**Individual Report — Charchil Singh (24311765)**

***Group 17: Harnessing Big Data to Improve Financial Integrity and Operational Efficiency***

**Tasks (5 Marks)**

My primary role in the project focused on **data exploration, anomaly detection modelling, and result interpretation**.  
I performed **Exploratory Data Analysis (EDA)** to understand the dataset’s composition and discovered that it represented a *Course × Unit × Field of Education (FOE)* structure rather than student-level data. This reinterpretation was crucial for accurate aggregation of **EFTSL (Equivalent Full-Time Student Load)** and funding amounts.

I then developed multiple **supervised learning models** — Logistic Regression, Random Forest, and XGBoost — to detect funding irregularities based on relationships between EFTSL, student contributions, and Commonwealth contributions.  
After comparing model performance through evaluation metrics (accuracy, precision, recall, F1, and ROC-AUC), XGBoost achieved the best results (ROC-AUC = 0.972).

Finally, I conducted **unsupervised clustering (K-Means with PCA visualization)** to reveal funding patterns across courses and FOEs.  
These findings were compiled into the **final modelling report**, including visualizations and recommendations for financial audit transparency.

**Methods and Results (15 Marks)**

**Problem Formulation**

The central challenge was to detect funding anomalies and ensure financial integrity within university funding records.  
The dataset combined complex variables such as EFTSL, student payments, and Commonwealth contributions, requiring domain understanding and precise aggregation.

I chose **supervised anomaly detection models** because the goal was to identify funding outliers based on known financial logic. These methods were selected based on knowledge from previous units like *CITS5508 Machine Learning*, which provided grounding in predictive modelling and pattern discovery.

The general methodology followed the **CRISP-DM framework** — involving data understanding, preparation, modelling, evaluation, and interpretation — which ensured a systematic and reproducible workflow.

**Problem Solving**

The modelling workflow was divided into three stages:

1. **Exploratory Data Analysis (EDA):**
   * Conducted descriptive statistics and correlation analysis to validate data integrity.
   * Identified that the dataset aggregates at the *Course × Unit × FOE* level.
   * Verified that recalculated totals matched government figures within 2%, confirming accurate understanding.
2. **Supervised Modelling (Anomaly Detection):**
   * Implemented **Logistic Regression**, **Random Forest**, and **XGBoost** to classify normal vs. anomalous funding records.
   * Tuned hyperparameters using cross-validation and evaluated performance using **accuracy**, **precision**, **recall**, **F1**, and **ROC-AUC** metrics.
   * **XGBoost** achieved the best performance (ROC-AUC = 0.972), demonstrating excellent capability to identify funding mismatches.
3. **Unsupervised Modelling (Clustering):**
   * Applied **K-Means clustering** to group records based on funding characteristics.
   * Determined **optimal k = 8** using silhouette analysis.
   * Visualized clusters using **PCA**, revealing distinct funding profiles — from high Commonwealth support to high student-fee structures.
4. **Visualization and Interpretation:**
   * Developed comparative plots to visualize anomalies per funding cluster and FOE categories.
   * Used boxplots and PCA scatter plots to interpret how funding patterns differ across clusters.

Through this approach, I achieved both **quantitative accuracy** (in anomaly detection) and **qualitative insight** (in cluster-based funding analysis).

**Ethical, Responsible AI and Broader Social Impact**

Given that the models operate on financial data linked to public funding, **ethical and responsible AI practices** were a core consideration.  
Incorrect anomaly predictions could result in legitimate funding being misclassified as irregular, leading to unnecessary audits or administrative actions.

To mitigate such risks, model validation was performed thoroughly using multiple algorithms and metrics to ensure fairness and reliability.  
Interpretability was enhanced through visualization of cluster trends rather than opaque model outputs, ensuring transparency for decision-makers.

This project aligns with the **WA Government’s AI Assurance Framework**, emphasizing **accountability, fairness, and transparency**.  
The approach demonstrates how data science can strengthen **financial governance** without replacing human oversight.  
Our findings support evidence-based decision-making while respecting the principle that **AI should assist, not replace, human judgment** in sensitive financial operations.

**Personal Reflection (5 Marks)**

This project was an important learning experience that deepened both my technical and analytical abilities.  
One of the major challenges was **understanding the true structure of the dataset** — initially assumed to be student-level — which required revisiting our assumptions and recalculating funding metrics at the aggregated level.  
This process taught me how critical **domain knowledge and data validation** are before applying any model.

Another challenge was managing **model complexity and interpretability**.  
While XGBoost produced excellent accuracy, it required careful feature scaling and interpretation.  
On the other hand, Logistic Regression provided a simpler baseline but failed to capture non-linear relationships.  
This reinforced the importance of **comparing models not just by metrics, but by contextual relevance**.

Time constraints also posed a challenge during visualization and fine-tuning, but teamwork and iterative experimentation helped overcome this.  
In future work, I would like to automate the anomaly detection workflow, expand the dataset to multiple years for **trend analysis**, and deploy a dashboard to communicate results interactively.

Overall, this project taught me how to connect **data science techniques to real-world decision-making**, balancing model accuracy, interpretability, and ethical responsibility — the core of responsible data practice.