Principles of Statistical Modeling

Project Term Paper

by

Sakshi Sharma

JACOBS
UNIVERSITY

Prof. Stefan Kettemann
Jacobs University Bremen

1 Introduction

Water is critical to a country's development, as it is not only used in agriculture but also for industrial development. Though Tanzania has access to a lot of water, the country still faces the dilemmas of many African countries where many areas have no reliable access to water .We are looking at the Dataset of water pumps in Tanzania to predict the operating condition of a water point.By finding which water pumps are functional,functional need repair,and non functional,the Tanzanian Ministry of water can improve the maintenance operations of the water pumps and make sure that clean, potable water is available to communities across Tanzania.

For me, this project serves as opportunity to utilize the knowledge I have gained through the semester in the Principles of Statistical Modelling course to explore the statistical concept behind the workings of an classification algorithm, namely: The Bayesian Classifier.

2 Data Description and Data Exploration

We are using the data from Taarifa and Tanzanian Ministry of Water to classify the water pumps. The data was collected using handheld sensor, paper reports, and user feedback via cellular phones. The dataset has features such as the location of the pump, water quality, source type, extraction technique used, and population demographics of pump location.The data for this comes from the Taarifa waterpoints dashboard, which aggregates data from the Tanzania Ministry of Water.The training set has 59,401 rows and 40 features including an output column. The output column specifies the status of the water pump in the category of functional, functional needs repairs, or non functional. Out of the 40 features in the data we have:

- 31 categorical variables
- 7 numerical variables
- 2 date variable.

The subsection below describes the dataset in correct mathematical formalism.

2.1 Data description in Mathematical Formalism

When statistical data are collected, the data collection process can be described using different components. Understanding these components of the data collection process helps us understand the dataset better and generate a mathematical abstraction of the dataset. The components that we will explore are as follows:

- **Universe and Elementary Events :** The Universe is a part of the real world from which we collect data. In mathematical terms the universe is represented by a set Ω. In the particular case of considering the Tanzanian dataset the Universe set is the collection of all the water pumps across the world.

Elementary events are elements of the universal set from which the observation data is collected. Mathematically the elementary events are represented by ω are elements of the universe set Ω. Following the definition of the universe set provided for the data, the elementary events from which the data was collected are the water pumps present in Tanzania Ω.

- **Data Value Space:** The data value space mathematically is also set which contains all the possible outcomes of an observation procedure acting on elementary events, which we also can call the "observation act". In particular when considering the Tanzanian data, each question present in the questionnaire and the user-feedback via cellular phones is an observation procedure and Government officials collecting answers to the questions is the observation act on an elementary event(the water pumps). A description about the water point that are reported by residents of a particular reigon in tanzania, shows that data values generated as answers to the questions are of three different types:
  - **categorical value :** The description about the water point that output a categorical value produces a data value space of finite sets. These finite sets can further be divided in to two types depending on the type of values the elements of the sets have.
    - **Nominal valued elements:** These sets are finite and have elements that have no natural order and no numeric value. For the purpose of this project we will represent sets with nominal valued elements by t_i where i represents the index of questions about water pumps that output Nominal values.One such details about the water pumps is the quality of the water whether it is salty,milky,coloured etc.,. The data value space generated by this particular question will be the finite set of 6 different description related to a particular water point. Figure 1 below shows the total count of waterpumps that belong to these 6 different quality groups. Next to make our task of defining the data value space easier we create a single nominal data value space represented by $\mathcal{C}$ by defining it to be the set product of t_i

$$\mathcal{C} = \times_i \, t_i \qquad (1)$$

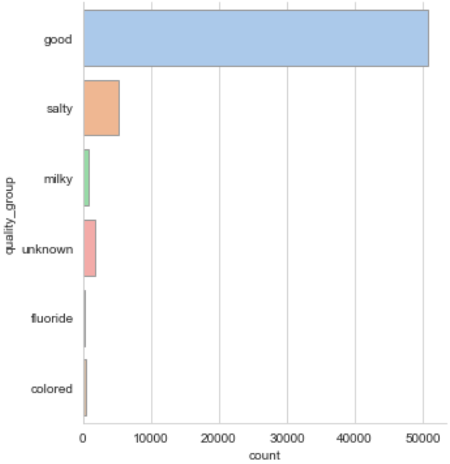


Figure 1: total count of waterpumps that belong to these 6 different quality groups

