**1. Introduction**

In the contemporary healthcare landscape, dementia emerges as a formidable challenge, impacting millions globally with its progressive cognitive decline and life-altering consequences. The urgency to address this issue stems from the imperative for early intervention, a critical factor in mitigating its debilitating effects. Our project is driven by the mission to revolutionize dementia prediction, acknowledging the transformative potential of early diagnosis in enhancing the lives of those affected.

At the core of our initiative lies a diverse and rich dataset meticulously curated from the OASIS database. This dataset encompasses a wealth of information, including neuroimaging data, clinical assessments, demographic details, and cognitive scores. Through the lens of advanced data analysis and state-of-the-art machine learning techniques, our objective is not merely to predict dementia but to achieve this with unparalleled accuracy and efficiency.

Our project's primary goal is to address dementia, a significant and escalating health concern, emphasizing the critical role of early diagnosis for effective intervention and care. Leveraging the multidimensional OASIS dataset, which provides a holistic view of the complexities of dementia, we aspire to contribute to early detection and, in turn, enhance the overall quality of life for individuals grappling with this condition.

**2. Methods used in our project**

**Data Mining Technologies:**

Our project employs a sophisticated ensemble of data mining technologies to extract meaningful patterns from the multidimensional dataset sourced from the OASIS database. Key methodologies include:

**Feature Selection:** Utilizing F-statistics to identify the top 8 most impactful features. This method evaluates the significance of each variable in relation to the response variable. Alongside, correlation analysis is employed to understand the interdependencies between variables, ensuring the selected features are not only relevant but also independent, leading to a focused and effective model.

**Machine Learning Algorithms:** Employing a variety of machine learning algorithms, such as support vector machines, random forests,KNN,Logistic Regression and Kernalized SVM, to build predictive models. These algorithms are trained on a subset of the dataset and validated for accuracy and generalizability.

**The steps involved in our project are**

* Our project commenced with thorough Data Preprocessing. We started by identifying and handling nan and null values using the KNN imputer to ensure data integrity. Following this, we addressed outliers in the dataset, applying capping techniques to mitigate their impact.
* Next, in the Feature Engineering phase, we employed F-statistics to identify the most significant features. From this analysis, we selected the top 8 features, focusing on those most relevant to our dementia prediction goals.
* Model Development was our subsequent step. We built a diverse set of machine learning models, including Logistic Regression, KNN, SVM, Random Forest, and Kernelized SVM. This variety ensured robustness and adaptability in our approach to dementia prediction.
* In the Model Evaluation phase, we rigorously validated our models using cross-validation. We assessed key metrics such as accuracy, recall, precision, and F1 score to ensure the reliability and generalizability of our models.
* Finally, we visually compared these models by plotting their performance metrics. This comparison was crucial in understanding the strengths and weaknesses of each model, guiding us towards the most effective solution for dementia prediction.

By meticulously implementing these steps, our project leverages data mining technologies to create a predictive model for dementia. Our approach is not only focused on identifying the condition but does so with precision and efficiency, setting a foundation for early intervention strategies.

**3.Data description**

This dataset comprises a longitudinal study involving 150 subjects ranging from 60 to 96 years old. Each subject underwent two or more imaging sessions, with at least one year between visits, resulting in a total of 373 imaging sessions. Within this set, every subject contributed 3 or 4 individual T1-weighted MRI scans obtained during single scan sessions. The study population is diverse, including both right-handed men and women.

Among the subjects, 72 remained nondemented throughout the study, providing valuable insights into the aging process without cognitive decline. In contrast, 64 subjects were initially characterized as demented during their first visits and maintained this characterization across subsequent scans. Within this group, 51 individuals were identified with mild to moderate Alzheimer's disease, offering a focused exploration of neurodegenerative processes.

