**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | **Data Preparation & Visualisation**  **Machine Learning for Data Analysis**  **Programming for Data Analytics**  **Statistics for Data Analytics** |
| **Assessment Title:** | **MSC\_DA\_CA1** |
| **Lecturer Name:** | **David McQuaid**  ***Muhammad Iqbal***  ***Sam Weiss***  ***Bharathi Chakravarthi*** |
| **Student Full Name:** | **Syed Asad Ailia** |
| **Student Number:** | **2023408** |
| **Assessment Due Date:** | **12th November,2023** |
| **Date of Submission:** | **12th November,2023** |

****

**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. |

**Abstract**

*In today's data-driven world, data analytics is becoming more and more significant, having a major impact on many facets of business, science, and society. In this report, we used data analytics techniques to clean insightful information from the mean and median hourly earnings across various economic sectors and employment statuses in Ireland. We carefully imported and checked the information to respond to specific questions. The development of a machine learning model for extracting output parameters from the validation dataset was the final step in this extensive process, which began with data preparation and continued with graphical representation using statistical techniques to identify trends. The Python framework was used to fulfil the programming requirements, and the entire project was recorded in a Jupyter Notebook.*

**Introduction**

In today's data-driven world, data analytics is an essential tool, particularly when looking at hourly salaries across various economic sectors and employment status. This analytical approach finds trends, patterns, and valuable information by methodically examining and interpreting data. Organizations, decision-makers, and scholars can gain a comprehensive grasp of the hourly wage market through the data analytics. Additionally, it offers data-driven insights into income inequality and job markets. Hourly earnings, a crucial economic metric, vary substantially between industries and job categories, reflecting the complexity of today's economy. Data analytics facilitates data-driven decision-making, highlights discrepancies, and points out causes, making it easier to navigate this complicated landscape. Our ability to analyze the data is further improved by statistical methods and machine learning models, which help us forecast future trends and the effects of policies in addition to describing the current state of affairs. When it comes to addressing economic disparities in a world where data-driven insights are essential, data analytics becomes a crucial ally that leads the way toward a more just and prosperous future. Here, we'll examine a few crucial data analytics pipelines to gain deeper understanding.

Data Preparation & Visualization

To prepare raw data for machine learning and statistical analysis, several important tasks involve data pre-processing, a fundamental step in the data analysis process. One of the first steps is data cleaning, which deals with the dataset's outliers, inconsistent values, and missing values. Imputation techniques can be used to address missing data, outliers can be identified and controlled to stop them from distorting the model, and standardization or normalization is frequently necessary to bring inconsistent data to a common scale. Another important component is data transformation, which includes feature engineering to produce new, informative attributes that can improve model performance, scaling of numerical features, and encoding of categorical variables. To reduce dimensionality and potentially improve model accuracy and computational overhead, data reduction techniques such as Principal Component Analysis (PCA) and feature selection methods are applied. To put it simply, data pre-processing is an essential step in ensuring that the data is prepared optimally for machine learning models to produce precise predictions.

A crucial step in the data analysis process is data visualization, which involves displaying data graphically to identify trends, relationships, and other patterns. Choosing the appropriate visualization techniques to match the goals and nature of the data is just as important to effective data visualization as designing visually appealing charts and graphs. There are many different types of visualizations that can be used, such as histograms for data distribution, scatter plots for relationships between variables, bar charts for comparisons, line charts for trends over time, and heatmaps for pattern recognition. Data visualization is an integral part of the data analysis process because its ultimate goal is to support data-driven decision-making and efficient dissemination of insights to stakeholders.

Machine Learning

Machine learning models are algorithms and mathematical constructs that empower computers to learn from data and make predictions or decisions without explicit programming. These models span a spectrum of techniques, with supervised learning focusing on labelled data, where the model is trained to make predictions based on input features, while unsupervised learning deals with unlabelled data to discover inherent patterns and structures. Within supervised learning, various algorithms like linear regression, decision trees, support vector machines, and deep neural networks are applied to tasks such as classification and regression. Unsupervised learning encompasses clustering techniques like K-means and dimensionality reduction methods like Principal Component Analysis (PCA) for tasks like data segmentation and feature reduction. Additionally, reinforcement learning is a branch of machine learning where agents learn to maximize cumulative rewards through a sequence of decisions in dynamic environments, serving purposes in areas like robotics and autonomous systems. The choice of the appropriate machine learning model depends on the problem at hand, and rigorous steps in model selection, hyperparameter tuning, and model evaluation are critical to ensure optimal model performance. It's essential to acknowledge that the quality of data and the effectiveness of data pre-processing and feature engineering significantly impact the overall success of machine learning models in real-world applications.

Wages in Ireland

When examining hourly earnings based on employment status, data analytics offers the capacity for a more granular and insightful perspective on income distribution. By segmenting data into employment categories such as full-time, part-time, and aggregating all employment types, it becomes possible to unravel the intricate variations in earnings and employment dynamics. This approach allows for a more comprehensive understanding of the labour market, revealing, for instance, whether certain sectors exhibit a higher prevalence of part-time employment and how these dynamic influences hourly wages. These insights are invaluable for labour market research, providing businesses and government entities with the tools to make well-informed decisions regarding labour policies, workforce management, and strategies for addressing income disparities and economic inequality.

The same pattern of analysis is used in this report as well, with an emphasis on mean and median hourly earnings in Ireland based on different employment statuses and economic sectors. The main goal is to give the public and government a thorough understanding of the shifting patterns in earnings and the role that employment status plays in maximizing wages. The data utilized is sourced from the year 2022, offering not only hourly earnings but also information on the number of hours worked during the week, which enhances the statistical insights. In the next section, we will further break down our analysis into distinct stages to obtain a more comprehensive and nuanced understanding of hourly earnings within Ireland's economic landscape. This segmentation will empower stakeholders with the insights needed to create policies, make decisions, and take actions that foster economic growth, ensure fair wages, and improve the well-being of the workforce while addressing the vital role of employment status in shaping earnings.