|  |  |  |
| --- | --- | --- |
| A picture containing text  Description automatically generated | **MSc Project Report** | |
| Student Name: Kirankumar Chaudhary | Matriculation Number: 2120051 | |
| Supervisor: Tiffany Young | Second Marker: Rob Lothian | |
| Course: MSc Data Science | | |
| Project Title: Energy Consumption Analysis and Optimisation for Scottish Councils using Big Data | | |
| Start Date: 07/02/2023 | Submission Date: 25/04/2023 | |
| **CONSENT**  I agree  That the University shall be entitled to use any results, materials or other outcomes arising from my project work for the purposes of non-commercial teaching and research, including collaboration.  **DECLARATION**  **I confirm:**   * **That the work contained in this document has been composed solely by myself and that I have not made use of any unauthorised assistance.** * **That the work has not been accepted in any previous application for a degree.** * **All sources of information have been specifically acknowledged and all verbatim extracts are distinguished by quotation marks.** | | |
| Student Signature: | | Date Signed: Click or tap to enter a date. |

# Acknowledgement

I would like to express my deepest gratitude to my project supervisor, Tiffany Young, for her unwavering support and guidance throughout the entirety of this project. Her expertise, patience, and dedication have been invaluable in helping me navigate the complexities of my research and address any doubts or questions that arose along the way. Thank you, Tiffany, for your unwavering support, and guidance.

# Abstract

I will write abstract here.

**Table of Contents**

[Acknowledgement](#_Acknowledgement)

[Abstract](#_Abstract)

Table of Figures

1. Introduction and Motivation
2. Literature Review
3. Project Specification
   1. Aim and Objectives
   2. Functional and Non-Functional Requirements
   3. Methodology and Plan
   4. Review of Legal, Ethical, Social and Professional issues as well as risks and safety
4. Design
5. Implementation
6. Evaluation of work (including comparison with objectives)
7. Conclusions and future works

References

Appendices

## Introduction and Motivation

## Literature Review

Energy consumption and its optimization have emerged as critical concerns in today's world, given the rising global energy demand, mounting environmental challenges, and the pressing need for sustainable energy systems. This project centres on collecting and analysing energy consumption data from the Scotland Council area spanning 2005 to 2020. The primary aim is to devise accurate predictive models for energy consumption and subsequently implement these models to optimize energy consumption patterns. This literature review delves into various aspects of energy consumption analysis, encompassing data collection and transformation, exploratory data analysis, business intelligence dashboards, predictive modelling, and optimization strategies. By thoroughly examining an array of studies and scholarly articles, this review elucidates the current state of research and best practices within the realm of energy consumption analysis and optimization.

**Are there any data sources that should be considered for a comprehensive understanding of energy consumption patterns in the Scotland Council area?**

The foundation of any successful energy consumption analysis and modelling lies in the accuracy and comprehensiveness of the data utilized. Numerous sources, including governmental and non-governmental organizations, provide open-source energy consumption data that can be leveraged for research and policy development. Notably, two key sources have emerged as crucial in the context of the Scotland Council area: the UK Department for Business, Energy & Industrial Strategy (BEIS) and the Scottish Government. These sources offer a wealth of data spanning various dimensions of energy consumption, providing researchers with ample opportunity to explore and analyse trends.

By accessing data from Scotland Government, one can gain insights into energy consumption patterns across different sectors, such as domestic, commercial, and public sector, as well as observe variations in usage within specific geographic locations spanning 2005 to 2020.

References

[1] **UK Department for Business, Energy & Industrial Strategy (BEIS)**, 2013. *Digest of UK Energy Statistics (DUKES)*. [Online]. Department for Energy Security and Net Zero and Department for Business, Energy & Industrial Strategy. Available from: <https://www.gov.uk/government/collections/digest-of-uk-energy-statistics-dukes> [Accessed Date].

[2] **Scottish Government**, 2014, *Energy Consumption*. [Online]. Scottish Government. Available from: <https://www.gov.scot/collections/energy-statistics/> [Accessed Date].

**Is there a dashboard available for the energy consumption of Scottish council areas from 2005 to 2020, and has exploratory data analysis been conducted on this energy consumption data**?

I could not find any specific dashboard or publication addressing energy consumption in Scottish council areas from 2005 to 2020. However, the UK Department of Energy & Climate Change has published factsheet named as "Sub-national total final energy consumption statistics," which provides energy consumption exploratory data analysis (EDA) for sub-national regions in the UK spanning from 2005 to 2013, including Scotland.

Additionally, one can refer to the Scottish government's official statistics on energy consumption. These sources provide comprehensive information on Scotland's energy consumption patterns and trends, but they not specifically focus on the council areas or cover the exact period from 2005 to 2020.

