**Logo, company name

Description automatically generated**

Hr. Assignment

Patrick Morgan Jones – 40251221

Mohammad Hamad – 40343069

Xuyang Liu – 40288851

Roisin McCusker – 40206122

Siddhesh Palav - 40339741

Table of Contents

[**1.0** **Introduction** 3](#_Toc91116928)

[**2.0** **Background** 4](#_Toc91116929)

[**2.1** **CRISP-DM** 4](#_Toc91116930)

[**2.2** **What factors cause employees to leave?** 5](#_Toc91116931)

[**2.3** **What factors predict when an employee is going to leave?** 6](#_Toc91116932)

[**3.0** **Methodology** 7](#_Toc91116933)

[**3.1** **Data Cleansing in KNIME** 7](#_Toc91116934)

[3.2 **Tableau: Data Visualizations and Dashboard.** 9](#_Toc91116935)

[**Visualization 1** 9](#_Toc91116936)

[**Visualization 2** 10](#_Toc91116937)

[**Visualization 3** 11](#_Toc91116938)

[**Visualization 4** 12](#_Toc91116939)

[**3.3 Tableau Dashboarding** 13](#_Toc91116940)

[**4.0 KNIME** 14](#_Toc91116941)

[**4.1 KNIME Models:** 14](#_Toc91116942)

[**Model 1** 14](#_Toc91116943)

[**Model 2: Salary and Compensation Variables** 15](#_Toc91116944)

[**Model 3: Survey Based variables. (Employee Perspective)** 16](#_Toc91116945)

[**Model 4: Performance-based variables.** 17](#_Toc91116946)

[**5.0 Comparison Dashboard** 19](#_Toc91116947)

[**6.0 Conclusion** 20](#_Toc91116948)

[**References** 21](#_Toc91116949)

[**Appendices** 22](#_Toc91116950)

[**Activity Log** 24](#_Toc91116951)

# **Introduction**

In Human Resource Management, attrition is a significant concern. It creates a anxiety and can even have a negative impact on cognitive organisational performance since it results in loss of useful information and expertise. Attrition has other adverse implications, such as increased expenditures. Hall (2019) outlines some of the expenses associated with attrition in an online article, admitting that there are direct costs, such as the cost of training. Hall also considers indirect costs, such as when superior employees are burdened with extra responsibilities from the loss. Employees may "wonder if they should likewise jump ship," according to Hall, because of indirect costs like this one. As a result, one employee's departure may have an impact on others, resulting in a downward spiral of attrition.

Resulting in negative consequences of attrition, it is critical for businesses with high levels of employee attrition to understand why this is the case. If the problem isn't remedied, a company will be severely harmed "both financially and emotionally" (Nelson, 2021). Human Resource Analytics is a valuable and modern approach to accomplish the main goal. This entails working together on study design, algorithms, and statistical techniques in order to "analyse employee data and translate outcomes into emotive reports" (Levenson, 2011, p.34). Predictive Analytics is a form of HR analytics. The building of models to explain developments and behaviours is included at this level, with the overarching goal of "recommending actions based on data" (Simbeck, 2019, p.1). This level of course is undoubtedly beneficial in predicting the reasons for employee attrition.

We will create an analytics solution for the Human resources department to assist in predicting which employees are most likely to quit. A KNIME workflow pulls employee data from HR (age, gender, and education, as well as salary, salary increments, business travel frequency, and so forth) and calculates the likelihood of each employee leaving the organisation.

The historic dataset has been recorded at a specified window during the entire workflow, based on the organization's operational approach. Outliers, erroneous values, or representational structure are then removed from the dataset using KNIME. Employee attrition datasets are typically unbalanced for the attrition category, necessitating reconciliation.

To update attrition probability, the workflow is then deployed on a KNIME Server and can be conducted on demand or using the scheduling option. The HR department uses the model output to create a dataset that can be used to analyse the profiles of employees who are likely to quit, such as how they are distributed amongst departments, years spent at the firm, years since last promotion, last wage increase, and so on. The results are exported directly to a Tableau dashboard using KNIME Analytics Platform's native Tableau integration. This gives business users rapid and easy access to informative and valuable visuals.

Business users can make parametric adjustments to evaluate their influence on the attrition mechanism using the KNIME. Users in the business world can also practice scenarios. For example, whether an employee's likelihood of staying grows when their income rises. This gives business users more control and encourages a data-driven decision-making culture.