- **Ordinal valued elements:** These sets are finite and have elements that have a natural order and a numeric value. The survey questions that output such values are questions that have true/false answer. The natural ordering given to the answer of such a question is most commonly true=1 and false =0. we represent such sets for the purpose of this project by $w_j = \{0, 1\}$ where j is is the index representing questions about water pumps that have a true/false answer. One such question in the questionnaire asks if the waterpump is permitted. The data value space generated by this particular question will be the finite set {0,1}. Figure 2 below shows the count of waterpumps if they are permitted or not. Similar to defining the nominal data value space, we can define a ordinal data value space represented by $\mathcal{D}$ as the set product of w_j.

$$\mathcal{D} = \times_j w_j \tag{2}$$

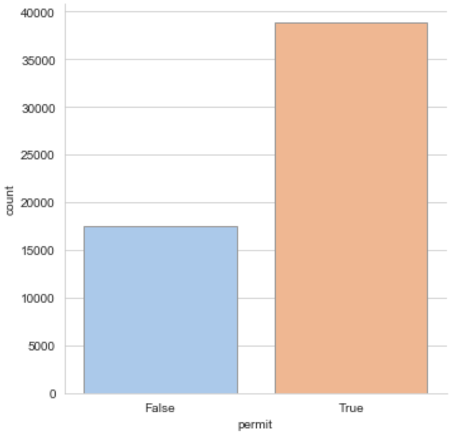
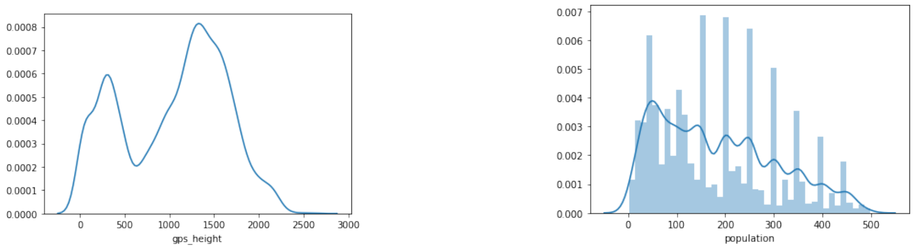


Figure 2: count of waterpumps if they are permitted or not.

- **Numerical values:** Some description output of a waterpump is numeric value. These description will create either a discrete data value space or a continuous data value space. The discrete valued data value space are again sets of various lengths that contain elements belonging to the real numbers. We will represent these sets by r_k where k is the index of question that output a discrete data value space. Then we can further define a discrete valued data space such that,

  $$\mathcal{G} = \times_k r_k \tag{3}$$

  on the other hand the continuous data value space contains different intervals of the real number line. To make the task of defining this continuous data value space we can merge the real line interval $[0, \infty]^n$ where n is the description of water pumps that require a continuous numerical answer. I have defined the upper limit of the line interval to be ∞ because, the data value space must contain all the possible outcome values, defining these intervals for each questions will require a long section and detailed discussion of the particular description therefore to make the task simpler, I have defined the interval with an upper bound of ∞. Most of the numerical values resulting from the descriptions are discrete and a few of them are continuous. One question which outputs a discrete numerical value is the population around the particular water pump. Similarly the altitude of the water pumps that are located in tanzania outputs a continuous numerical value. Figure 3 below shows plots of the distribution of the poplulation and gps height, based on the tanzania data



(a) distribution of altitude of waterpump (continuous numerical)

(b) distribution of population(discrete numerical)

Figure 3: distribution of the variables gps height and population the y axis has been normalized to make the area under the distribution curve 1