Given the available sources, I may need to conduct your EDA and create a customized dashboard to analyse and visualize energy consumption trends in Scottish council areas during the specified period i.e., from 2005 to 2013.

References:

[3] **UK Department of Energy & Climate Change,** 2012. *Sub-national total final energy consumption statistics: factsheet*. [Online]. Department of Energy & Climate Change. Available from: <https://www.gov.uk/government/statistics/sub-national-total-final-energy-consumption-statistics-2010-factsheet> [Accessed Date]

[4] **Scottish Government**, n.d. *Energy statistics*. [Online]. Energy and Climate Change Directorate. Available from: <https://www.gov.scot/collections/energy-statistics/>

**What are the key factors that influence the choice between Power BI and Tableau for building dashboard for energy consumption data?**

Power BI and Tableau are both powerful data visualization and business intelligence tools. They have their unique strengths and limitations, which makes them suitable for different use cases. Power BI and Tableau offer various features for creating visually appealing and interactive dashboards. Tableau is known for its advanced visualization capabilities, while Power BI is appreciated for its ease of use, integration with Microsoft products, and lower cost (Biswal 2023).

One advantage of Power BI over Tableau is its seamless integration with other Microsoft products, such as Excel, SharePoint, and Azure. This makes it easier for users already familiar with the Microsoft ecosystem to adopt Power BI for creating dashboards. Furthermore, Power BI provides a more consistent user experience across different platforms and devices (Biswal 2023).

Power BI also has a more intuitive and user-friendly interface, making it easier for non-technical users to create dashboards without advanced coding or data analysis skills. The learning curve for Power BI is considered less steep compared to Tableau, which can make it more accessible for a wider audience. Another aspect where Power BI shines is in terms of cost-effectiveness. Power BI offers a more affordable pricing structure compared to Tableau, making it a better option for small businesses or organizations with limited budgets (Biswal 2023).

In conclusion, after considering factors such as seamlessness, my familiarity with the tool, ease of use, and cost-effectiveness, I have decided that Power BI is a better choice for me to create dashboards for this project.

References:

[5] Biswal A., 2023. *Power BI Vs Tableau: Difference and Comparison.* [Online]. Simplilearn. Available from: <https://www.simplilearn.com/tutorials/power-bi-tutorial/power-bi-vs-tableau> [ Accessed Date]

**Which programming language and software is best suited for performing exploratory data analysis (EDA)?**

When performing exploratory data analysis (EDA) in energy consumption data for the Scotland Council, there are several programming languages and software options available. Two popular options are Python and R, which are both widely used in data analysis tasks.

Python is widely used for EDA due to its versatility, readability, and the availability of a vast ecosystem of libraries specifically designed for data analysis and visualization. Some of the most notable libraries include pandas, NumPy, and Matplotlib, which offer robust data manipulation and visualization capabilities, making the EDA process efficient and accessible (McKinney 2017).

Additionally, Python's IPython and Jupyter Notebooks provide an interactive computing environment, enabling users to write, run, and visualize code all in one place, which is particularly useful for EDA. In "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython," Wes McKinney provides a comprehensive guide for utilizing Python's capabilities to perform EDA. This book covers techniques and tools that enable users to efficiently clean, transform, and visualize data, thereby streamlining the process of discovering patterns and insights within the data (McKinney 2017).

R, on the other hand, is a programming language specifically designed for statistical computing and graphics (Ihaka and Gentleman 1996). It provides a wide range of statistical and graphical techniques, with packages like ggplot2 and dplyr that cater specifically to EDA tasks

Based on above findings, I have decided to use Python and Jupyter Notebook for EDA due to my familiarity with the language and the abundance of tools and libraries available for data analysis and visualization tasks.

Reference:

[8] McKinney, W., 2017. *Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython.* 2nd ed. Sebastopol, CA: O'Reilly Media, Inc.

[9] Ihaka, R. and Gentleman, R., 1996. R: A Language for Data Analysis and Graphics, *Journal of Computational and Graphical Statistics*, 5:3, pp. 299-314.

**How do different machine learning techniques, such as linear regression, support vector machines, artificial neural networks, and ensemble methods, compare in terms of predictive accuracy, complexity, and computational requirements when applied to energy consumption data?**

Different machine learning techniques have varying strengths and weaknesses when applied to energy consumption data, considering their predictive accuracy, complexity, and computational requirements.

Linear regression is a simple and interpretable model, making it easy to understand and implement. However, its simplicity limits its ability to capture complex relationships in the data, potentially leading to suboptimal predictive accuracy (James et al. 2013). Despite this limitation, linear regression may still perform well if the relationships in the energy consumption data are predominantly linear.

Support Vector Machines (SVM) are versatile models capable of handling both linear and nonlinear relationships by employing kernel functions. Compared to linear regression, SVMs may provide better predictive accuracy on complex datasets. However, they tend to be more computationally demanding, especially with large datasets, and their results may be more challenging to interpret (Cortes and Vapnik 1995).

