DELIVERABLE 3 PART 2 MVP

[**Prototype / MVP Document 4**](#_y9ive56okq1d)

[1. Introduction 4](#_76bi7k3n5vqc)

[2. Prototype Overview 4](#_45rhujerjs7i)

[2.1 Prototype concept 4](#_qamvaq39k1oc)

[2.2 Key features 4](#_sqkae9mq3wcf)

[3. Current project Codebase & Dataset 5](#_nzxf0shsuxrt)

[3.1 Codebase 5](#_8qp0bxd3vqwy)

[3.1.1 Assets folder 6](#_u0f8fqsif4x1)

[3.1.2 Source code folder 6](#_5jb9qtqsdqr2)

[3.1.3 Documentation 6](#_224sfwwhphqd)

[3.2 Dataset 7](#_e840eoganv66)

[Sponsor meeting, feedback and response to feedback 7](#_fkqxnyajnkkl)

[Prototype Feedback Adjustment 7](#_3x41c17c9lao)

[**1. Feature engineering: 8**](#_lzfgmtf9iad0)

[**(deciding what you are looking for and how to go about it) - Adam 8**](#_qdr5niyh7mnu)

[1.1 Trends and ranges 8](#_bf9lmhi2r7so)

[1.2 Data characteristics and pipeline 9](#_3pxj91upzvfc)

[1.3 Feature definition 12](#_eo4n0cp4jp4n)

[**2. Solution Architecture: 13**](#_ryj7yw870g23)

[**(choice of macro architecture / pipeline) 13**](#_gs7zi48otr1z)

[2.1 Current architecture 13](#_3xgo0cs9b1dv)

[Cleaning Stage 13](#_t4x6398b1eh)

[Enriching Stage 13](#_2kqd6blhi7wt)

[Synthesizing Stage 14](#_msghfuvvlpjw)

[Merging Stage 14](#_fz234khsx8b5)

[2.2 Cloud architecture (not yet deployed) 15](#_k0kki89gevit)

[**3. Algorithms / models methods: 16**](#_l2mjyumkiy8h)

[**(detail what is in each part of the solution architecture, including models used and initial conditions / config settings) - Ninuri 16**](#_bwc0r8p50qq6)

[3.1. KNN Model 17](#_h43q0yu2c5dr)

[3.1.1. Configuration Settings 18](#_e1rhfjojw0ub)

[3.1.2. Reasons for Model selection 18](#_egrx6mswua1e)

[3.2. Random Forest Models 19](#_qvxe909pruj9)

[3.2.1. Configuration Settings 19](#_m2rb7z33gedi)

[3.2.2. Reasons for Model selection 20](#_f2liiara9pq)

[3.3. Why were these models chosen over other approaches? 20](#_lmxr7f5dtxej)

[**4. Detailed Data descriptions: 21**](#_gv4z3eqx89tg)

[4.1 Original Raw Datasets 21](#_vw5e3ri0812r)

[4.2. Columns in Raw Datasets 22](#_619wrttzrtw8)

[Raw Dataset 1: 22](#_550boq51ih0z)

[Raw Dataset 2: 25](#_c2hcic46y9mz)

[Raw Dataset 3: 26](#_g5jzp4l0szj3)

[4.3. Cleaning & Enriching 28](#_dsq15dkeog8x)

[Dataset 1 & Dataset 2 28](#_1ma3k47mi0tu)

[Cleaned & Enriched Dataset 1: 28](#_e1ywewqaifz9)

[Cleaned & Enriched Dataset 2: 30](#_dafjelwky24w)

[Dataset 3 31](#_q2r1g1f2ddpe)

[Cleaned & Enriched Dataset 3: 31](#_2k96w7qjkx9d)

[4.4. Generating Data: 32](#_2jh8fpqb5hmu)

[Method 1: 32](#_1k0o3efn6idt)

[Method 2: 34](#_m5s659ehbp43)

[Merging Datasets: 36](#_75744aujprf7)

[**1. Model evaluation: - Tash 42**](#_cg04p2hhhvi1)

[1.1. Model Evaluation and Comparison 42](#_3z1wmobrz3b0)

[Data Cleaning, Synthesisation, and Imputation: 42](#_a6b7yqmtd1vm)

[1.1.1. Correlation Matrix and ANOVA Analysis: 42](#_obe1pjidf3wk)

[1.1.2. Validation of Synthesised Data: 43](#_3guxe941bnqt)

[1.1.3. Imputation Techniques: 43](#_3pao0cpto3a6)

[For Numerical Columns: 43](#_6cw8iziz8a34)

[For Categorical Columns: 43](#_269wiewt7wn)

[1.2. How we compared and evaluated Models and decided what to choose 44](#_5gw1fjam8azd)

[In our project, we used several models for both imputation and classification, including K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Random Forest. While the choice of model is essential, the performance evaluation metrics we used to assess these models played an even more significant role in guiding our final decisions. These evaluation techniques allowed us to assess the models' ability to accurately impute missing data and classify fraudulent and non-fraudulent claims. While also comparing and contrasting the models objectively, ensuring we selected the most suitable approach for each specific task. 44](#_569gz4fo7stb)

[1.2.1. R² (Coefficient of Determination) 44](#_loviq9tc2461)

[1.2.2. Accuracy 44](#_rnxiahg0558x)

[1.2.3. Mean Squared Error (MSE) and Mean Absolute Error (MAE) 45](#_wu6vztkmaiur)

[1.3. Overall 46](#_449qy26g8h8x)

[**2. Performance evaluation results: - Adam/Noor 46**](#_c8ue64c9px0s)

[References 47](#_gyoqyjck1ap0)

**Prototype / MVP Document**

# **1. Introduction**

The MVP document offers a detailed overview of the essential components of our data processing pipeline, covering everything from infrastructure setup to the generation of the final dataset. It provides insights into our approach to data management and processing, aimed at delivering a high-quality dataset. The document outlines our infrastructure design, the codebase, and the data pipeline. It also presents the resulting prototype dataset, consisting of 19 features that have been thoroughly cleaned, synthesised, and enriched.

# **2. Prototype Overview**

## **2.1 Prototype concept**

The prototype consists of a detailed presentation that highlights two critical components:

1. **Codebase**: A comprehensive explanation of our GitHub repository, showcasing the infrastructure design, data processing scripts, and automation workflows that constitute the data pipeline.
2. **Final Dataset:** A version of the enriched dataset containing 19 key features essential for effective fraud detection in insurance claims.

### 

## **2.2 Key features**

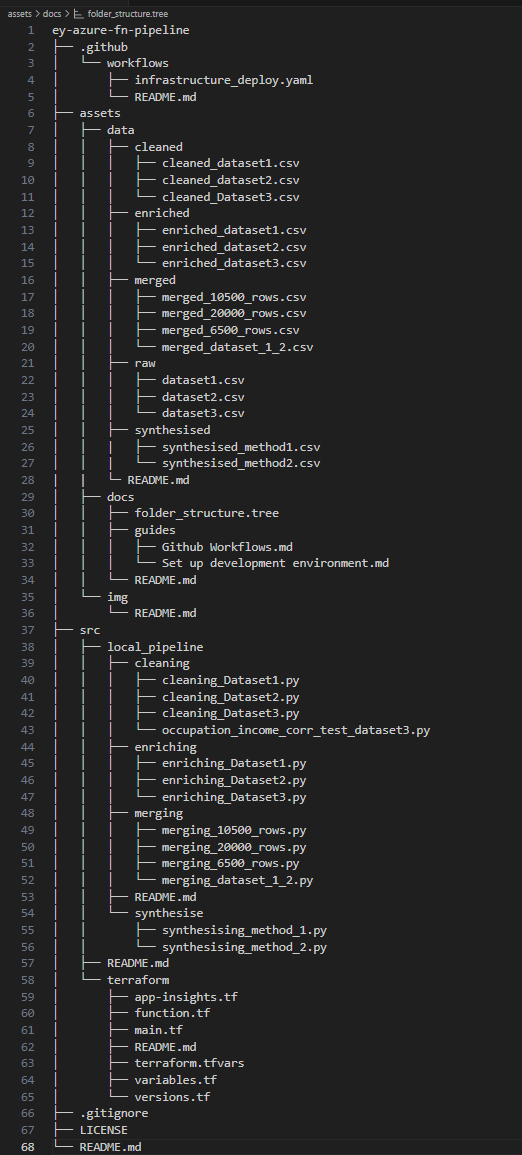
1. **Infrastructure as Code:** Utilising Terraform to define and manage Azure infrastructure, ensuring reproducibility and manageability.
2. **Automated Deployment:** Implementing GitHub Actions scripts for seamless deployment of infrastructure and Azure Functions to the cloud.
3. **Data Loading:** Ensuring efficient and secure data transfer with PowerShell and Azure Data Lake Gen2 REST API to ingest raw data.
4. **Data Processing Pipeline:** The pipeline includes scripts for data cleaning (such as renaming and removing columns), synthesising, enriching, and merging data from multiple sources. At this stage, the code is not yet deployed on Azure. Instead, we are developing and testing it locally, simulating the pipeline that will eventually be deployed on the Azure platform. All scripts are version-controlled on GitHub and are currently being executed manually.
5. **Final Dataset:** A dataset with essential columns that will be used by the AI team to train their initial prototype.

# **3. Current project Codebase & Dataset**

## **3.1 Codebase**

The most up to date version of the codebase can always be found at this link: <https://github.com/AlanDataPortfolio/ey-azure-fn-pipeline>.

The folder structure cannot be shown with a screenshot so here is the best way to visualise the current codebase.



### **3.1.1 Assets folder**

The 'assets' folder stores data at various stages of the transformation pipeline. Within the 'data' subfolder, the data is further partitioned to reflect its raw, enriched, synthesised, and merged states. Additionally, documentation related to the project will be stored in this folder. At the time the screenshot was taken, the metadata for the dataset in the gold layer had not yet been completed. This document will be continuously updated to reflect the latest metadata for datasets shared with the AI team. The 'img' folder contains images that will be used in README.md files and other project documentation.

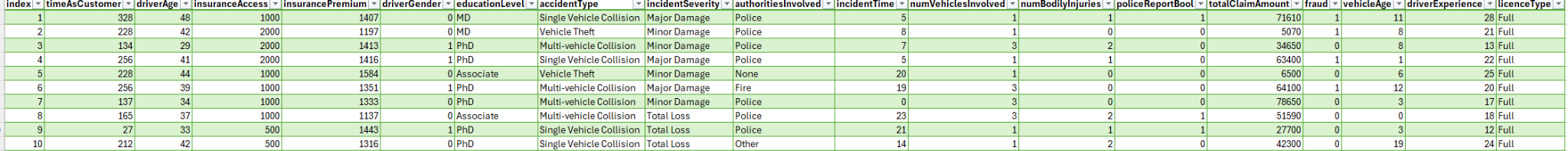
### **3.1.2 Source code folder**

The most crucial component of this prototype is the contents of the 'src' folder, which houses both the source code for setting up the infrastructure and the code that forms the transformation pipeline. As shown in the screenshot above, the folder is organised into two sections: 'terraform,' which is responsible for setting up the cloud infrastructure, and 'local\_pipeline,' which is used to build the transformation pipeline. The pipeline consists of four stages: cleaning, enriching, synthesising, and finally merging the data.

### **3.1.3 Documentation**

Lastly, documentation is not limited to the 'assets' folder but is distributed throughout the codebase within the README.md files. This approach allows for more effective management of the content in each README.md file, ensuring the information remains concise and does not overwhelm the reader. This method also ensures that relevant details are accessible in context, making it easier for team members to reference specific information without needing to navigate through separate documents.

## **3.2 Dataset**



The dataset, which consists of 20,876 rows of data in CSV format, will be elaborated on in more detail later in the document.