These different data value spaces created depending upon description of the water pump can be merged together into a big data value space $\mathcal{S}$ which is mathematically represented as

$$\mathcal{S} = \mathcal{C} \times \mathcal{D} \times \mathcal{G} \times [0, \infty]^n \tag{4}$$

- **Random Variables:** are functions which map the elementary events from the universe set into data values in a data value space. The random variables are the mathematical formalism of representing the observation procedure. This follows that each description about the waterpump, $q_l \mid l \in \{1, ...n\}$ where n=41 is the total number of description about the waterpump, is a random variables that maps the elementary events into the data value space. These individual random variables q_l can be represented by a compound random variable $\mathcal{Q}$, where:

$$\mathcal{Q} := \bigotimes_{i=1}^{41} q_l \tag{5}$$

This compound random variable $\mathcal{Q}$ is what translates water pump characteristics into the a mathematical set i.e. the data value space S.

$$\mathcal{Q} : \Omega \rightarrow \mathcal{S} \tag{6}$$

This Mathematical formalization the data value space and the random variables shows us how the 41 columns in the data set are generated and how they mathematically represent water pump characteristics.

Fig4 below shows

2.2 Dataset Exploration

In this section we plot various graphs to explore and identify presence of distinct relationships between the waterpump characteristics and the labels of the waterpump (characterised by the ”functional”,”functional need repair”,and ”non functional” labels).

For numerical variables we can plot the distribution ,pearson correlation and scatter plot on a single scatter plot matrix of features. Figure below shows the pearson correlation (upper half of the scatter plot matrix), Desnisty plot (diagonal) and scatter plots (lower half of the scatter plot matrix) between the 3 numerical features in our dataset.

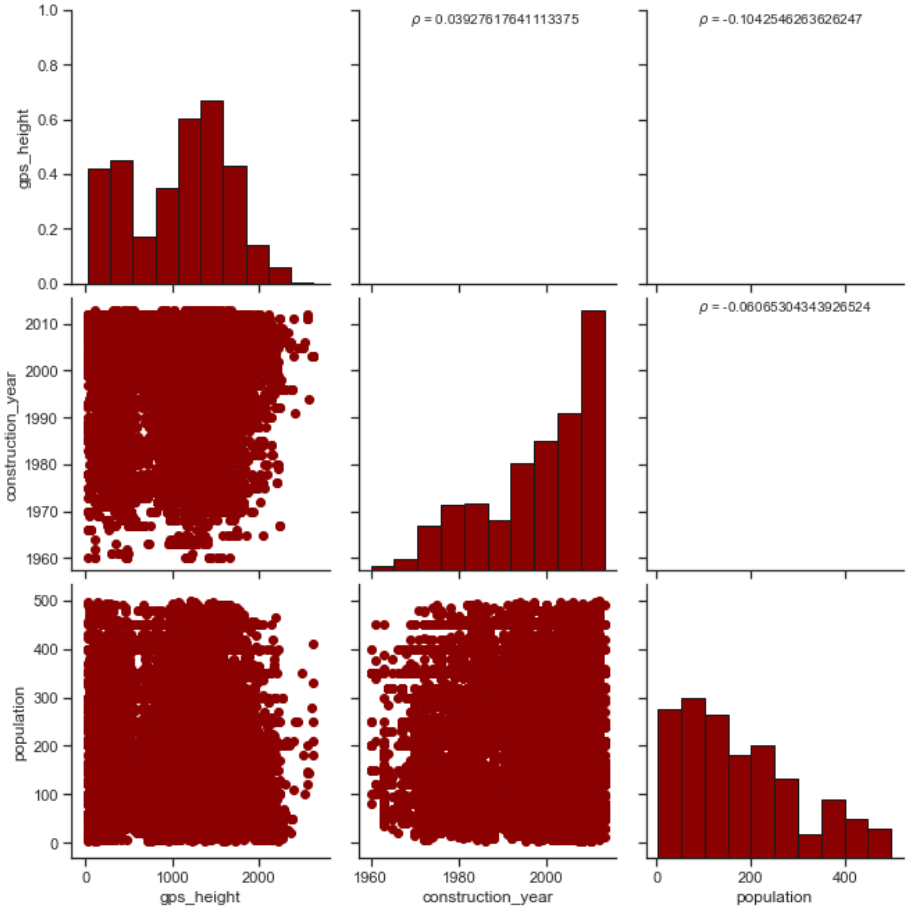


Figure 4: Scatter plot matrix of numerical features based on pearson correlation,density and scatter plot

