**A logo for college computing

Description automatically generated**

**Assessment Cover Page**

|  |  |
| --- | --- |
| ***Student Full Name: Conor Kirwan*** |  |
| ***Student Number: 24130*** |  |
| ***Module Title: Programming for AI*** |  |
| ***Assessment Title: CA1*** |  |
| ***Assessment Due Date: 31/Oct/2024*** |  |
| ***Date of Submission: 31/Oct/2024***  ***GitHub –*** [***sba24130***](https://github.com/sba24130/sba24130---CA1--Programming-for-AI-)  ***Videos -*** [***Drive***](https://drive.google.com/drive/u/0/folders/1SGMIvl_OmYrgw8c4JM69jys1Xrf3O79i) |  |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

**Contents**

[Question 1 3](#_Toc181309413)

[Question 2 3](#_Toc181309414)

[Question 3 4](#_Toc181309415)

[Question 4 4](#_Toc181309416)

[Question 5 5](#_Toc181309417)

[References 7](#_Toc181309418)

[Appendix - Q1 Code 8](#_Toc181309419)

[Appendix – Q2 Code 9](#_Toc181309420)

[Appendix – Q3 Code 12](#_Toc181309421)

[Appendix – Q4 Code 15](#_Toc181309422)

[Appendix – Q5 Code 19](#_Toc181309423)

[Appendix - Q5 Figures 25](#_Toc181309424)

# Question 1

A critical step in machine learning and artificial intelligence is data preprocessing, as it ensures that the data is in a suitable format for further analysis and modeling (Serey *et al.*, 2021) . By identifying the data types of features, which is particularly important in most machine learning algorithms that require input data to be in numerical form, and as such they cannot directly process categorical or string data (*Applied Predictive Modeling | SpringerLink*, 2013)

A Python function to categorize the columns of a dataset into numeric and categorical types was implemented in this question, without using libraries. The function iterates over each column and checks the data type of its values, using Python's built-in instance function to distinguish between integers (numeric) and strings (categorical) (*Learning Python, 5th Edition*, 2013) This categorization is crucial because it allows for appropriate data manipulation and encoding techniques to be applied to the different data types (*(PDF) Data Preprocessing for Supervised Learning, 2016*)

While the implemented function effectively categorizes columns containing integers as numeric and strings as categorical, it has a limitation: it does not account for other data types, such as characters or floating-point numbers. In real-world datasets, features may consist of various data types, and a more comprehensive approach to data type identification may be required (*Applied Predictive Modeling | SpringerLink*, 2013).

Moreover, the function's approach of iterating over each value in a column can be computationally intensive, especially for large datasets. More efficient methods, such as using vectorized operations or sampling techniques, could be explored to improve performance (*Applied Predictive Modeling | SpringerLink*, 2013).

However, the implemented function provides a good baseline for subsequent data preprocessing tasks, such as encoding categorical variables, scaling numerical features, and handling missing values. These preprocessing steps are essential for ensuring that the data is properly formatted and prepared for various machine learning algorithms and techniques, ultimately improving the performance and accuracy of AI models.

# Question 2

Data preprocessing is a fundamental step in machine learning and data analysis pipelines. It involves transforming raw data into a format suitable for further analysis and modelling (Serey *et al.*, 2021) This task emphasizes the importance of data concatenation, removal of duplicates, and correlation analysis, all important aspects of data preprocessing.

Concatenating datasets is a common operation when working with data from multiple sources or different formats. The use of Pandas provides a powerful tool for concatenating Data Frames along rows or columns, ensuring painless integration of data, even though in this assignment the removal of duplicate rows was unsuccessful. The removal of duplicate rows is critical for data cleaning and ensuring data integrity, as duplicates can introduce bias, create bad date, and possibly lead to inaccurate analysis (*Applied Predictive Modeling | SpringerLink*, 2013).

In the NumPy section, calculating the correlation matrix is a vital step in understanding the relationships between features. The identification a high correlation feature aids in feature selection and dimensionality reduction, potentially improving model performance and interpretability (*Applied Predictive Modeling | SpringerLink*, 2013).

