Final Project Presentation

Team Members:-

1.Mukul Bichkar

2. Sushil Thasale

Steps Followed:-

1. **Pre-Processing :-**

a. <u>Understand Domain</u>

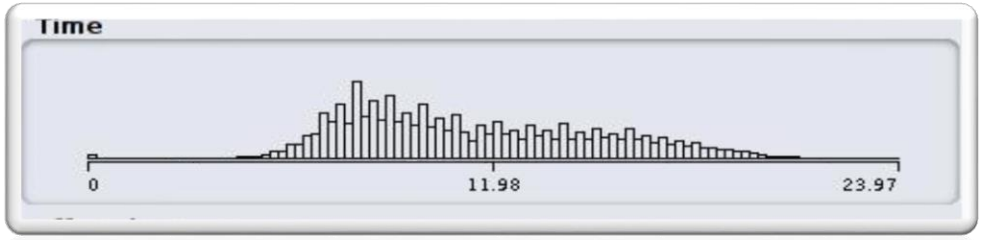
Online study for Red-Winged birds.

Found some important parameters

- location
- State
- Population of the area
- temperature
- Time (diurnal nature)
- water bodies
- ecology

Pre-Processing

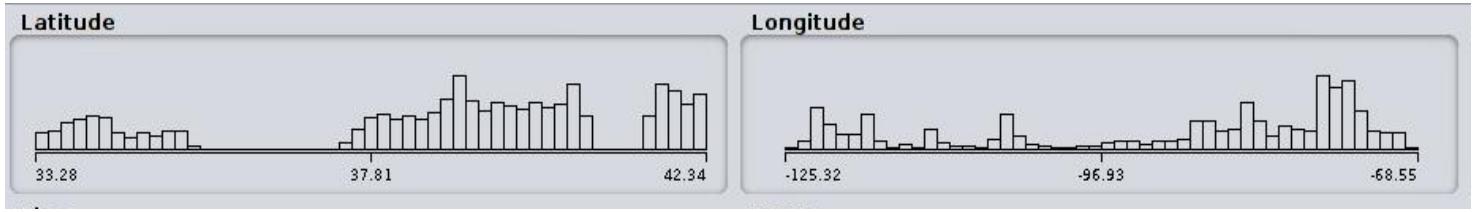
- To verify claims we sampled the labelled data at different sampling rates (1%, 4% and 20%). Processed this data to extract all the fields which we thought could affect the classification. Then this data was fed to Weka Explorer and following graphs were obtained.
- Time



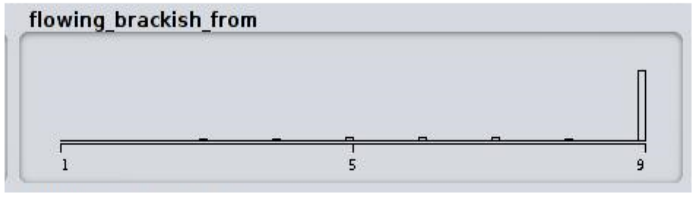
- Housing Density

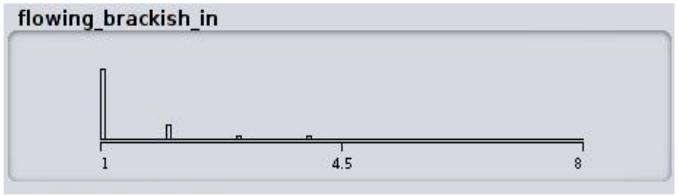
Pre-Processing

- Latitude and Longitude



- Flowing Brackish Water (Distance from and in)





- Population per Square Mile

Building and Training the Model

- Job – 1

Mapper(record):

a. Read labelled data and create a custom Object (SamplingEventDetails) with all the desired field i.e. the one's that would affect training the model.

b. Generate a random number

c. emit(random number, SamplingEventDetails)

Reducer(List[SamplingEventDetails]):

a. Initialize Training Set

b. Add each value to Training Set

c. Train Three Models (Naïve Bayes, Decision Tree, Random Tree) for given training set

d. write models to disk

e. emit null

Job-2 Validating the Model

Mapper(record):

a. Read record from unlabeled data and build custom SamplingEventDetails object

b. Generate a random number

c. emit(random number, SamplingEventDetails)

Reducer(List[SamplingEventDetails]):

a. Load all the trained models from disk

b. For each s in List([SamplingEventDetails])

c. For each model

Determine the probability of spotting the bird

d. Output a string containing Sampling Id and probability

Accuracy

Sr. No.	Model	Accuracy for Sample Data	Attributes used
1	Naïve Bayes	71 %	month, time, housing vacant, population per sq mile, distance from flowing and standing fresh and brackish water
2	RandomTree	75 %	month, time, housing density, population per sq mile, distance from flowing and standing fresh and brackish water
3	Random Forest [depth=15, trees=15, features = 19]	76 %	month, time, caus_temp_avg, housing density, housing vacant, population persq mile, distance from flowing and standing fresh and brackish water
4	All three combined i.e. NB, RandomTree and Random Forest. [depth=15, trees=15, features = 19]	72%	month, time, caus_temp_avg, housing density, housing vacant, population persq mile, distance from flowing and standing fresh and brackish water

Questions ?