# **4. Sponsor meeting, feedback and response to feedback**

We had a prototype presentation on Tuesday, 24th September 2024, at EY’s office. After the presentation, one of the mentors questioned our choice of tools, specifically suggesting the use of another library called 'AutoML,' which is essentially a library designed to automate the process of building machine learning models. He mentioned that with this library, the tasks we are handling in the transformation pipeline could be completed in just five lines of code and asked whether we had considered this option. In the next section, **Prototype Feedback Adjustment**, we will explore how we plan to evaluate and potentially incorporate the AutoML library into our workflow to optimise the transformation process.

## **4.1. Prototype Feedback Adjustment**

During the presentation, we outlined two critical objectives to focus on for the next prototype delivery: deploying the data pipeline to Azure and further refining the transformation pipeline. The mentor's feedback has been integrated into the second objective, where we plan to assign 1-2 team members to research the potential of using ‘AutoML’ and explore its application in our pipeline. Our goal is to recreate the pipeline—if not with the exact same methods, at least following the same steps, such as cleaning and synthesising the data. Following this, a team review will be conducted using predefined standards to assess which approach yields the better dataset. Additionally, we will incorporate feedback from the AI team to further determine the most effective method.

**Analysis, Design + Testing Documentation for DATA SCIENCE projects**

# **1. Feature engineering:**

# *(deciding what you are looking for and how to go about it) - Adam*

· From the data, are there any trends / ranges to look for?

e.g. if you are analysing time-based movement data, what accelerometer and gyroscope ranges / trends do you want to look for to identify someone that has fallen / collapsed?

e.g. if you are looking at financial data, what does an "upward sale trend" look like?

e.g. if you are looking at population data, or education rates, what "characteristics" would you want to look at and why?

· Describe what data characteristics are being looked for, and how your data pipeline is being processed to generate these features.

· Give each different "feature" or "characteristic" a name and then attribute some form of data ranges / statistical definition.

## **1.1 Trends and ranges**

Identifying the trends and ranges is crucial for feature engineering as the patterns will help to better understand the dataset by identifying relationships between variables, detecting outliers, and evaluating the data quality. Correlations between variables can be checked to identify potential relationships.Through this approach, our raw data can be converted to a more meaningful format which will be well structured and easier to interpret by the model, therefore leading to a better performance of the model. In our case, when detecting fraudulent claims, it will be key to look at any anomalies in the data. For example, if the ‘totalClaimAmount’ is high and the ‘incidentSeverity’ is minor, it is most likely to be a fraudulent claim. The ‘totalClaimAmount’ column itself can be a good indicator of a fraudulent claim. If the amount is too high and is significantly more than the average claim amount, that claim will be flagged for a fraudulent claim.

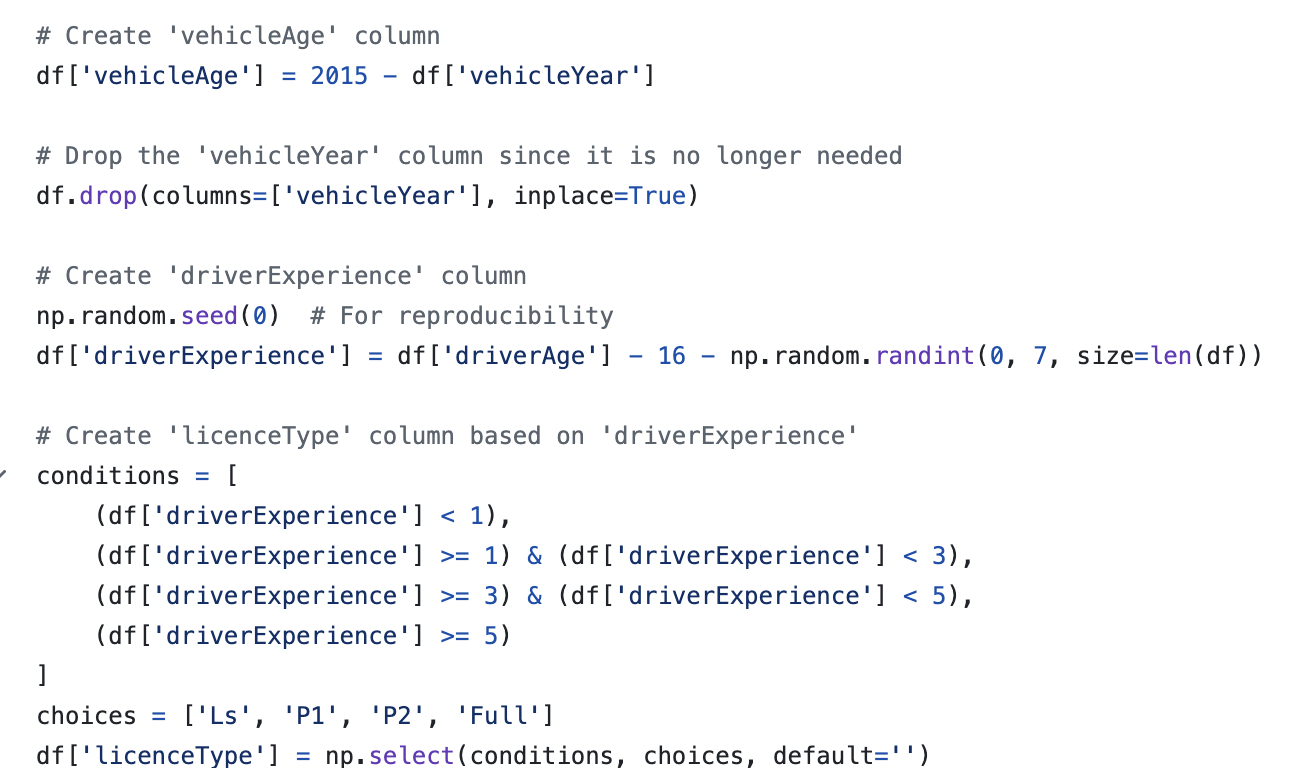
Columns such as ‘driverAge’, ‘driverExperience’, and ‘licenseType’ can give us information about the driver. These columns can also provide insights on the driver behaviour. For example, younger drivers, who are less experienced, are more likely to be driving rough and engage in speeding. Whereas, middle-aged drivers are more experienced and are more likely to drive safely (Keall & Frith). Therefore, these columns can help to identify the type of driver and help in interpreting whether the claim is fraudulent or not.

## **1.2 Data characteristics and pipeline**

Fraud indicators must be considered because the dataset is designed to be used to train an AI model that would identify fraudulent claims. Therefore, we would need a fraud column and columns that can help explain a fraudulent claim. Information relating to the incident and the customer details will also be needed to determine a false claim.

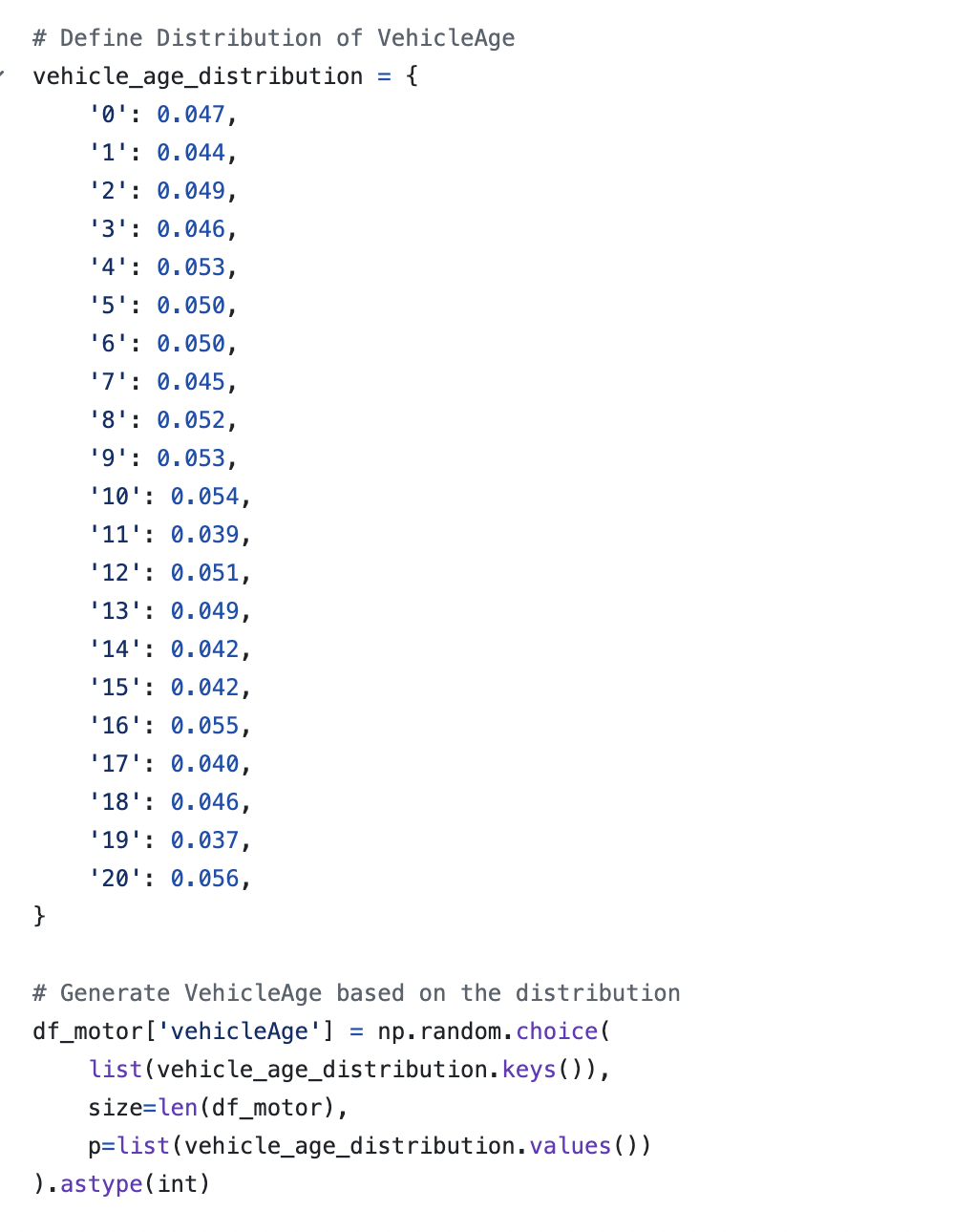
Our data pipeline consists of cleaning, enriching, synthesising and merging to get the combined dataset. We are using our dataset 1 as our foundational dataset. In data cleaning, the columns were renamed to match camel casing. All three datasets had the same column names and followed camel casing. For dataset 1, the missing values were handled. For dataset 2, ‘EducationLevel’ was filled in assuming the lowest degree which is ‘High School’. For dataset 3, the missing columns were filled in using various imputation methods such as KNN, Random Forest, GridSearchCV, etc.