The combined use of Pandas and NumPy libraries streamlines data preprocessing tasks, enabling efficient data manipulation, cleaning, and analysis. This task highlights the importance of mastering these libraries for effective data preprocessing in machine learning and data analysis workflows.

# Question 3

The task of creating a NumPy array with random elements and taking the mean of every 5-sample window is a common data processing operation in various fields, including signal processing and time-series analysis (Smith, 2003). By calculating the mean over a sliding window, this algorithm helps smooth out fluctuations and noise in the data, revealing underlying patterns and trends (Oppenheim & Schafer, 2014).

The determination of minimum and maximum values of the window means is useful for understanding the range and variability of the data. This information can be valuable for data normalization, outlier detection, and feature scaling ( *Wes McKinney - Google Books*, 2017.) Furthermore, finding the difference between absolute min & max means can provide a measure of the overall spread or dispersion of the data, which can be crucial for selecting appropriate analysis techniques or tuning algorithm parameters (*Applied Predictive Modeling | SpringerLink*, 2013).

This task demonstrates the versatility of NumPy in handling array operations and highlights the importance of understanding fundamental data processing techniques. The applications are often employed as preprocessing steps in machine learning pipelines, ensuring that the data is properly prepared for further analysis or modeling (Serey *et al.*, 2021)

# Question 4

Relational Database Management Systems (RDBMS) and SQL (Structured Query Language) play an important role in the field of AI. They allow for data storage, retrieval, preprocessing, and integration, which are essential for developing and deploying AI applications.

AI systems rely on large quantities of data for training machine learning models to make accurate predictions (Serey *et al.*, 2021). RDBMS, such as MySQL, provide a secure, scalable solution for organizing, managing & manipulating this data. They offer a structured way to store and manipulate data, ensuring data integrity, consistency, and security (*Database System Concepts - 7th edition*, 2020)

Data preprocessing is a critical step in AI & ML, and SQL is a capable tool for cleaning, transforming, and integrating data from multiple sources (Domingos, 2012.) Data querying and manipulation in SQL enables tasks such as filtering, sorting, joining, and aggregating data (Serey *et al.*, 2021). These are common in AI applications, making it useful when dealing with large datasets.

Moreover, AI systems often need to continuously update their knowledge base or retrain models with new data. RDBMS provide a centralized and organized way to store and manage this data, ensuring data integrity and consistency, which is essential for maintaining the accuracy and reliability of AI models over time.

The process of connecting to a MySQL database typically involves establishing a connection using a Python library like “mysql.connector”. Once connected, SQL statements can be executed to create tables, insert data, and perform various operations (Silberschatz et al., 2020). Please see Appendix for example.

By employings RDBMS and SQL, AI systems can make use of existing data sources, reducing the need for manual data collection and entry, which improves the overall efficiency and effectiveness of the system (Domingos, 2012.) Additionally, SQL's optimization capabilities make it highly efficient for querying and analyzing large datasets, which is essential for AI systems that deal with big data.

# Question 5

(*FISHER - 1936* ) introduced The Iris dataset, and it is a widely-used benchmark dataset in machine learning and data analysis. It consists of 150 iris flowers with a variety of numerical features including sepal length, sepal width, petal length, and petal width. The target variable, which indicating the species (setosa, versicolor, or virginica) is often used for training machine learning model.

A critical step in the development of AI systems is Exploratory Data Analysis (EDA), as it helps to the learner to gain a comprehension into the underlying patterns and characteristics of the data (*Pyle, D. (1999.*) The first step taken was examine the structure of the dataset. The “.shape” attribute reveals that the dataset has 150 rows and 5 columns (4 features and 1 target variable, as outlined above). The “.info()” method shows that all features are numerical, and there are no missing values, as confirmed by “.isnull().sum().” Additionally, “.duplicated().sum()” reveals 1 duplicate instances in the dataset (*Applied Predictive Modeling | SpringerLink*, 2013).

