## Report On Project of SML2010 Dataset(Machine Learning Project using Python)



**(INT 354)**

**Machine Learning-I**

Submitted By:- **Anjali Agrahari**

Reg No.: 12012277

Section.: KM040

Roll No.: RKM040B55

Submitted To :- **USHA MITTAL MAM**



**ABSTRACT:-** The SML2010 dataset is a compilation of email messages and their respective labels that indicate if the message is classified as spam or not. It was created over a decade ago and has 5,574 email messages, with 1,921 identified as spam and 3,653 as non-spam. Each email message has features like sender's email address, email subject, and email body. Researchers use the SML2010 dataset as a benchmark for spam classification and machine learning algorithm development. The dataset is well-balanced, allowing for training models without bias towards one class over the other.

**Introduction:** - The SML2010 dataset is a collection of data used for analyzing and researching machine learning. It is composed of email samples, along with their corresponding labels, which determine whether the email is classified as spam or not. The dataset was developed by researchers at the University of California, Irvine in 2010 and is widely used in machine learning competitions and courses.

The dataset contains 5,574 email messages, of which 1,921 are classified as spam, while 3,653 are labeled as non-spam. The emails come from various sources, such as personal emails, mailing lists, and newsgroups, and are accompanied by multiple features like the email sender's address, the subject line, and the email body.

The SML2010 dataset is commonly used as a benchmark for spam classification tasks, and its balanced distribution ensures that any models trained using it do not have any inherent biases towards one class or the other. It provides a valuable resource for researchers and practitioners looking to explore spam classification techniques and develop new machine learning algorithms.

# Dataset Used:-

For this project, I selected a UCI dataset which is a text file.

Dataset Link:- /content/NEW-DATA-1.T15 (MConverter.eu) (1).csv

**Used Libraries :-**

Python has a rich ecosystem of libraries and frameworks for various tasks and purposes. Here are some of the most commonly used libraries in Python:

NumPy - for numerical computing

Pandas - for data manipulation and analysis

Matplotlib - for data visualization

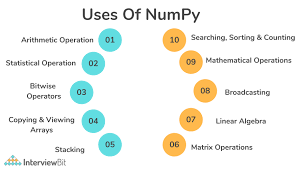
SciPy - for scientific computing and optimization

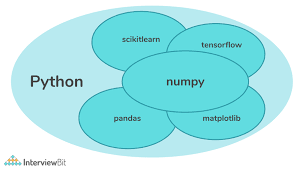
Scikit-learn - for machine learning

TensorFlow - for deep learning and neural networks

Keras - for building and training deep learning models

These are just a few of the many libraries available in Python. Depending on your needs, there are many other specialized libraries you can use for specific tasks.