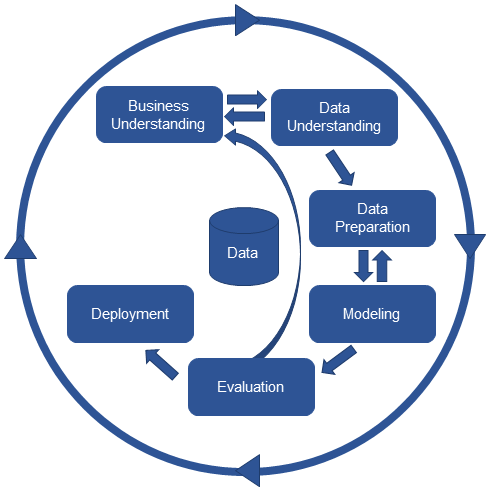
# **Background**

## **CRISP-DM**

It is essential to acknowledge the methodology we adopt whilst performing the task. Cross-industry standard process for data mining (CRISP-DM) is a popular methodological approach for constituting many data mining projects. It is a set of directives that help to “plan, organize and execute” a data-analysis project (Manasson, 2019). There are six steps involved within this process.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

The below diagram provides a clear outlook on these steps. (Cited from Taylor (2017)).



This methodology has proven to be useful by several authors and researchers. In an article by Wirth and Hipp (2000), the target was to stipulate and agree on a process which could be used efficiently by people of various skill levels in a response modelling application process. The results from the project displayed that CRISP-DM contains both the suppleness and the structure to fit the requirements of both high- and low-skilled people.

## **What factors cause employees to leave?**

Amongst the literature there is a wide range of inferential studies which attempt to generalize factors that may cause an employee to leave an organization. Selesho and Naile (2014) conduct a study on high employee attrition amongst university staff across different universities in South Africa. A central factor observed from the data was in relation to many employees being unsatisfied with salaries in their current positions. Selesho and Naile claim that being unhappy with salaries has “provoked many academic staff to abandon their profession” they further state that as a result of this there is a “loss of quality” for university teaching and administration (p.300). By identifying such factors, the authors can provide recommendations towards these issues, e.g. universities will have to certify that there are “unambiguous, comprehensible promotional guidelines” (p.302). Factors that influence employees to leave are also identified in a call center by Modau (2018). The results from the study displayed that there are precise retention aspects that have a huge impact on an employee’s plan on leaving. Some aspects include lack of supervisor support and career advancement. Like the study on university staff, this too also provides recommendations to help address the issues.

## **What factors predict when an employee is going to leave?**

Whilst it is worthwhile for organizations to understand what causes employees to leave, what is even more valuable, is the ability to predict when an employee will leave.

Wang and Zhi (2021) state that it is most important to “develop an intelligent system” of which can “accurately predict the likelihood of ET” (p.351). They adopted a Machine Learning analytical framework. This was adopted to examine an automated approach aimed at constructing a classification model of which predicts the likeliness of an employee leaving an organization (similarly to our collaborative task). The framework was used within two datasets, a final model was produced for each set. The authors claim that the employed framework has the aptitude to be utilized to produce “intelligent decision-making systems in other business domains” (368). Listed by the authors are domains such as inventory management systems. This brief description of some of the steps taken by Wang and Zhi demonstrate that the development of a predictive model in HR Analytics involves a lot of attention, time, and knowledge. As well as this, there are a selection of base models to help with decision making.

# **Methodology**

## **Data Cleansing in KNIME**

Essentially, it was important that we addressed the data quality issues within the dataset. This was done in KNIME through the method of data-cleansing. However, we agreed that we should go lightly on the data cleansing and try and keep as much as we require.

To avoid any problems, we first decided to remove any outliers from the data using column filter. Firstly, “over 18” was dropped as all the employees within the data set are over the age of eighteen. Similarly, we dropped “standard hours” because each employee’s value remains at eighty. Having these variables within the dataset were of no significance, so removing them made the data we had much more significant. We additionally used Rule Engine to remove over 40 working years, as no employee had worked this duration. The Rule Engine was also adopted to remove the impossible age of 443, this was most likely a typo. We additionally agreed that both “date of birth” and “education filed” were insignificant variables and so agreed to drop them as well. However, we did not completely abandon education as we instead kept the level of education column. Row filters were adopted to remove rows with N/A from the dataset as it only appeared five times amongst a huge dataset and so keeps the integrity of the data.

