ACTIVE LEARNING

Prepared for:

Dr. Navneet Goyal

Instructor In-charge, Machine Learning (BITS F464)

Department of Computer Science & Information Systems, BITS Pilani

Prepared by:

Namit Shrivastava	**2020A2PS1767P**
Aryan Raj	**2020A2PS1777P**
Yashvardhan Prasad	**2020A2PS1775P**

ABSTRACT

Large volumes of data are necessary for the successful training of most supervised machine learning models. And while that may sound naive, the truth is that most businesses have a hard time giving their data scientists access to the data they need, especially tagged data. The latter is essential for training any supervised model and can quickly become a data team's biggest bottleneck. Data scientists are typically tasked with training high-quality models using large, unlabeled data sets. Whenever there is a big volume of data, manual labeling becomes impractical, making it difficult for data teams to train good supervised models. Labeling data with the highest potential impact on training a supervised model first is an example of active learning. When there is too much data to classify manually and a priority must be set to categorize the data intelligently, active learning can be utilized.

ABOUT THE DATASET

Identifying different types of forest cover using simply cartographic data (no remotely sensed data). By utilising information from the US Forest Service's Region 2 Resource Information System (RIS), we were able to identify the specific type of forest cover present at a given observation (a 30 by 30 metre cell). The independent variables were constructed using information that was initially collected by the US Geological Survey (USGS) and the United States Forest Service (USFS). Quantitative independent variables are represented as binary (0 or 1) columns in the raw (unscaled) data (wilderness areas and soil types).

Northern Colorado's Roosevelt National Forest is home to four separate wilderness regions that make up the focus of this research. These regions are representative of undisturbed forests, where the forms of forest cover seen today are the product of natural processes rather than human forest management.

Here's some context for those four unspoiled spots: Of the four protected regions, Neota (area 2) likely has the greatest mean elevation value. Cache la Poudre (area 4) would have the lowest mean elevation if the other three areas of Rawah (area 1), Comanche Peak (area 3), and Cache La Poudre (area 2) were combined.

While spruce/fir (type 1) would predominate in Neota, lodgepole pine (type 2) would be the dominant species in Rawah and Comanche Peak, followed by spruce/fir (type 1) and aspen (type 2). (type 5). Type 3 Ponderosa pine, Type 6 Douglas-fir, and Type 7 cottonwood/willow are typical of the forests found in Cache la Poudre (type 4).

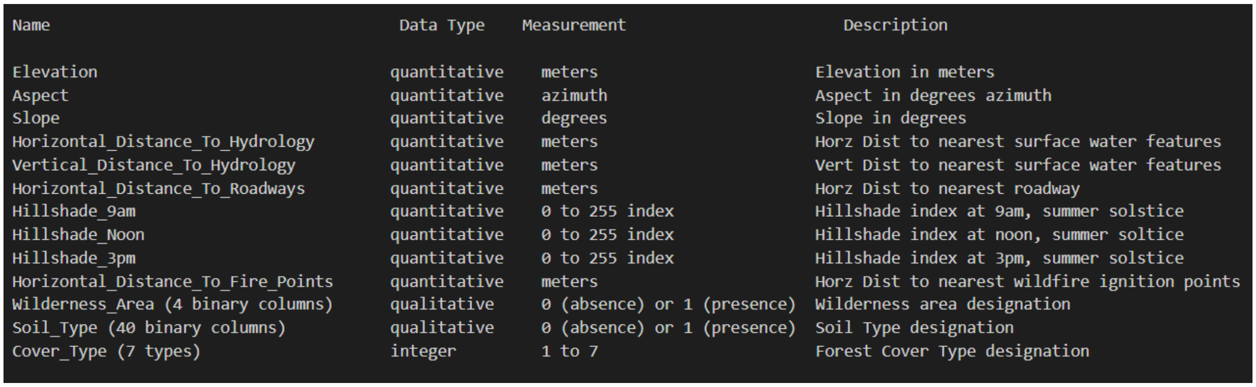
Both the Rawah and Comanche Peak regions have a diverse mix of tree species and a wide variety of predictive variable values, making them more representative of the whole dataset than either the Neota or Cache la Poudre regions (elevation, etc.) Because of its lower elevation range and hence different species composition, Cache la Poudre is likely to be distinct from the others.