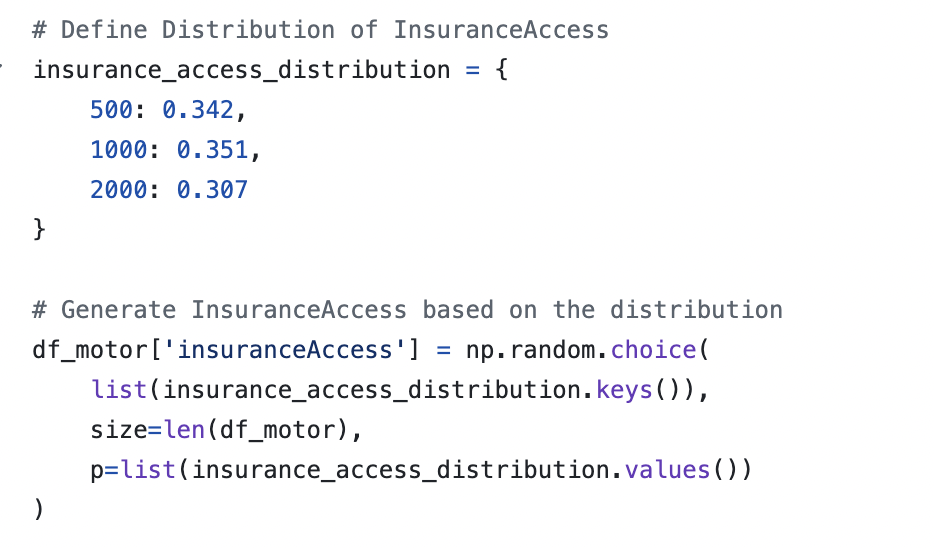
In enrichment, ‘True’ ‘False’, ‘YES’ ‘NO’, ‘MALE’ ‘FEMALE’ were all converted to numeric 0’s and 1’s. Here, the categorical variables that could have been converted have been converted to numerical 0’s and 1’s to make it more understandable to the model. Here, the unwanted columns were also dropped in all three datasets. Some columns were generated such as the ‘vehicleAge’ column. As our dataset was based in 2015, we subtracted 2015 from the ‘vehicleYear’ column. The ‘driverExperience’ column was also generated using the ‘driverAge’ column. The ‘driverAge’ is subtracted by 16 as it is the minimum driving age and is further subtracted by a random number from 0 to 7 as we want some variability. The ‘licenseType’ column is also generated based on the ‘driverExperience’ column. If the ‘driverExperience’ is less than 1 year, it was an ‘L’ licence, if the ‘driverExperience’ is more than 1 year and less than 3 years, it was a ‘P1’ licence, if the ‘driverExperience’ is more than 3 years and less than 5 years, it was a ‘P2’ licence and finally if the ‘driverExperience’ is more than 5 years, it was a ‘Full’ licence. The code to generate these columns is provided below:



For dataset 2, a few columns had to be generated based on dataset 1 as dataset 2 did not have those columns. The ‘AccidentType’ column had to be enriched based on the distribution of that column in dataset 1. Using that distribution, ‘AccidentType’ was then randomised and generated. For ‘NumVehiclesInvolved’, the same method was used to generate it. If the ‘AccidentType’ was ‘Multi-Vehicle Collision’, the column was randomly generated based on the normal distribution of ‘Multi-Vehicle Collision’ in dataset 1. Otherwise, the column returned 1 as the rest of the accident types were involving a single vehicle. The ‘VehicleAge’ column was also generated based on the distribution of it in dataset 1. The weights were assigned to the different ages and were randomly generated once again. A similar process was also used to generate the ‘InsuranceAccess’ column. The code for the columns generated is provided below:







In all of the datasets, Fraud was converted to 1 and Non-Fraud was converted to 0. The same processes are used to generate the features in dataset 3. The 2 datasets were then merged and based on the merged dataset, 2 synthesization methods were used to generate 8000 rows of data.

## **1.3 Feature definition**

## 

A lot of features can be created following the process we will discuss later on in this section. However, these are some of the important ones. For ‘driverExperience’, we can name a feature called ‘experienceCategory’ which will classify the driver experience into different categories. This will help to determine the information about the experience of the driver. Ranges for this feature can be: <3 years = ‘Low Experience’, 3-6 years = ‘Moderate Experience’, >6 years = ‘High Experience’.

The same is partially done for ‘licenseType’. A feature called ‘licenseCategory’ can be named which will classify the licence type of the driver. In our case, if the ‘driverExperience’ was less than 1 year, then the licence was a Learners or ‘L’. If the ‘driverExperience’ was more than 1 year and less than 3, then the licence was a Provisional P1 or ‘P1’. If the ‘driverExperience’ was more than 3 years and less than 5, then the licence was a Provisional P2 or ‘P2’. If the ‘driverExperience’ was more than 5 years, then the licence was a Full License or ‘Full’.

The ages for the driver can also be categorised into a feature called ‘driverAgeCategory’. The ranges for that category can be: <25 years = ‘Young’, 25-50 years = ‘Middle-Aged’, >50 years = ‘Old’.

The same can be done for ‘vehicleAge’. A new feature called ‘vehicleAgeCategory’ can be created. The ranges for that category can be: <3 years = ‘New’, 3-10 years = ‘Moderate’, >10 years = ‘Old’.

A feature called ‘timeCategory’ can be created to describe the time the incident took place. The range for that category can be, 0-6 = ‘Late Night’, >6-12 = ‘Morning’, >12-18 = ‘Afternoon’, >18-24 = ‘Night’.

# **2. Solution Architecture:**

# *(choice of macro architecture / pipeline)*

### 2.1 Current architecture

### 

### **Cleaning Stage**

We have three raw data sources in CSV format from Kaggle, each provided by different organisations. As a result, these sources do not share the same columns and exhibit varying levels of quality. The initial task in the cleaning stage is to prepare these datasets by renaming similar columns to a predefined standard and addressing any missing values. We have developed several imputation methods for this purpose, ranging from simple techniques—such as using the median and mean for numerical data and the mode for categorical data—to more advanced algorithms like Support Vector Regression and Random Forest Classifier for imputing other columns. Given that the combined raw data consists of fewer than 15,000 rows and amounts to approximately 3MB, the computational resources required for this stage are minimal, allowing for efficient processing without the need for extensive infrastructure.

### 

### **Enriching Stage**

The enriching stage focuses on creating more meaningful columns within the datasets. One significant feature to be added is **Driver Experience**, which will be inferred from the distribution of the Australian legal driving age and the customer’s age. However, we face limitations in the amount of data available, which is insufficient for the AI team to effectively train their model. As a result, we have prioritised data enrichment over synthesis to maintain the reliability of our final dataset. This approach enhances the quality of the data, allowing our synthesis methods to learn from the enriched data's features and patterns. Notably, Dataset 3 has numerous missing columns necessary for enrichment. Therefore, we will focus on enriching only Datasets 1 and 2, ensuring that the enrichment process does not exceed one minute in running time due to the small size of the dataset. It is crucial that the enriched data remains realistic and not randomly fabricated, maintaining the integrity and reliability of the final dataset.

### 

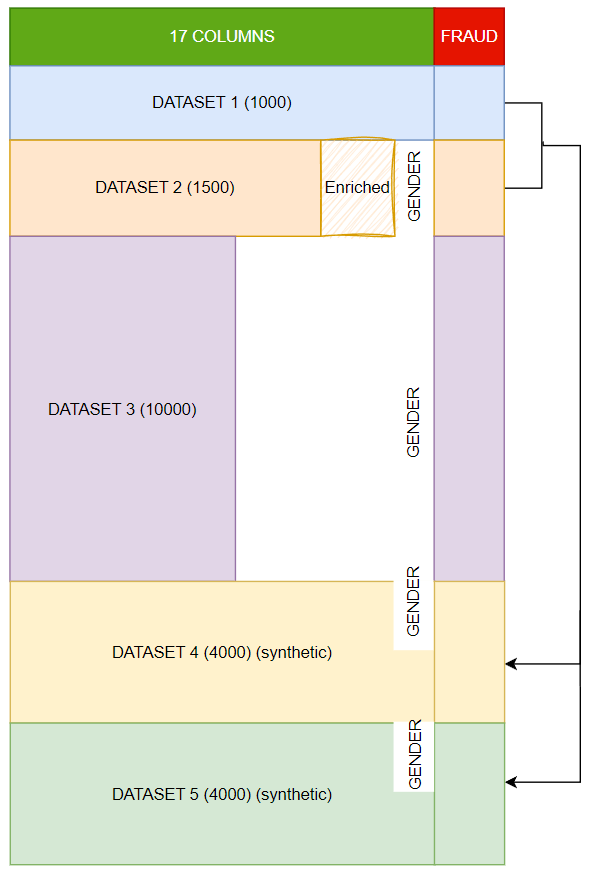
### **Synthesizing Stage**

The synthesising stage involves generating new data to address the limitations of our existing datasets. We will employ mathematical remapping techniques combined with sampling and Conditional Generative Adversarial Networks (CTGAN) to synthesise additional data. To ensure the reliability of the synthesised data, we will validate the output against known distributions and patterns from the enriched datasets. However, we recognize that CTGAN can be time-consuming, and Azure Functions have a time limit of approximately 10 to 15 minutes per execution. If the process exceeds this time frame, we will split the code into multiple runs to accommodate the time constraints.

### 

### **Merging Stage**

In the merging stage, we will integrate the synthesised datasets (1 and 2) with the enriched dataset (3). The merging will be conducted based on common keys or criteria defined during the cleaning and enrichment stages. Our expected output is a cohesive dataset with a reasonable number of null values, acknowledging that some null values are acceptable given the inherent imperfections of real-world data. However, we will prioritise creating a realistic dataset, as established in the enriching stage, ensuring it serves the intended purpose effectively.

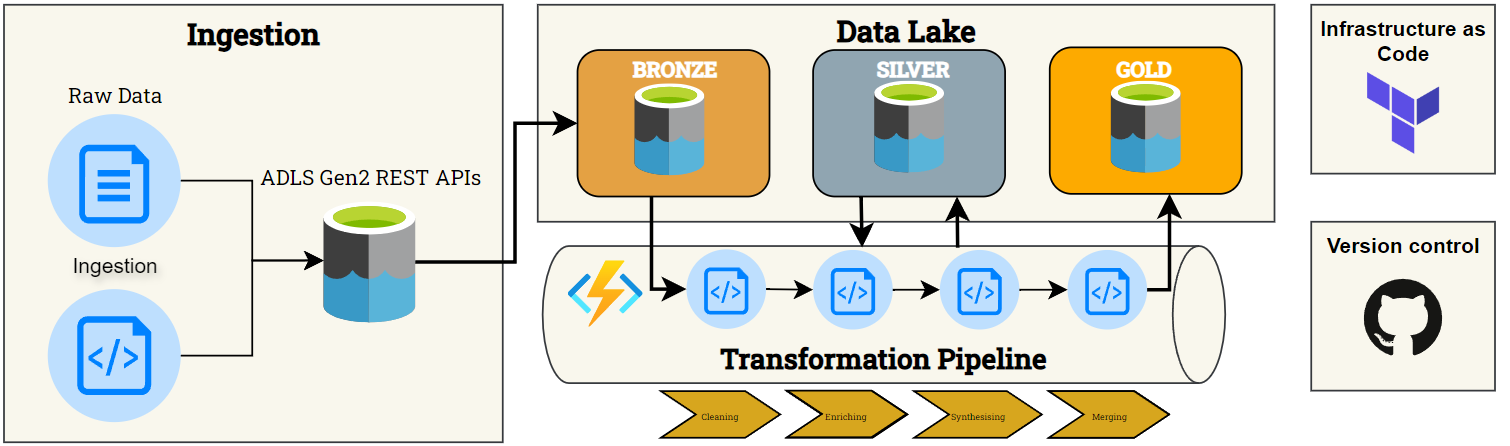


The final dataset is a combination of enriched dataset 1 and 2, synthesised dataset 4 and 5 from 2 methods and dataset 3.

### 

### 2.2 Cloud architecture (not yet deployed)

### 



The pipeline's order closely resembles our current architecture. While we have successfully implemented the pipeline, it is currently not automated and operates solely on our local machines. The key distinction in this iteration is our use of Terraform for Infrastructure as Code. All data will be stored in Azure Data Lake Gen2, which will be organised into three distinct layers. The Bronze layer will exclusively contain our raw data, while the Silver layer will store data at various stages of transformation, including after cleaning, merging, synthesising, and further processing. Once the dataset is finalised, it will be moved to the Gold layer. These layers serve as an abstraction, facilitating efficient management of our data processes.

Additionally, we will utilise Azure Functions to execute Python scripts, automating processes whenever modifications are made to our scripts. Azure Functions is a cost-effective solution for small to medium-sized workloads, offering the first 400,000 GB/s of execution and up to 1,000,000 executions free of charge. Given our initial dataset, which comprises approximately 10,000 rows and around 5MB of data for the Bronze layer, costs will only be incurred if our data volume or execution frequency increases. Even in such cases, the projected expenses will remain minimal, considering the scale of this project.

