SURP 23b2444 Harsh kumar singh

PREDICTING MOLECULAR PROEPRTIES

DATA PROCESSING

In this assignment we start with data processing where we read csv file for SMILE string of molecules and storred it in dataframe.Then we converted the smiles into morgan fingerprints vector using Rdkit (i used 512 bit because most molecules were smaller and we can have compact represenations).

EDA(EXPLORATORY DATA ANALYSIS)

We visulaized the data (shape,density of 4.11%) and examined fingerprint bit frequecnies .basically it shows **how common each fingerprint feature is across your molecules**. **Tanimoto similarity** was also used in code to measure how similar two molecules are based on their fingerprint. I didnt perform exploratory data analysis.

MODEL DEVELOPMENT

Dataset is randomly split into test:train of 25:75 where x is morganfingerprints and y is log of measured solubility in moles per litre. Then i used fingerprints smiles and performed simpel random forest regression but on testing we get R2 of 0.67(which must be near 1.0 for good model)so this model didnt perform very well .

We then use descriptors on smiles of dataset to get more information like molecular wright, no. of h bonds , rotable bonds etc. These descroiptors+morgan footprints help in making model which is better than before.

SVM is supervised learning algorithm which i aslo used as baseline but R2 was lesser and RSME was high.(Uses **kernels** to handle nonlinear relationships by mapping input features into higher dimensions.)

I then moved to ensemble models like random forest,xgboost and lightgbm….

—> RANDOM FOREST regressor

i used random forest model which is based on principal that many decision tree working independently and paralley generating geberalised output(bagging based) .R2 was higher than before now..

—>XGBOOST regressor

Gradient boosting algorithm where decision trees work sequentially and it divided at each level This model performed well with R2 of 0.88 which is nice.

—->LIGHTGBM regressor

As the name suggest the light gradient boosting algorithm uses sequencial boosting of decision trees but herre tree divided at a leaf node and not every level. This makes it work faster than other models (R2 was also around 0.87)

We used morgan fingerprints and descriptors for them which helped in increasing accuracy overall for a model.

```
from rdkit.Chem import Draw
```

We used this package to draw 2D structure of molecules as well for top 10 molecules. Helped in visalization.

```
from torch_geometric.utils import to_networkx
```

This package was used to visialize the molecule as node and edge diagram where each atom is node and edge are bonds.

Graph neural network GNN

GNN outperforms them by learning directly from molecular graph structure, capturing spatial and relational information between atoms.Input: Molecular graphs derived from SMILES.