METADATA

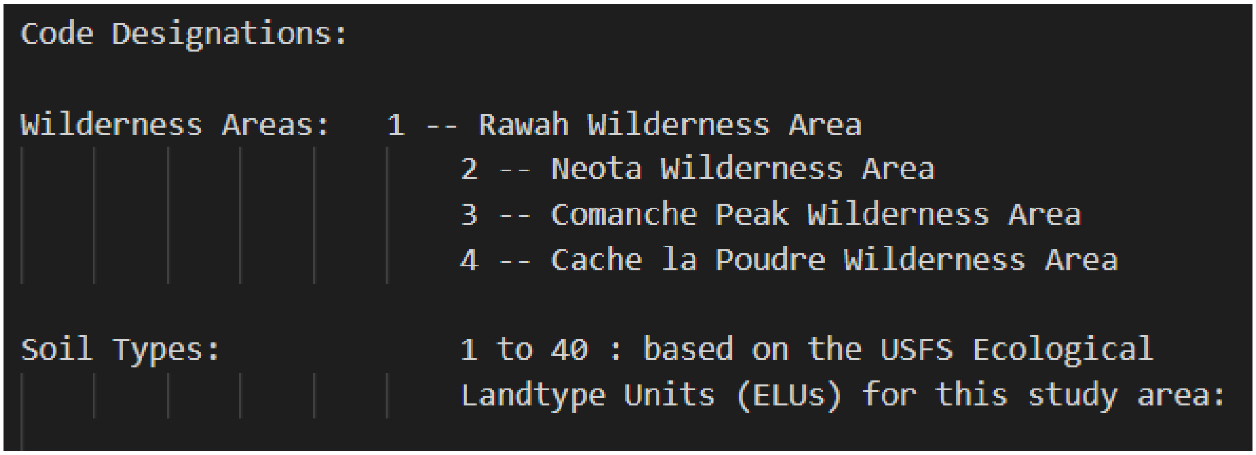
Number of instances (observations): 581,012

Number of Attributes: 12 measures, but 54 columns of data
(10 quantitative variables, 4 binary
wilderness areas and 40 binary
soil type variables)

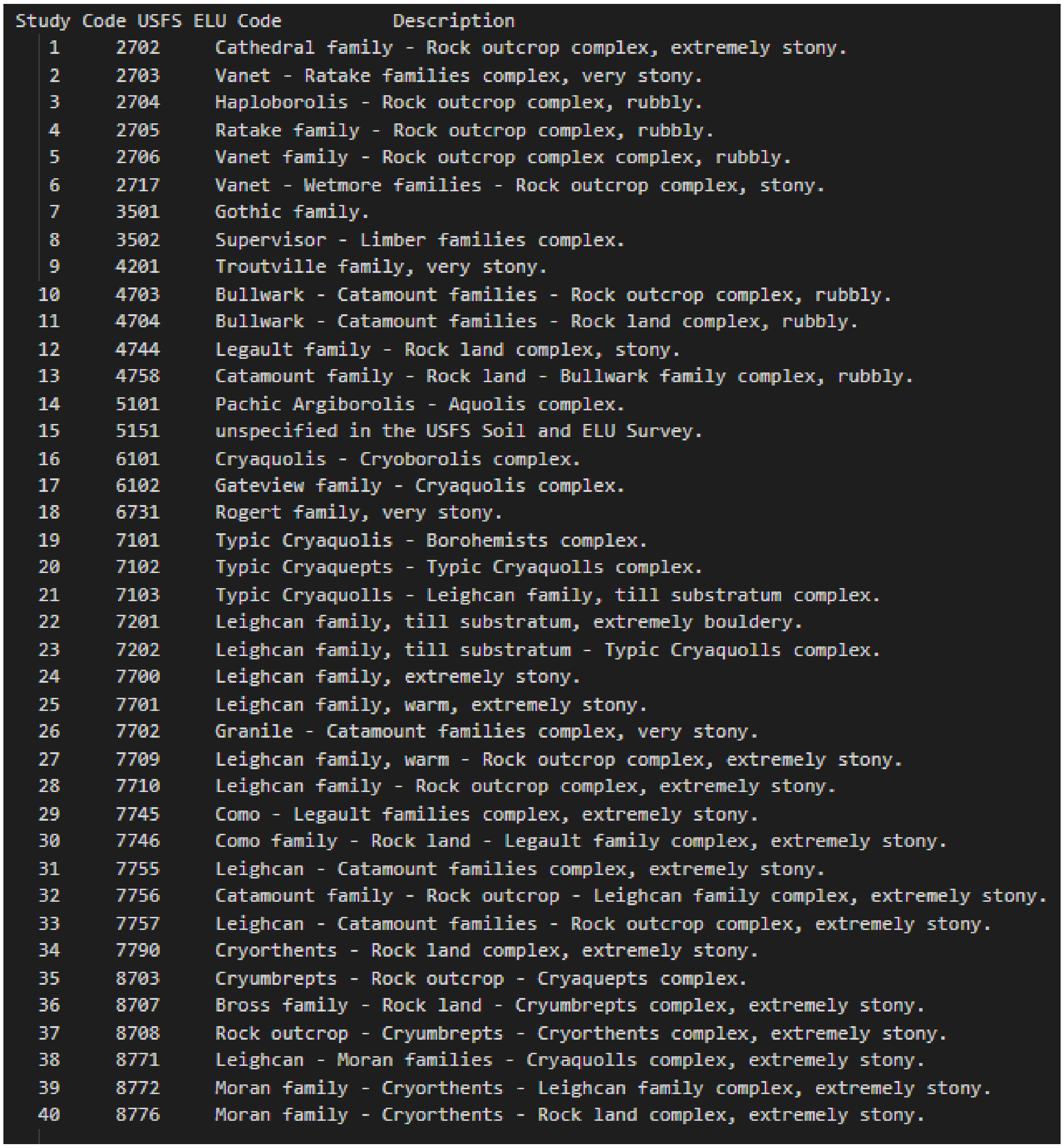
Attribute information:
The given information includes the attribute's name, type, measurement unit, and a brief description. The forest cover type is the challenge of categorisation. This listing's order conforms to the sequence of numbers along the rows of the database.