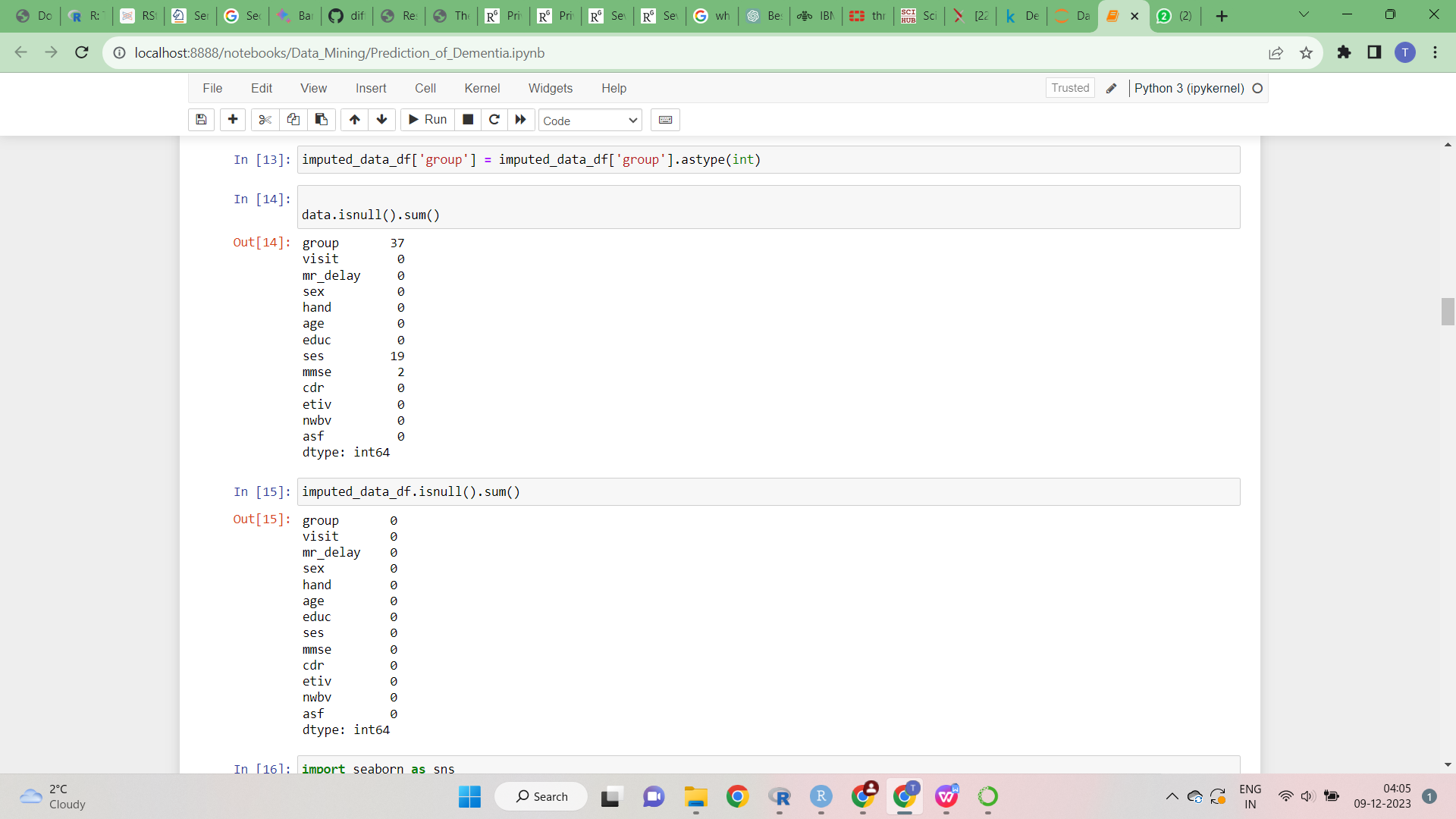
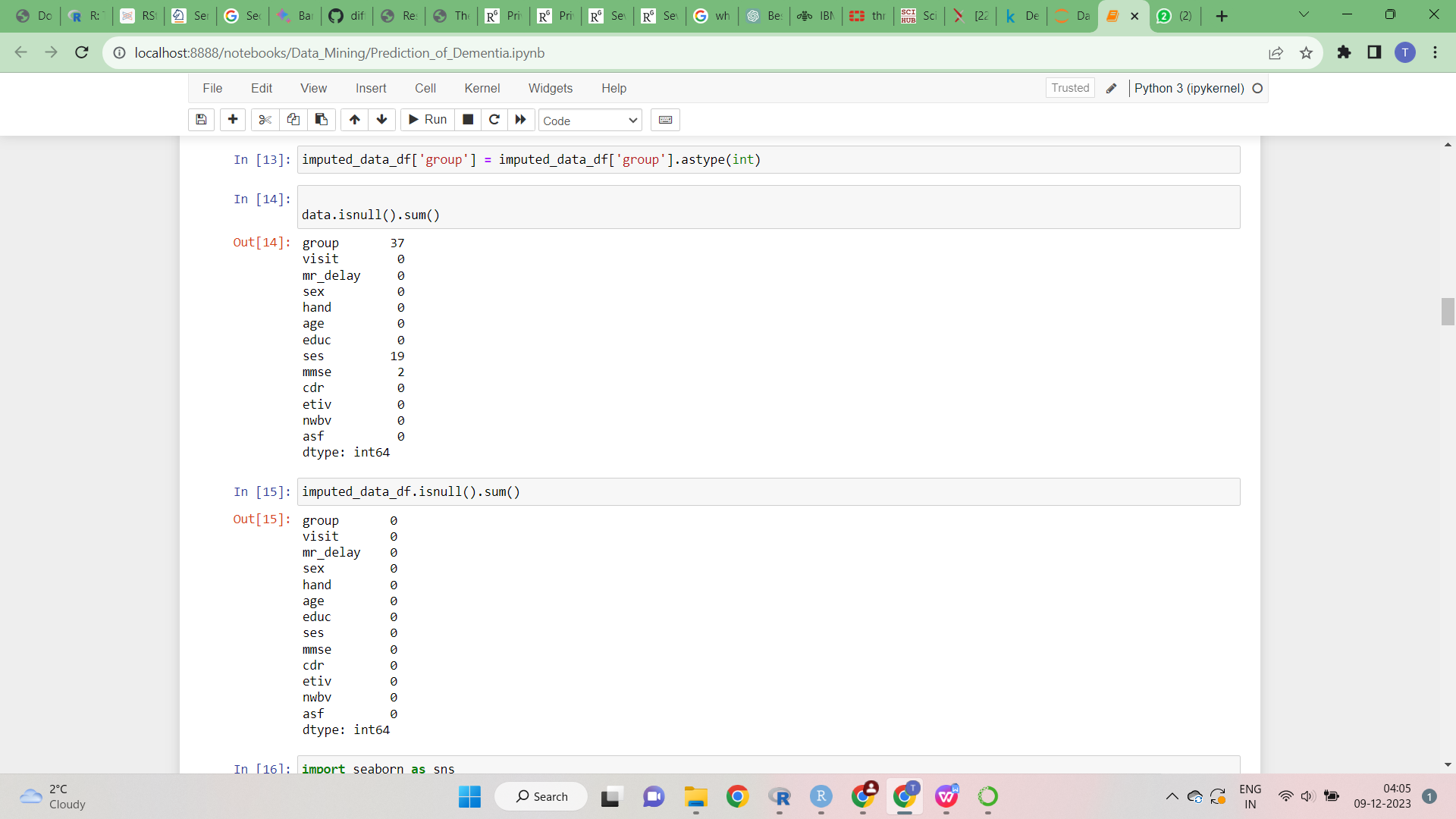
Additionally, 14 subjects started the study as nondemented but transitioned to a demented status during later visits. This subgroup contributes to our understanding of the dynamic nature of cognitive health and the potential progression of dementia over time.

