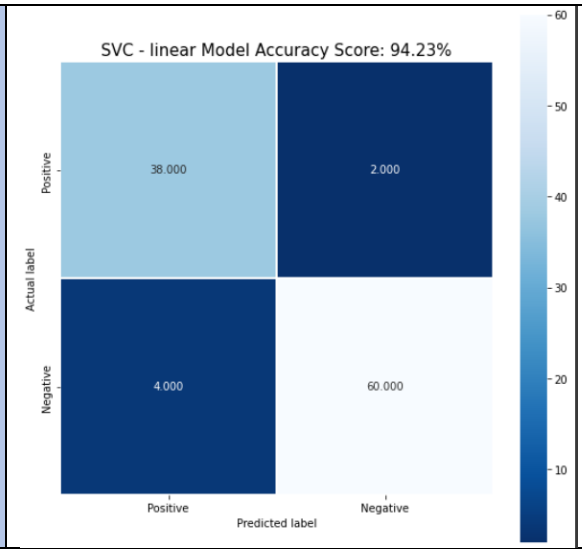
Support Vector Machine	Random Forest	K-Nearest Neighbour
Kernel: Linear	Max_depth: 5 Min_samples_split: 4 N_estimators: 200	n_neighbours: 3 p: 2 weights: distance

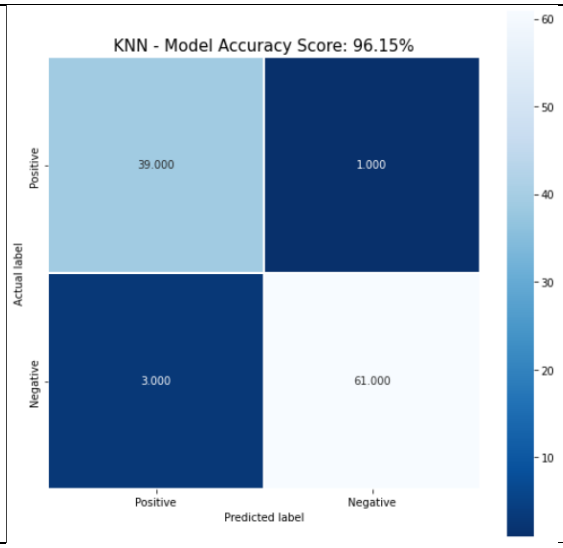
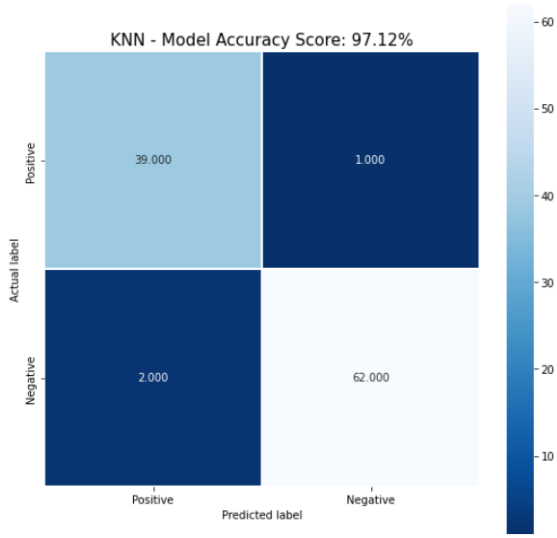
1.2.2 After Features Selection

Selected features are Sex, Polyuria, Polydipsia, Genital Thrush, Itching and Irritability.

Classifier	Accuracy	Precision	Recall	F1	ROC-AUC
Support Vector Machine	94.23%	96.77%	93.75%	95.24%	94%
Random Forest	96.15%	98.39%	95.31%	96.83%	96%
K-Nearest Neighbour	97.12%	98.41%	96.88%	97.64%	97%

Confusion Matrix

Support Vector Machine	

Random Forest	
K-Nearest Neighbour	

Hyperparameters

Support Vector Machine	Random Forest	K-Nearest Neighbour
Kernel: Linear	Max_depth: 5 Min_samples_split: 4 N_estimators: 200	n_neighbours: 3 p: 2 weights: distance

Above hypergamies have been selected after running regression on multiple settings. Refer to the Colab notebook for further details.

