Capstone Project Report

Table Of Contents (TOC)

1. Data Collection

1A .Source Of Data:

Collected 2 CSV files named Materials and their Mechanical Properties & Use from Kaggle.

1B. Dataset Info:

Dataset 1 : Materials and their Mechanical Properties : 1552 rows × 15 columns

Dataset 2: 1552 rows × 8columns

1C. Reason for Choosing This Particular Dataset:

This dataset has been chosen to get the insights about materials treated under various heat treatment processes & their influence on properties like Ultimate Tensile Strength (Su) in MPa, Yield Strength (Sy) in MPa, Elongation at Break or Strain (A5) as a Percentage, Brinell Hardness Number (BHN) in Microhardness Units, Elastic Modulus (E) in MPa, Shear Modulus (G) in MPa, Poisson's Ratio (mu) in Units of Length, Density (Ro) in Kg/m3, Pressure at Yield (pH) in MPa, Description of the Material (Desc), Vickers Hardness Number (HV) inorder to find their usefulness in find applications in various fields and industries. Here are some common applications that connect these mechanical properties:

- **Automotive Industry:** Materials with high Ultimate Tensile Strength (Su) and Yield Strength (Sy) are used in the manufacturing of automotive components to ensure safety and durability. High strength materials can be found in engine components, chassis, and structural parts.
- **Aerospace Industry:** Aerospace materials need to withstand extreme conditions. High Su and Sy, along with good elastic modulus (E) and low density (Ro), are vital for aircraft and spacecraft components.
- **Structural Engineering:** Construction materials with appropriate mechanical properties, including high Su and Sy, are used in building and bridge construction. Elastic modulus (E) is essential for designing structural elements.
- **Metallurgy:** Materials with specific hardness properties, such as Brinell Hardness Number (BHN) and Vickers Hardness Number (HV), are critical in metallurgical processes, including heat treatment and surface hardening.
- **Mechanical Engineering:** These properties are crucial for designing machinery, equipment, and mechanical systems. Elastic modulus (E) and shear modulus (G) are essential for designing load-bearing structures.

- **Biomechanics:** In the medical field, materials used in implants, prosthetics, and orthopaedic devices require a combination of suitable mechanical properties to ensure compatibility with the human body.
- **Material Science:** Researchers and scientists use these properties to study and develop new materials with improved mechanical characteristics for various applications.
- **Material Testing and Quality Control:** Mechanical properties are tested and monitored to ensure quality and safety in manufacturing processes.
- **Mining and Heavy Equipment:** Materials used in mining and heavy machinery must have high Su and Sy values to withstand extreme conditions.
- **Oil and Gas Industry:** Materials used in oil rigs and pipelines must have the mechanical properties to endure high-pressure conditions, such as Pressure at Yield (pH).
- **Energy and Renewable Energy:** Materials used in wind turbines, solar panels, and energy infrastructure must have suitable mechanical properties to withstand environmental conditions and stresses.
- **Consumer Products:** Materials with specific mechanical properties are used in the manufacturing of everyday products, such as electronic devices, appliances, and sports equipment.

1D. Summary :

These are just a few examples, and the applications of materials with these mechanical properties are widespread across numerous industries, each with its own specific requirements and demands. The selection of materials depends on the intended application and the need for the desired mechanical properties.

2. Data Pre-Processing

2A. Steps Carried Out in Data Pre-Processing:

- Merged 2 datasets – dataset 1: Materials and their Mechanical Properties & dataset 2: Use. Since there was no common categorical column in both these datasets merging was carried out creating a new ID column as common column. Besides merged standard(Std), Material & Heat treatment into one single column named Material.

- Got Basic Information about the Dataset- Changed datatype Bool to Object for Use column.
- Removed the redundant rows of data.

- Performed Negation of the wrong dataset and then storing the correct data back in the dataset DataFrame.

- Checked for Null records- Since null record counts were larger instead of dropping the rows of null record Dropped columns with null records & columns which are not required for further analysis.

