NOVA
NOVA SCHOOL OF BUSINESS & ECONOMICS

Responsible AI – Part 3

Milton de Sousa

Let’s build an AI solution!

Goal: To create an AI-powered system within Google AI Studio or Chat GPT that assists in the initial screening of CEMS master's program applications, specifically analysing CVs to identify potentially strong candidates (strong,weak)… **responsibly!**

Note: Traditional Predictive ML would be better suited for this task!

Step 1 - Defining the Scope and Purpose (Planning & Requirements Gathering)

- Define selection criteria
- Identify Data Sources for training and validation
- Define success metrics (precision, or reduction in review time)
- Outline Workflow with the AI solution integrated into the existing application review process
- Determine Human Oversight

Step 1 – Responsible AI checklist

- Purpose and Values Alignment (with CEMS)?
- Stakeholder Identification and Involvement?
- Potential Impacts Assessment (benefits and risks)?
- Fairness Considerations (outcome, features, representation, access)?
- Transparency about the use of AI in the application process to applicants?
- Human oversight and correction of errors made by the AI?

Step 2 - Data Preparation and Collection (Ethical Sourcing & Annotation)

- Data Collection of a representative dataset of CVs
- Obtain proper consent if using data from past applicants, and ensure compliance with privacy regulations (e.g., GDPR, CCPA). Anonymize or redact personally identifiable information.
- Synthetic data: consider using synthetic CVs which may mitigate privacy concerns and allow better control of the training data (**USE THIS FOR THE PURPOSE OF THE ASSIGNMENT!**)
- Data Cleaning: Remove irrelevant characters, standardize format, and handle missing data.
- Data Splitting: Divide your dataset into training (your own), validation & test sets (from other groups)
- Data Annotation (Labelling) to train your AI system. Manually label your training data to indicate which applications are "strong“ or "weak" based on your defined criteria.

Step 2 – Responsible AI checklist

- Data Provenance: Is the data representative of the target population (future applicants)?
- Data Privacy: Have you anonymized or redacted to protect applicant privacy?
- Data Bias Detection: Have you analyzed the data for potential biases?
- Data Quality: Is the data accurate, complete, and consistent?
- Annotation Guidelines: Are the annotation guidelines clear, consistent, and unbiased?
- Inter-Annotator Agreement: Measure agreement between multiple annotators to ensure quality.
- Data Versioning: Are you tracking changes to the dataset?

Step 3 - Model Development (Design and Training)

- Select a Model Type (Google AI Studio and Chat GPT offers various models).
- Prompt Design: Craft effective prompts to guide the AI model.
- Model Training: Use your labelled data to train the model
- Monitor the training process and adjust parameters as needed.
- Experiment with different prompts, parameters, and training data to improve the model's performance.

Step 3 – Responsible AI checklist

- Model Selection Justification?
- Bias Mitigation?
- Explainability?
- How are you tuning the model's parameters to optimize for both accuracy and fairness?
- Model Documentation: Document the model's architecture, training process, and performance metrics.

Step 4 - Model Evaluation and Validation (Testing and Refinement)

- Evaluate the performance of your trained model using the validation & test datasets (from other groups).
- Performance Metrics: Measure accuracy, precision, recall, and F1-score on the validation and test sets.
- Fairness Metrics: Evaluate the model's fairness across different subgroups (calculate F1)
- Analyze the types of errors the model is making. Are there specific sub-groups with poor performance?
- Human Review: Have human reviewers evaluate a sample of the model's predictions.
- Iterative Refinement: Based on the evaluation results, refine the model by adjusting the training data, prompts, or model parameters.

Step 4 – Responsible AI checklist

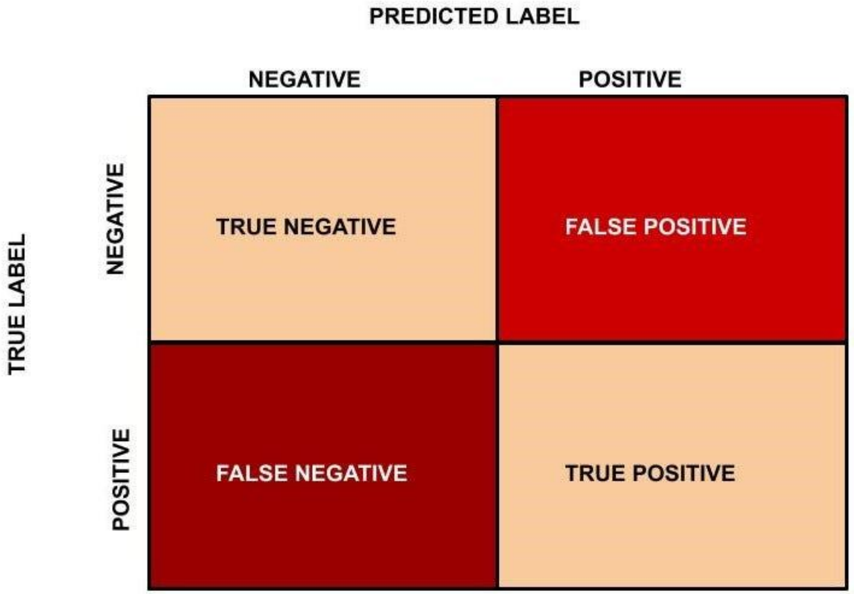
- Comprehensive Evaluation of the model on diverse datasets that are representative of the population?
- Precision? Recall? F1 score?
- Fairness Assessment of the model across different subgroups? F1 score for different subgroups?
- Bias Mitigation Refinement steps to mitigate them and re-evaluate the model?
- Robustness Testing: performance under different conditions or with slightly modified inputs?
- Adversarial Testing: Can the model be easily fooled by adversarial attacks?
- Error Documentation of the types of errors the model is making and the potential consequences?