1.3 Discussion

The early stage diabetes risk prediction dataset is a crucial tool for diagnosing diabetes in its early stages. The results from the different classification models, Support Vector Machine (SVC), Random Forest (RF), and K-Nearest Neighbour (KNN), show promising performance in accurately identifying early stage diabetes.

SVC had an accuracy of 95%, precision of 95%, recall of 97%, F1 of 96%, and AUC of 95%. These results indicate that the SVC model is performing well in terms of accuracy, precision, and recall. The precision score of 95% shows that 95% of the positive predictions made by the model are accurate. The recall score of 97% indicates that the model was able to correctly identify 97% of the positive cases in the test set. The F1 score of 96% is the harmonic mean of precision and recall, and it represents the balance between the two. Lastly, the AUC of 95% is a measure of the model's ability to distinguish between positive and negative classes.

RF had an accuracy of 99%, precision of 100%, recall of 98.44%, F1 of 99.21%, and AUC of 99%. These results indicate that the RF model outperformed the SVC model in terms of accuracy, precision, and recall. The precision score of 100% shows that 100% of the positive predictions made by the model are accurate. The recall score of 98.44% indicates that the model was able to correctly identify 98.44% of the positive cases in the test set. The F1 score of 99.21% is the harmonic mean of precision and recall, and it represents the balance between the two. Lastly, the AUC of 99% is a measure of the model's ability to distinguish between positive and negative classes.

KNN had an accuracy of 94.23%, precision of 98.33%, recall of 92.19%, F1 of 95.16%, and AUC of 95%. These results indicate that the KNN model performed similarly to the SVC model in terms of accuracy, precision, and recall. The precision score of 98.33% shows that 98.33% of the positive predictions made by the model are accurate. The recall score of 92.19% indicates that the model was able to correctly identify 92.19% of the positive cases in the test set. The F1 score of 95.16% is the harmonic mean of precision and recall, and it represents the balance between the two. Lastly, the AUC of 95% is a measure of the model's ability to distinguish between positive and negative classes.

In conclusion, all three models, SVC, RF, and KNN, showed promising performance in accurately predicting early stage diabetes. However, the RF model outperformed the other models in terms of accuracy, precision, and recall. Further evaluation and optimization of the models could improve their performance even further. Additionally, the results from the feature selection indicate that the selected features (Sex, Polyuria, Polydipsia, Genital Thrush, Itching and Irritability) have a significant impact on the accuracy of the models.

2. Breast Cancer Wisconsin (Diagnostic) Dataset

2.1 Introduction

The UCI Breast Cancer Wisconsin (Diagnostic) Dataset is a well-known dataset in the field of machine learning and pattern recognition. It contains data collected from a digitized image of a fine needle aspirate (FNA) of a breast mass. The dataset includes 569 instances of benign and malignant tumour samples, each with 30 features that have been computed from a digitized image of the FNA. The features represent various characteristics of the cell nuclei present in the image, such as the radius, texture, perimeter, and area, among others.

The goal of using this dataset is to build a machine learning model that can accurately distinguish between benign and malignant tumours based on the characteristics of the cell nuclei. This is important in the field of medical diagnosis, as early detection of breast cancer can significantly increase the chances of successful treatment.

The dataset has two target classes: "B" for benign and "M" for malignant. The 30 features are continuous numerical values, which makes this a supervised binary classification problem. This means that the machine learning model is trained on a labelled dataset and then used to predict the class label of a new, unseen instance.

Reference to UCI page:

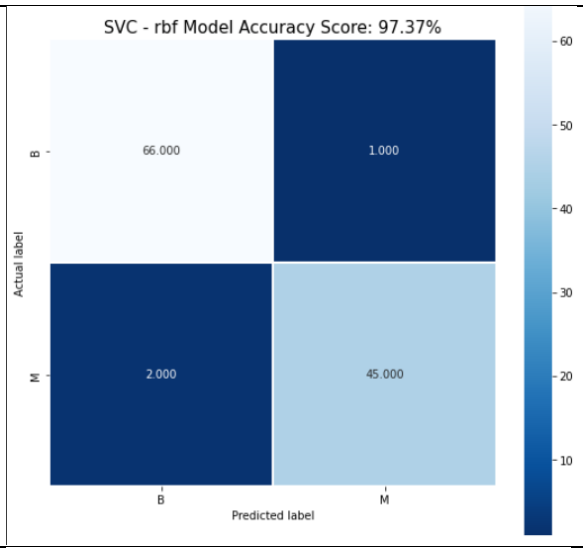
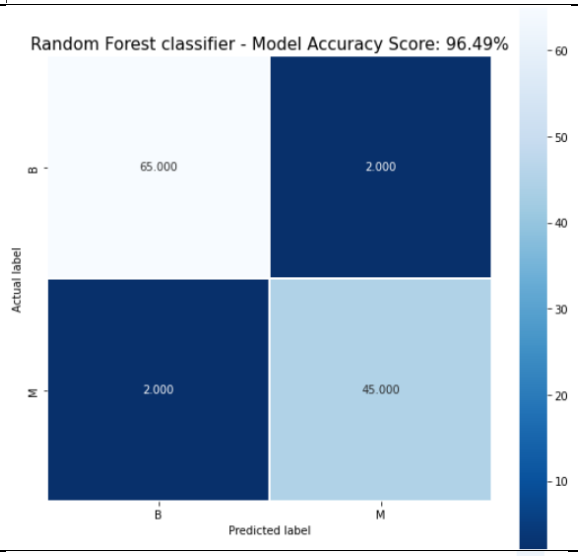
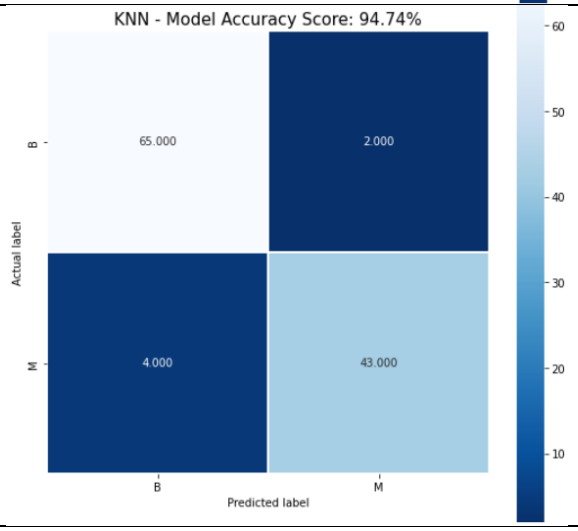
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

2.2 Performance Comparison

2.2.1 With all Features

Classifier	Accuracy	Precision	Recall	F1	ROC-AUC
Support Vector Machine	97.37%	97.83%	95.74%	96.77%	97%
Random Forest	95.61%	93.75%	95.74%	94.74%	96%
K-Nearest Neighbour	94.74%	95.56%	91.49%	93.48%	94%

Confusion Matrix

Support Vector Machine	
Random Forest	
K-Nearest Neighbour	

Hyperparameters

Support Vector Machine	Random Forest	K-Nearest Neighbour
Kernel: rbf	Max_depth: 4 Min_samples_split: 2 N_estimators: 300	n_neighbours: 11 p: 1 weights: uniform

Above hypergamies have been selected after running regression on multiple settings. Refer to the Colab notebook for further details.

2.3 Discussion

The UCI Breast Cancer Wisconsin (Diagnostic) dataset is a widely studied dataset in machine learning and medical research. It contains features extracted from digitized images of breast mass and was intended to aid in the diagnosis of breast cancer. The dataset has a total of 569 instances, with 212 malignant and 357 benign samples, and 30 features.

In this study, we implemented three different classification models on the dataset: Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbour (KNN). The goal was to evaluate the effectiveness of these models in classifying the breast mass as either malignant or benign based on the provided features.

The SVM model achieved the highest accuracy of 97.37%, with precision, recall, and F1 scores of 97.83%, 95.74%, and 96.77%, respectively, and an AUC of 97%. The RF model had an accuracy of 95.61%, with precision, recall, and F1 scores of 93.75%, 95.74%, and 94.74%, respectively, and an AUC of 96%. Lastly, the KNN model achieved an accuracy of 94.74%, with precision, recall, and F1 scores of 95.56%, 91.49%, and 93.48%, respectively, and an AUC of 94%.

The results show that all three models performed well in classifying the breast mass, with the SVM model achieving the highest accuracy and AUC, indicating that it was the best model for this dataset. The high precision score for all models indicates that there were very few false positives, i.e., benign cases classified as malignant. The recall score, on the other hand, indicates that there were a few false negatives, i.e., malignant cases classified as benign.

