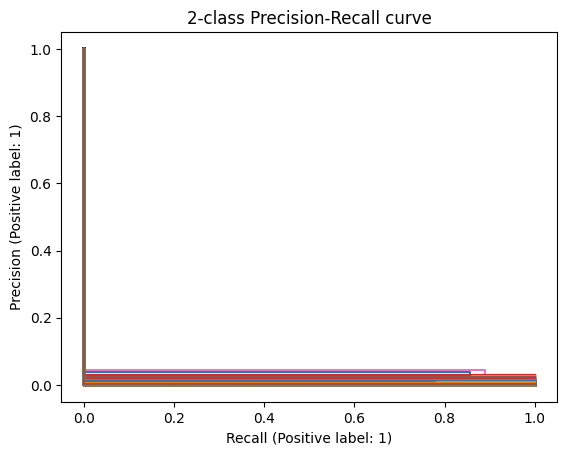
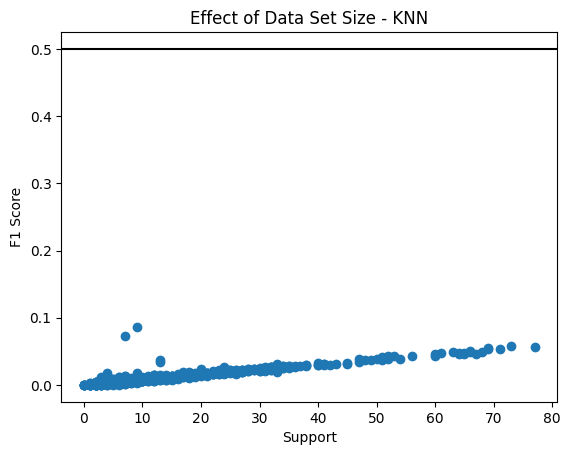
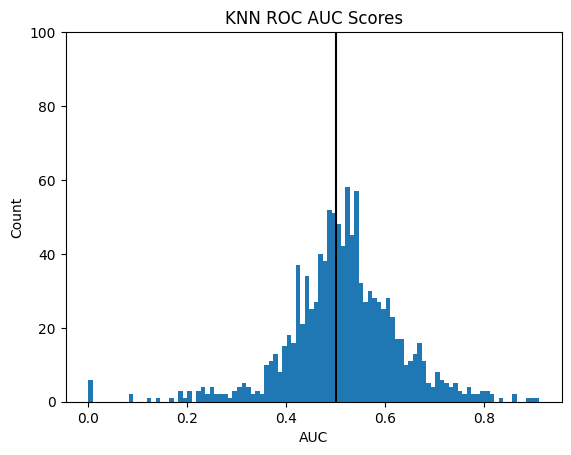
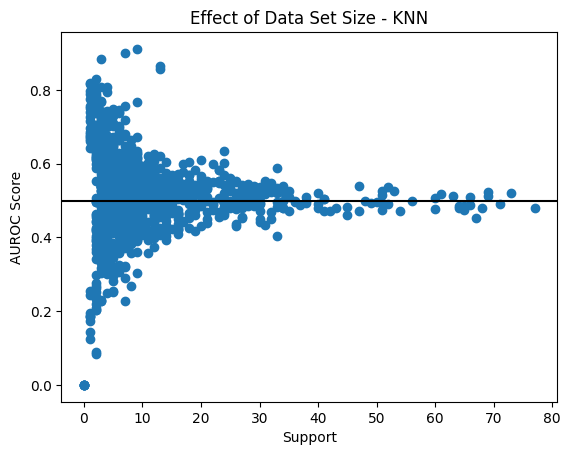
KNN

The KNN classifier from sci-kit learn was used. The parameters for the model were through hyper tuning which found the best results were obtained using a k equal to 2, the euclidean distance metric, and a uniform weighting of neighbors. The average AAROC obtained with these parameters 50.6% which means the model was on average randomly guessing. To try to improve this metric a combination of oversampling and undersampling was performed. The synthetic minority oversampling technique from the Imbalanced learn package was used to create an equal number of positive and negative labels for each GO term after undersampling of the negative class was performed. The mean AUROC saw a slight improvement to 51.5%.

The precision recall curves show that the model was unable to effectively predict GO terms and any improvements from optimizations were minimal.



The GO terms with the highest F1 score by ascending rank are

GO:0001819 - "positive regulation of cytokine production":

GO:0006811 - "ion transport":

GO:0006952 - "defense response":

GO:0045184 - "establishment of protein localization":

GO:0006511 - "ubiquitin-dependent protein catabolic process":

Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. (2009) QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics. 2009; 25(22):3045-6.