Descriptive statistics, obtained “using .describe()”, provide insights into the central tendency and spread of the features. An example would be the mean sepal width which is 3,05 cm, with a standard deviation of 0.43 cm. Visualizing the unique values of the target variable using “.unique()” reveals the three distinct species: setosa, versicolor, and virginica. The data types were checked with “.dtypes” and it was seen that all features are stored as numerical values.

To gain an insight into distributions of the features, the use Seaborn's “sns.histplot()” and “sns.boxplot()”. The histograms highlight the differences in feature distributions among the species, with all species being clearly separable from the other two species based on petal length and width ( Muller & Guido, 2016).) The “boxplots” further also observes this, as there is minimal overlap between all species for these features as seen in the Appendix.

To investigate relationships between features and the target variable, Seaborn's “sns.countplot(“) and “.groupby()” were used. The “countplot” shows an equal distribution of instances across the three species. By combining the data by species and plotting the mean values of each feature using “.groupby()” a visual representation is created of the differences in feature values among the species.

Finally, a correlation matrix was generated using Pandas' “.corr()” method and it was visualized using Seaborn's “sns.heatmap.

EDA is a critical machine learning principles & their application in AI, as it helps to identify potential issues with the data, such as missing values, outliers, or imbalanced classes, which can adversely affect the performance of AI models (*Pyle, D. (1999*). This is also a common theme noticed through all questions above. Another thing of note is EDA enables data scientists to gain valuable insights into the data, which can inform feature engineering, feature selection, and the choice of appropriate algorithms and modeling techniques (Serey *et al.*, 2021).

Throughout the EDA process, it is essential to follow best practices in data exploration and visualization, as outlined by (‘Exploratory Data Analysis’, 2008) to gain valuable insights and prepare the data for further analysis or modeling.

# References

*Applied Predictive Modeling | SpringerLink* (no date). Available at: https://link.springer.com/book/10.1007/978-1-4614-6849-3 .

*Database System Concepts - 7th edition* (no date). Available at: https://db-book.com/.

Domingos, P. (2012) ‘A few useful things to know about machine learning’, *Commun. ACM*, 55(10), pp. 78–87. Available at: https://doi.org/10.1145/2347736.2347755.

‘Exploratory Data Analysis’ (2008) in *The Concise Encyclopedia of Statistics*. New York, NY: Springer, pp. 192–194. Available at: https://doi.org/10.1007/978-0-387-32833-1\_136.

*Introduction to Machine Learning with Python[Book]* (no date). Available at: https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/ .

*Learning Python, 5th Edition[Book]* (no date). Available at: https://www.oreilly.com/library/view/learning-python-5th/9781449355722/ .

*(PDF) Data Preprocessing for Supervised Learning* (no date). Available at: https://www.researchgate.net/publication/228084519\_Data\_Preprocessing\_for\_Supervised\_Learning .

*Pyle, D. (1999) Data Preparation for Data Mining. Morgan Kaufmann Publishers, California. - References - Scientific Research Publishing* (no date). Available at: https://www.scirp.org/reference/referencespapers?referenceid=1863384 .

*Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython - Wes McKinney - Google Books* (no date). Available at: https://books.google.ie/books/about/Python\_for\_Data\_Analysis.html?id=2BYfvgAACAAJ&redir\_esc=y .

Serey, J. *et al.* (2021) ‘Artificial Intelligence Methodologies for Data Management’, *Symmetry*, 13(11), p. 2040. Available at: https://doi.org/10.3390/sym13112040.

Smith, S. (2003) *Digital Signal Processing: A Practical Guide for Engineers and Scientists*. Newnes.

*THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS - FISHER - 1936 - Annals of Eugenics - Wiley Online Library* (no date). Available at: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1936.tb02137.x.