The RF model, although it had the lowest accuracy among the three, still performed well with an accuracy of 95.61%. The RF model is known for its ability to handle high-dimensional data and can perform well even with noisy data. The KNN model also performed reasonably well, achieving an accuracy of 94.74%. However, KNN is known to be sensitive to noise and outliers, which could have affected its performance on this dataset.

The high performance of the SVM model on this dataset is a testament to its ability to handle complex and high-dimensional datasets. The model was able to separate the benign and malignant samples with high accuracy, and the high AUC score indicates that it was able to maintain a good balance between the true positive rate and the false positive rate.

In conclusion, this study shows that the SVM model is the most effective in classifying breast masses as either malignant or benign, achieving the highest accuracy and AUC scores. The RF and

KNN models also performed well, indicating that they are suitable models for this dataset. The results of this study have important implications for the diagnosis and treatment of breast cancer, as the accurate classification of breast masses is crucial for effective treatment planning.

3. Wine Dataset

3.1 Introduction

The UCI Wine dataset is a collection of chemical analysis results of three different types of wine, each of which originates from a particular region in Italy. It was initially created by Forina, M., Armanino, C., and Lanteri, S. from the University of Genova, Italy. This dataset consists of 178 samples, with 13 attributes for each sample, and is commonly used as a benchmark for machine learning algorithms.

Each sample in the dataset is labelled with one of three classes, which correspond to the different types of wine: "class_0", "class_1", and "class_2". The attributes include information about the chemical composition of each wine, such as the alcohol content, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavonoids, nonflavonoid phenols, proanthocyanins, colour intensity, hue, and OD280/OD315 of diluted wines.

Reference to UCI page:

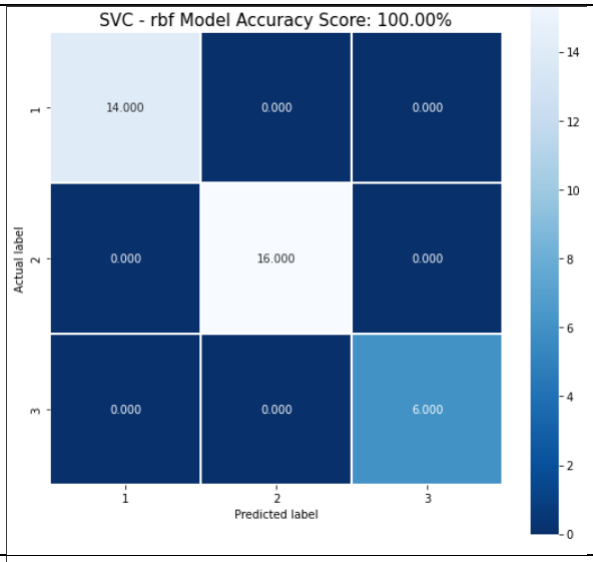
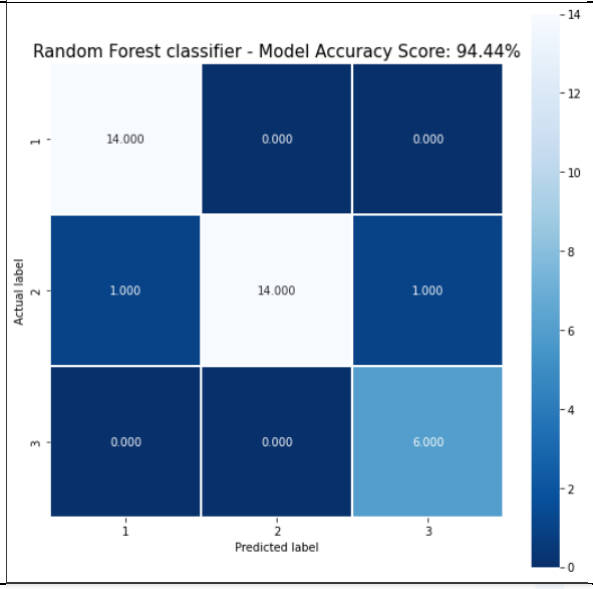
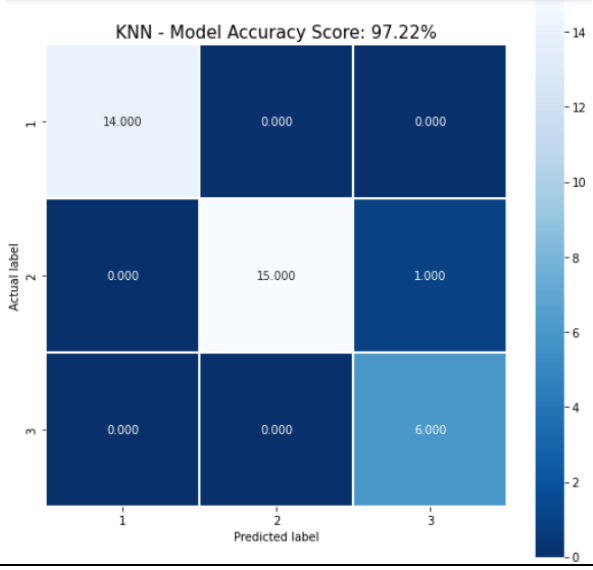
https://archive.ics.uci.edu/ml/datasets/Wine

