Diabetes Prediction Report

Comparative Report: XGBoost vs. LLM-based In-Context Learning for Diabetes Prediction

1. Introduction

This report presents a comprehensive comparison between a traditional supervised learning model, XGBoost, and a large language model (LLM)-based in-context learning (ICL) approach for predicting diabetes progression scores using the Diabetes dataset. The goal is to assess the predictive capabilities of both techniques and analyze how prompt design and hyperparameter tuning impact LLM performance.

2. Dataset Description

The study uses the classic sklearn.datasets.load\_diabetes() dataset. This dataset contains 442 samples with 10 normalized features such as age, BMI, blood pressure, and others, along with a target numerical score representing disease progression one year after baseline.

3. Experiment Overview

* - Traditional model: XGBoost (n\_estimators=100)
* - LLM: Qwen2-7B-Instruct via API, few-shot in-context learning
* - Prompt types: standard vs. structured
* - Temperatures tested: 0.0, 0.3, 0.5, 0.7
* - Shots tested: 5 and 10
* - Metrics: MSE, MAE, R²

4. Prompt Design

Standard Prompt: "Input: [0.0381, 0.0507, ...], Output: 151.00"

Structured Prompt: "Input: [age: 0.0381, bmi: 0.0507, ...], Output: 151.00"

5. Experimental Results

XGBoost → MSE: 3351, MAE: 46.39, R²: 0.37

Best LLM (Structured-T0-S5) → MSE: 4411, MAE: 56.07, R²: 0.17

Worst LLM (Standard-T0.7-S10) → MSE: 8540, MAE: 70.78, R²: -0.61

6. Key Insights

* - Structured prompts improve accuracy significantly.
* - Higher temperature worsens performance in numeric tasks.
* - More shots do not necessarily improve ICL results.
* - LLMs underperform supervised models in numeric regression.

7. Comparative Analysis

| Feature | XGBoost | LLM (ICL) |

|----------------|---------|----------------|

| Accuracy | High | Low-Moderate |

| Speed | Fast | Slow |

| Training need | Yes | No |

| Stability | High | Prompt-sensitive |

8. Conclusion

While LLMs can mimic regression through prompt design, traditional models like XGBoost remain more reliable and accurate in numeric prediction. Structured prompts and low temperature help improve LLM outputs.

## Future Work

* - Evaluate GPT-4, Claude for regression
* - Automate prompt engineering
* - Combine LLM with supervised models