🧬 Genome-Dx

AI-powered Disease Prediction from Gene Mutations

Genome-Dx is a genomics-driven machine learning project that predicts the **pathogenicity** (benign / pathogenic / uncertain) of gene mutations
and provides **probable disease associations** with **LLM-generated explanations** for better interpretability.

🚀 Project Overview

Component	Description
Objective	Predict whether a gene mutation is *pathogenic* or *benign* and infer the most probable disease association.
ML Model	Supervised classification model trained on encoded genomic features: `Chromosome_Encoded` , `Gene_Symbol_Encoded` , `IS_SNP` , `IS_INDEL` .
Vector DB (Qdrant)	Integrated **Qdrant** vector database to semantically search disease descriptions — currently limited to *pathogenic* clinical significance embeddings.
LLM Integration	Uses **Gemini LLM** through **DSPy** to generate natural-language reasoning and human-readable explanation text.
Data Source	Processed ClinVar-derived dataset (gene mutations, clinical significance, and disease annotations).
Interface	Streamlit app for interactive prediction, explanation, and variant lookup.

⚙️ Workflow

Raw ClinVar Data
↓
Preprocessing → Encoding (chromosome, gene, variant type)
↓
ML Model Training (Pathogenic vs Benign)
↓
Prediction
↓
Qdrant Semantic Lookup (Pathogenic Diseases Only)
↓
LLM (Gemini + DSPy) → Explanation
↓
Streamlit UI Display

🧩 Features

- 🧠 **ML-based Pathogenicity Prediction** — trained on numeric genomic features.
- 🔍 **Probable Disease Lookup (via Qdrant)** — semantic disease retrieval for *pathogenic* variants only.
- 💬 **LLM-Generated Explanations** — contextual text via Gemini + DSPy for interpretability.
- 🧾 **Hybrid Results** — combines deterministic lookup with AI reasoning and vector search.
- 🌐 **Streamlit Interface** — user-friendly input/output layout for experimentation.

📊 Tools & Technologies

Category	Stack
Language	Python 3.10
Frameworks	scikit-learn, pandas, numpy
LLM / AI	Google Gemini LLM + DSPy

Category	Stack
Vector DB	Qdrant
Visualization / UI	Streamlit
Data Handling	joblib, csv, numpy
Version Control	Git / GitHub

🧠 Future Extensions

- Expand Qdrant embeddings to include *benign* and *uncertain* variants.
- Add **multi-task model** to jointly predict disease + pathogenicity.
- Include **phenotype and clinical notes** for richer context.
- Expand dataset with **OMIM / HPO mappings** for rare diseases.

🧬 Example Input

Input: Gene = BRCA1, IS_INDEL = 1
Model Prediction: Pathogenic
Probable Disease (Qdrant search): Hereditary breast and ovarian cancer
Explanation (Gemini via dspy):
Based on ClinVar and model inference, this mutation in *BRCA1* is classified as pathogenic and is most frequently associated with hereditary breast and ovarian cancer (HBOC).

Models & Embeddings

- `pritamdeka/S-BioBert-snli-multinli-stsb` – biomedical sentence embeddings (CC BY-NC 3.0)
- `sentence-transformers` – base embedding framework (Apache 2.0)
- `Qdrant` – vector database for semantic search (Apache 2.0)

🧾 License

Genome-Dx — *bridging genomics and AI to explain disease risk from gene mutations.*