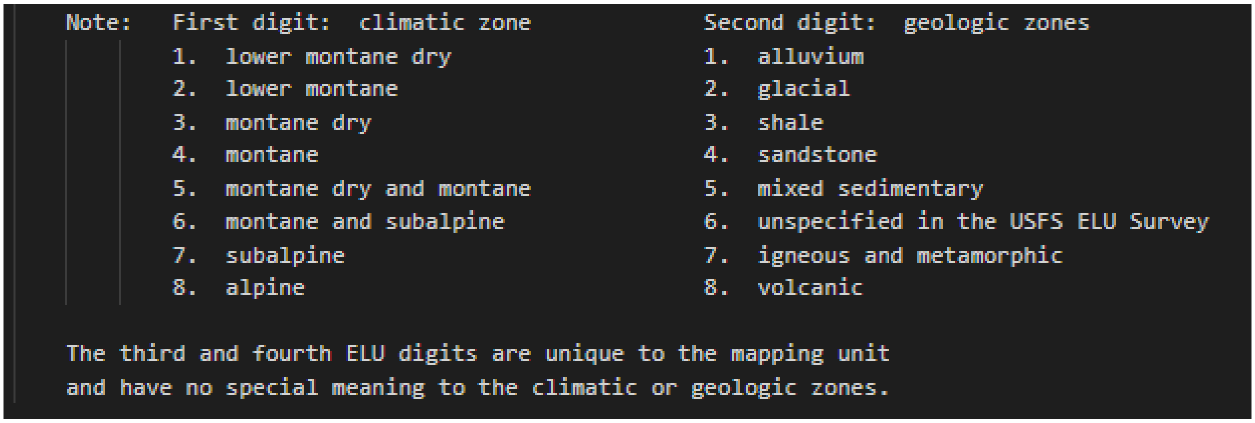
```
Name                                Data Type    Measurement                  Description

Elevation                          quantitative    meters                    Elevation in meters
Aspect                             quantitative    azimuth                   Aspect in degrees azimuth
Slope                              quantitative    degrees                   Slope in degrees
Horizontal_Distance_To_Hydrology   quantitative    meters                    Horz Dist to nearest surface water features
Vertical_Distance_To_Hydrology     quantitative    meters                    Vert Dist to nearest surface water features
Horizontal_Distance_To_Roadways    quantitative    meters                    Horz Dist to nearest roadway
Hillshade_9am                      quantitative    0 to 255 index            Hillshade index at 9am, summer solstice
Hillshade_Noon                     quantitative    0 to 255 index            Hillshade index at noon, summer soltice
Hillshade_3pm                      quantitative    0 to 255 index            Hillshade index at 3pm, summer solstice
Horizontal_Distance_To_Fire_Points quantitative    meters                    Horz Dist to nearest wildfire ignition points
Wilderness_Area (4 binary columns) qualitative     0 (absence) or 1 (presence) Wilderness area designation
Soil_Type (40 binary columns)      qualitative     0 (absence) or 1 (presence) Soil Type designation
Cover_Type (7 types)               integer         1 to 7                    Forest Cover Type designation
```