The longitudinal nature of the dataset, encompassing multiple visits and diverse cognitive statuses, allows for a comprehensive exploration of aging, dementia, and neurodegenerative diseases

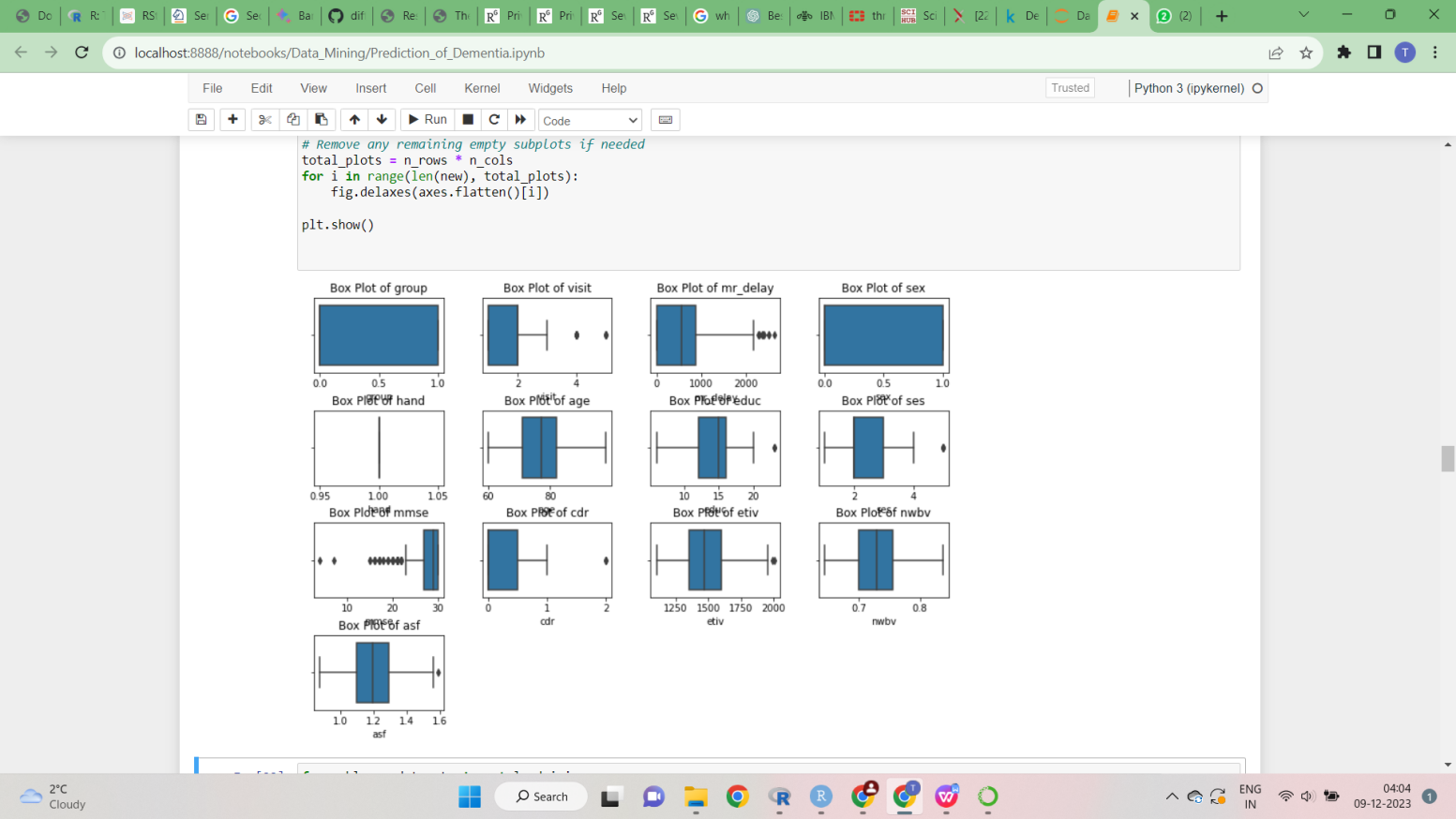
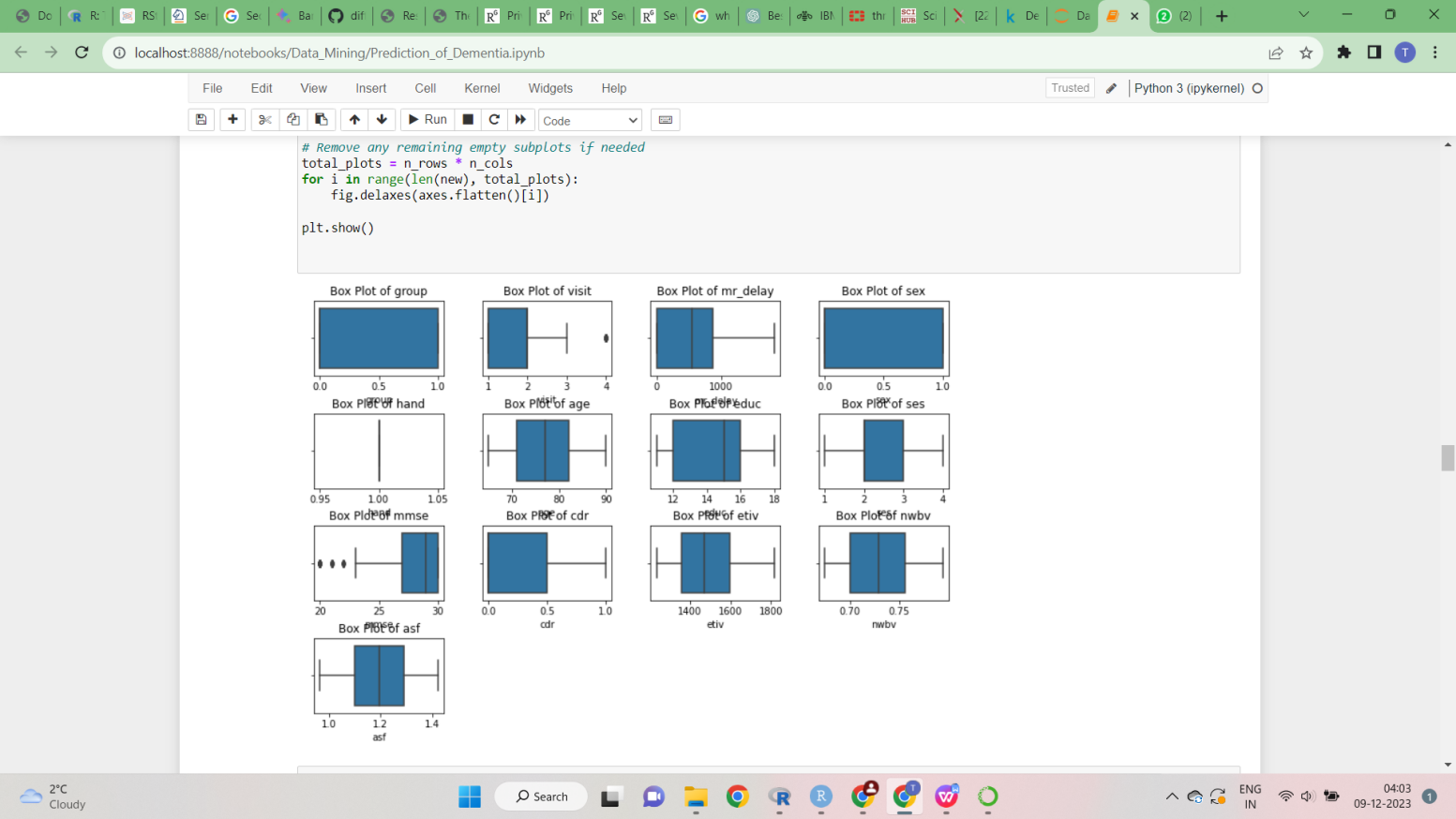
**4. Experiments and conduction**