To avoid any team members feeling confused, we additionally renamed column “travel frequency” to “business travel” to align with the data dictionary. This was done through column rename.

We thought it would be important to keep the self-reported values of the variables, such as Marital Status and Relationship satisfaction. We also agreed to keep “years at company”, “current managers”, “years in current role.” These variables were additionally kept as they are longstanding employees who have stayed for the duration of the organization. We did not want to remove these variables initially as we felt that they were valid employees. However, if our model was of low accuracy, we could then drop them and run our models again.

We additionally edited some of the variables that were being measured numerically instead to string as they were simply not numerical values, e.g., environment satisfaction, job involvement, job level, education, job satisfaction. However, we later noticed within the Business Travel variable there was a remaining outlier that we had just noticed as we had both “none” and “0”, obviously these equate to the same thing. Therefore, we had to rename a variable to rename the “0” variable to “none” to combine the two. This reflects CRISP-DM as a continuous process throughout the project.

A picture containing timeline

Description automatically generatedOnce we had completed our data pre-preparation, we needed to transport the data into a tableau format to begin the visualizations. To begin visualizations, we created visualizations and a dashboard.

**Diagram 1- Data Cleansing in KNIME.**

## **Tableau: Data Visualizations and Dashboard.**

As previously discussed, we exported the data out of KNIME using the native tableau writer node to ensure that our data was transferred over to tableau in the correct format so that no errors could occur, e.g., ensuring all the numeric adjustments that had been instead changed to string, had stayed. We began looking at different variable combinations to gain insights about different models. We agreed to use a combination of numeric and string variables.

### **Chart, bar chart Description automatically generatedVisualization 1**

**Diagram 2- Attrition with Income and Education**

From this visualization we observed that generally, the higher level of education an employee has, the higher their income is. We can see that the employees who have remained in the organization earn a higher salary per month than those who have left. This may be a reason why the employees left, as for example, we noticed from the ex-employee table that the average doctor who has left the organization earns less than someone who is educated at below university level individual that has remained in the organization.

It is therefore straightforward why a doctor would leave the organization, as not only are they earning less than other doctors within the organization but additionally, employees of lower educational levels. Doctors are experts in their fields and so their skills are of high demand, this makes leaving a job much easier for people within their level of expertise. Overall, from this visualization we observed a correlation between lower paid employees and attrition, and so this is likely a central factor that can influence attrition.

### **Visualization 2**

Chart, scatter chart

Description automatically generated

***Diagram 3-* Attrition with Working Years and Income.**

From this visualization, we observed a general upwards trend which is the longer an employee has worked at the organization, the higher their monthly income. We observed that whilst the employees first few years at the company are slowly progressing, there appears to be a major salary hike in around five to six years. We noticed that the employee’s salary remains uniform for another few years with another major salary hike. Over the years there seems to be a regular increase of this process.

From the difference in colours throughout the visualization, we observed that attrition seems to follow a similar upwards trajectory. Therefore, the inference we have made is that although some employees are gaining an increase in salary, there are still employees leaving the organization towards the top of the visualization. Therefore, a pay increase will not prevent these certain employees from leaving. Thus, there are most likely other workforce environment factors that could instead be causing attrition. Additionally, the graph displays that there are employees who have worked for the organization for a much longer period than other employees yet are earning the same salary. This would negatively affect the longer employed individual and cause them to leave the organization.

### **Visualization 3**

**Chart, line chart

Description automatically generated*Diagram 4-* Training effect on earning by department.**

This graph displays how training has affected individual departments throughout the organization. For Human Resource Management, it is evident the training has had a phenomenal influence on the department, as before training they were earning on average £14,442 for the organization and due to a high level of training over the year, the figure has jumped to £23,965. Therefore, the approach adopted to help with HR training is clearly working for the organization. Whilst the increase in average monthly rate is not as high for Research and Development, there is still a clear increase in how much money they have made for the organization by £3,151. Whilst this training has benefitted this department, in comparison to the outcomes of the HRM department, the organization may be disappointed. Therefore, adjustments could be made to tailor the training approach for the department which may lead to better future outcomes.

However, a rather different pattern is observed in the sales department as the training has caused a steep decline in the money earned by this department. Sales have gone from earning the organization an average of £18,669 to an average of £13,902. This visualization therefore provides the organization with an insight to what is working for them and what is not. This therefore allows them to focus on the adjustment of the training approach to this department to help them increase their earnings.