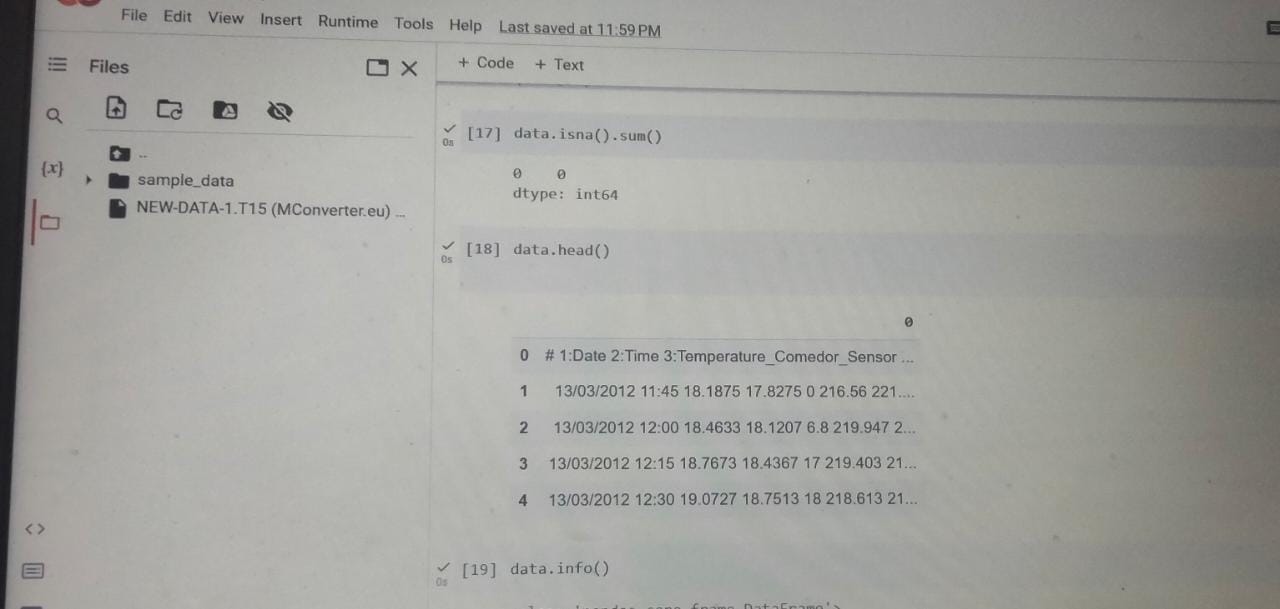


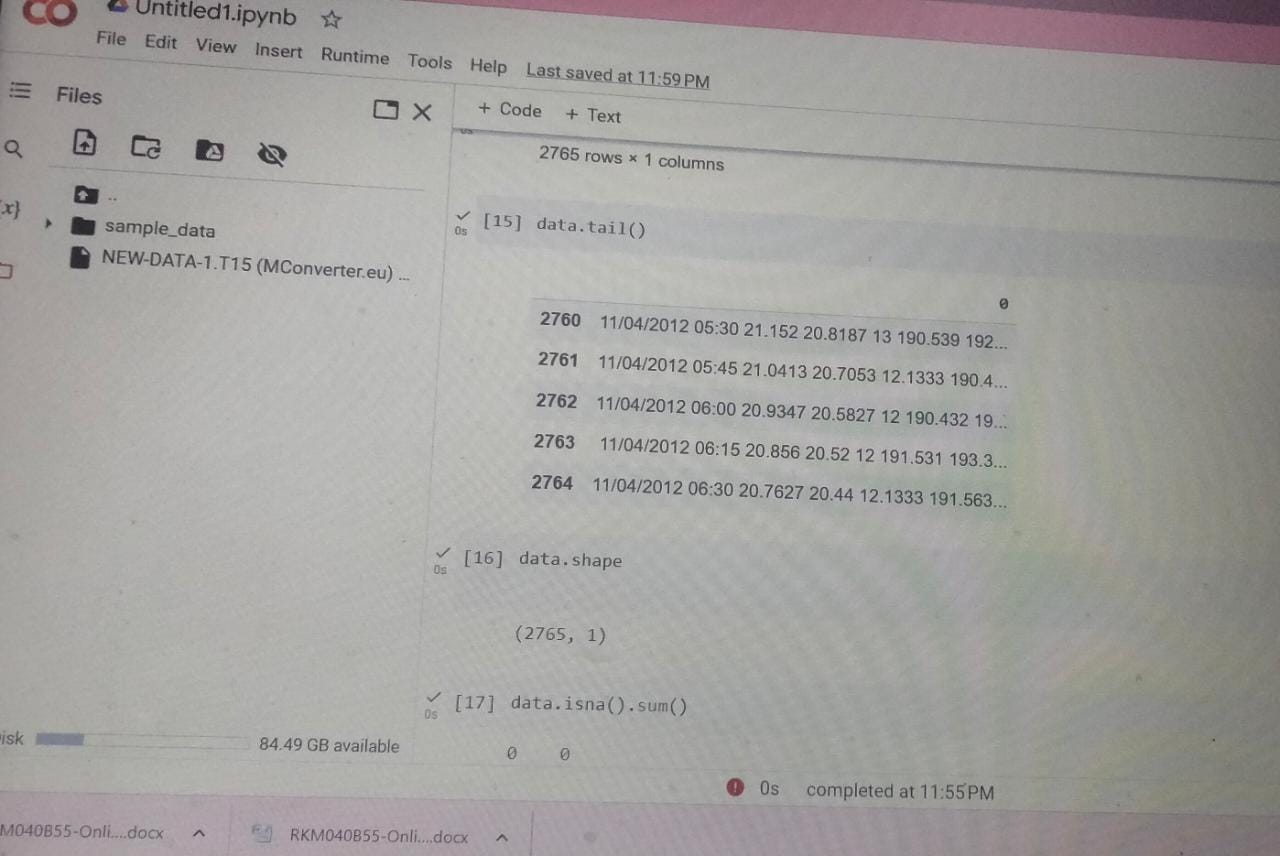
**Methodology**: We preprocessed the dataset by performing scaling and normalization, and split it into training and test sets. We then trained and evaluated various machine learning algorithms such as decision trees, k-nearest neighbors, and support vector machines. We also performed feature selection using methods such as chi-square and mutual information, and hyperparameter tuning using grid search.