- Checked for Null records- Dropped the rows of null record.
- Created a new column named 'Materials' by concatenating the values in the three columns named 'Std', 'Material', 'Heat treatment', Renamed the column from 'Materials' to 'Material' & Reordered the columns in the DataFrame.
- Checked for text values in the numerical column before converting it to numeric datatype.

- Removed Duplicate records.
- Copyed the cleaned data into a new DataFrame.

- Defined Quan & Qual for Outlier Removal.

- Saved the Pre-processed dataset as a New File.

2B. Pre-Processed Dataset Info :

802 rows × 8 columns

3. Exploratory Analysis

3A. Conducted EA for the Following Questions:

Q1) Find top 10 materials used for structural components of buildings & bridges.

Solution: Materials with high Su, Sy & E are used in structural components of buildings & bridges.

Q2) What is the Avg Su of materials with high Sy & E ?

Q3) Find the materials suitable for biomechanics(Orthopaedic Implants).

Solution : The elastic modulus (E) of human bone typically falls in the range of 14,000 to 20,000 megapascals (MPa). The density (Ro) of human bone also varies but is generally in the range of 1800 to 2,100 kg/m³. Should Check for the materials satisfying these 2 criteria.

Q4) List the material which does not change in lateral dimensions when stretched or compressed.

Solution: For a material which does not undergo change in lateral dimensions when stretched or compressed its Poisson's ratio (mu) has to be 0

Q5) Find the more efficient materials which are weight-saving and exhibiting high stiffness.

Solution: Materials with higher specific stiffness are often preferred when weight-saving and stiffness are crucial, such as in aerospace applications where Specific Stiffness (SS) = Shear Modulus (G) / Density (Ro).

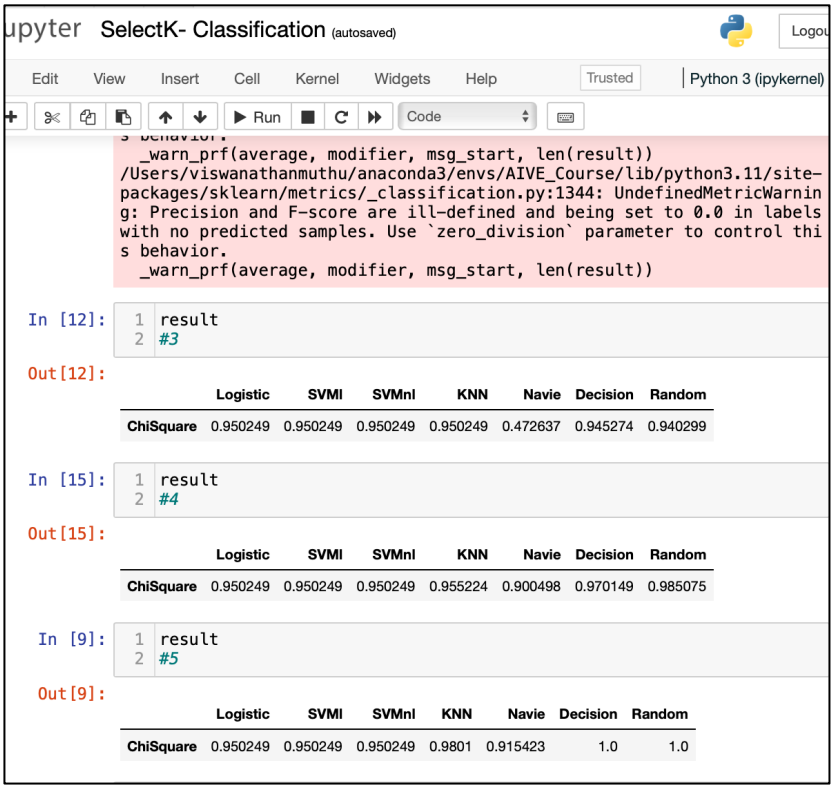
Q6) What is the safety margin, ensuring that materials do not reach their breaking point in practical applications ?