### **Visualization 4**

**Graphical user interface, application

Description automatically generated**

**Diagram 5- Parameter.**

Graphical user interface, application, table, Word

Description automatically generatedThis visualization shows a breakdown between male and female employees keeping gender and attrition as the base parameter. The measures we have included are age, monthly rate, monthly income, percent salary hike, years at company.

**Diagram 6- Data Parameters showing the difference between monthly rate and monthly billable.**

The main insight we can draw from this visualization is that whilst all employees earn the same billable rate, current employees have a higher monthly income. Whilst the employees who left were earning significantly less than the current employees, they had the same billables. This complements our other visualizations that display that there is a correlation between income and attrition.

As we have seen across our visualizations, lower income employees are likely to leave the organization, and a salary percentage hike may not be the solution to retain the employees.

# **3.3 Tableau Dashboarding**

Application

Description automatically generated with medium confidence

**Diagram 7 - Dashboard**

**Link to interactive dashboard:**

<https://public.tableau.com/app/profile/patrick1412/viz/Hrassignment2/Dashboard1>

The above diagram displays our four data visualizations under a data summary header. All the visualizations and summaries are interactive and can be used to filter the dashboard by several variables, such as gender, attrition, department etc. This interactive dashboard can be accessed via the link provided and aids intuitive data-based decision making. This is a valuable tool for HR managers that can be used in the board room as part of a wider presentation.

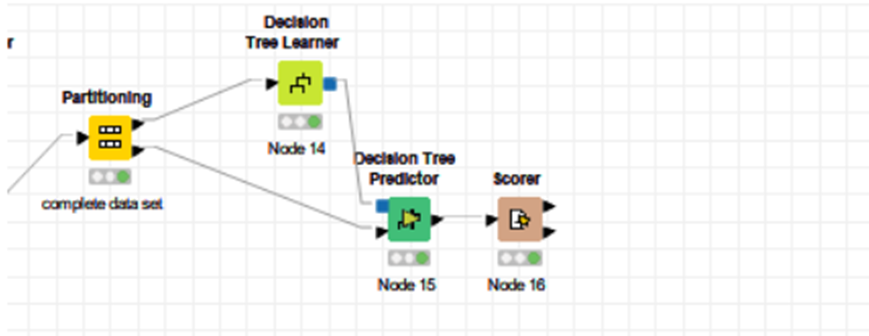
# **4.0 KNIME**

Using the partitioning node, we split the data with stratified sampling with attrition as our target column, so that within our split we had the same weightage. We went for an 80:20 split, in which 80% was training, and 20% was testing. A seed was created “123456789”, so that there was the same split amongst all partitioning nodes.

## **4.1 KNIME Models:**

### **Model 1**

On the first model we used a decision tree learner to include all the data variables to use this model as a baseline starting point. The accuracy score from this model was 81.6%. Whilst this is a great level of accuracy, we wanted a selection of models to choose from which had less variables. We then decided to breakdown and categorize our variables into Salary and Compensation Variables, Survey-based variables, and Performance-based variables. As well as testing to achieve a good accuracy, the below three models are additionally from different HR perspectives.



**Diagram 8 – Complete Data Model**

### **Model 2: Salary and Compensation Variables**

Graphical user interface, table

Description automatically generated

The variables listed above are included within the second model. We felt like salary and compensation variables would be an insightful category to choose from as across the wider literature and in the visualizations, salary is evidently a huge factor when it comes to attrition. Additionally, billable rate and monthly rate reflect on the performance of an employee and so may affect attrition. However, when we ran the model, the accuracy had dropped to just 74.5%. We agreed that our desired model would be above 75% and so continued to test different categories of variables.

Graphical user interface

Description automatically generated with low confidence

**Diagram 9: Compensation Model**

### **Model 3: Survey Based variables. (Employee Perspective)**

Graphical user interface, application, table

Description automatically generated

A picture containing table

Description automatically generatedWe felt that by including variables based on the perspectives of employees would be an ideal approach in relation to attrition, as although variables such as distance from home and satisfaction with the working environment do not relate to salary, they still have an important influence on attrition. The accuracy score for this model is higher than the second model as the outcome is 79.3%. However, this score is still relatively lower than the first model which included all the variables and, there are more than double the number of variables in this model compared to model two.