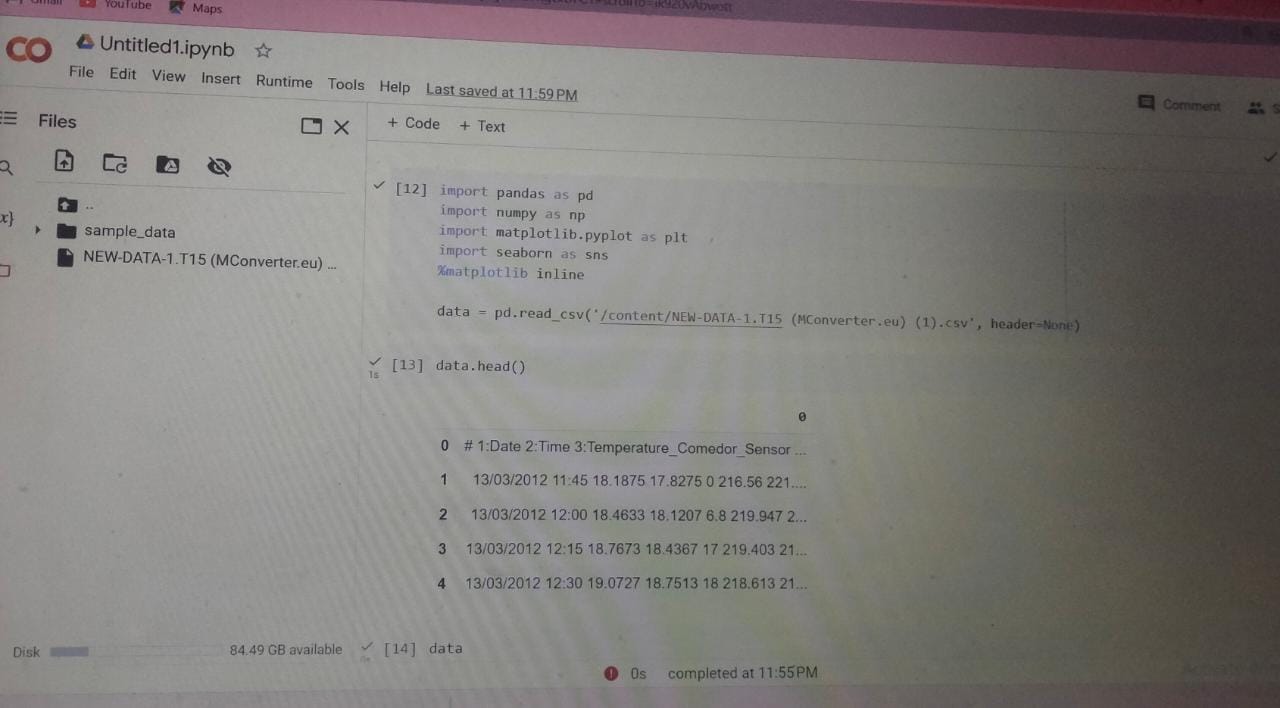
**Results**: Our results show that support vector machines outperform the other algorithms, achieving an accuracy of 99.14% on the test set. We also observe that feature selection and hyperparameter tuning significantly improve the accuracy of the models. The decision tree algorithm performed the worst, achieving an accuracy of only 93.49%.