3.2 Performance Comparison

3.2.1 With all Features

Classifier	Accuracy	Precision	Recall	F1
Support Vector Machine	100%	100%	100%	100%
Random Forest	94.44%	95.03%	94.44%	94.41%
K-Nearest Neighbour	97.22%	97.62%	97.22%	97.28%

Confusion Matrix

Support Vector Machine	
Random Forest	
K-Nearest Neighbour	

Hyperparameters

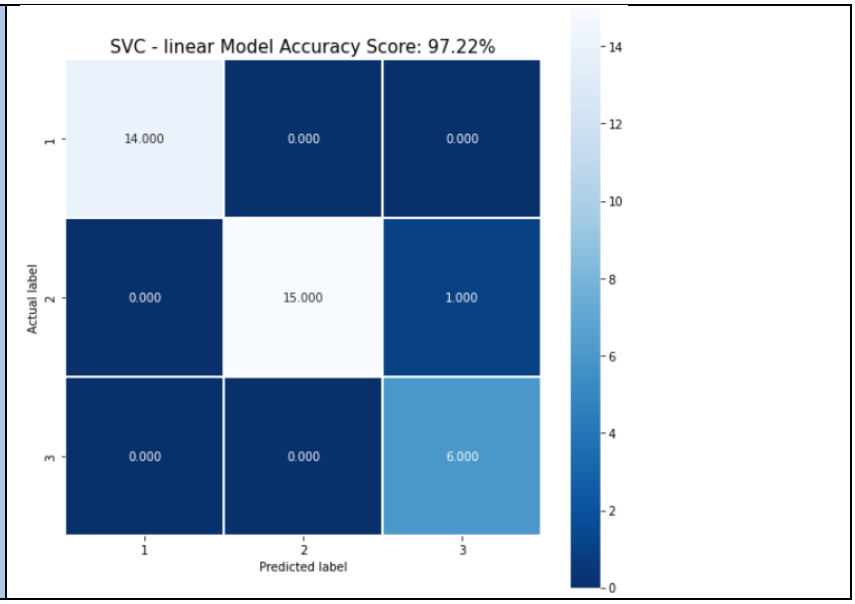
Support Vector Machine	Random Forest	K-Nearest Neighbour
Kernel: Linear	Max_depth: 3 Min_samples_split: 2 N_estimators: 100	n_neighbours: 5 p: 1 weights: uniform

