Byron Washington

CDS 302

April 7, 2025

Revised Proposal + ERD

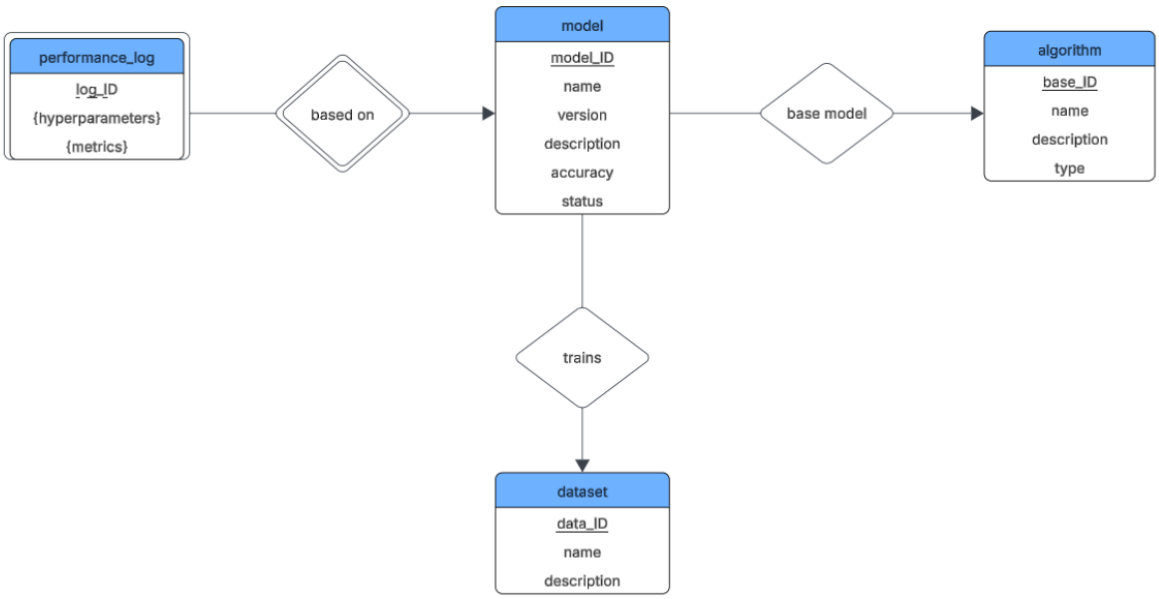

Figure 1. ER Diagram

My database would be a repository for machine learning models. This would allow users to store, track, and compare different models. The database would include models, their base model (algorithm), performance logs which include output, and the datasets that they were trained on. The intended users would be professional or amateur data scientists/analysts.

A function that is expected by users is registering different models. Users would register a new model and include its algorithm, name, version (if it's using the same algorithm and dataset), a description, its accuracy (in %), and its status (active or inactive). In addition to those base attributes, the users would have to enter the algorithm used and dataset it was trained on. These attributes hold the relationship between the algorithm table and dataset table respectively.

In the "base model" relationship in Figure 1., each algorithm can be a base model for multiple models, but each model only has one algorithm as a base model. For the "trains" relationship in Figure 1., each dataset can be used to train multiple models, but each model can only be trained on one dataset. These relationships ensure that every model must have exactly one algorithm and dataset attached to it.

For the “based on” relationship in Figure 1., each model can have many performance logs, but each performance log is based on one model only. Additionally, *performance_log* by itself is dependent on *model* making its primary key both log_id and model_id, The *performance_log* table would allow users to track performance of a model with certain metrics and hyperparameters.

The dataset table allows users to enter the datasets that are being used by models. An expected function could be if a dataset has an issue, then models that were trained on it could be set to inactive, letting users know not to use or trust their outputs.

The algorithm table allows users to enter the algorithms being used for each model. An expected function would be a user wanting to see what kind models are used for classification for instance. This would allow the user to go through those classification models and decide which one(s) would work best for their specific problem.