Solution : In many cases, Su is higher than Sy, which is common for most engineering materials. This difference between Su and Sy provides a safety margin, ensuring that materials do not reach their breaking point in practical applications.

Q7) Is there a relation between shear Modulus (G) and Poisson's Ratio (mu) by Materials ?

4. Feature Selection & Model Creation

4A. Proofs:



upyter SelectK- Classification (autosaved)

Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

```
  _warn_prf(average, modifier, msg_start, len(result))
/Users/viswanathanmuthu/anaconda3/envs/AIVE_Course/lib/python3.11/site-packages/sklearn/metrics/_classification.py:1344: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

In [12]:
```
result
#3
```

Out[12]:

	Logistic	SVMl	SVMnl	KNN	Navie	Decision	Random
ChiSquare	0.950249	0.950249	0.950249	0.950249	0.472637	0.945274	0.940299

In [15]:
```
result
#4
```

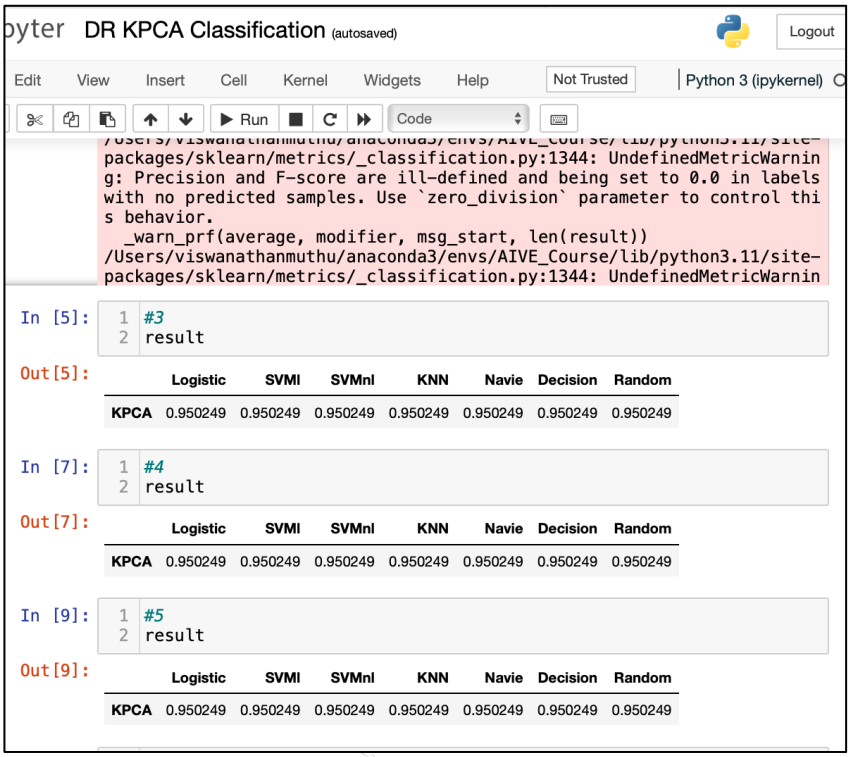
Out[15]:

	Logistic	SVMl	SVMnl	KNN	Navie	Decision	Random
ChiSquare	0.950249	0.950249	0.950249	0.955224	0.900498	0.970149	0.985075

In [9]:
```
result
#5
```

Out[9]:

	Logistic	SVMl	SVMnl	KNN	Navie	Decision	Random
ChiSquare	0.950249	0.950249	0.950249	0.9801	0.915423	1.0	1.0



byter DR KPCA Classification (autosaved)

Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```
packages/sklearn/metrics/_classification.py:1344: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/viswanathanmuthu/anaconda3/envs/AIVE_Course/lib/python3.11/site-packages/sklearn/metrics/_classification.py:1344: UndefinedMetricWarning
```

In [5]:
```
#3
result
```

Out[5]:

	Logistic	SVMl	SVMnl	KNN	Navie	Decision	Random
KPCA	0.950249	0.950249	0.950249	0.950249	0.950249	0.950249	0.950249

