Feature Clustering for the SN_DREAMS Diabetic Retinopathy (DR) dataset

Aadit Deshpande

Nature of Data and Dataset

- 13 Risk Factors for Diabetic Retinopathy (DR)
- 1555 Patients
- Categorical and Continuous Columns
- 14^{th} DR Column Indicates Diabetic Retinopathy
- Expert-Labelled Primary and Secondary Risk Factors
- Data points (Patients) with DR = 1 are sparse => Imbalanced Data
- Pre-processing includes SMOTE-ENN, Standardization

Motivation for Feature Clustering

- Identifying Features as primary and secondary indicators of DR
- Dimensionality Reduction + Ranking Features
- Comparing ML vs Expert-labelled Feature Clusters

Primary Features	Secondary Features
Hypertension_cat	Gen_cat
Insulin_Treatment_cat	IHD_cat
BP_sys	Fam_His_Diab_cat
HbA1c_grade_cat	Physical_AS_cat
DurYrs	Ann_Family_In
FBS_Grp_cat	AgeGrp_cat
	BMI_cat

Clustering (Attempt 1)

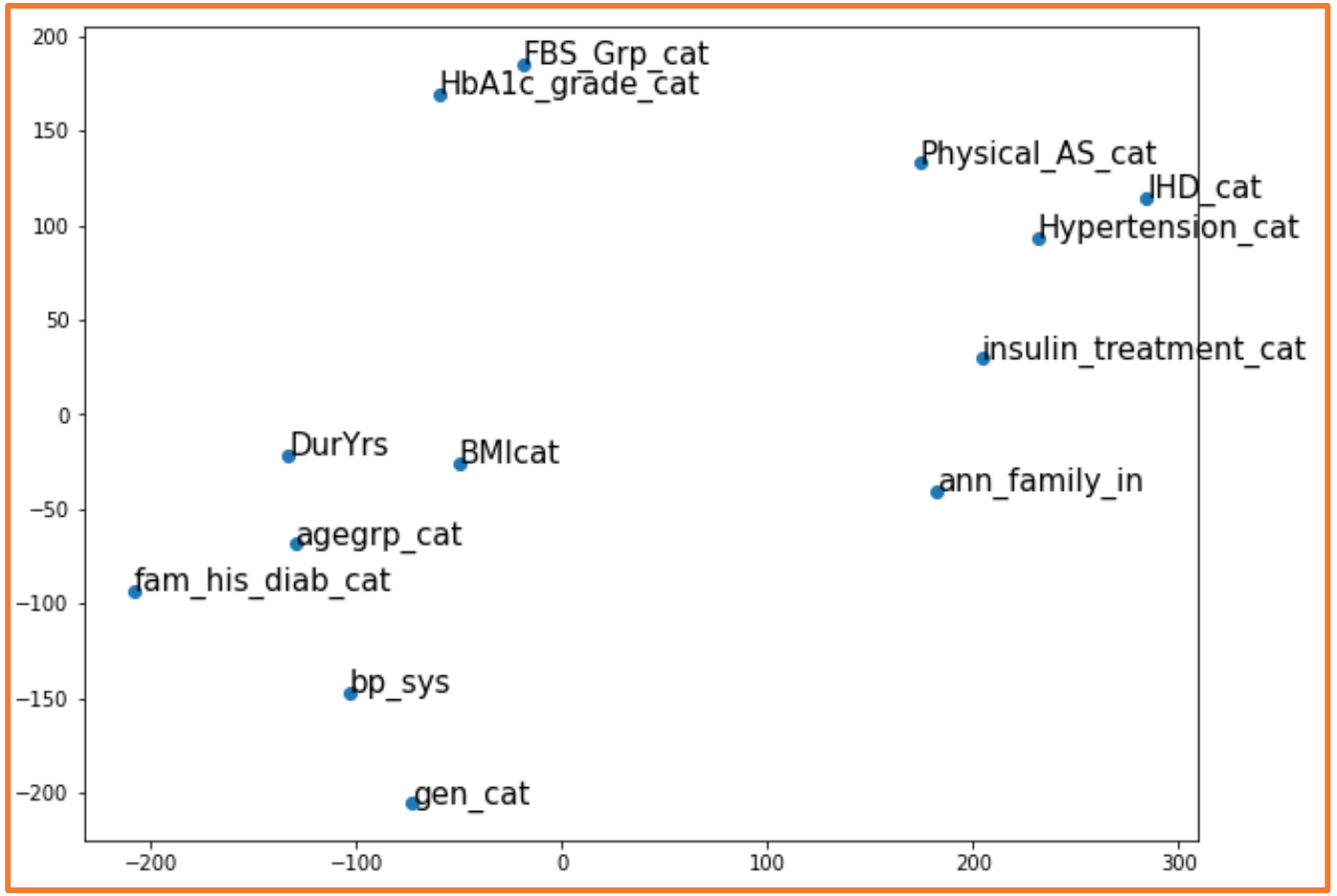
- Transpose Matrix (Cluster by feature, not patient)
- Standard Scaling
- K-Means Clustering (k=2, k-means++)
- Agglomerative Clustering (Hierarchical)
- [Neither matched the real Clusters]
- T-SNE Visualization

```
Cluster  0  Attributes are:
  gen_cat
  bp_sys
  HbA1c_grade_cat
  agegrp_cat
  DurYrs
  FBS_Grp_cat

Cluster  1  Attributes are:
  Hypertension_cat
  IHD_cat
  insulin_treatment_cat
  fam_his_diab_cat
  Physical_AS_cat
  ann_family_in
  BMIcat
```