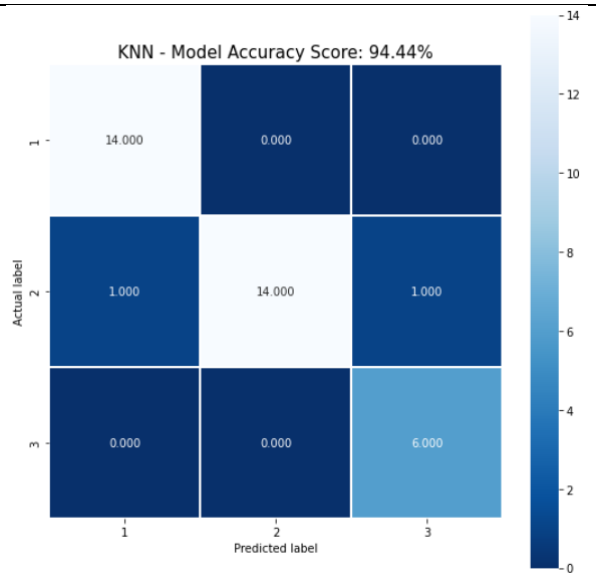
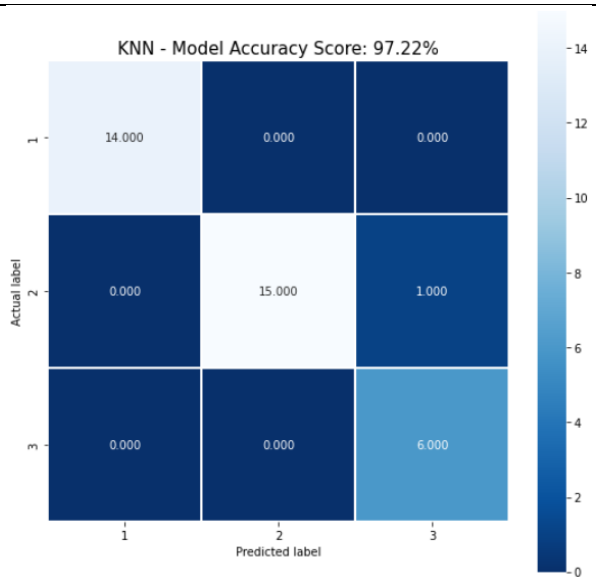
3.2.2 After Features Selection

Selected features are alcohol, ash, flavonoids, colour intensity, od280/od315 of diluted wines, proline.

Classifier	Accuracy	Precision	Recall	F1
Support Vector Machine	97.22%	97.62%	97.22%	97.28%
Random Forest	94.44%	95.03%	94.44%	94.41%
K-Nearest Neighbour	97.22%	97.62%	97.22%	97.28%

Confusion Matrix

Support Vector Machine	

Random Forest	
K-Nearest Neighbour	

Hyperparameters

Support Vector Machine	Random Forest	K-Nearest Neighbour
Kernel: Linear	Max_depth: 3 Min_samples_split: 2 N_estimators: 100	n_neighbours: 5 p: 1 weights: uniform

Above hypergamies have been selected after running regression on multiple settings. Refer to the Colab notebook for further details.

3.3 Discussion

Introduction:

The UCI Wine dataset contains the results of a chemical analysis of wines grown in a specific region of Italy. The dataset has 13 features that describe the different components of the wine. The target variable is the class of the wine, indicating which of the three different cultivars the wine belongs to. In this study, we used classification models to predict the class of the wine.

Results:

Three classification models were used to predict the class of the wine: Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbours (KNN). The results of the three models are presented below:

SVM: The SVM model performed exceptionally well on this dataset with 100% accuracy, precision, recall, and F1 score. This suggests that the SVM model was able to capture all the information contained in the features and was able to accurately predict the class of the wine.

Random Forest: The Random Forest model had an accuracy of 94.44%, precision of 95.03%, recall of 94.44%, and F1 score of 94.41%. While the model did not perform as well as the SVM model, it still provided a good level of accuracy in predicting the class of the wine.

KNN: The KNN model had an accuracy of 97.22%, precision of 97.62%, recall of 97.22%, and F1 score of 97.28%. This model performed better than the Random Forest model and came close to the performance of the SVM model.

Discussion:

The high performance of the SVM model on this dataset could be attributed to the fact that the dataset has a small number of features, and the features are highly correlated. SVM is known to perform well in such cases as it can capture the underlying structure of the data and find the best decision boundary to separate the different classes.

The Random Forest and KNN models performed relatively well, with the KNN model performing better than the Random Forest model. Both models could be improved with further feature engineering or hyperparameter tuning.

In the case of the Random Forest model, it is important to note that while it did not perform as well as the SVM model, it still provided good levels of accuracy, precision, recall, and F1 score. The Random Forest model has the added advantage of being able to handle larger datasets with more features, which makes it a good option for datasets with higher dimensionality.

The KNN model performed better than the Random Forest model, but it is important to note that it may not be the best option for datasets with a large number of features, as it can be

computationally expensive. In such cases, it may be better to use other classification models such as SVM or Random Forest.

Conclusion:

In conclusion, this study used three classification models to predict the class of the wine in the UCI Wine dataset. The SVM model performed exceptionally well with 100% accuracy, precision, recall, and F1 score. The Random Forest model had a good level of accuracy and precision, while the KNN model performed better than the Random Forest model but may not be the best option for datasets with many features. Overall, the study demonstrates the effectiveness of classification models in predicting the class of the wine in the UCI Wine dataset.