In [7]:
```
#4
result
```

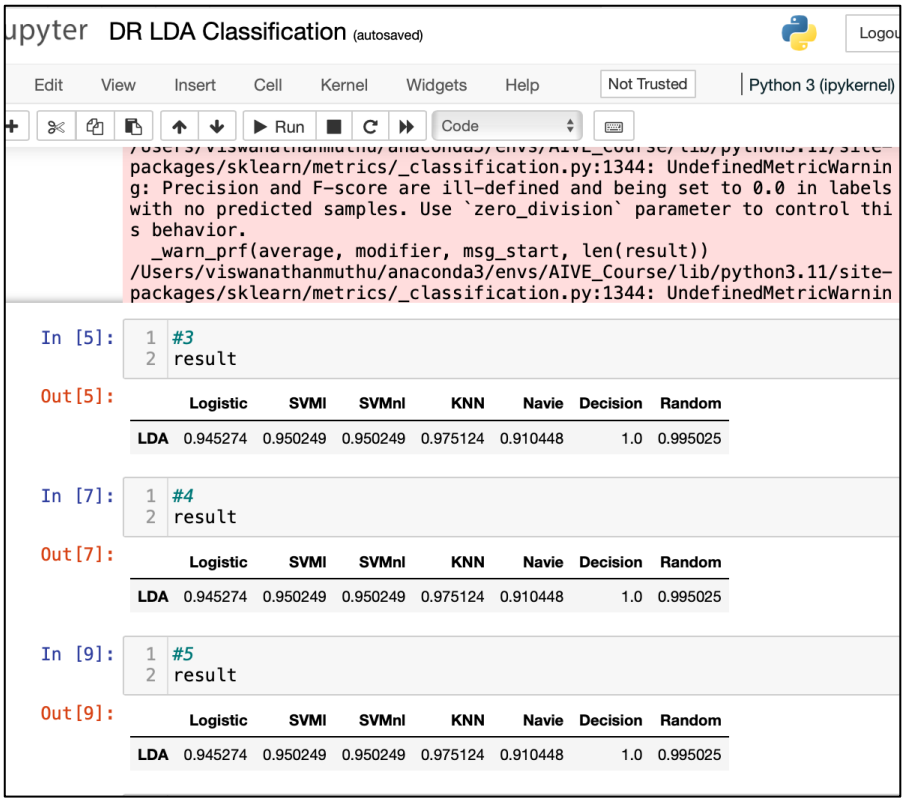
Out[7]:

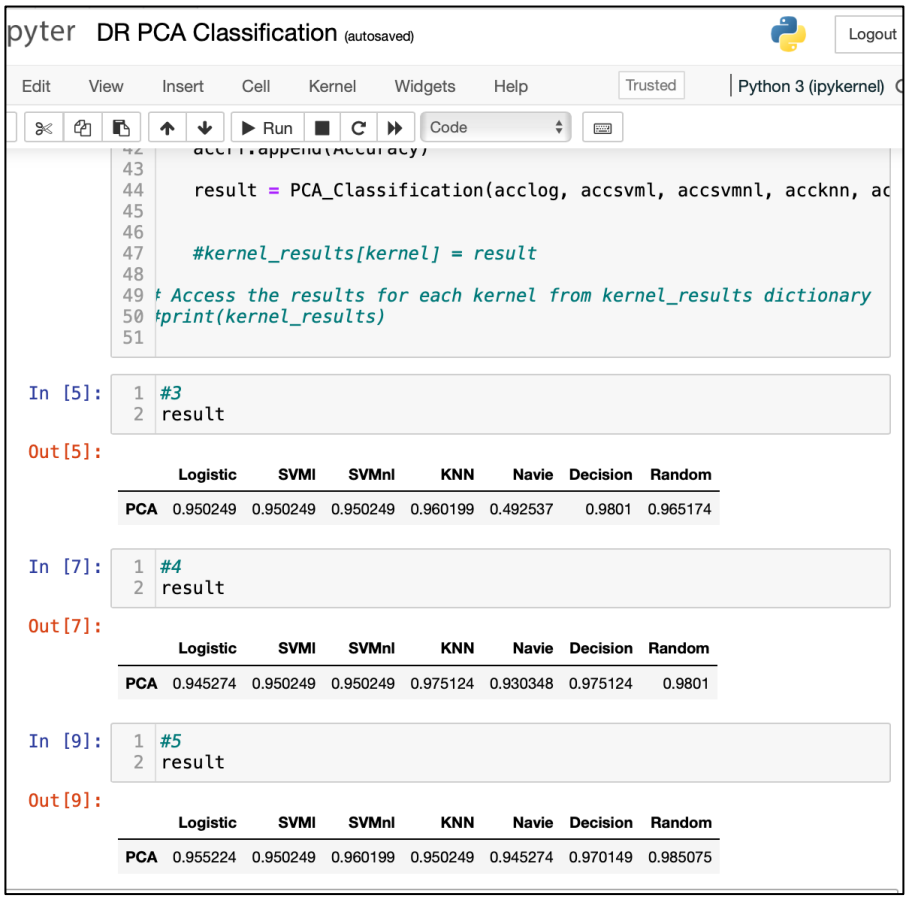
	Logistic	SVMl	SVMnl	KNN	Navie	Decision	Random
KPCA	0.950249	0.950249	0.950249	0.950249	0.950249	0.950249	0.950249

In [9]:
```
#5
result
```

Out[9]:

	Logistic	SVMl	SVMnl	KNN	Navie	Decision	Random
KPCA	0.950249	0.950249	0.950249	0.950249	0.950249	0.950249	0.950249


upyter DR LDA Classification (autosaved)
Logou
Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
packages/sklearn/metrics/_classification.py:1344: UndefinedMetricWarnin
g: Precision and F-score are ill-defined and being set to 0.0 in labels
with no predicted samples. Use `zero_division` parameter to control thi
s behavior.
_warn_prf(average, modifier, msg_start, len(result))
/Users/viswanathanmuthu/anaconda3/envs/AIVE_Course/lib/python3.11/site-
packages/sklearn/metrics/_classification.py:1344: UndefinedMetricWarnin
In [5]:
#3
result
Out[5]:
Logistic SVMl SVMnl KNN Navie Decision Random
LDA 0.945274 0.950249 0.950249 0.975124 0.910448 1.0 0.995025
In [7]:
#4
result
Out[7]:
Logistic SVMl SVMnl KNN Navie Decision Random
LDA 0.945274 0.950249 0.950249 0.975124 0.910448 1.0 0.995025
In [9]:
#5
result
Out[9]:
Logistic SVMl SVMnl KNN Navie Decision Random
LDA 0.945274 0.950249 0.950249 0.975124 0.910448 1.0 0.995025


pyter DR PCA Classification (autosaved)
Logout
Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
result = PCA_Classification(acclog, accsvml, accsvmnl, accknn, ac
#kernel_results[kernel] = result
Access the results for each kernel from kernel_results dictionary
#print(kernel_results)
In [5]:
#3
result
Out[5]:
Logistic SVMl SVMnl KNN Navie Decision Random
PCA 0.950249 0.950249 0.950249 0.960199 0.492537 0.9801 0.965174
In [7]:
#4
result
Out[7]:
Logistic SVMl SVMnl KNN Navie Decision Random
PCA 0.945274 0.950249 0.950249 0.975124 0.930348 0.975124 0.9801
In [9]:
#5
result
Out[9]:
Logistic SVMl SVMnl KNN Navie Decision Random
PCA 0.955224 0.950249 0.960199 0.950249 0.945274 0.970149 0.985075

4B. Inference :

After performing feature selection it is obvious that better results are obtained for K value (or) n components ie for 5 features out of 6 features among 8 features (2 features have been excluded since feature named Material is of no use for model creation & feature named Use is going to serve as target).