Artificial Neural Networks (ANN) are powerful models inspired by biological neural networks. They can model complex relationships and nonlinearities in the data, often leading to improved predictive accuracy compared to simpler methods. However, ANNs come with higher complexity and increased computational requirements, and their results may lack interpretability (Goodfellow et al. 2016).

Ensemble methods, such as Random Forests, Gradient Boosting Machines (GBM), and eXtreme Gradient Boosting (XGBoost), combine multiple base models to improve predictive accuracy. These methods can handle complex data relationships and provide robust predictions, often outperforming single models. However, ensemble methods tend to be more computationally demanding and may require more extensive parameter tuning to achieve optimal performance (Hastie et al. 2009).

In summary, each machine learning technique has its advantages and disadvantages when applied to energy consumption data. The choice of the best method depends on the specific dataset, available computational resources, and the desired level of interpretability. To find the most suitable technique for a given problem, it is often helpful to perform cross-validation and compare the performance of different models on the data at hand.

References:

[10] James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013. *An Introduction to Statistical Learning: with Applications in R*. New York: Springer.

[11] Cortes, C., and Vapnik, V., 1995. Support-vector networks. *Machine Learning*, 20(3), pp. 273-297.

[12] Goodfellow, I., Bengio, Y., and Courville, A., 2016. *Deep Learning*. Cambridge Massachusetts: MIT Press.

[13] Hastie, T., Tibshirani, R., and Friedman, J., 2009. *The Elements of Statistical Learning*. 2nd ed. New York: Springer.

**How do different evaluation metrics and criteria impact the selection of the most suitable predictive model for energy consumption analysis in varying contexts?**

The choice of evaluation metrics and criteria can significantly influence the selection of the most appropriate predictive model for energy consumption analysis in different contexts. Metrics such as mean absolute error (MAE), mean squared error (MSE), and the coefficient of determination (R²) provide different perspectives on model performance [15]. For example, MAE and MSE measure the average deviation of predicted values from actual values, with MSE giving more weight to larger errors. R², on the other hand, reflects the proportion of variance in the dependent variable explained by the model. In the study conducted by Willmott et al. [16], multiple linear regression models and ANN models were compared in predicting energy consumption in residential buildings. The researchers found that ANN models outperformed linear regression models, as indicated by lower MAE, MSE values, and higher R² scores. This underscores the importance of employing multiple evaluation metrics to provide a comprehensive assessment of model performance and to make informed decisions when selecting the most suitable predictive model for a given context.

**Practices should be considered when deploying machine learning models for energy consumption analysis in real-world applications?**

The deployment of machine learning models is a critical phase in utilizing trained algorithms to make predictions and recommendations. It is crucial to integrate these models into production systems, where they can process new data and generate valuable insights [27]. When deploying machine learning models, several factors must be considered, including model performance, scalability, maintainability, and security [28]. In the context of energy consumption, the deployment of predictive models facilitates real-time monitoring, forecasting, and optimization. Research demonstrates the significance of deploying machine learning models in various energy-related applications such as demand response management [29], smart grid optimization [30], and building energy management systems [31]. The successful deployment of these models in production environments enables stakeholders to make data-driven decisions and implement energy-saving strategies.

Challenges in deploying machine learning models in real-world applications include ensuring model performance, scalability, maintainability, and security [28]. Researchers need to account for factors such as model interpretability, computational requirements, and model robustness to maintain optimal performance in production environments. Implementing best practices, such as continuous monitoring and updating of models, can address potential shifts in data distribution and evolving energy consumption patterns [32]. In the realm of energy consumption analysis, effective model deployment allows stakeholders to leverage predictive analytics' power, leading to informed decisions and optimized energy consumption.

Conclusion

In conclusion, a comprehensive understanding of energy consumption patterns in the Scotland Council area can be achieved by utilizing data from key sources like BEIS, the Scottish Government, and the National Grid. Performing exploratory data analysis using Python and Jupyter Notebook allows for effective exploration and interpretation of the data, while deploying suitable machine learning techniques helps to build accurate predictive models for energy consumption. Evaluating these models with appropriate metrics and considering the challenges associated with model deployment in real-world applications, such as performance, scalability, maintainability, and security, enables stakeholders to make data-driven decisions and implement effective energy-saving strategies. By employing a robust analytical process, one can contribute to more sustainable energy management and consumption practices in the Scotland Council area and beyond.

## Project Specification

* 1. Aim
  2. Objectives
  3. Functional and Non-Functional requirements.
  4. Methodology and Plan
  5. Review of Legal, ethical, social, and professional issues as well as risks and safety.

## Design

## Implementation

## Evaluation

## Conclusions and Future of work

# Appendices

# Referencing