Final Report: **Language Detection (Kyrgyz vs Other Languages)**

### **1. Project Objective**

**The goal of this project was to create a text classification model capable of identifying whether a given phrase is written in Kyrgyz or belongs to any other language.  
 We performed fine-tuning on the Whisper Small Kyrgyz model from Hugging Face, using a custom multilingual dataset containing simple phrases in Kyrgyz, Russian, and English.**

### **2. Dataset Description**

* **Total samples: 40,000**
* **Languages:**
  + **Kyrgyz 20000**
  + **Russian 10000**
  + **English 10000**

**Structure:**

| **Text** | **Label** |
| --- | --- |
| **"Салам"** | **1** |
| **"Hello, how are you?"** | **0** |
| **"Как дела?"** | **0** |

**Label Explanation:**

* **1 → Kyrgyz**
* **0 → Other languages**

**We have trained the model for 20 minutes with 1 epoch because when we tried to make more epochs it gave the retrained model.**

### **4. Training Results**

**After fine-tuning, the model achieved perfect classification results on validation data:**

| **Metric** | **Value** |
| --- | --- |
| **Training Loss** | **0.0032** |
| **Validation Loss** | **0.0000** |
| **Accuracy** | **1.0000** |
| **F1 Score** | **1.0000** |

**Note: Since the dataset contained very simple and distinct phrases, the model quickly achieved 100% accuracy on training and validation sets.**

### **6. Conclusion**

**This project successfully demonstrated how a multilingual dataset and fine-tuning of an existing Whisper Kyrgyz model can be used to build an efficient language classifier for distinguishing Kyrgyz text from other languages.**

**The high accuracy is due to the simplicity of the phrases and the clear separation between language labels. In future work, a larger, more diverse dataset with complex phrases could be used for further improvement and robustness.**