Setting up and deploying Azure Functions is straightforward, making it a suitable choice for small-scale projects like ours. The pay-as-you-go model further enhances cost efficiency, as we only pay for the resources utilised. For our storage solution, we will employ Azure Data Lake Storage Gen2 (ADLS2) in the first MVP, leveraging its foundation on Azure Blob Storage. This option is not only economical but also capable of storing virtually limitless amounts of data, ensuring that both cost and scalability remain manageable.

# **3. Algorithms / models methods:**

# *(detail what is in each part of the solution architecture, including models used and initial conditions / config settings) - Ninuri*

Cleaning the dataset is a crucial step in the project as the quality of the dataset directly affects the accuracy and reliability of the outputs from the AI model. This project with NRMA involves collaborating with Team 13 to create an AI solution to detect fraudulent claims. Our team is responsible for delivering a dataset to Team 13 that is clean, complete and comprehensive so it can be used to train the AI model to make accurate predictions and generalise the data properly generating reliable predictions. To do this the raw dataset contains missing and inconsistent values using machine learning models to impute and standardise the values.

The main method of addressing the missing values was predictive supervised models due to their ability to make accurate predictions in relation to patterns in the data. Compared to other imputation methods like mean or mode, predictive supervised models are able to use the underlying relationships between features resulting in the imputed values being more context-aware. As the dataset contain both numerical and categorical data appropriate models were selected to handle each type of data

* **Numerical data:** we decided to use K-Nearest Neighbors (KNN) and Random Forest Regressor to predict missing numerical values. KNN is able to capture local relationships between data points, thus effective for data that can be predicted based on similar observations in the dataset. While Random Forest Regressor is able to capture complex relationships in the dataset and is robust against overfitting, this is effective for data that is more prone to overfitting and has more complex relationships.
* **Categorical data:** we decided to use Random Forest Classifier to predict the missing categorical values.

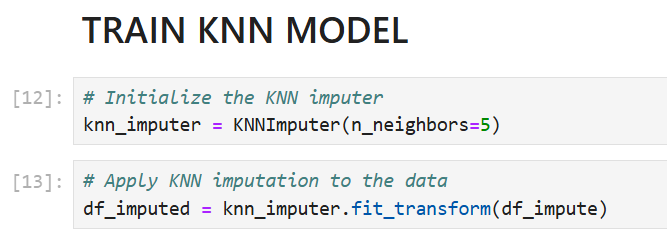
Through the combination of these predictive models, the dataset was clean so that the integrity and patterns of the original dataset is preserved and enables accurate training for the AI model.

Missing values:

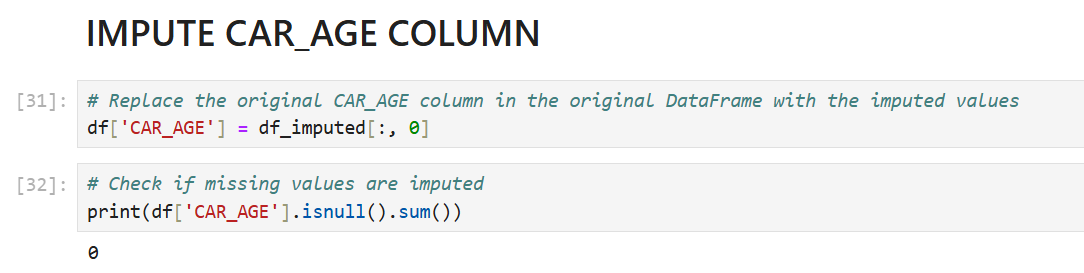
* Dataset 1: police\_report\_available, authorities\_contacted
* Dataset 2: address\_line2 (8505), city (54), customer\_education\_level (529), authorities\_contacted (1945), vendor\_id (3245)
* Dataset 3: age (7), yoj (548), income (570), home\_val (575), occupation (665), car\_age (639)

## **3.1. KNN Model**

KNN was used to impute the missing numerical values in the dataset. This is done by finding similar data on other columns and imputing the missing value based on the average of its nearest neighbour. KNN was used on columns like age, income, home\_val and car\_age in dataset 3 to impute the missing values from that column.



*Figure #: Training the KNN model*



*Figure #: Impute values into the ‘CAR\_AGE’ column*

### **3.1.1. Configuration Settings**

* n\_neighbours = 5: balances computations accuracy and complexity

### **3.1.2. Reasons for Model selection**

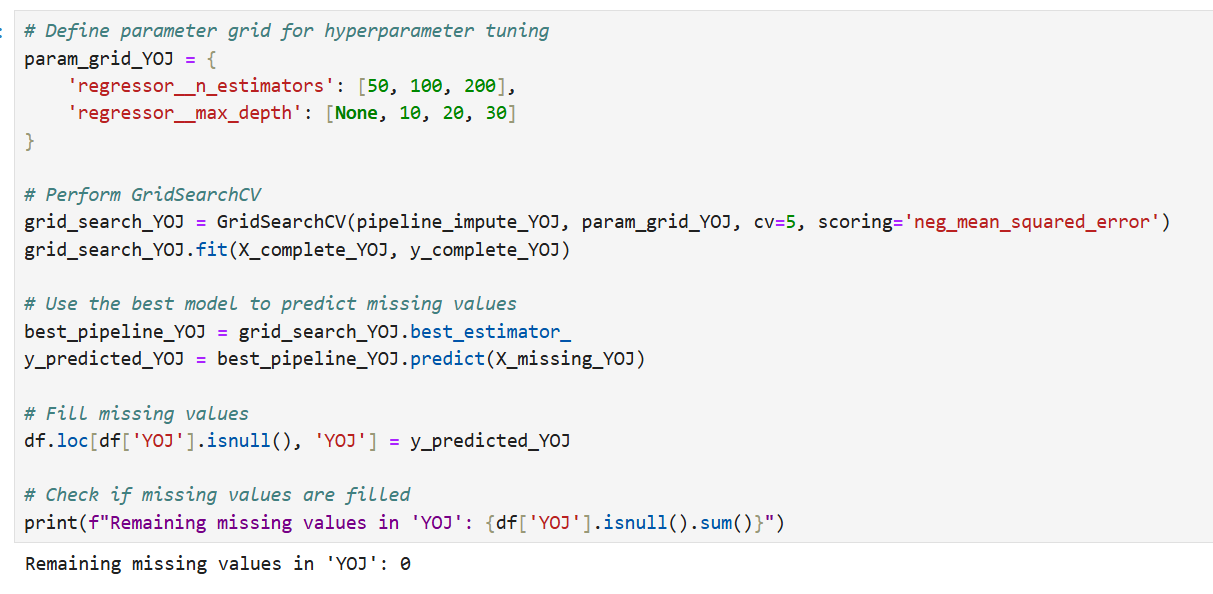
* Takes into account the similarity of records which is important when values relate to other columns.
* Ensures data completeness and standardisation
* By considering the relationship between other records it helps preserve the integrity of the dataset
* KNN is highly intuitive and simple to implement. The idea of using neighbours for predictions is easy to understand and visualise, making it a user-friendly model, especially for imputation tasks.
* KNN is versatile and works effectively with both numerical and categorical data, making it suitable for a wide range of imputation tasks across different types of datasets.
* KNN is non-parametric, meaning it does not make assumptions about the underlying data distribution, making it flexible in handling complex and varied data patterns.
* For imputation, KNN easily fills in missing values based on the average or most common value of similar data points (neighbours), without the need for complex pre-processing or assumptions about data structure.
* While KNN can be computationally intensive on very large datasets due to its reliance on distance calculations, its simplicity makes it efficient on moderately-sized datasets. Optimization techniques (like KD-Trees) can improve scalability.
* KNN inherently captures non-linear relationships since it relies on proximity in feature space rather than any functional mapping, which gives it an edge over linear models for data with non-linear dependencies.
* Key hyperparameters like the number of neighbours (k) can be easily tuned to optimise model performance, improving accuracy in various settings.

## **3.2. Random Forest Models**

* Random Forest Regressor:
* Random Forest Classifier:

### **3.2.1. Configuration Settings**

* GridSearchCV: hyperparameter tuning used to improve accuracy
* RandomisedSearchCV: hyperparameter tuning used to improve accuracy
* n\_estimators = 100: number of trees was 100, which provided a good balance between both computation and performance time
* Max\_depth = none: allows the tree to fully grow increasing the ability of the model to capture complex data patterns



****

### **3.2.2. Reasons for Model selection**

* Handles continuous target variables and categorical target variables
* Can handle imbalance classes
* Performance is robust even in incomplete datasets
* Though more complex than KNN, Random Forest is still easy to interpret. Each tree in the forest is a simple decision tree, and the final prediction is an average (for regression) or a majority vote (for classification), which is relatively straightforward.
* Random Forest handles both numerical and categorical data effortlessly. It is particularly useful in datasets with mixed types of features, as it can split on categorical variables without needing to encode them.
* Like KNN, Random Forest is non-parametric and does not assume any underlying distribution in the data. This flexibility allows it to model complex, non-linear relationships between features and the target variable effectively.
* Random Forest can handle missing data internally by building multiple decision trees that rely on different feature subsets. It does not require pre-imputation of missing values, which simplifies the preprocessing pipeline.
* Random Forest is highly scalable and can handle large datasets efficiently. It works well with both high-dimensional data and large sample sizes due to its ensemble approach, which reduces variance and prevents overfitting.
* Random Forest is excellent for capturing non-linear relationships in data. Each decision tree can split non-linearly at different thresholds, making Random Forest a robust model for tasks where nonlinear interactions are important.
* Random Forest offers several hyperparameters (e.g., number of trees, maximum depth, minimum samples per split) that can be tuned to optimise performance. This flexibility allows it to adjust to various datasets and improve predictive accuracy.
* For classification tasks, Random Forest can deal with class imbalances effectively through techniques like adjusting class weights or oversampling the minority class during training. This makes it a strong choice when the dataset has skewed class distributions.
* Since Random Forest is an ensemble of decision trees, it significantly reduces the risk of overfitting compared to individual decision trees. It builds trees on random subsets of features and data, which improves generalisation.

## **3.3. Why were these models chosen over other approaches?**

3.3.1. Predictive Models

* Linear regression: is a model that predicts missing values through linear relationships between target and independent variables, however, are not able to capture complex interactions in the data.
* Decision Trees: however, are prone to overfitting on small datasets like dataset 1 and dataset 2 and may not perform effectively with non-linear relationships.

3.3.2. Supervised Learning Models

* Neural networks: are able to predict missing values through the analysis of complex patterns in the dataset however, are computationally expensive and require significant tuning which, furthermore, they require large datasets for accurate training.
* Gradient Boosting: However, this is also computationally expensive and requires significant tuning while also having a slower performance.

3.3.3. Unsupervised Learning Models

* K-means: has the potential to group similar data points and impute missing values based on the mode or mean of that cluster however, it method mainly focuses on clustering hence not ideal for getting accurate prediction of missing values
* Principal Component Analysis: can be used to reduce dimensionality and use the dataset’s principal components to fill in missing values; however, this method will result in simple imputations failing to capture patterns in the data.

3.3.4. Classifiers

* Logistic Regression: can be used to predict missing categorical values through modelling binary outcomes however, due to this it is not suited for imputing missing values.
* Naive Bayes: uses probability distributions to impute missing values

3.3.5. Chosen Models

* KNN:
* Random Forest

# **4. Detailed Data descriptions:**

*Data being used, data being generated, data being stored, as well as any summaries and/or reports.*

## **4.1 Original Raw Datasets**

The raw datasets used in this project consist of three datasets sourced from Kaggle, each containing historical data on automobile insurance claims. These datasets were selected primarily because they include a column that explicitly indicates whether a claim was fraudulent, which is essential for training machine learning and AI models for fraud detection.

The original datasets are in a raw, unclean state, with many missing or non-standardized values. A thorough cleaning and enrichment process was applied to prepare them for analysis.

