README: Molecular Property Prediction Model

This document outlines the process, assumptions, and detailed explanations of the steps involved in developing a molecular property prediction model using machine learning techniques. The goal is to predict molecular properties, such as docking scores, based on molecular descriptors and fingerprints derived from the SMILES representation of molecules.

Dataset Preparation

Data Loading and Cleaning

- Excel Data Import: The initial dataset is loaded from an Excel file into a pandas DataFrame. This step assumes that the Excel file contains relevant molecular data, including SMILES strings and possibly other properties like docking scores.
- Duplicate Removal: Duplicate entries are removed to ensure the uniqueness of each molecule in the dataset.
- Handling Missing Values: Columns with more than 80% missing values are dropped. This threshold is chosen to retain columns with sufficient data while removing those that are largely incomplete.

Feature Engineering

- SMILES Conversion: The SMILES (Simplified Molecular Input Line Entry System) strings are used to generate two types of features: molecular descriptors and fingerprints.
    - Molecular Descriptors: Calculated using RDKit, including molecular weight, LogP, number of hydrogen donors, and number of hydrogen acceptors.
    - Fingerprints: ECFP (Extended Connectivity Fingerprints), specifically Morgan fingerprints with a radius of 2 and 2048 bits, are computed for each molecule.
- Data Cleaning for SMILES: Entries with invalid or missing SMILES strings are filtered out to ensure that only valid molecular data is processed.

Data Preprocessing

- Numeric Feature Selection and Imputation: Numeric columns, including the generated descriptors and fingerprints, are selected. Missing values in these columns are imputed using the median value, a strategy chosen to handle missing data without introducing bias.
- Normalization: Features are normalized to a [0, 1] scale using MinMaxScaler to ensure that all features contribute equally to the model training. If present, the target variable (e.g., docking score) is also normalized.

Model Development and Evaluation

- Model Architecture: A deep learning model is constructed using TensorFlow and Keras, with layers designed to handle the input feature set's complexity.
    - Input Layer: Matches the number of input features.
    - Hidden Layers: Two dense layers with ReLU activation to introduce non-linearity and handle complex relationships in the data.
    - Output Layer: A single neuron without an activation function for regression tasks.
- Training and Validation Split: The dataset is split into training, validation, and test sets with proportions of 60%, 20%, and 20%, respectively, to evaluate model performance and generalize to unseen data.
- Model Training: Includes early stopping, model checkpoints, and learning rate scheduling to optimize performance and prevent overfitting.
- Evaluation: The model's performance is assessed using the validation set and finally on a separate test set to ensure it generalizes well. Metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R^2 score.

Assumptions and Considerations

- Data Integrity: It's assumed that the SMILES strings and any other molecular properties provided in the dataset accurately represent the molecules of interest.
- Feature Representation: The choice of molecular descriptors and fingerprints assumes these features are sufficient to capture the relevant chemical space for the prediction tasks.
- Model Capacity: The model architecture assumes that a deep neural network is appropriate for capturing the complex relationships between molecular features and the properties being predicted.

- Normalization and Imputation: The steps for data normalization and imputation assume these techniques are suitable for the dataset and the prediction task, aiming to improve model training and prediction accuracy.

Conclusion

This document provides a comprehensive guide to developing a molecular property prediction model. Each step, from data preparation through model evaluation, is designed to build a robust and accurate model for predicting molecular properties based on chemical structure represented by SMILES strings. The process involves careful data handling, feature engineering, model design, and evaluation, with considerations for data quality, feature relevance, and model optimization strategies.