# Appendix - Q1 Code

A screenshot of a computer program

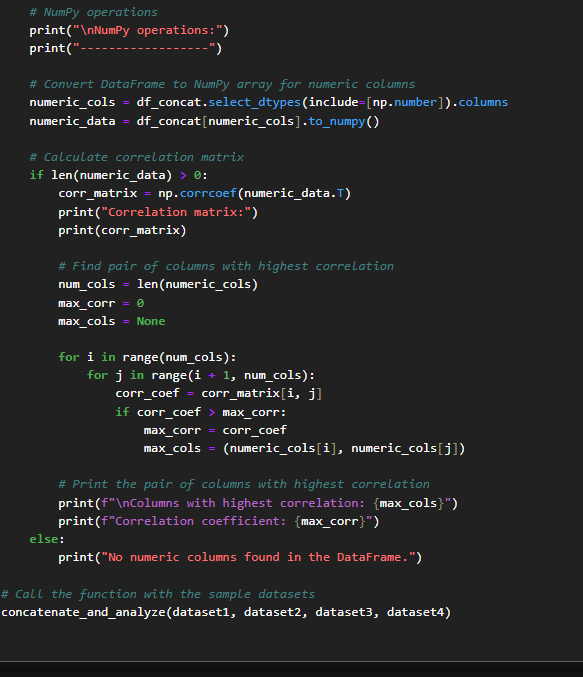
Description automatically generated

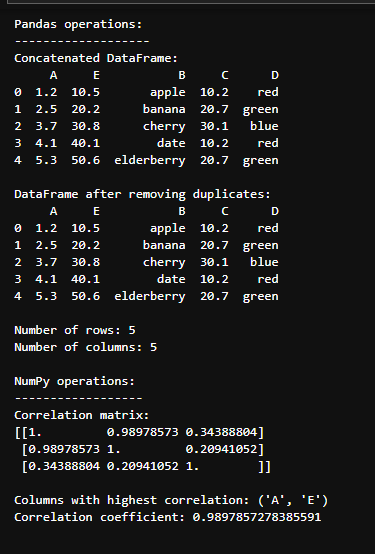
A screenshot of a computer

Description automatically generated

# Appendix – Q2 Code







# Appendix – Q3 Code

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer program

Description automatically generated

# Appendix – Q4 Code

A computer screen shot of a black screen

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated A screenshot of a computer screen

Description automatically generated

# Appendix – Q5 Code

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

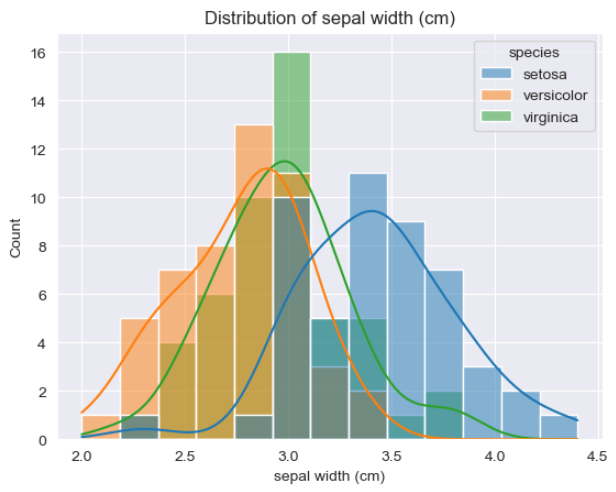
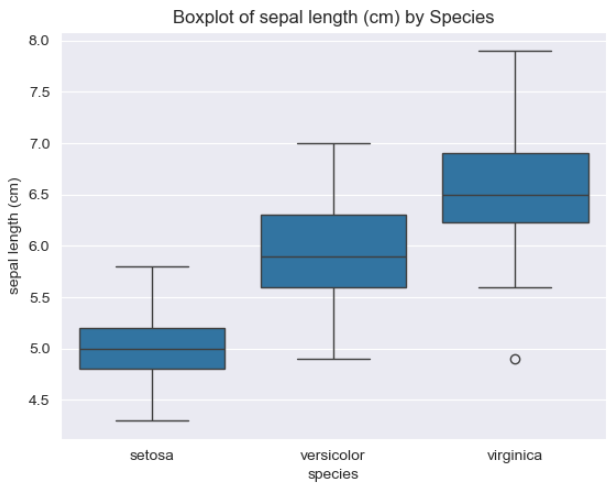
A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

# Appendix - Q5 Figures



A graph of different colored lines

Description automatically generated