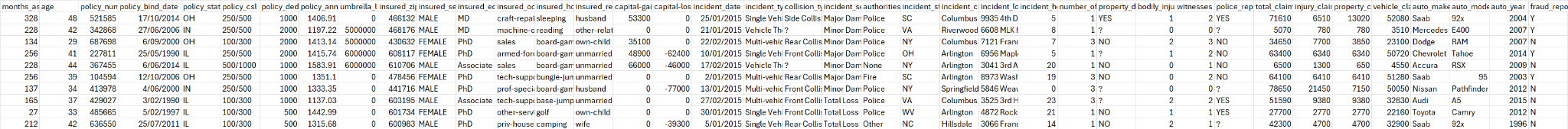
* Dataset 1: This dataset, containing 1,000 rows, is well-structured and widely available across various websites. It is frequently used in insurance fraud analysis due to its quality and completeness. The dataset was sourced from Kaggle, as it is an open-source platform. The dataset is hosted on GitHub at: [GitHub Link](https://github.com/mwitiderrick/insurancedata/blob/master/insurance_claims.csv), and the Kaggle source is available here: [Kaggle Link](https://www.kaggle.com/datasets/sumansuhag/insurance-dataset-csv).
* Dataset 2: This dataset is also quite well-structured and originally contains 10,000 rows but includes data on various types of fraud, not limited to automobile insurance. After filtering for automobile fraud, approximately 1,575 relevant rows were extracted. This dataset offers a broader range of fraud types but required focused extraction for the target use case of automobile insurance fraud detection. The dataset can be accessed on Kaggle: [Kaggle Link](https://www.kaggle.com/datasets/mastmustu/insurance-claims-fraud-data?select=insurance_data.csv).
* Dataset 3: This dataset, containing 10,303 rows, is the most challenging due to its high level of missing, blank, and non-standardized values. It required significant cleaning efforts to make it usable for analysis. Despite its issues, it provides valuable data for automobile insurance fraud detection once processed. The dataset can be found on Kaggle: [Kaggle Link](https://www.kaggle.com/datasets/xiaomengsun/car-insurance-claim-data).

These datasets form the foundation for the final dataset to be used by the fraud detection model. Providing this historical claims data allows the machine learning algorithms to learn patterns and anomalies indicative of fraudulent claims. The presence of a fraud indicator in each dataset ensures that they are particularly well-suited for supervised learning in fraud detection.

## 

## **4.2. Columns in Raw Datasets**

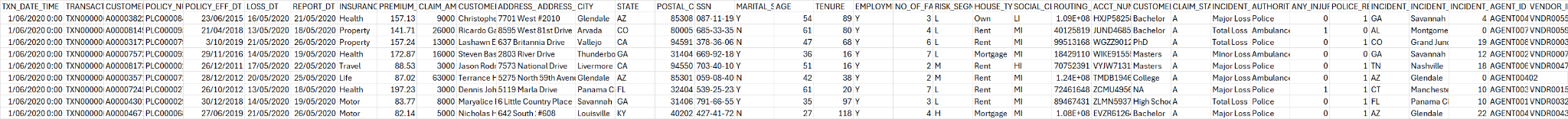
### **Raw Dataset 1:**



| NO. | Column Name | Column Letter | Data Type | Description |
| --- | --- | --- | --- | --- |
| 1 | months\_as\_customer | A | Numerical | Number of months the individual has been a customer |
| 2 | age | B | Numerical | Age of the customer |
| 3 | policy\_number | C | Categorical | Unique identifier for the insurance policy |
| 4 | policy\_bind\_date | D | Date | The date when the policy was created/bound |
| 5 | policy\_state | E | Categorical | The state where the policy was issued |
| 6 | policy\_csl | F | Categorical | Combined Single Limit for the policy |
| 7 | policy\_deductable | G | Numerical | The amount of deductible in the policy |
| 8 | policy\_annual\_premium | H | Numerical | The annual premium paid for the insurance policy |
| 9 | umbrella\_limit | I | Numerical | The umbrella limit in the policy |
| 10 | insured\_zip | J | Categorical | Zip code of the insured person |
| 11 | insured\_sex | K | Categorical | Gender of the insured person |
| 12 | insured\_education\_level | L | Categorical | Education level of the insured person |
| 13 | insured\_occupation | M | Categorical | Occupation of the insured person |
| 14 | insured\_hobbies | N | Categorical | Hobbies of the insured person |
| 15 | insured\_relationship | O | Categorical | Relationship status of the insured person |
| 16 | capital-gains | P | Numerical | Capital gains of the insured person |
| 17 | capital-loss | Q | Numerical | Capital loss of the insured person |
| 18 | incident\_date | R | Date | Date of the incident |
| 19 | incident\_type | S | Categorical | Type of incident |
| 20 | collision\_type | T | Categorical | Type of collision |
| 21 | incident\_severity | U | Categorical | Severity of the incident |
| 22 | authorities\_contacted | V | Categorical | Whether authorities were contacted |
| 23 | incident\_state | W | Categorical | State where the incident occurred |
| 24 | incident\_city | X | Categorical | City where the incident occurred |
| 25 | incident\_location | Y | Categorical | Specific location of the incident |
| 26 | incident\_hour\_of\_the\_day | Z | Numerical | Hour of the day when the incident occurred |
| 27 | number\_of\_vehicles\_involved | AA | Numerical | Number of vehicles involved in the incident |
| 28 | property\_damage | AB | Categorical | Whether property damage occurred |
| 29 | bodily\_injuries | AC | Numerical | Number of bodily injuries |
| 30 | witnesses | AD | Numerical | Number of witnesses to the incident |
| 31 | police\_report\_available | AE | Boolean | Whether a police report is available |
| 32 | total\_claim\_amount | AF | Numerical | The total amount claimed |
| 33 | injury\_claim | AG | Numerical | Claim amount for injuries |
| 34 | property\_claim | AH | Numerical | Claim amount for property damage |
| 35 | vehicle\_claim | AI | Numerical | Claim amount for vehicle damage |
| 36 | auto\_make | AJ | Categorical | Make of the vehicle involved |
| 37 | auto\_model | AK | Categorical | Model of the vehicle involved |
| 38 | auto\_year | AL | Numerical | Year of the vehicle involved |
| 39 | fraud\_reported | AM | Boolean | Whether fraud was reported |

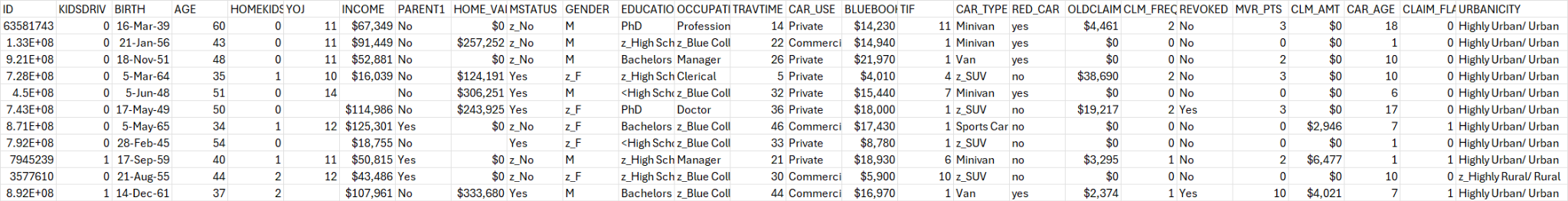
### 

### **Raw Dataset 2:**



| NO. | Column Name | Column Letter | Data Type | Description |
| --- | --- | --- | --- | --- |
| 1 | TXN\_DATE\_TIME | A | Date/Time | The date and time of the transaction |
| 2 | TRANSACTION\_ID | B | Categorical | Unique identifier for the transaction |
| 3 | CUSTOMER\_ID | C | Categorical | Unique identifier for the customer |
| 4 | POLICY\_NUMBER | D | Categorical | Unique identifier for the policy |
| 5 | POLICY\_EFF\_DT | E | Date | The policy effective date |
| 6 | LOSS\_DT | F | Date | Date when the loss occurred |
| 7 | REPORT\_DT | G | Date | Date when the incident was reported |
| 8 | INSURANCE\_TYPE | H | Categorical | Type of insurance |
| 9 | PREMIUM\_AMOUNT | I | Numerical | Premium amount for the insurance policy |
| 10 | CLAIM\_AMOUNT | J | Numerical | Amount claimed |
| 11 | CUSTOMER\_NAME | K | Categorical | Name of the customer |
| 12 | ADDRESS\_LINE1 | L | Categorical | Address line 1 of the customer |
| 13 | ADDRESS\_LINE2 | M | Categorical | Address line 2 of the customer |
| 14 | CITY | N | Categorical | City of the customer |
| 15 | STATE | O | Categorical | State of the customer |
| 16 | POSTAL\_CODE | P | Categorical | Postal code of the customer |
| 17 | SSN | Q | Categorical | Social Security Number |
| 18 | MARITAL\_STATUS | R | Categorical | Marital status of the customer |
| 19 | AGE | S | Numerical | Age of the customer |
| 20 | TENURE | T | Numerical | Number of years the customer has been with the insurance company |
| 21 | EMPLOYMENT\_STATUS | U | Categorical | Employment status of the customer |
| 22 | NO\_OF\_FAMILY\_MEMBERS | V | Numerical | Number of family members |
| 23 | RISK\_SEGMENTATION | W | Categorical | Risk category assigned to the customer |
| 24 | HOUSE\_TYPE | X | Categorical | Type of house the customer owns or rents |
| 25 | SOCIAL\_CLASS | Y | Categorical | Social class of the customer |
| 26 | ROUTING\_NUMBER | Z | Categorical | Customer’s bank routing number |
| 27 | ACCT\_NUMBER | AA | Categorical | Customer’s bank account number |
| 28 | CUSTOMER\_EDUCATION\_LEVEL | AB | Categorical | Education level of the customer |
| 29 | CLAIM\_STATUS | AC | Categorical | Status of the claim |
| 30 | INCIDENT\_SEVERITY | AD | Categorical | Severity of the incident |
| 31 | AUTHORITY\_CONTACTED | AE | Boolean | Whether the authorities were contacted |
| 32 | ANY\_INJURY | AF | Boolean | Whether any injury was reported |
| 33 | POLICE\_REPORT\_AVAILABLE | AG | Boolean | Whether a police report is available |
| 34 | INCIDENT\_STATE | AH | Categorical | State where the incident occurred |
| 35 | INCIDENT\_CITY | AI | Categorical | City where the incident occurred |
| 36 | INCIDENT\_HOUR\_OF\_THE\_DAY | AJ | Numerical | Hour of the day when the incident occurred |
| 37 | AGENT\_ID | AK | Categorical | Unique identifier for the insurance agent |
| 38 | VENDOR\_ID | AL | Categorical | Unique identifier for the vendor |

### **Raw Dataset 3:**



| NO. | Column Name | Column Letter | Data Type | Description |
| --- | --- | --- | --- | --- |
| 1 | ID | A | Categorical | Unique identifier for the customer |
| 2 | KIDSDRIV | B | Numerical | Number of kids driving in the household |
| 3 | BIRTH | C | Date | Birth date of the customer |
| 4 | AGE | D | Numerical | Age of the customer |
| 5 | HOMEKIDS | E | Numerical | Number of kids living in the household |
| 6 | YOJ | F | Numerical | Years on the job |
| 7 | INCOME | G | Numerical (Currency) | Income of the customer |
| 8 | PARENT1 | H | Categorical | Whether the customer is a single parent |
| 9 | HOME\_VAL | I | Numerical (Currency) | Value of the home |
| 10 | MSTATUS | J | Categorical | Marital status |
| 11 | GENDER | K | Categorical | Gender of the customer |
| 12 | EDUCATION | L | Categorical | Education level of the customer |
| 13 | OCCUPATION | M | Categorical | Occupation of the customer |
| 14 | TRAVTIME | N | Numerical | Travel time to work |
| 15 | CAR\_USE | O | Categorical | Whether the car is used for private or commercial purposes |
| 16 | BLUEBOOK | P | Numerical (Currency) | Value of the car |
| 17 | TIF | Q | Numerical | Time in force of the insurance policy |
| 18 | CAR\_TYPE | R | Categorical | Type of car |
| 19 | RED\_CAR | S | Boolean | Whether the car is red |
| 20 | OLDCLAIM | T | Numerical (Currency) | Amount claimed in prior incidents |
| 21 | CLM\_FREQ | U | Numerical | Frequency of claims |
| 22 | REVOKED | V | Boolean | Whether the driver's licence has been revoked |
| 23 | MVR\_PTS | W | Numerical | Motor Vehicle Record points |
| 24 | CLM\_AMT | X | Numerical (Currency) | Amount claimed in the current incident |
| 25 | CAR\_AGE | Y | Numerical | Age of the car |
| 26 | CLAIM\_FLAG | Z | Boolean | Whether a claim was filed |
| 27 | URBANICITY | AA | Categorical | Urban/rural classification |