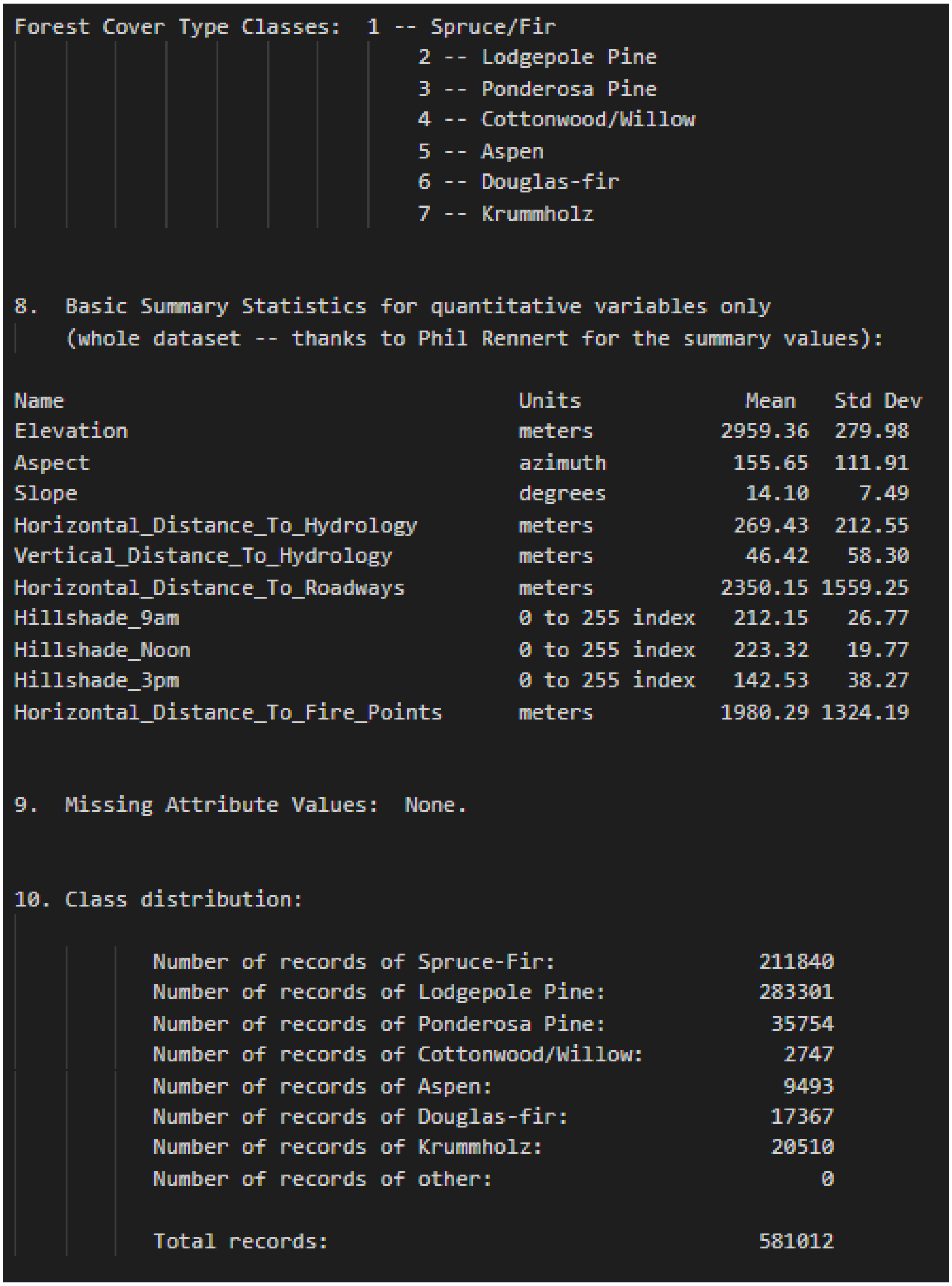
```
Code Designations:

Wilderness Areas:   1 -- Rawah Wilderness Area
                        2 -- Neota Wilderness Area
                        3 -- Comanche Peak Wilderness Area
                        4 -- Cache la Poudre Wilderness Area

Soil Types:             1 to 40 : based on the USFS Ecological
                        Landtype Units (ELUs) for this study area:
```



```
Study Code USFS ELU Code          Description
    1     2702     Cathedral family - Rock outcrop complex, extremely stony.
    2     2703     Vanet - Ratake families complex, very stony.
    3     2704     Haploborolis - Rock outcrop complex, rubbly.
    4     2705     Ratake family - Rock outcrop complex, rubbly.
    5     2706     Vanet family - Rock outcrop complex complex, rubbly.
    6     2717     Vanet - Wetmore families - Rock outcrop complex, stony.
    7     3501     Gothic family.
    8     3502     Supervisor - Limber families complex.
    9     4201     Troutville family, very stony.
   10     4703     Bullwark - Catamount families - Rock outcrop complex, rubbly.
   11     4704     Bullwark - Catamount families - Rock land complex, rubbly.
   12     4744     Legault family - Rock land complex, stony.
   13     4758     Catamount family - Rock land - Bullwark family complex, rubbly.
   14     5101     Pachic Argiborolis - Aquolis complex.
   15     5151     unspecified in the USFS Soil and ELU Survey.
   16     6101     Cryaquolis - Cryoborolis complex.
   17     6102     Gateview family - Cryaquolis complex.
   18     6731     Rogert family, very stony.
   19     7101     Typic Cryaquolis - Borohemists complex.
   20     7102     Typic Cryaquepts - Typic Cryaquolls complex.
   21     7103     Typic Cryaquolls - Leighcan family, till substratum complex.
   22     7201     Leighcan family, till substratum, extremely bouldery.
   23     7202     Leighcan family, till substratum - Typic Cryaquolls complex.
   24     7700     Leighcan family, extremely stony.
   25     7701     Leighcan family, warm, extremely stony.
   26     7702     Granile - Catamount families complex, very stony.
   27     7709     Leighcan family, warm - Rock outcrop complex, extremely stony.
   28     7710     Leighcan family - Rock outcrop complex, extremely stony.
   29     7745     Como - Legault families complex, extremely stony.
   30     7746     Como family - Rock land - Legault family complex, extremely stony.
   31     7755     Leighcan - Catamount families complex, extremely stony.
   32     7756     Catamount family - Rock outcrop - Leighcan family complex, extremely stony.
   33     7757     Leighcan - Catamount families - Rock outcrop complex, extremely stony.
   34     7790     Cryorthents - Rock land complex, extremely stony.
   35     8703     Cryumbrepts - Rock outcrop - Cryaquepts complex.
   36     8707     Bross family - Rock land - Cryumbrepts complex, extremely stony.
   37     8708     Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony.
   38     8771     Leighcan - Moran families - Cryaquolls complex, extremely stony.
   39     8772     Moran family - Cryorthents - Leighcan family complex, extremely stony.
   40     8776     Moran family - Cryorthents - Rock land complex, extremely stony.
```