K-Means Clusters

```
Cluster  0  Attributes are:
  gen_cat
  fam_his_diab_cat
  bp_sys
  HbA1c_grade_cat
  agegrp_cat
  BMIcat
  DurYrs
  FBS_Grp_cat

Cluster  1  Attributes are:
  Hypertension_cat
  IHD_cat
  insulin_treatment_cat
  Physical_AS_cat
  ann_family_in
```

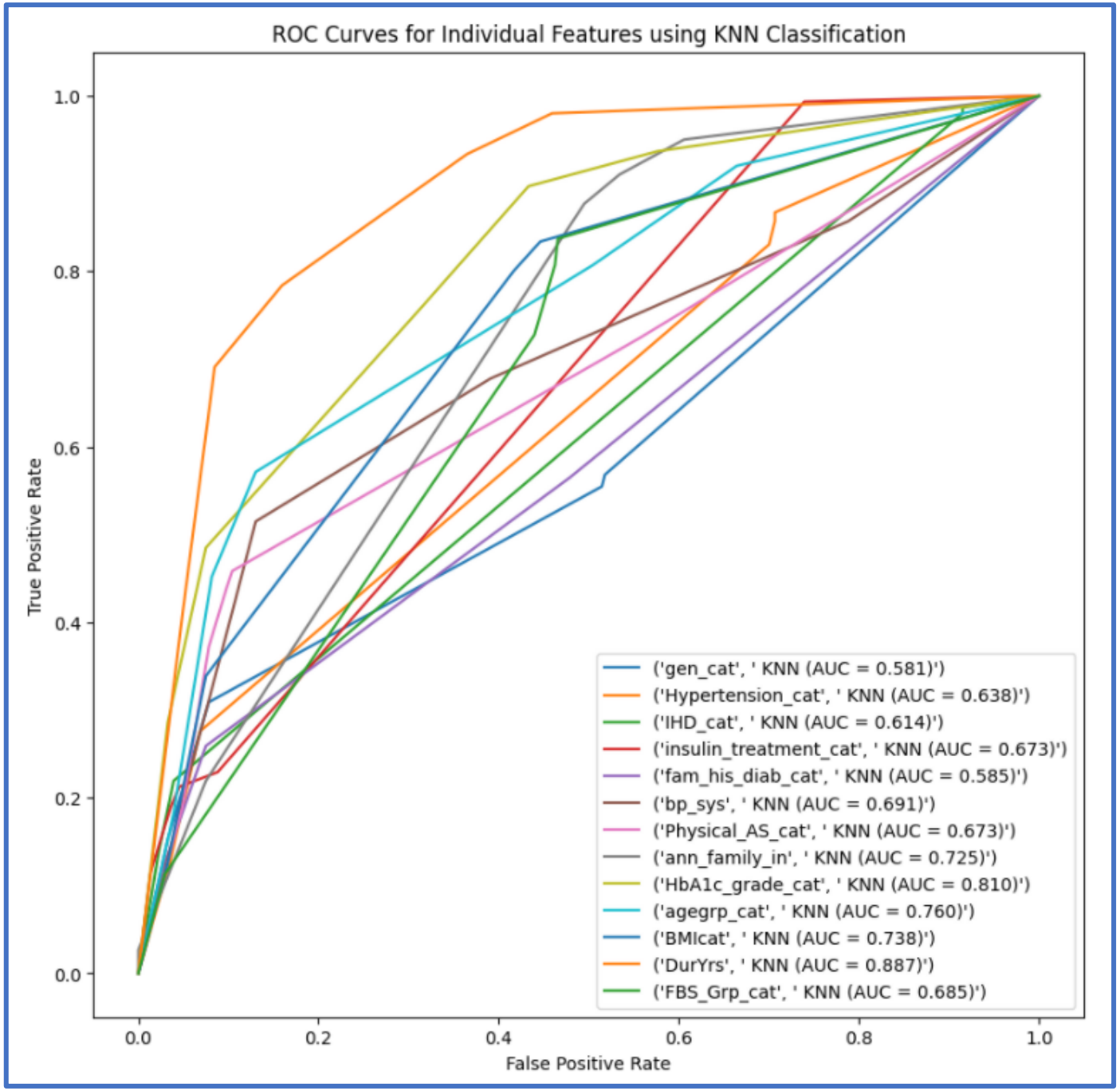
Agglomerative Clusters



TSNE Visualization

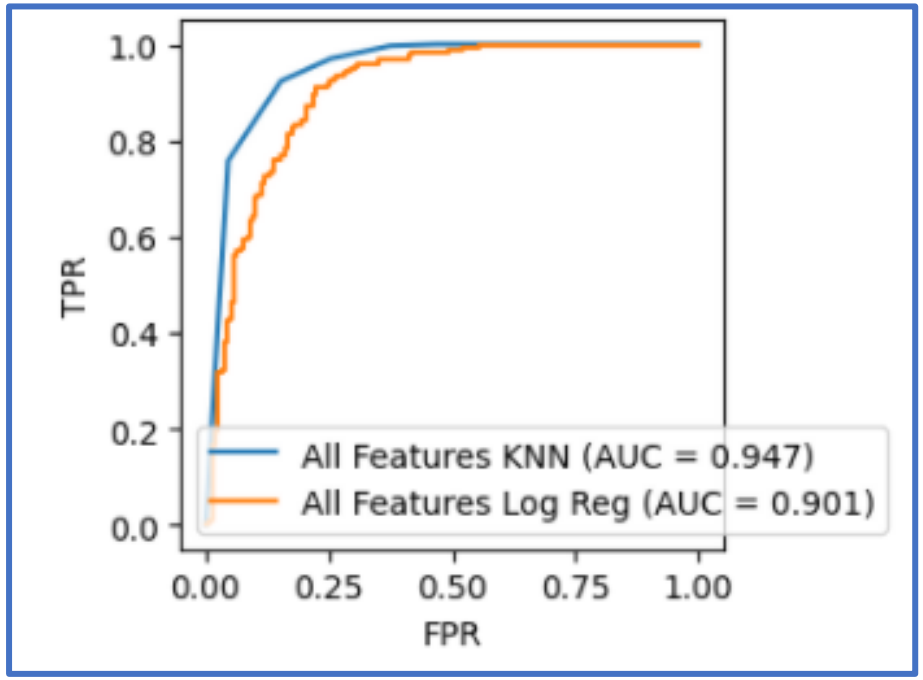
Classification (Attempt 2)