## 

## **4.3. Cleaning & Enriching**

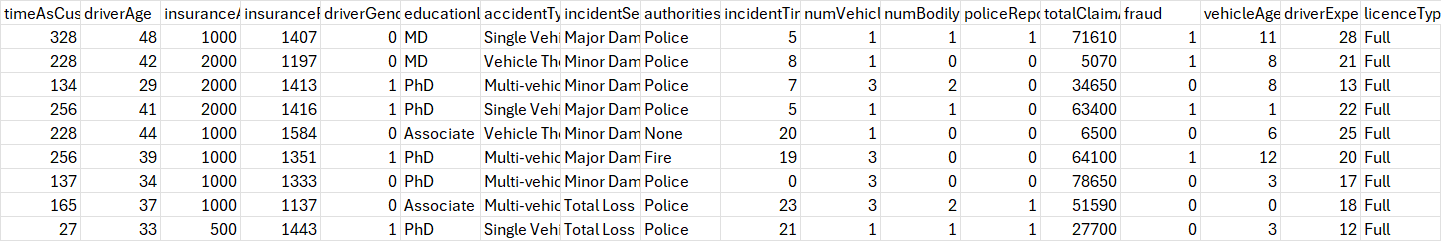
### **Dataset 1 & Dataset 2**

After the cleaning and enrichment process, Dataset 1 (1,000 rows) and Dataset 2 (approximately 1,500 rows) were refined by selecting 16 key columns and generating 2 additional columns using the original data. These newly created columns, Driver Experience and Licence Type, were calculated using straightforward mathematical formulas.

* Driver Experience: This column is derived from the driver's age, using the formula:  
  driver experience = driver age - 16 + R,  
  where R is a random variable between 0 and 6. The subtraction of 16 accounts for the minimum age to obtain a licence, while the random variable represents the variability in the time drivers might wait before getting their licence.
* Licence Type: The licence type is determined by a simple set of conditional statements based on the driver’s experience:
  + Less than 1 year of experience: Ls
  + 1 to 2 years of experience: P1
  + 2 to 4 years of experience: P2
  + More than 4 years of experience: Full

For Dataset 2, no gender data was available, so the driverGender column remains blank for that dataset, ensuring consistency without introducing any erroneous or bias inducing data.

### **Cleaned & Enriched Dataset 1:**



| NO. | Column Name | Column Letter | Data Type | Description |
| --- | --- | --- | --- | --- |
| 1 | timeAsCustomer | A | Numerical | Number of months the individual has been a customer |
| 2 | driverAge | B | Numerical | Age of the driver |
| 3 | insuranceAccess | C | Numerical | The level of access to insurance |
| 4 | insurancePremium | D | Numerical | The premium amount for the insurance policy |
| 5 | driverGender | E | Boolean | Gender of the driver (0 = Male, 1 = Female) |
| 6 | educationLevel | F | Categorical | Education level of the driver |
| 7 | accidentType | G | Categorical | Type of accident |
| 8 | incidentSeverity | H | Categorical | Severity of the incident |
| 9 | authoritiesInvolved | I | Categorical | Whether authorities were involved |
| 10 | incidentTime | J | Numerical | Time of the incident |
| 11 | numVehiclesInvolved | K | Numerical | Number of vehicles involved in the accident |
| 12 | numBodilyInjuries | L | Numerical | Number of bodily injuries |
| 13 | policeReportBool | M | Boolean | Whether a police report was available (0 = No, 1 = Yes) |
| 14 | totalClaimAmount | N | Numerical | The total claim amount |
| 15 | fraud | O | Boolean | Whether the claim was fraudulent (0 = No, 1 = Yes) |
| 16 | vehicleAge | P | Numerical | Age of the vehicle |
| 17 | driverExperience | Q | Numerical | Number of years of driving experience |
| 18 | licenceType | R | Categorical | Type of licence held by the driver |

### 

### **Cleaned & Enriched Dataset 2:**

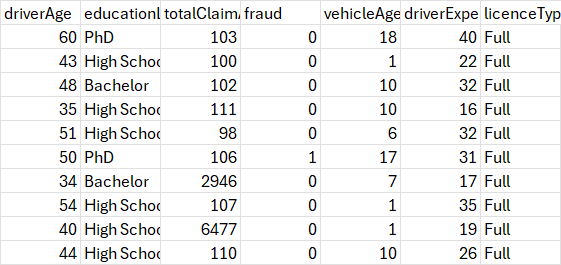
| NO | Column Name | Column Letter | Data Type | Description |
| --- | --- | --- | --- | --- |
| 1 | timeAsCustomer | A | Numerical | Number of months the individual has been a customer |
| 2 | driverAge | B | Numerical | Age of the driver |
| 3 | insuranceAccess | C | Numerical | The level of access to insurance |
| 4 | insurancePremium | D | Numerical | The premium amount for the insurance policy |
| 5 | driverGender | E | Boolean (0 = Male, 1 = Female, empty) | Gender of the driver (0 = Male, 1 = Female) EMPTY |
| 6 | educationLevel | F | Categorical | Education level of the driver |
| 7 | accidentType | G | Categorical | Type of accident |
| 8 | incidentSeverity | H | Categorical | Severity of the incident |
| 9 | authoritiesInvolved | I | Categorical | Whether authorities were involved |
| 10 | incidentTime | J | Numerical | Time of the incident |
| 11 | numVehiclesInvolved | K | Numerical | Number of vehicles involved in the accident |
| 12 | numBodilyInjuries | L | Numerical | Number of bodily injuries |
| 13 | policeReportBool | M | Boolean (0 = No, 1 = Yes) | Whether a police report was available (0 = No, 1 = Yes) |
| 14 | totalClaimAmount | N | Numerical | The total claim amount |
| 15 | fraud | O | Boolean (0 = No, 1 = Yes) | Whether the claim was fraudulent (0 = No, 1 = Yes) |
| 16 | vehicleAge | P | Numerical | Age of the vehicle |
| 17 | driverExperience | Q | Numerical | Number of years of driving experience |
| 18 | licenceType | R | Categorical | Type of licence held by the driver |

### **Dataset 3**

Dataset 3, which is the largest dataset (10000) rows, was missing many of the columns present in the other datasets. Instead of synthesising all the columns it was missing, it was decided to just get the columns that match those and leave the rest empty. This was to prevent such a large dataset becoming biassed and skewed from data from smaller datasets. It increased variation and variety of data, and would assist in helping the AI model learn what to do if not all rows of data were provided.

Thus, Dataset 3’s selected columns are (included LicenceType and DriverExperience which were enriched using same formula as previous datasets):

### **Cleaned & Enriched Dataset 3:**



1. driverAge: Age of the driver (Numerical).
2. educationLevel: Education level of the driver (Categorical).
3. totalClaimAmount: The total claim amount (Numerical).
4. fraud: Whether the claim was fraudulent (Boolean: 0 = No, 1 = Yes).
5. vehicleAge: Age of the vehicle (Numerical).
6. driverExperience: Number of years of driving experience (Numerical).
7. licenceType: Type of licence held by the driver (Categorical).

| No. | Column Name | Column Letter | Data Type | Description |
| --- | --- | --- | --- | --- |
| 1 | driverAge | A | Numerical | Age of the driver |
| 2 | educationLevel | B | Categorical | Education level of the driver |
| 3 | totalClaimAmount | C | Numerical | The total claim amount |
| 4 | fraud | D | Boolean | Whether the claim was fraudulent (0 = No, 1 = Yes) |
| 5 | vehicleAge | E | Numerical | Age of the vehicle |
| 6 | driverExperience | F | Numerical | Number of years of driving experience |
| 7 | licenceType | G | Categorical | Type of licence held by the driver |

## **4.4. Generating Data:**

The dataset consists of 4,000 rows of synthetic data generated using Method 1 and an additional 4,000 rows generated using Method 2. Method 1 involved using statistical remapping and random sampling techniques to synthesise data based on trends in the original dataset. Method 2 leveraged advanced machine learning techniques, such as Generative Adversarial Networks (GANs), to create more realistic synthetic data by modelling the underlying patterns of the original dataset.

This dual approach allowed us to create a more diverse and comprehensive dataset for analysis and machine learning tasks.

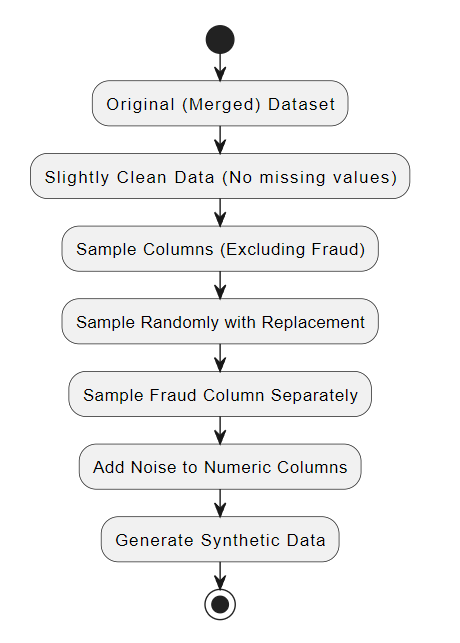
### **Method 1:**

Synthetic Data Method 1 generates additional data by sampling from the original dataset, maintaining the same distribution and characteristics of the real data. This method is used to expand the dataset, especially when more data is needed for training machine learning models, but without introducing random or irrelevant data.

The synthetic data is generated by first cleaning any missing or blank values (such as replacing blanks in authoritiesInvolved with "none"). Then, we sample the original dataset to create new rows. The important point here is that the fraud column is kept reliable and unchanged. While other columns are sampled and noise is added to numeric fields to make the data varied, the fraud column is handled separately to ensure that its distribution remains consistent with the original data.

For the fraud column, we directly sample from the original fraud values and then add them to the synthetic data without any modification or randomization. This ensures that the synthetic data reflects the same patterns of fraudulent and non-fraudulent cases found in the original dataset, maintaining its integrity and reliability.

By preserving the fraud column’s original structure and distribution, this method ensures that synthetic data can be confidently used for analysis or machine learning without introducing artificial patterns into the fraud labels.



Steps:

1. Original Dataset: The raw data with columns like driverAge, accidentType, fraud, etc.
2. Clean Data: Any missing values (like in authoritiesInvolved) are replaced with "none".
3. Sample Columns (Excluding Fraud): Randomly sample all columns except fraud.
4. Sample Fraud Column Separately: The fraud column is sampled independently.
5. Add Noise to Numeric Columns: Add small noise to numeric columns to vary values.
6. Generate Synthetic Data: Final synthetic dataset output.

### 

### 

### **Method 2:**

Synthetic Data Method 2 uses a more advanced approach compared to Method 1. It leverages a generative model called CTGAN (Conditional Tabular GAN), which is designed to generate highly realistic synthetic data by learning the underlying patterns and relationships in the original dataset. This method is particularly effective when dealing with complex datasets that include both categorical and continuous variables.

The process begins by loading the original dataset and defining its metadata. The CTGANSynthesizer is then used to train a model on the original data. This model learns the data's distribution and structure over several epochs (1,800 in this case), ensuring that the synthetic data it generates closely mimics the patterns in the original dataset.