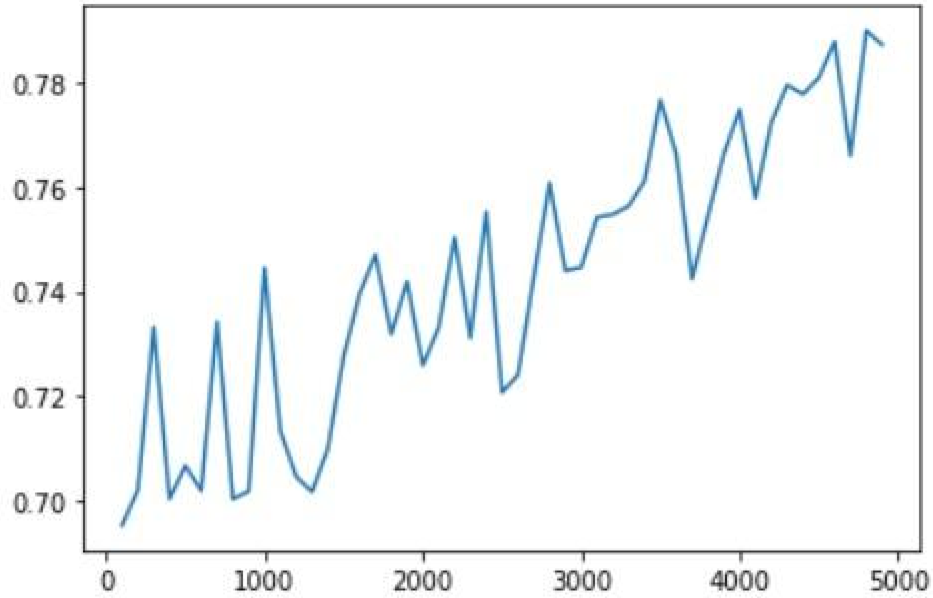
```
Note:   First digit:  climatic zone                Second digit:  geologic zones
        1.  lower montane dry                      1.  alluvium
        2.  lower montane                          2.  glacial
        3.  montane dry                            3.  shale
        4.  montane                                4.  sandstone
        5.  montane dry and montane                5.  mixed sedimentary
        6.  montane and subalpine                  6.  unspecified in the USFS ELU Survey
        7.  subalpine                              7.  igneous and metamorphic
        8.  alpine                                 8.  volcanic

The third and fourth ELU digits are unique to the mapping unit
and have no special meaning to the climatic or geologic zones.
```