**Diagram 10 – Survey Based Model**

### **Model 4: Performance-based variables.**

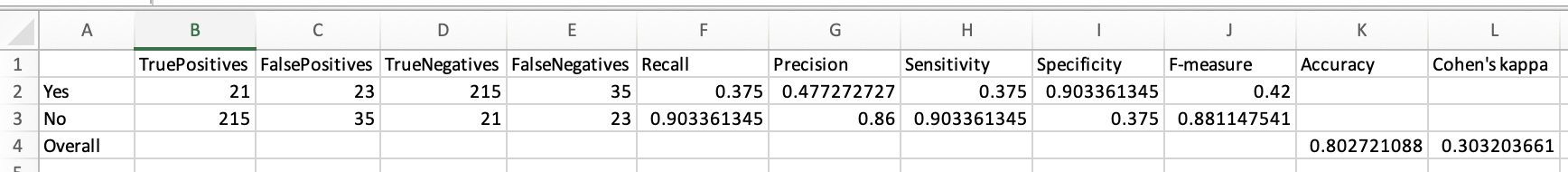
Graphical user interface, application

Description automatically generated

A picture containing chart

Description automatically generatedAgain, we tried to create a model from a different perspective, therefore deciding to look at variables which affect an employee’s performance. Variables such as education, training times, and work-life balance are undoubtedly going to affect an employee’s performance, and performance will affect overall attrition. When we ran the model, the accuracy was 80.3%. We were happy with this level of accuracy as it was the highest out of the three categorized models.

**Diagram 11 – Performance Based Model**

Whilst all our models were of a high accuracy, only one could be brought forward for dash boarding. We decided that the most accurate model to take forward would be model number four which was based on performance rating variables. Whilst this has the second-best accuracy score it is the best to adopt. It has very few variables in comparison to model one, yet still delivers a relatively high accuracy score. Additionally, whilst the first model we ran was higher, it is only higher by around 1.3%. This means the chosen model provides better insights.

***Diagram 12-* Confusion Matrix from Model Four.**

The above diagram is the confusion matrix from model four. The diagram shows, the accuracy and number of false positives and negatives.

# **5.0 Comparison Dashboard**

A screenshot of a computer

Description automatically generated

**Diagram 13**

**Link to interactive dashboard:**

<https://public.tableau.com/views/Predicteddata/Dashboard1?:language=en-GB&publish=yes&:display_count=n&:origin=viz_share_link>

After creating our model, we exported the predicted data from model four into tableau, again using the native tableau writer node. This dashboard allows us to have a clear visualization of the confusion matrix and predicted attrition. In our scatter plot, we can see that the attrition model is highly accurate as the number of false readings is low. We can see that the predictions from our model are mostly the same as the original data which makes our model reliable. We can also see that the false negative rate is 9.8%. Therefore, the risk of the model returning “no” when the employee is likely to leave is low but not inconsequential.

# **6.0 Conclusion**

Through this project we were able to create several predictions for the organization regarding employee attrition. We can with confidence predict the factors that are likely to lead to employee attrition and what factors will be relevant or irrelevant at retaining the employee(s). The model we decided to take forward incorporated a small number of variables that were key in understanding and predicting attrition as well as performance. The two created dashboards allow HR managers in the organization to predict and compare the effectiveness of the model using the original data and the predictive data. Going forward, we would recommend the organization also records the reason for employee attrition to ascertain whether there are patterns or failings within their organization

# **References**

Hall, John. (2019). The Cost Of Turnover Can Kill Your Business And Make Things Less Fun, Forbes. Accessed: 20th November 2021. URL: <https://www.forbes.com/sites/johnhall/2019/05/09/the-cost-of-turnover-can-kill-your-business-and-make-things-less-fun/?sh=6ac8b08a7943>

Manasson, Alexander (2019). Why using CRISP-DM will make you a better data scientist, Towards Data Science. Accessed: 20th November 2021. URL: <https://towardsdatascience.com/why-using-crisp-dm-will-make-you-a-better-data-scientist-66efe5b72686>

Modau, F. D., Dhanpat, N., Lugisani, P., Mabojane, R., & Phiri, M. (2018). Exploring employee retention and intention to leave within a call centre. *SA Journal of Human Resource Management*, *16*(1), 1-13.