- SMOTE-ENN to correct *DR=1 Sparseness (*Data Imbalance)
- 70:30 Train Test Split
- kNN Classification (k=5, Minkowski Distance (p=2))
- Classify using the 13 risk factors individually (DR is the dependent Variable)
- Metrics: ROC-AUC Score and Classification Accuracy
- Sort into 2 halves based on AUC Scores
- Clusters are better but still don’t match exactly



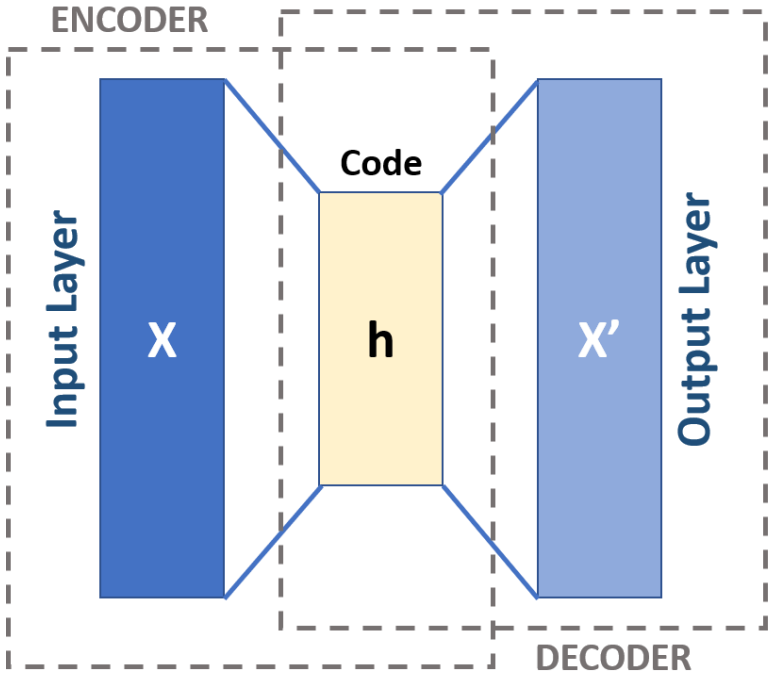
kNN with Individual Features (Ranked by AUC Scores)

kNN with All Features (AUC Score Upper Bound)



Autoencoder (Attempt 3)