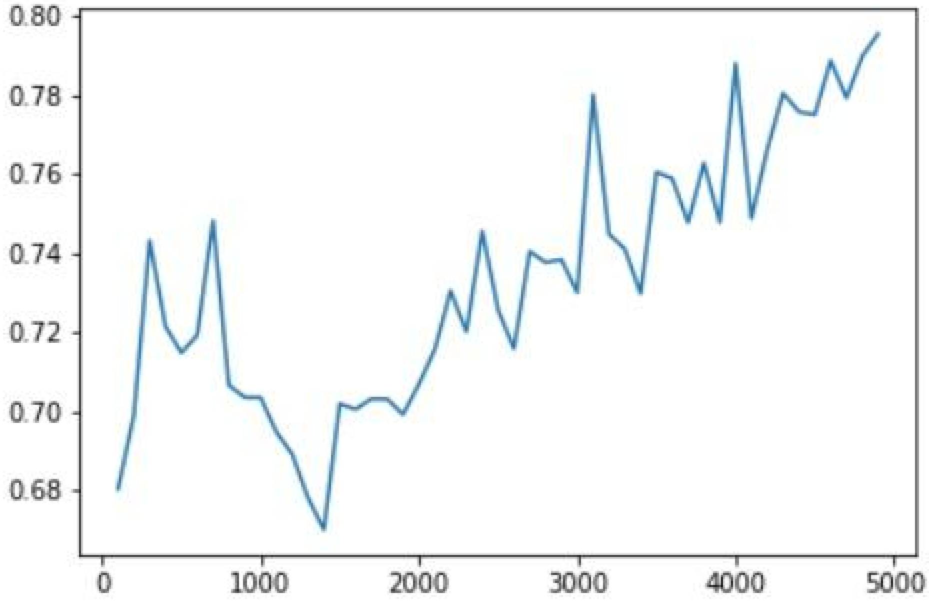
```
Forest Cover Type Classes:  1 -- Spruce/Fir
                                2 -- Lodgepole Pine
                                3 -- Ponderosa Pine
                                4 -- Cottonwood/Willow
                                5 -- Aspen
                                6 -- Douglas-fir
                                7 -- Krummholz


8.  Basic Summary Statistics for quantitative variables only
    (whole dataset -- thanks to Phil Rennert for the summary values):

Name                                    Units             Mean   Std Dev
Elevation                               meters          2959.36  279.98
Aspect                                  azimuth          155.65  111.91
Slope                                   degrees           14.10    7.49
Horizontal_Distance_To_Hydrology        meters           269.43  212.55
Vertical_Distance_To_Hydrology          meters            46.42   58.30
Horizontal_Distance_To_Roadways         meters          2350.15 1559.25
Hillshade_9am                           0 to 255 index   212.15   26.77
Hillshade_Noon                          0 to 255 index   223.32   19.77
Hillshade_3pm                           0 to 255 index   142.53   38.27
Horizontal_Distance_To_Fire_Points      meters          1980.29 1324.19


9.  Missing Attribute Values:  None.


10. Class distribution:

           Number of records of Spruce-Fir:                211840
           Number of records of Lodgepole Pine:            283301
           Number of records of Ponderosa Pine:             35754
           Number of records of Cottonwood/Willow:           2747
           Number of records of Aspen:                       9493
           Number of records of Douglas-fir:                17367
           Number of records of Krummholz:                  20510
           Number of records of other:                          0

           Total records:                                  581012
```