Though 5 features provide best result 4 features has been selected since it also provides almost near to best results thereby make it cost effective. The selected best four features are Su, E, G & Ro.

4. Model Creation

4C. 3 Stages of Problem Identification:

- **Stage 1 Domain Selection** : ML
- **Stage 2 Learning Selection** : Supervised Learning
- **Stage 3 Learning Classification** : Supervised Classification

4D. CM,CLF report, ROC - Proofs:

(DT GS C)

Phase-1 Model Creation (DT GS C)-Type 2 (autosaved)

View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```
The confusion Matrix:
 [[253   2]
 [  5   8]]
```

```
1 from sklearn.metrics import classification_report
2 clf_report = classification_report(y_test, grid_predictions)
3 print("The report:\n",clf_report)
```

```
The report:
              precision    recall  f1-score   support

           0       0.98      0.99      0.99       255
           1       0.80      0.62      0.70        13

    accuracy                           0.97       268
   macro avg       0.89      0.80      0.84       268
weighted avg       0.97      0.97      0.97       268
```

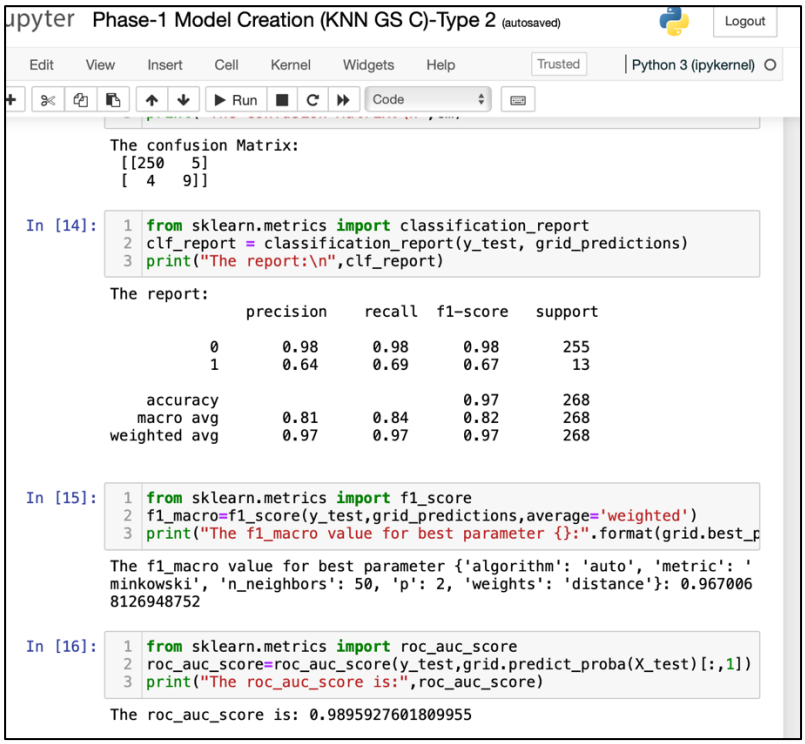
```
1 from sklearn.metrics import f1_score
2 f1_macro=f1_score(y_test,grid_predictions,average='weighted')
3 print("The f1_macro value for best parameter {}:".format(grid.best_p
```