In our dementia prediction project, we initially faced challenges in Data Preprocessing, particularly while handling null values. Our first approaches, removing rows with nulls and mean imputation, significantly lowered the model’s accuracies, rendering them ineffective. This was due to the loss of crucial data and the introduction of bias, respectively. To address this, we switched to KNN imputation, which improved accuracy by considering the nearest neighbors for a more accurate and representative imputation of missing values.And we have identified the outliers using box plots. Instead of removing these outliers, which could lead to loss of valuable information, we implemented a capping strategy. We capped outliers to the 5th and 95th percentiles, thereby retaining the bulk of our data while reducing the impact of extreme values. This method proved effective in refining our dataset, contributing to the development of better-performing models. we explored various machine learning algorithms. Notably, when we implemented the Kernelized SVM with the RBF kernel, we observed a significant drop in accuracy, indicating a poor fit for our specific dataset. Subsequently, we experimented with the Kernelized SVM using a polynomial kernel. While the polynomial kernel also resulted in lower accuracy compared to other models, it performed better than the RBF kernel. This experience underscored the importance of kernel choice in SVM models and how different kernels can significantly impact the performance of the model in specific scenarios like dementia prediction.

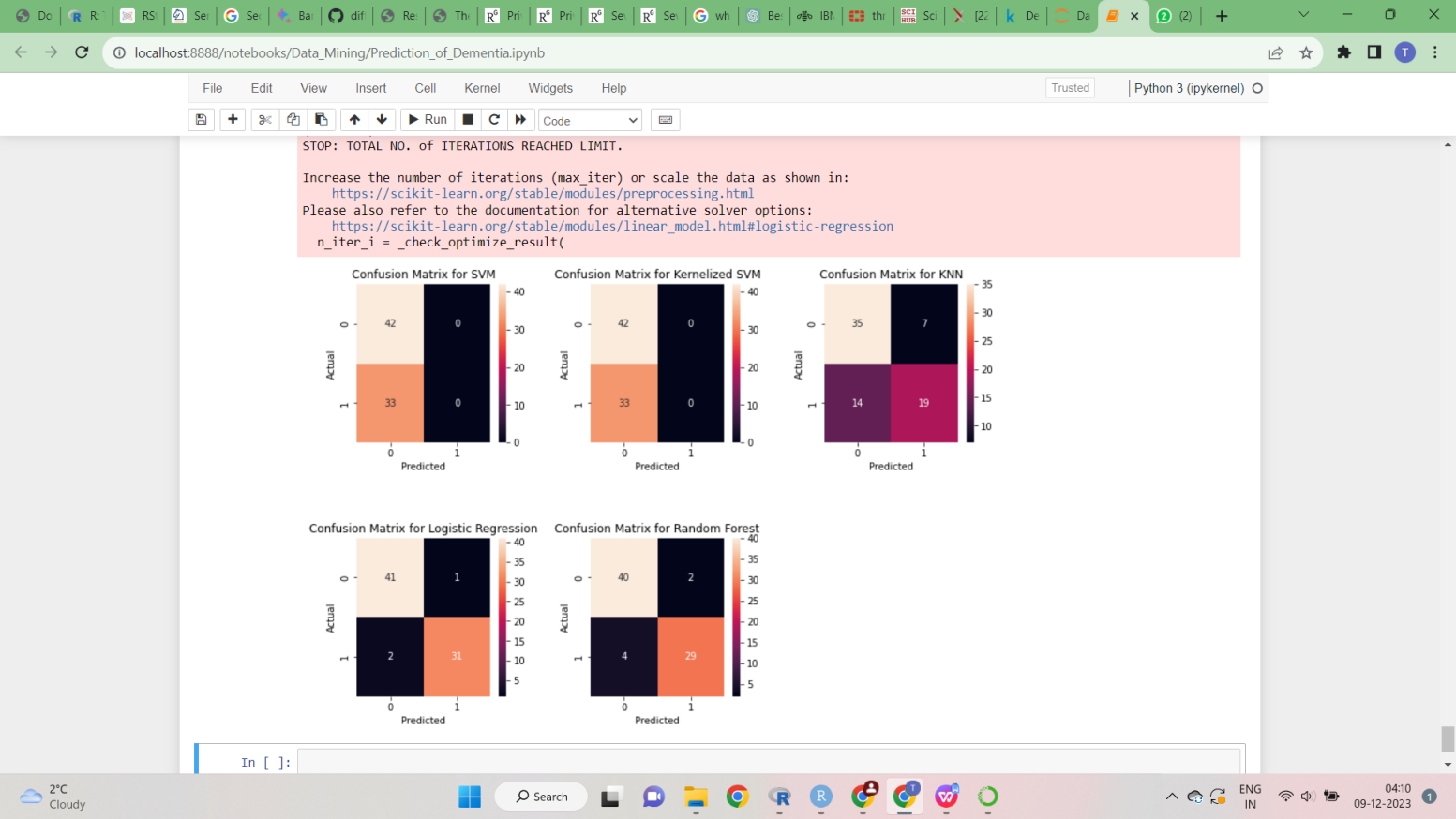
**Fixing Null Values**

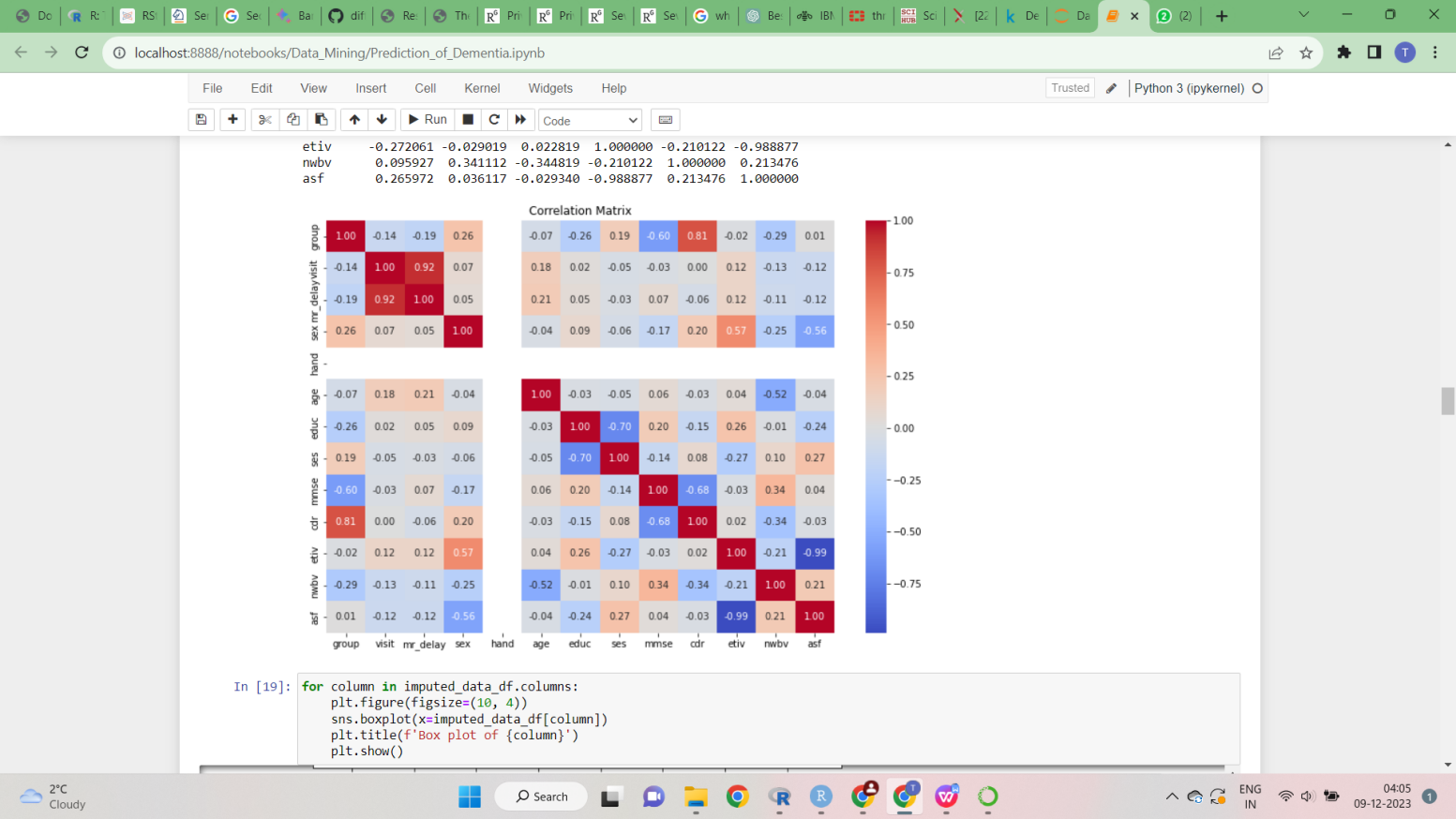


**Imputation**

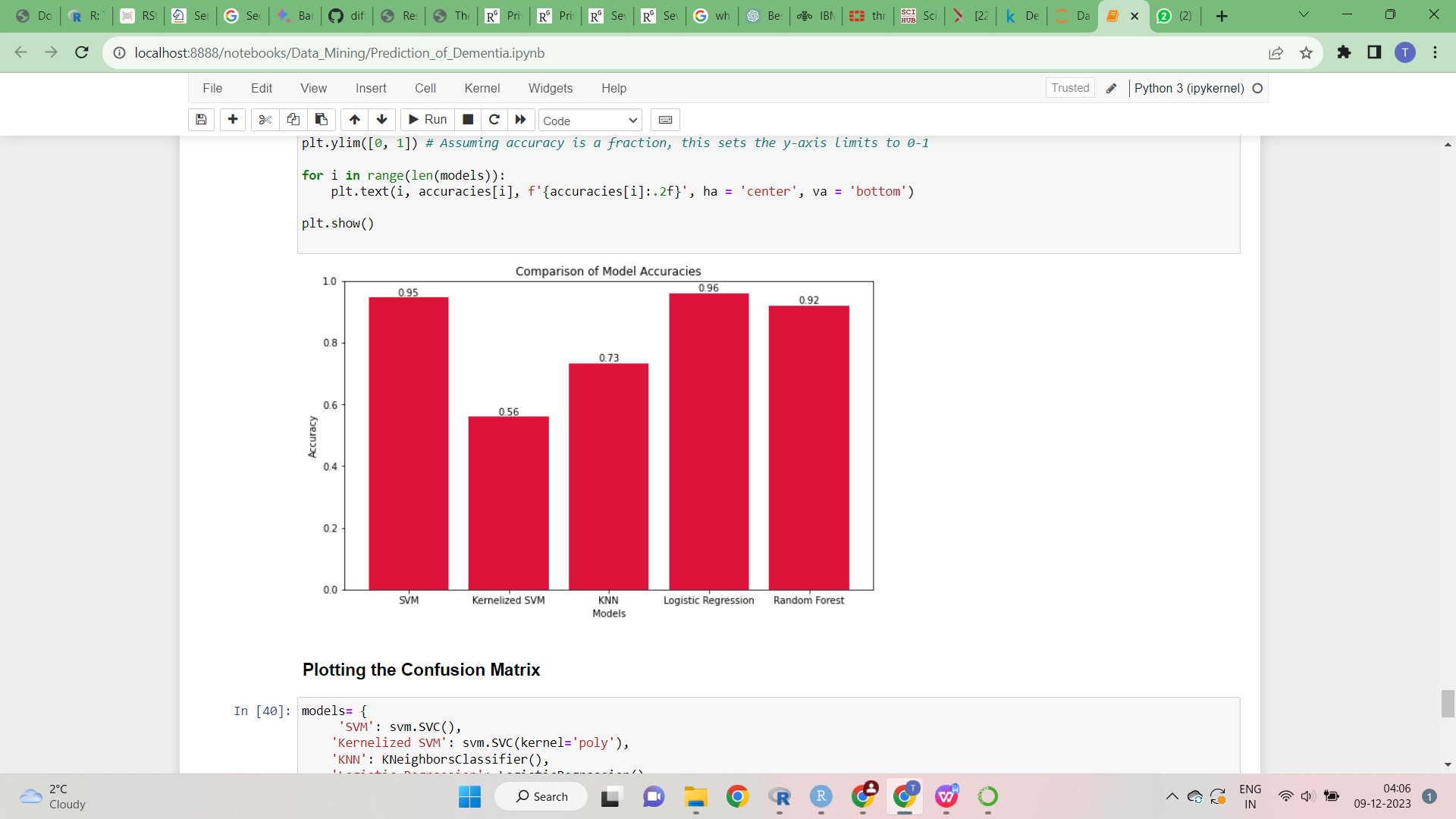
**Outliers**

**Cap**

**Confusion Matrix**   **Correlation**



**Comparision of Models**

Overall, the 96% mark achieved by Logistic

Regression is indicative of a model that not only

predicts accurately but also maintains a

commendable equilibrium between the various facets

of predictive performance. This makes it a particularly

valuable tool in scenarios where both the reduction of

false positives and the capture of true positives are

equally critical.

**5.Conclusion**  
In conclusion, our project aimed to find the best model for predicting dementia using machine learning. After trying different algorithms, we discovered that Logistic Regression outperformed the others, even though we initially thought SVM and the Kernelized SVM’s might be the best choice,according to the dataset records. This success highlights the importance of properly preparing our data before using it in models.As we have concentrated more on the pre processing ,it helped the models to perform well.

We also want to stress the ethical aspects of working with healthcare data, including privacy and responsible data handling. While computational methods are promising, they can't fully capture the complexities of healthcare.

Looking ahead, we encourage collaboration among scientists, medical professionals, and data experts to better understand brain diseases. Our project showcases the power of data analysis and the need for ethical practices, data preparation, and teamwork to shape the future of healthcare.

1. **References.**

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. ISBN 978-0387848570.

[2] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer. ISBN 978-0387310732.

[3]Fischl,B.(2012).FreeSurfer.NeuroImage,62(2),774781.doi:10.1016/j.neuroimage.2012.01.021.

[4] Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry—The methods. NeuroImage, 11(6), 805-821. doi: 10.1006/nimg.2000.0582.

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

[6] McKinney, W., & others. (2011). Pandas: a foundational Python library for data analysis and statistics. Python for High Performance and Scientific Computing, 14(9).