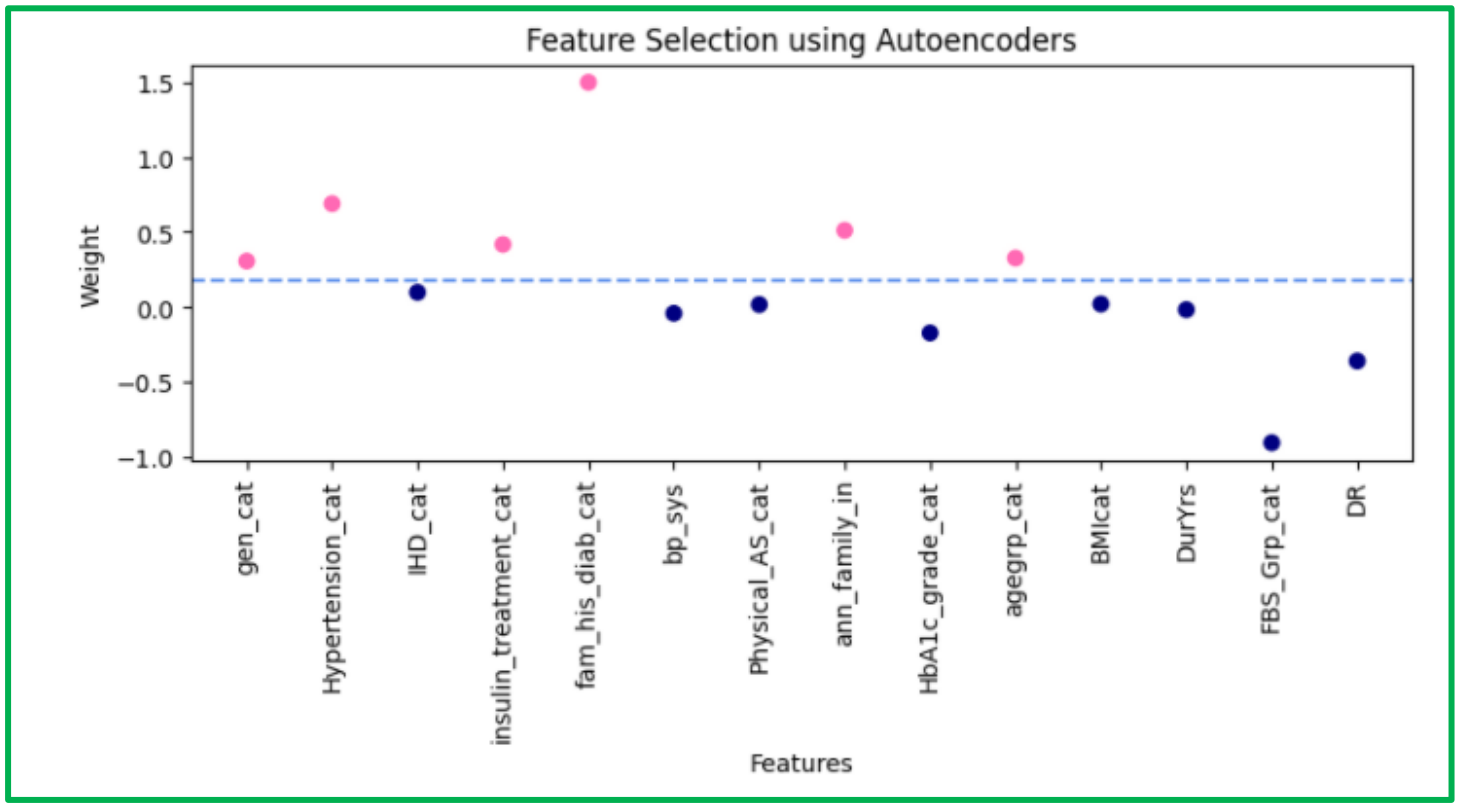
- 70:30 Train Test Split
- Standard Scaler + Normalization
- 2-layer Autoencoder
  - Densely Connected ***Code Layer*** (7 neurons)
  - Dense ***Output Layer*** (13 neurons)
  - 202 Parameters
- Training
  - Mean Absolute Error
  - Adam Optimizer
  - 15 epochs
- Use Code Layer Weights used to score features
- Much Better Clusters



```
Model: "autoencoder"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 Code_Layer (Sequential)     (None, 7)                 105

 Output_Layer (Sequential)   (None, 14)                112

=================================================================
Total params: 217
Trainable params: 217
Non-trainable params: 0
_________________________________________________________________
```

Autoencoder Model Structure

Feature Selection
(Code Layer Weights threshold)



References

- https://www.tensorflow.org/tutorials/generative/autoencoder
- https://scikit-learn.org/stable/modules/classes.html#module-sklearn.cluster
- https://towardsdatascience.com/dimensionality-reduction-pca-versus-autoencoders-338fcaf3297d
- Khalid, S., Prieto-Alhambra, D. Machine Learning for Feature Selection and Cluster Analysis in Drug Utilisation Research. *Curr Epidemiol Rep* **6,** 364–372 (2019).