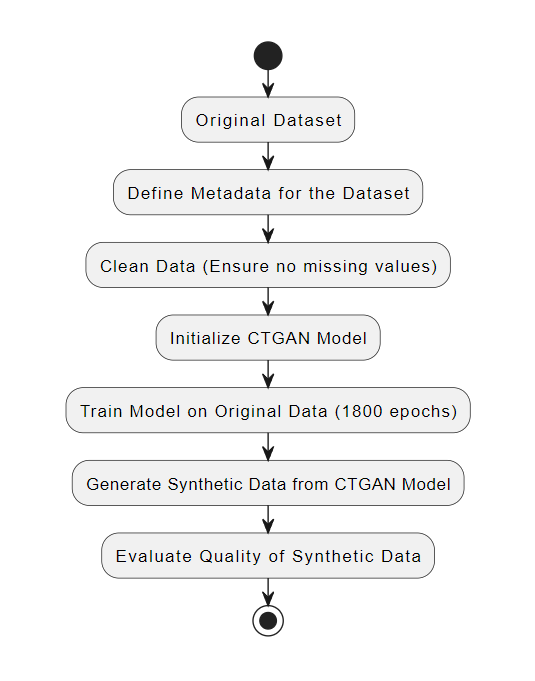
One of the key strengths of Method 2 is that it not only captures basic distributions, but it also models the relationships between different columns. This means that the generated data retains important dependencies between variables, making it more robust for tasks like machine learning or data analysis.

Fraud Column Reliability:

Similar to Method 1, the fraud column in Method 2 is kept reliable. Since the CTGAN model learns from the actual data distribution, it maintains the integrity of the fraud column and generates fraud labels in line with the original dataset. Unlike random sampling, this method ensures that the synthetic data generated by the GAN reflects the real-world proportions and relationships between fraudulent and non-fraudulent cases.

Once the synthetic data is generated, it is evaluated for quality using a tool that compares it to the real data. This ensures that the synthetic data is both accurate and usable for further analysis. Finally, the data is saved, with the index reset to start from 1, ensuring a clean output.

In summary, Synthetic Data Method 2 provides an advanced and reliable approach for generating synthetic data by using CTGAN. It captures the complex relationships in the dataset while ensuring that the fraud column remains accurate and trustworthy.

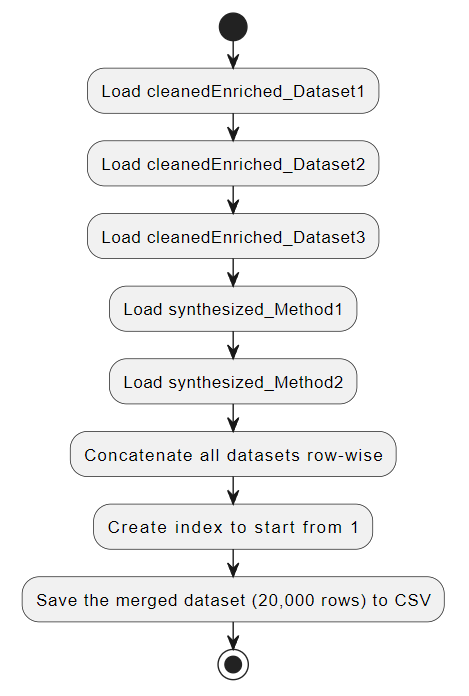


Steps:

1. Original Dataset: Load the original dataset with columns such as driverAge, accidentType, and fraud.
2. Define Metadata: The structure and relationships of the columns are defined in the metadata, which helps guide the model's learning.
3. Clean Data: Ensure that missing values are handled, for example, replacing empty values in authoritiesInvolved with "none".
4. Initialize CTGAN Model: The CTGAN model is initialised with hyperparameters such as the number of training epochs.
5. Train Model on Original Data: The model is trained on the original dataset, learning the underlying patterns and relationships.
6. Generate Synthetic Data: After training, the model generates new synthetic data that mirrors the original data in terms of structure and distribution.
7. Evaluate Quality: The generated synthetic data is evaluated against the original data to ensure quality and reliability, particularly for sensitive columns like fraud.

### **Merging Datasets:**

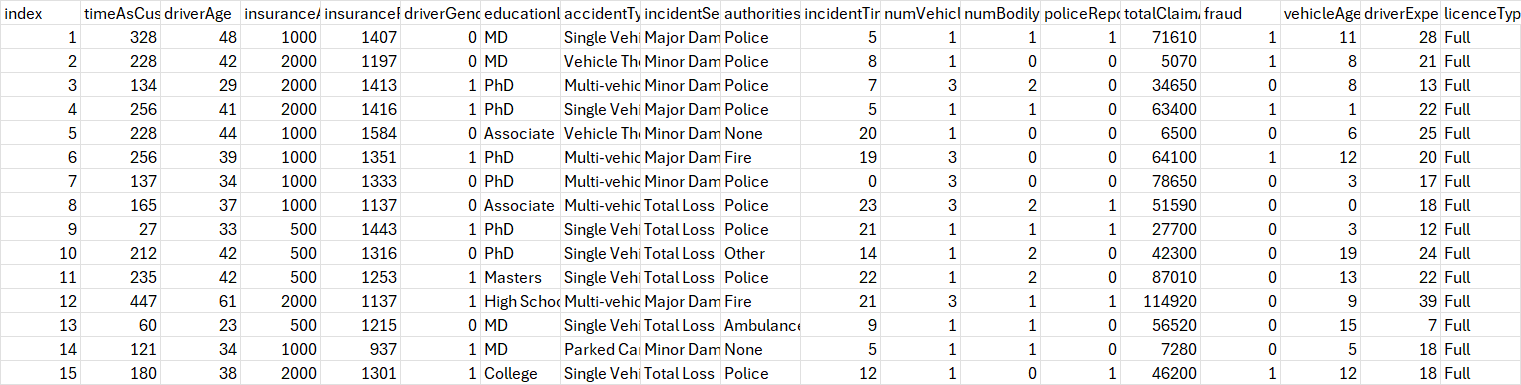
The dataset merging process involves combining three cleaned and enriched datasets with two synthesised datasets into a single dataset containing ~ 20,500 rows. This process ensures the data is consolidated, structured, and ready for further analysis or machine learning tasks.



Steps in the Dataset Merger:

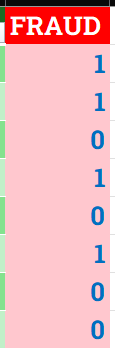
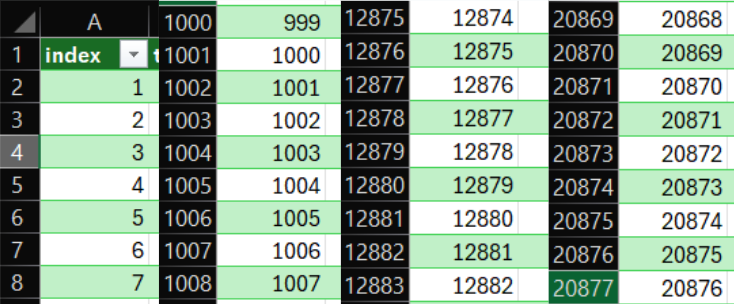
1. Loading the Datasets:
   * Three cleaned and enriched datasets (cleanedEnriched\_Dataset1, cleanedEnriched\_Dataset2, and cleanedEnriched\_Dataset3) are loaded. These datasets have undergone data cleaning and enrichment processes to ensure completeness and quality.
   * Two synthesized datasets (Synthesized\_Method1 and Synthesized\_Method2) are also loaded, representing synthetic data generated through advanced methods.
2. Concatenating the Datasets:
   * The datasets are concatenated row-wise, combining rows from all five datasets into one. This ensures that the merged dataset includes a complete set of data points from all the cleaned, enriched, and synthesised sources.
3. Resetting the Index:
   * A new index is generated for the merged dataset, starting from 1. This provides a unique identifier for each row in the dataset, ensuring that it is well-structured and organised.
4. Saving the Merged Dataset:
   * The final merged dataset, containing 20,000 rows, is saved to a CSV file. This dataset is now ready for further use, such as in machine learning models or data analysis.

Merged Dataset (20877 rows):

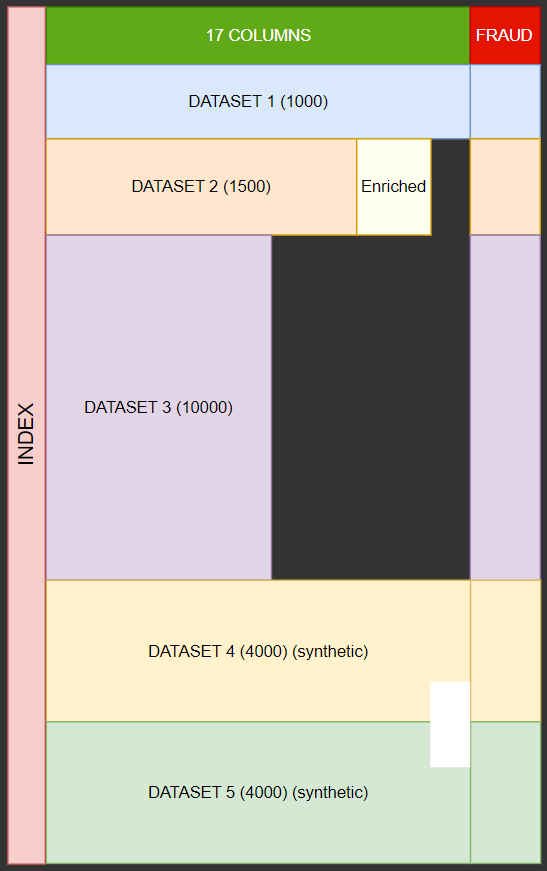


| NO. | Column Name | Column Letter | Data Type | Description |
| --- | --- | --- | --- | --- |
| 1 | index | A | Numerical | Row index for tracking purposes |
| 2 | timeAsCustomer | B | Numerical | Number of months the individual has been a customer |
| 3 | driverAge | C | Numerical | Age of the driver |
| 4 | insuranceAccess | D | Numerical | The level of access to insurance |
| 5 | insurancePremium | E | Numerical | The premium amount for the insurance policy |
| 6 | driverGender | F | Boolean | Gender of the driver (0 = Male, 1 = Female) |
| 7 | educationLevel | G | Categorical | Education level of the driver |
| 8 | accidentType | H | Categorical | Type of accident |
| 9 | incidentSeverity | I | Categorical | Severity of the incident |
| 10 | authoritiesInvolved | J | Categorical | Whether authorities were involved |
| 11 | incidentTime | K | Numerical | Time of the incident |
| 12 | numVehiclesInvolved | L | Numerical | Number of vehicles involved in the accident |
| 13 | numBodilyInjuries | M | Numerical | Number of bodily injuries |
| 14 | policeReportBool | N | Boolean | Whether a police report was available (0 = No, 1 = Yes) |
| 15 | totalClaimAmount | O | Numerical | The total claim amount |
| 16 | fraud | P | Boolean | Whether fraud was reported (0 = No, 1 = Yes) |
| 17 | vehicleAge | Q | Numerical | Age of the vehicle |
| 18 | driverExperience | R | Numerical | Number of years of driving experience |
| 19 | licenceType | S | Categorical | Type of licence held by the driver |

Index: Fraud:



Structure of resulting dataset after merging:



**TEST SPECIFICATION**

## 

# **1. Model evaluation: - Tash**

*· how will any data / outputs be compared / tested / evaluated for correctness and accuracy?*

*· If you are choosing between models, how will the models be compared / contrasted to see which one has a better performance (e.g. if you are comparing different classifier models, on what basis are you comparing them? Detail each comparison.*

### 

## **1.1. Model Evaluation and Comparison**

#### **Data Cleaning, Synthesisation, and Imputation:**

In order to confirm the precision and trustworthiness of our data, a comprehensive evaluation was conducted on the datasets, including comparisons between original, cleaned, and synthesised data. This step was essential in verifying the reliability and continuity of the data sets utilised in our research.