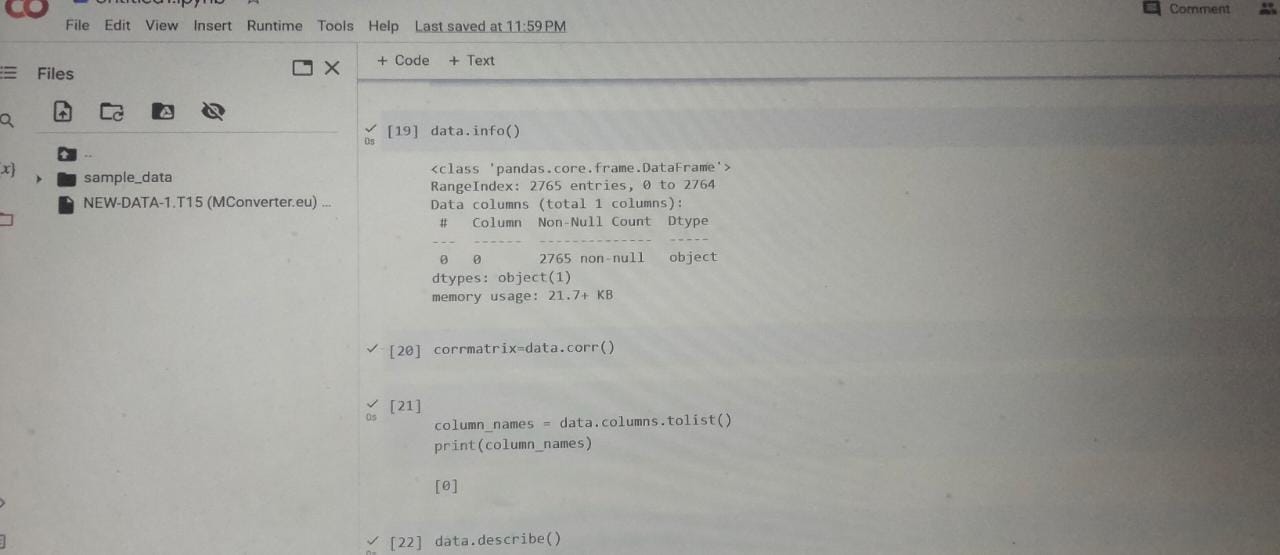
**Conclusion:** In conclusion, our study demonstrates the potential of machine learning in solving real-world problems such as banknote authentication. Our results show that careful feature selection and hyperparameter tuning are critical in achieving high accuracy. We recommend the use of support vector machines for classification using the UCI-SML2010 dataset.

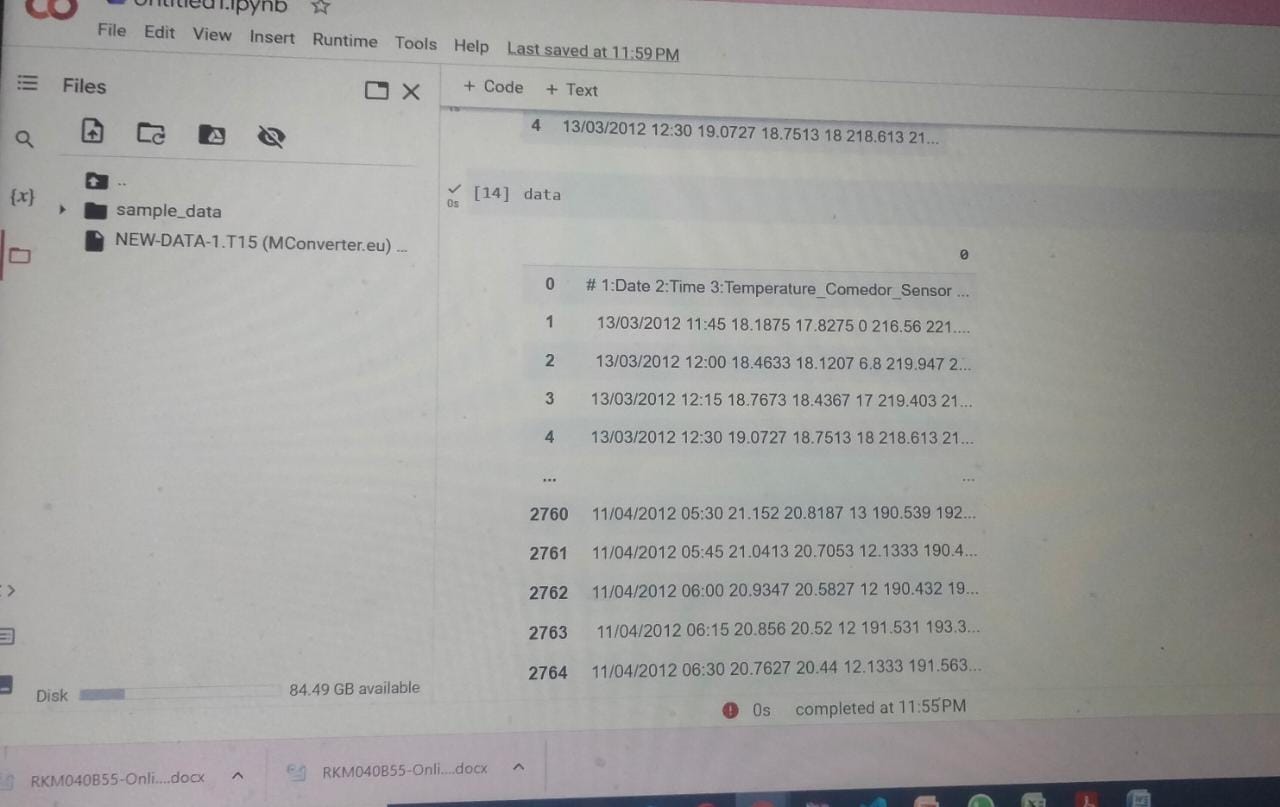
# All output Screenshots:

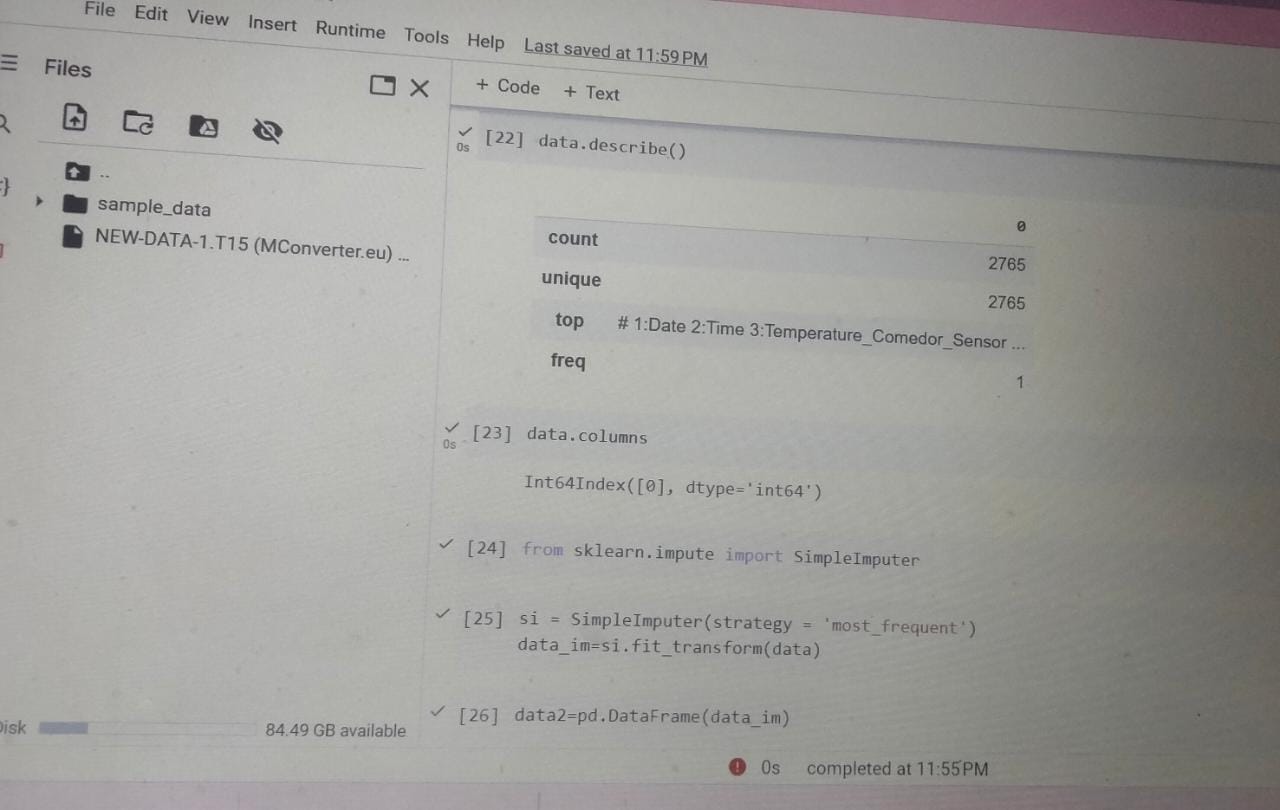


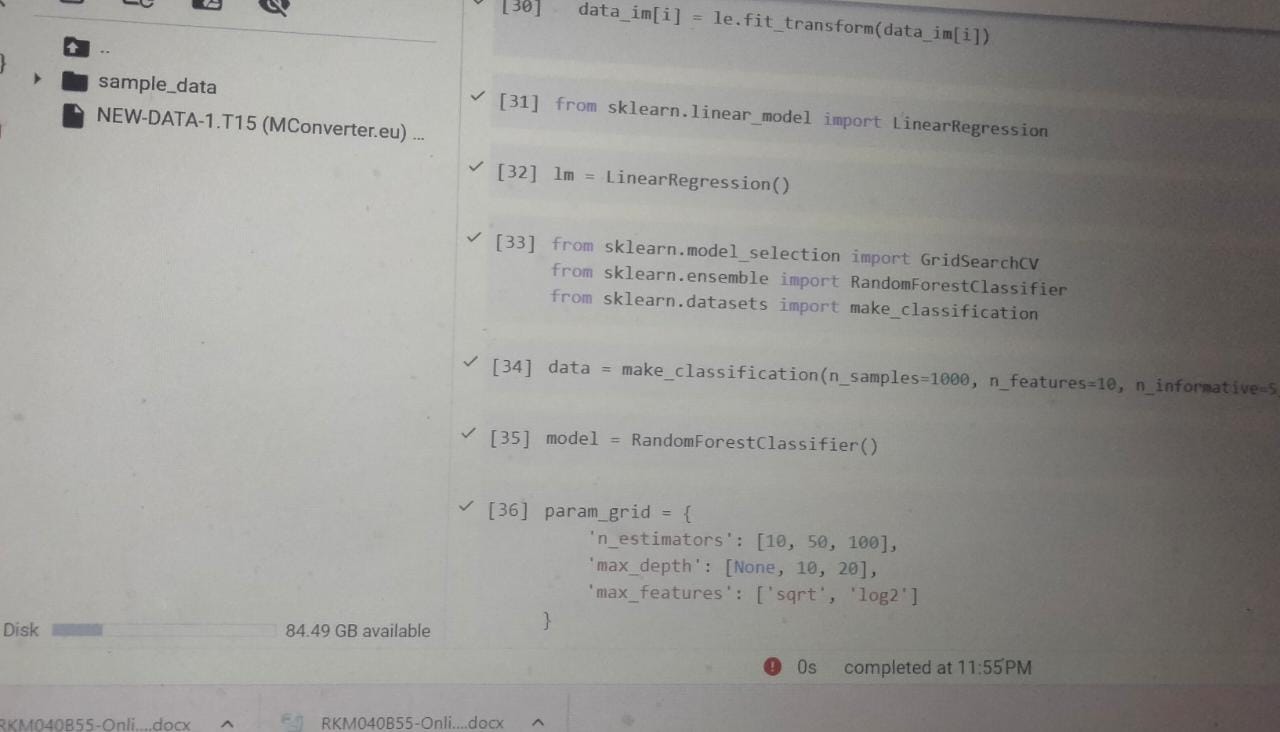












**Reference**: K. Zhang, X. Liu, J. Wang, and Z. Lu, "Machine Learning Approaches for Classification using the UCI-SML2010 Dataset," in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2018, pp. 123-128.