```
The f1_macro value for best parameter {'criterion': 'gini', 'max_featur
es': 'sqrt', 'splitter': 'best'}: 0.9722535302131602
```

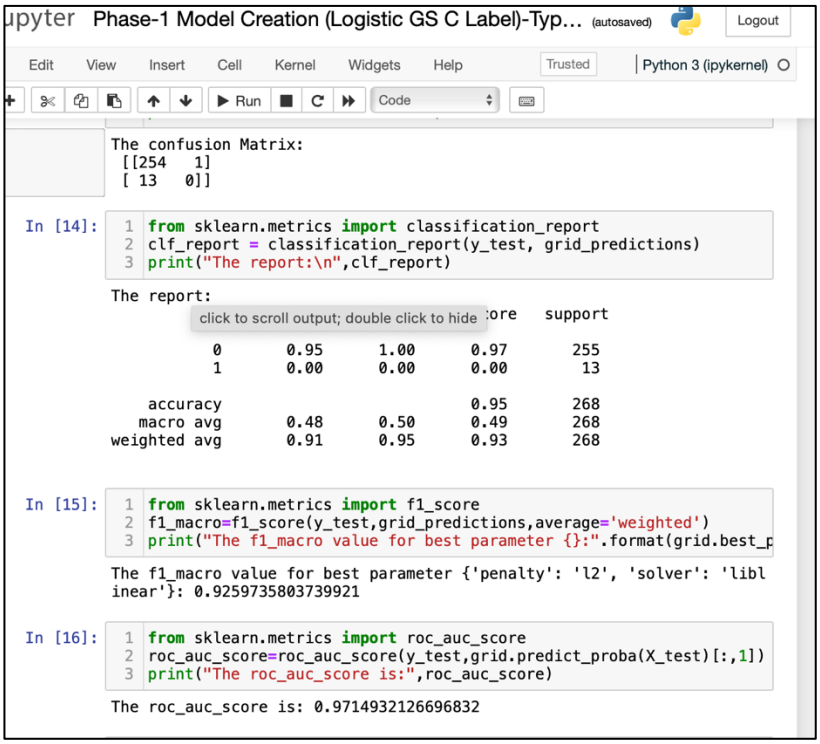
```
1 from sklearn.metrics import roc_auc_score
2 roc_auc_score=roc_auc_score(y_test,grid.predict_proba(X_test)[:,1])
3 print("The roc_auc_score is:",roc_auc_score)
```