### **1.1.1. Correlation Matrix and ANOVA Analysis:**

**Correlation Matrix Comparison:** Upon completing our data cleaning processes and synthesising data sets for analysis purposes. We examined correlation matrix comparisons to assess any alterations, in the strength of relationships among numerical features before and after cleaning our datasets and between the cleaned data and synthesised data sets. By scrutinising these correlation matrices for changes resulting from our data enhancement procedures and ensuring the consistency of variable relationships, across datasets remained intact. We were able to validate the effectiveness of our preprocessing steps effectively.

**ANOVA Tests:** We utilised ANOVA to compare the values across groups, in the datasets aiming to examine any variations present among them. Our team conducted ANOVA analyses to determine the discrepancies between the data distributions in the datasets compared to the initial dataset and to contrast the distributions between polished data and generated data. This statistical examination aided us in confirming that preprocessing procedures did not introduce any biases or notable alterations, in the distribution of data.

**Consistency Check:** In both synthesised datasets (generated using different methods),the overall count of fraudulent cases is tallied and matched up for comparison purposes. The number of fraud instances, in both scenarios hovered close to 400 points in line with the dataset figures. This alignment in the fraud case counts implies that both methods effectively mirror the distribution of activities. Validation through Distribution: Given the similarity in fraud counts across the synthesised datasets, it is highly unlikely that such results are coincidental. This suggests that the models are accurately capturing the underlying patterns of fraud found in the data set and confirming the effectiveness of the synthetic generation techniques.

### **1.1.2. Validation of Synthesised Data:**

**Fraud Count Consistency:** Both synthetic datasets generated approximately 400 instances of fraud, providing strong validation for the synthesis process. Given the rarity of fraud cases, it is improbable that both synthesised datasets would have this coincidence unless the process successfully replicated the underlying distribution patterns of the real data. This consistency was used as an additional checkpoint to ensure that the synthetic data reflected the real world occurrences correctly.

##### 

### **1.1.3. Imputation Techniques:**

Missing data presents a challenge that can significantly impact the quality and reliability of model predictions. We explored multiple imputation methods for both numerical and categorical columns. Given the complexity of our dataset and the need for more advanced handling of missing values, we decided to use K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Random Forest models for imputation. We used several key metrics to evaluate the accuracy and effectiveness of our imputation methods for both numerical and categorical columns.

##### 

##### **For Numerical Columns:**

The primary metrics used for evaluating imputation in numerical columns were:

* **Mean Squared Error (MSE):** MSE measures the average squared difference between the imputed values and the actual values (where known). Our target MSE was for it to be as low as possible ideally <0.10
* **Mean Absolute Error (MAE):** MAE is similar to MSE but focuses on the absolute differences between the imputed values and the actual values, making it less sensitive to large errors.Our target MSE was for it to be as low as possible ideally <0.05.
* **R-squared (R²):** R² measures how well the imputed values explain the variance in the dataset. Our target for R2 was to be as close to 1 ideally >0.80.

These metrics provided insight into how well the imputed values aligned with the true values in the dataset, allowing us to optimise and refine our approach.

##### 

##### **For Categorical Columns:**

For categorical columns, accuracy metrics are more appropriate for evaluating model performance. We focused on:

* **Accuracy:** Measures the percentage of correctly imputed categories compared to the total number of categories.Our aim for imputation accuracy was to >75%
* **Consistency Checks**: For instance, checking if the imputed values for columns like "Occupation" and "Education Level" logically fit the other related columns (e.g., if a person with a PhD is categorised under an occupation like "Clerical" might indicate poor imputation).

#### 

## **1.2. How we compared and evaluated Models and decided what to choose**

In our project, we used several models for both imputation and classification, including K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Random Forest classifier/ regressor. While the choice of model is essential, the performance evaluation metrics we used to assess these models played an even more significant role in guiding our final decisions. These evaluation techniques allowed us to assess the models' ability to accurately impute missing data and classify fraudulent and non-fraudulent claims. While also comparing and contrasting the models objectively, ensuring we selected the most suitable approach for each specific task.

### **1.2.1. R² (Coefficient of Determination)**

The **R²** metric is particularly useful for continuous variables. It measures how well the model explains the variability in the target variable. An R² closer to 1 indicates that the model captures most of the variability, whereas an R² closer to 0 means the model explains little of the variance.

**Impact on model choice:** For tasks like imputing continuous columns, we relied heavily on R² to understand how well the model was capturing the underlying patterns. A low R² suggested the model was underperforming, prompting us to try different models or adjust the features being used. On the other hand, a higher R² indicated the model was a good fit for the data, leading us to keep and refine those models.

For the YOJ (Years on Job) column, which is continuous data, R² became the most critical. The first model we applied gave us an R² score of 42%, indicating that the model explained only 42% of the variance in the target variable. This was relatively low, and it informed us that the model wasn't capturing the complexity of the data. Therefore we used it to refine the model. With such a low R², we explored other models (KNN, SVR, and Random Forest) to determine if we could achieve better performance. For continuous data like YOJ, a higher R² is crucial because it shows how well the model is capturing the real-world variations in the data. After hyperparameter tuning, we managed to improve the R² score to 99%, justifying our selection of the final model.

### **1.2.2. Accuracy**

Accuracy is a key metric for evaluating classification tasks. It simply measures the percentage of correctly predicted outcomes out of the total predictions made. While accuracy is often the most intuitive metric, it can sometimes be misleading, especially when dealing with imbalanced datasets.

**Impact on model choice:** In our tasks, accuracy helped us identify models that performed well overall. However, we also recognized its limitations, especially in cases where certain categories or classes were underrepresented. For example, if only a small portion of the data represented a specific outcome (e.g., fraudulent claims), a model might have high accuracy while still failing to predict the minority class effectively. As a result, we supplemented accuracy with other metrics (like precision and recall) to ensure the model was suitable for all parts of the data.

When imputing the 'Occupation' column, accuracy was used to evaluate how well each model predicted missing values. Initial attempts with one model resulted in low accuracy scores, indicating that it struggled to capture the relationships in the data. By analysing the accuracy, we identified that tweaking the hyperparameters and switching models improved performance.

For the 'Education Level' column, the accuracy score again played a crucial role. When accuracy was low, we either adjusted the model or added more features to improve predictions. The feedback from accuracy scores ensured that we did not settle on a model unless it provided reasonable confidence in the imputed values. While accuracy was a useful initial measure, we avoided making decisions just based on it, because we didn't always capture the nuances in our dataset.

#### 

### **1.2.3. Mean Squared Error (MSE) and Mean Absolute Error (MAE)**

Both MSE and MAE are fundamental metrics for evaluating the performance of models, especially for tasks like imputing missing values or handling continuous data. These metrics help quantify the prediction errors, guiding us toward selecting models that minimise these errors.

**MSE**: Measures the average squared difference between actual and predicted values. Because it squares the differences, it penalises larger errors more heavily. This makes MSE especially useful when outliers need to be handled with care. Impact on model choice: In cases where larger errors are undesirable, we prioritised models with lower MSE scores, knowing they would provide more reliable predictions, particularly for continuous data.For example, our results for the HOMEVAL column was an MSE of 0.179 suggests that, on average, the value is relatively low meaning the imputation predictions are close to the true values, with small errors.

**MAE**: Represents the average of the absolute differences between actual and predicted values. Unlike MSE, it doesn’t overly penalise larger errors, making it more robust to outliers. Impact on model choice: When MSE indicated that certain models were too sensitive to outliers, we referred to MAE to see if a model could still offer stable performance. Models with consistently low MAE scores were prioritised for their robustness in handling diverse data points. By comparing MSE and MAE, we balanced error sensitivity and the need for stability, selecting models that provided the best trade-off between the two metrics.For example, our results for the HOMEVAL column was a MAE of 0.075 suggests that, on average, the predictions deviate from the actual values by about 0.115 units. This is a low error, signifying good predictive performance.

For the YOJ (Years on Job) column MSE and MAE were used for Error Minimisation. Alongside R², we used MSE and MAE to measure the precision of our predictions. The models that minimised these errors were retained, while those with high MSE or MAE were discarded. Random Forest, in particular, performed well in minimising these errors, making it a suitable choice for YOJ imputation. R² was a crucial indicator for assessing models’ ability to generalise and provide accurate predictions for continuous data, influencing which models we selected for such tasks.

### 

## **1.3. Overall**

Ultimately, the evaluation metrics like MSE, MAE, accuracy, and R² guided our decision-making far more than the models themselves. These metrics provided detailed insights into the strengths and weaknesses of each model, allowing us to choose models that performed best for specific tasks.

Instead of focusing solely on which model to use, we focused on how these evaluation techniques informed us about model performance. This approach ensured that we selected models based on how well they fit the data and how effectively they minimised errors or maximised predictive power, rather than relying purely on the theoretical strengths of each model.

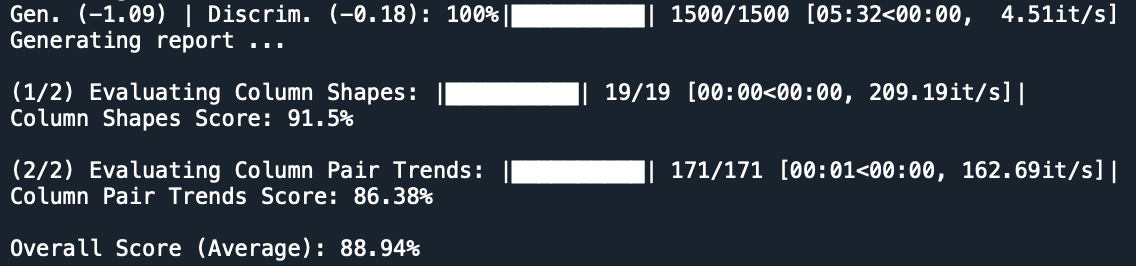
# **2. Performance evaluation results: - Adam/Noor**

*What are the results of any tests run so far and what are the future planned tests for future iterations? (e.g. if this is Deliverable 3, what is planned for Deliverable 4? IF this is Deliverable 4, what is planned before the handover?)*

Refer to the rubric / marking scheme (“All Teams: Deliverable 3 and 4 marking schemes”) for sections to be included in the design and testing documentation.

It is possible to know that the synthesising process is reliable and accurate because both synthesising methods and processes were different and carried out independently. However, their output results were similar, indicating that both need to have been correct to have gotten similar outputs. I.e. it couldn’t have been a coincidence.

An indicator of their similar output is the occurrence of fraud in the fraud column. 466 occurrences of fraud in method 1, and 417 occurrences in method 2. The total rows synthesised by each method is 4000, so in 4000 rows being this close to each other certainly means that both methods of synthesising are valid and reliable.



The screenshot above shows the results of a test which was run to evaluate the quality of the synthetic data generated using the synthetic method 2 which is Conditional Tabular Generative Adversarial Networks (CTGAN). This test is run using a function in the Synthetic data vault which is a library in python called ‘evaluate\_quality’.

The column shapes score indicates how closely the distribution of each column in the synthetic data matches with the dataset that it was trained on. A score of 91.5% indicates that the distribution of the synthetic data was similar to the real data.

The column pair trends score evaluates the relationships between the columns in the synthetic dataset compared to that of the dataset that it was trained on. This score will indicate the pairwise trends. A score of 86.38% indicates that most column pairs show a similar pattern to that of the dataset it was trained on.

The overall score indicates the actual accuracy of the synthetic method. A high score of 88.94% indicates that this method is good and most of the data synthesised closely mimics the actual dataset which is exactly what we want.

### References

Keall, M.D. and Frith, W.J. *OLDER DRIVER CRASH RATES IN RELATION TO TYPE AND QUANTITY OF TRAVEL*. Australasian Transport Research Forum.

<https://australasiantransportresearchforum.org.au/wp-content/uploads/2022/03/2003_Keall_Frith.pdf>