A Pearson correlation is a number between -1 and +1 that indicates to which extent 2 variables are linearly related. The correlation value in the plot above is very less, so these features do not show distinct separation between them.

Since each water pump have been labeled as functional,functional need repair,and non functional so we can further plot another scatter plot matrix, with the distribution in the diagonal separated by the class labels().

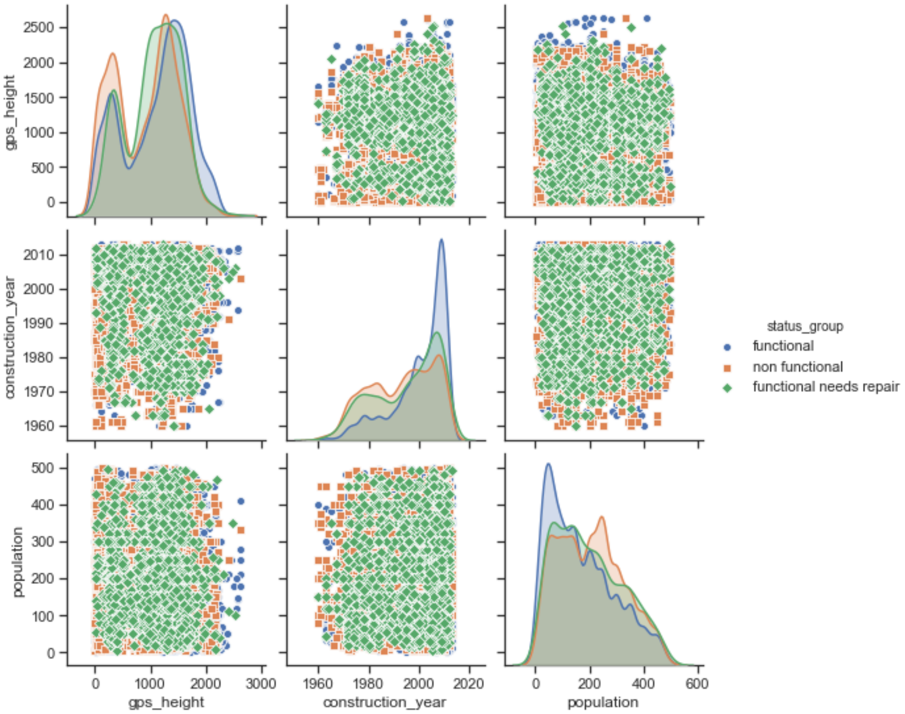


Figure 5: Scatter plot matrix of numerical features between functional, functional needs repair and non-functional responses

Comparing these distributions is interesting, because if the class distributions were separable, it would hint if the particular variable is useful in as a predictor for classification.

For categorical variable we can use Multiple Correspondance Analysis(MCA).Since most of our features in the dataset are catagorical,MCA would help us identify if any categories of the different features have strong relationship with the catagories of the waterpump labels. Figure below shows the MCA plot on all the categorical features and labels present in our dataset.