LEARNING CURVES

- **Margin Sampling**

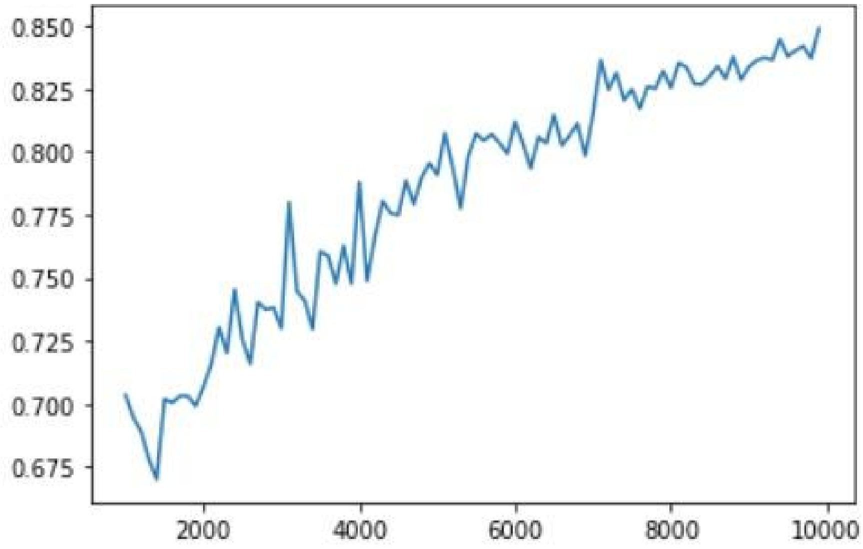


- **Least Confidence Sampling**



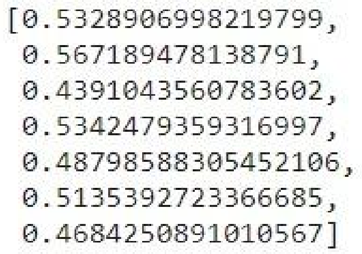
- **Entropy Sampling**

```
[<matplotlib.lines.Line2D at 0x7f0014b59590>]
```



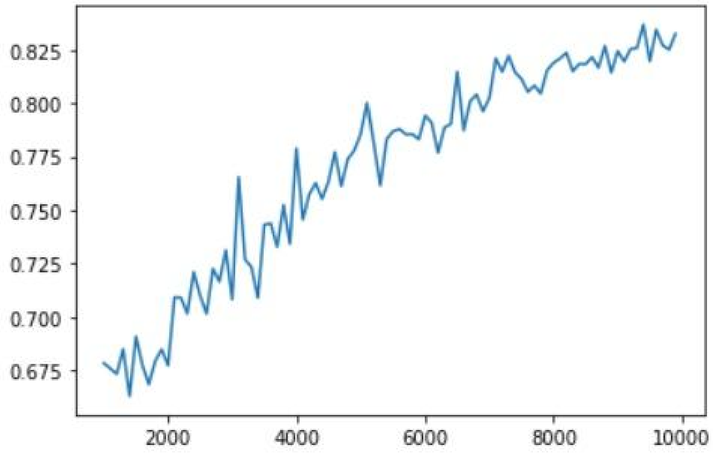
- **Vote Entropy**

  Committee used: Decision Trees (Random Forest)



```
[0.5328906998219799,
 0.567189478138791,
 0.4391043560783602,
 0.5342479359316997,
 0.48798588305452106,
 0.5135392723366685,
 0.4684250891010567]
```

```
[<matplotlib.lines.Line2D at 0x7f22f48c1a10>]
```