Nelson, Nikki. (2021). Identifying and addressing employee turnover issues, Wolters Kluwer. Accessed: 20th November 2021. URL: <https://www.wolterskluwer.com/en/expert-insights/identifying-and-addressing-employee-turnover-issues>

Selesho, J. M., & Naile, I. (2014). Academic staff retention as a human resource factor: University perspective. *International Business & Economics Research Journal (IBER)*, *13*(2), 295-304

Simbeck, K. (2019). HR analytics and ethics. *IBM Journal of Research and Development*, *63*(4/5), 9-1.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1). London, UK: Springer-Verlag.

# A screenshot of a computer Description automatically generated with medium confidence**Appendices**

**KNIME Workflow**

Diagram

Description automatically generated**Decision Tree (Model 4)**

# **Activity Log**

|  |  |
| --- | --- |
| **Date of Meeting:** | 15.11.21 |
| **Location:** | Microsoft Teams |
| **Attendees:** | Roisin, Xuyang, Mohammad, Patrick, and Siddhesh |
| **Discussion:** | |
| * Everyone agreed we needed to become more familiar with KNIME as well as Tableau. This is because it was only introduced properly in Week 8 and so needed more experience with it for the assignment. | |
| **Plan for next meeting:** | |
| * We agreed that we would meet again on Thursday the 25th of November at 4pm to have another meeting to discuss where we would go next with the assignment and to bring together wider literature based on the issues of the assignment, e.g., high staff turnover. | |

|  |  |
| --- | --- |
| **Date of Meeting:** | 25.11.21 |
| **Location:** | Queens Graduate School |
| **Attendees:** | Roisin, Xuyang, Mohammad, Patrick, and Siddhesh |
| **Discussion:** | |
| * We agreed to go forward with chosen material for introduction/literature review. | |
| **Plan for next meeting:** | |
| * Next meeting goes through data cleansing * Start writing introduction/literature review | |

|  |  |
| --- | --- |
| **Date of Meeting:** | 6.12.21 |
| **Location:** | Queen’s graduate School. |
| **Attendees:** | Roisin, Xuyang, Mohammad, Patrick, and Siddhesh |
| **Discussion:** | |
| * Data Cleansing and first brief model. | |
| **Plan for next meeting:** | |
| * Each member agreed to go away and try to create some models before the next meeting. | |

|  |  |
| --- | --- |
| **Date of Meeting:** | 9.12.21 |
| **Location:** | Queen’s graduate school. |
| **Attendees:** | Roisin, Xuyang, Mohammad, Patrick, and Siddhesh |
| **Discussion:** | |
| * Brought the models together and decided on four models. | |
| **Plan for next meeting:** | |
| * Prepare tableau visualizations to bring to the next meeting. | |

|  |  |
| --- | --- |
| **Date of Meeting:** | 13.12.21 |
| **Location:** | Queen’s graduate school. |
| **Attendees:** | Roisin, Xuyang, Mohammad, Patrick, and Siddhesh |
| **Discussion:** | |
| * Picked our visualizations and dashboards. * We also produced some of the methodology during this meeting. | |
| **Plan for next meeting:** | |
| * Work around on few more predictor variables for visualisation for accurate model | |

|  |  |
| --- | --- |
| **Date of Meeting:** | 15.12.21 |
| **Location:** | Queen’s graduate school. |
| **Attendees:** | Roisin, Xuyang, Mohammad, Patrick, and Siddhesh |
| **Discussion:** | |
| * Complete tableau dashboard. * Complete methodology. | |
| **Plan for next meeting:** | |
| * Everyone agreed to begin looking at findings and prepare for second dashboard. | |

|  |  |
| --- | --- |
| **Date of Meeting:** | 16.12.21 |
| **Location:** | Queen’s graduate school. |
| **Attendees:** | Roisin, Xuyang, Mohammad, Patrick, and Siddhesh |
| **Discussion:** | |
| * Bring the entire project together * Create final dashboard * Finalize and submit. | |
| **Plan for next meeting:** | |
|  | |

Declaration of Equal Contribution Signatures 16/12/21:

|  |  |
| --- | --- |
| Mohammad Hamad | **Text, letter  Description automatically generated** |
| Patrick Morgan-Jones | Text  Description automatically generated |
| Roisin McCusker | A picture containing icon  Description automatically generated |
| Xuyang Liu | Text, letter  Description automatically generated |
| Siddhesh Palav | A picture containing text  Description automatically generated |