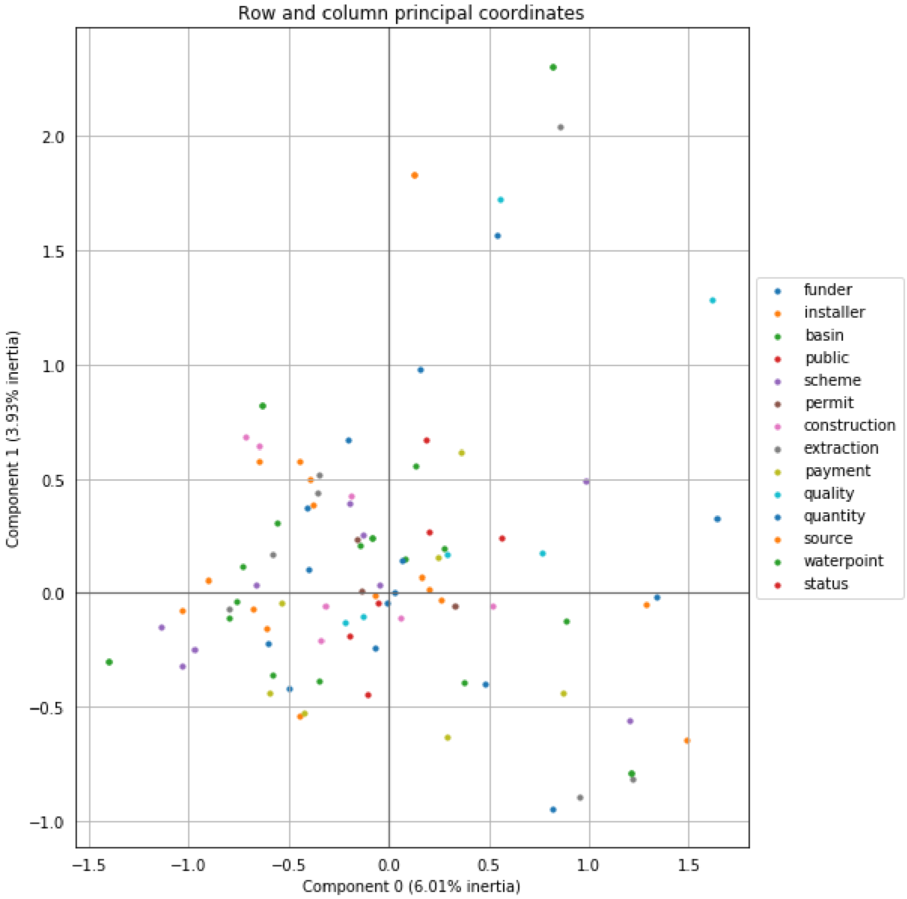


Figure 6: Multiple Correspondance Analysis(MCA) of categorical features between functional, functional needs repair and non-functional responses

From the above plot it is really difficult to understand which features and their categories, have strong relationship with the categories in the 7labels. so, to simplify the feature selection we used chi-square test.In Chi-Square goodness of fit test, categorical data is divided into intervals. Then the numbers of points that fall into the interval are compared, with the expected numbers of points in each interval. Figure below shows best 25 categorical features that are best fit for classifying between 3 labels.

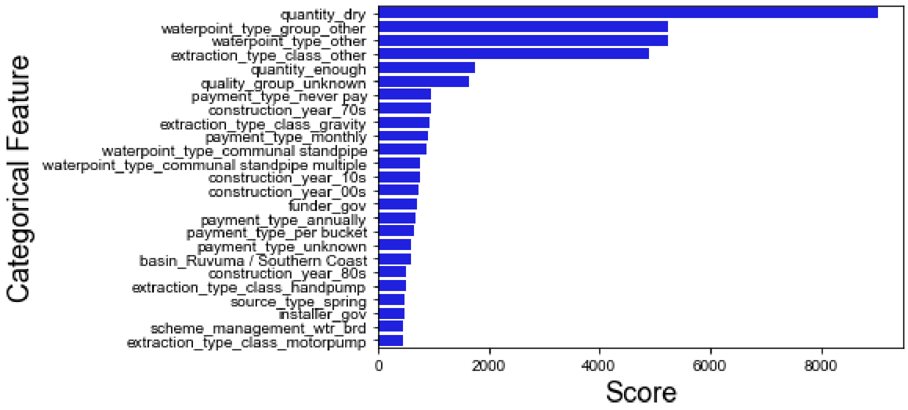


Figure 7: chi2 scores of best 25 categorical feature categories

The way to interpret the above chi2 scores is that categorical features with the highest values for the chi-squared stat indicate higher relevance and importance in predicting labels and may be included in a predictive model development.