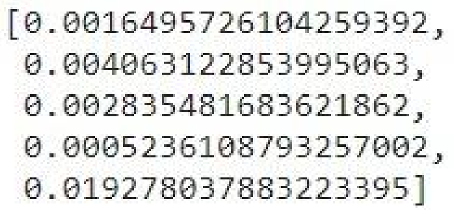


❖ KL Divergence

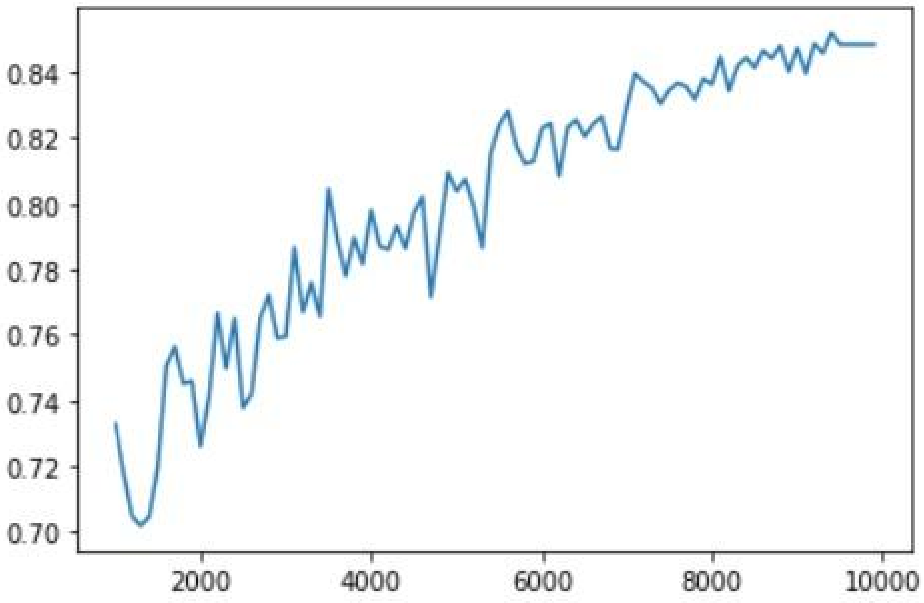
The Kullback-Leibler divergence of Q from P is defined as

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\,\ln\frac{P(i)}{Q(i)}.$$

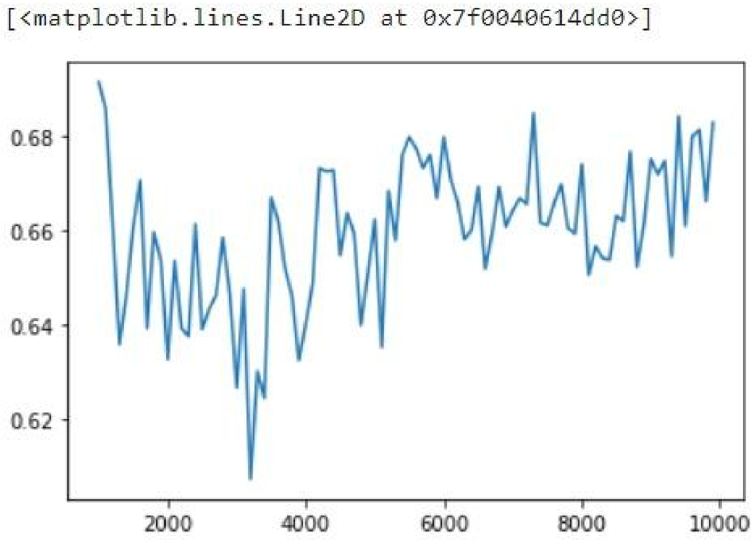
This KL divergence measures the amount of information lost when Q is used to approximate P. In the active learning context, Q is the average prediction probability of the committee, while P is the prediction of a particular committee member.



```
[0.0016495726104259392,
 0.004063122853995063,
 0.002835481683621862,
 0.0005236108793257002,
 0.019278037883223395]
```



- **Random Sampling**



```
[<matplotlib.lines.Line2D at 0x7f0040614dd0>]
```

The best results are obtained using Entropy sampling, as shown by the graphs of Uncertainty sampling techniques.

The model with the highest vote entropy wins out over KL Divergence in QBC.