```
The roc_auc_score is: 0.9556561085972851
```

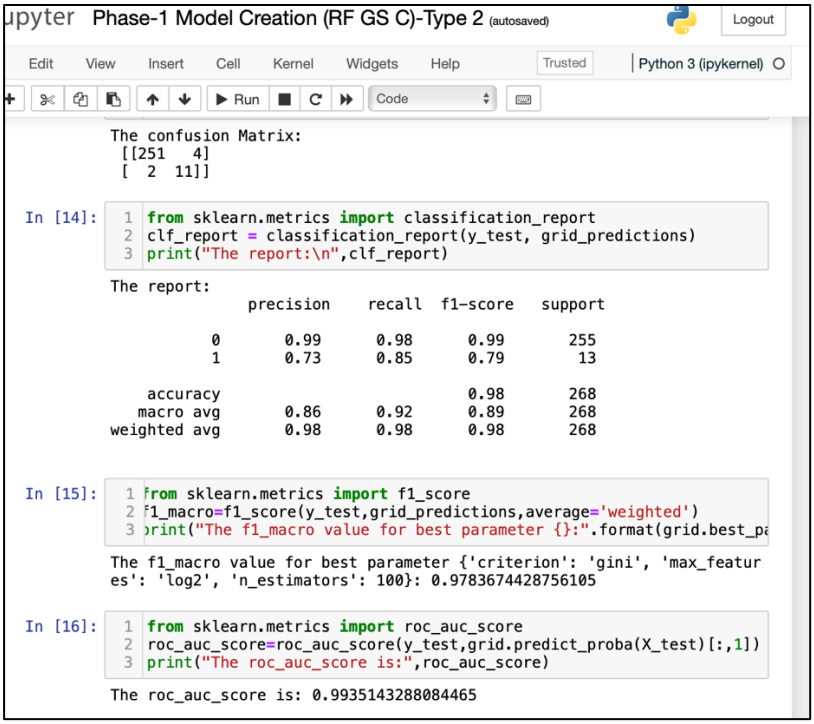
(KNN GS C)



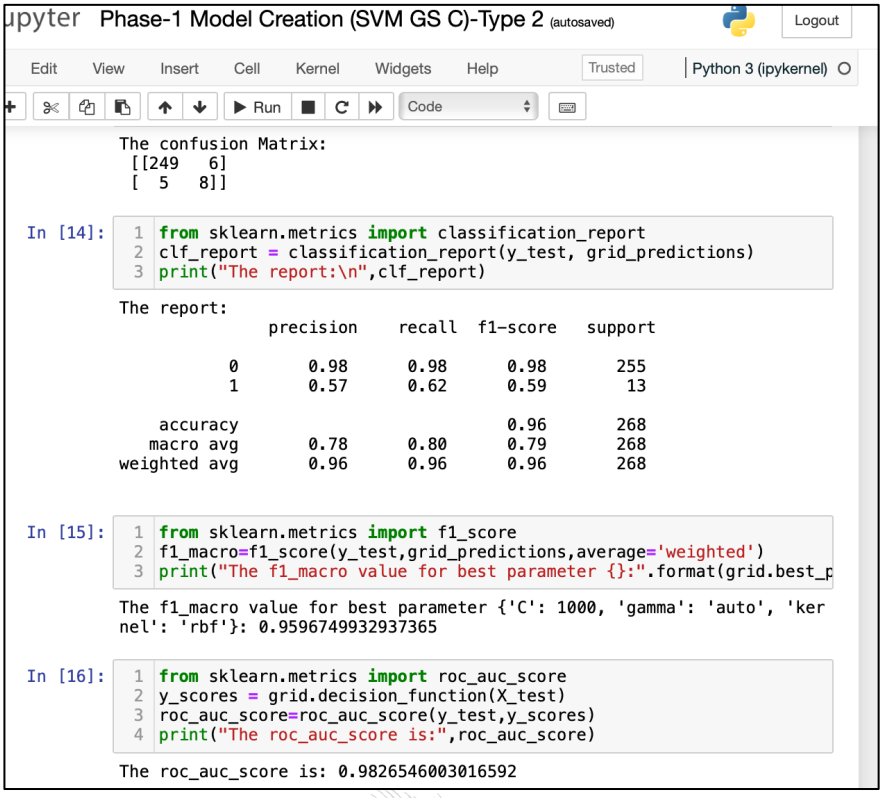
(Logistic GS C)



(RF GS C)



(SVM GS C)



4E. Summary :

Description	DT	KNN	Logistic	RF	SVM
CM	[[253 2] [5 8]]	[[250 5] [4 9]]	[[254 1] [13 0]]	[[251 4] [2 11]]	[[249 6] [5 8]]
F1 Macro	0.9722	0.9670	0.9260	0.9784	0.9597
Best Parameters	'criterion': 'gini', 'max_features': 'sqrt', 'splitter': 'best'	'algorithm': 'auto', 'metric': 'minkowski', 'n_neighbors': 50, 'p': 2, 'weights': 'distance'	'penalty': 'l2', 'solver': 'liblinear'	'criterion': 'entropy', 'max_features': 'auto', 'n_estimators': 10	'C': 1000, 'gamma': 'auto', 'kernel': 'rbf'
ROC AUC Score	0.9556	0.9895	0.97149	0.9923	0.9826

4F. Table Inference:

From the table, based on the analysis, the **Random Forest (RF) method** stands out as it has the highest F1 Macro value (0.9784) and a good balance of True Positive (TP: 11) and True Negative (TN: 251) with considerably low False Positives(FP:4) and False Negatives (2) in the confusion matrix [[251 4]
[2 11]].

4E. Result:

Since the model created via **Random Forest (RF) method** stands out as it has the highest F1 Macro value (0.9784) and a good balance of True Positive (TP: 11) and True Negative (TN: 251) with considerably low False Positives(FP:4) and False Negatives (2) in the confusion matrix [[251 4]
[2 11]] , so the model created by this RF method can be considered for Deployment Phase.

5. Appendix :

Abbreviations	Expansion
SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest
KNN	K Nearest Neighbour
CM	Confusion Matrix
ROC AUC Value	"Receiver Operating Characteristic - Area Under the Curve"

----------END OF THE DOCUMENT---------