3 Naive Bayes Classifier for classifying the waterpumps

Classification is a supervised learning task, where the dataset used must be represented as a labelled pair:

$$(x_i, y_j) \tag{7}$$

where x_i are the different features that represent a case in the dataset that has been labelled using a class y_j. In the Tanzania dataset we have each waterpump represented by features (the waterpump characteristics, collected from the resident feedback) and a class label represented by the status of the waterpump i.e functional,functional need repair,and non functional. A classification algorithm is also an estimator, which takes as input a dataset and produces an class estimate for objects in the dataset. Using classification algorithm to classify objects in a dataset requires two steps, the first step which we classically call the ”training” step and the second step the ”testing” step. The training step is where we input a subset of our dataset as training-data into the classifier and allow the algorithm to ”learn” from the training data. The objective of the learning process is to obtain a ”well learnt” algorithm that when fed with the test dataset generalizes in a meaningful way what it has ”learnt” from the training sample. The ”learning” task looked upon mathematically, is trying to estimate the conditional distribution:

$$P(Y = y_j \mid X = x_i) \tag{8}$$

3.1 Naive Bayes Classifier algorithm

A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task.They are extremely fast and simple classification algorithms that are often suitable for very high-dimensional datasets.

3.1.1 Naive Bayes Explanation and Theory

As discussed above, The objective of the learning process of an classification task looked upon mathematically, is to find an estimate of the conditional distribution:

$$P(Y = y_j \mid X = x_i) \tag{9}$$

One way to estimate the conditional distribution $P(Y = y_j \mid X = x_i)$, is to utilize Bayes' theorem, mathematically is written as:

$$P(Y = y_j \mid X = x_i) = \frac{P(Y = y_j)P(X = x_i \mid Y = y_j)}{\sum_j P(Y = y_j)P(X = x_i \mid Y = y_j)} \tag{10}$$

The denominator can be removed and a proportionality can be introduced.

$$P(Y = y_j \mid X = x_i) \propto P(Y = y_j) \prod_{i=1}^{n} P(X = x_i \mid Y = y_j) \tag{11}$$

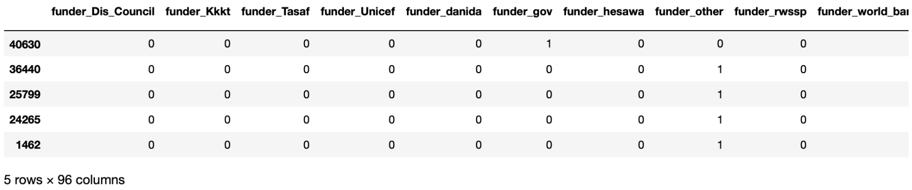
Finally, the class label predicted is the class for which the estimated value of $P(Y = y_j \mid X = x_i)$ is maximum, i.e.

$$Y_{pred} = \ argmax_Y P(Y) \prod_{i=1}^{n} P(x_i \mid Y) \tag{12}$$

The classifier can estimate the values for each $P(Y = y_j)$ and $P(X = x_i \mid Y = y_j)$ from the dataset during the training process. The value of $P(Y = y_j)$ is easily estimated by computing the ratio of data points that fall in each class y_j. Further, the modeller also needs to make important assumptions regarding the conditional distribution $P(X = x_i \mid Y = y_j)$. according to the dataset, the modeller can assume any distribution, and compute the distribution parameters from the dataset. Using the Naive Bayes classifier function from Skit learn, allows us to choose between three different widely used distributions. These are

- Normal (Gaussian) distribution
- Multivariate distribution
- Bernoulli distribution