Step 4 – F1 score

- The **F1 Score** is a metric in machine learning that provides a balanced measure of a model's precision and recall.
- **Precision**: The accuracy of positive predictions. The number of **true** positive predictions is divided by the **total number** of positive predictions (true positives + false positives).
- **Recall** (Sensitivity or True Positive Rate): Recall represents how well a model can identify actual positive cases. Defined as the number of **true** positive predictions divided by the total number of **actual** positive instances (true positives + false negatives).
- We want a model that performs well on both metrics. The F1 combines precision and recall into a single harmonious mean metric to provide a more comprehensive evaluation of model performance.
- **F1 = 2 x (precision x recall) / (precision + recall)**

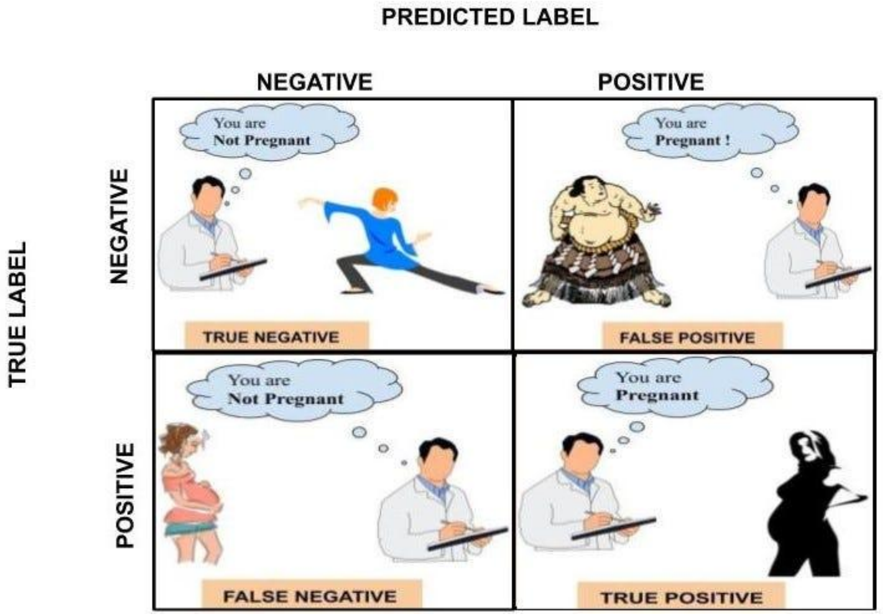
Step 4 – F1 score

- Confusion matrix
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1 = 2 x (P x R) / (P + R)



PREDICTED LABEL

TRUE LABEL		NEGATIVE	POSITIVE
	NEGATIVE	TRUE NEGATIVE	FALSE POSITIVE
	POSITIVE	FALSE NEGATIVE	TRUE POSITIVE



Step 5 - Deployment and Monitoring (Implementation and Ongoing Oversight)

- Integration: Integrate the AI solution into your application management system.
- Human-in-the-Loop: Implement a human-in-the-loop system where human reviewers can review and override the AI's predictions. Flag uncertain or borderline cases for human review.
- Real-time Monitoring: Monitor the AI's performance in real-time. Track metrics like accuracy, fairness, and user feedback.
- Alerting: Set up alerts to notify you of any performance degradation or fairness issues.
- Feedback Mechanism: Provide a mechanism for applicants and reviewers to provide feedback on the AI system.
- Regular Audits: Conduct regular audits of the AI system to ensure it is performing as expected and that it is not creating unintended biases

Step 5 – Responsible AI checklist

- Transparency Communication: Are you clearly communicating the use of AI?
- Explanation Provision: Are you providing applicants with explanations of the AI's decisions (where appropriate)?
- Feedback Collection: Are you collecting feedback from applicants and reviewers on the AI system?
- Performance Monitoring: Are you continuously monitoring the AI's performance and fairness in production?
- Bias Drift Detection: Are you monitoring for bias drift over time?
- Incident Response Plan: Do you have a plan for responding to incidents involving the AI system (e.g., errors, biases, privacy violations)?