We do not assume the conditional distribution $P(X = x_i \mid Y = y_j)$ to be Gaussian distribution because Random variables that have a Gaussian distribution must take up a continuous data value space. In the data set we use, there are only three features that are continuous. Instead, before running the model we will drop these three variables. Similarly, if we do not assume that the conditional distribution $P(X = x_i \mid Y = y_j)$ is a Multivariate distribution because multinomial distribution describes the probability of observing counts among a number of categories, and thus multinomial naive Bayes is most appropriate for features that represent counts or count rates. In the data set we use, there are no such features that represent counts. Therefore,We will build our classifier on the assumption that the conditional distribution $P(X = x_i \mid Y = y_j)$ is a Bernoulli distribution.Bernoulli implements the naive Bayes training and classification algorithms according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable. As, this method requires samples to be represented as binary-valued feature vector we have created dummy variable for all the categorical features present in our cleaned processed dataset. This final dataset has a total of 99 variables, each taking a boolean data value space. Figure below shows a snapshot of the first five rows of the final dataset with the dummy features.



	funder_Dis_Council	funder_Kkkt	funder_Tasaf	funder_Unicef	funder_danida	funder_gov	funder_hesawa	funder_other	funder_rwssp	funder_world_ba
40630	0	0	0	0	0	1	0	0	0	
36440	0	0	0	0	0	0	0	1	0	
25799	0	0	0	0	0	0	0	1	0	
24265	0	0	0	0	0	0	0	1	0	
1462	0	0	0	0	0	0	0	1	0	

5 rows × 96 columns

Figure 8: dummy variable dataset

4 Results

After creating the dummy variable for our features, the dataset has total 99 features.Using all the 99 features might lead to the over fitting of the data. So, to avoid the over fitting we need to figure out what are the best number of features to be used.Therefore, we first split the data set into training data and the test data and run the different models for different number of feature ranging from 5 to 99. Below plot shows the testing accuracy of Bernoulli Naive Bayes for features ranging from 5 to 99.

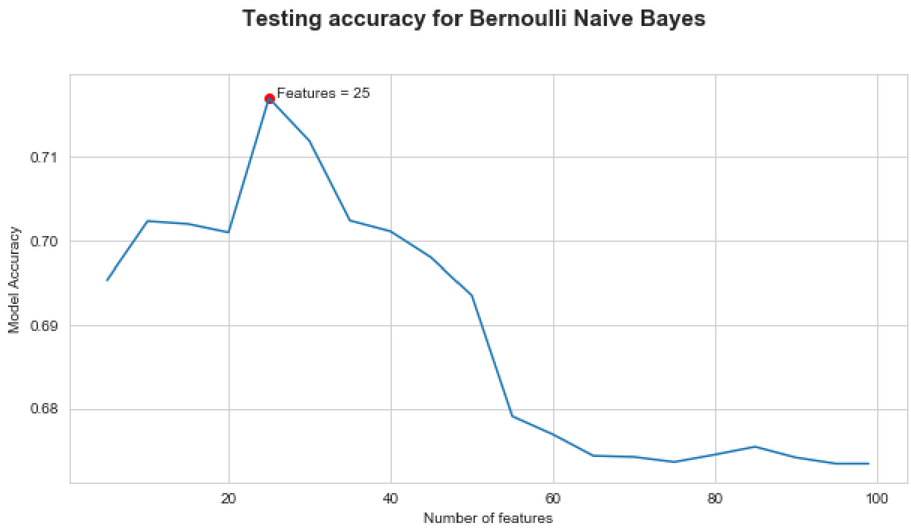
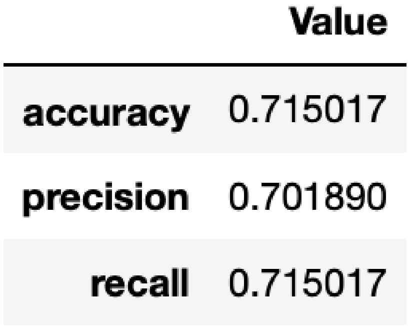


Figure 9: Bernoulli Naive Bayes Model Accuracy for different number of variables

From the plot, we can determine that the model with 25 features has the maximum accuracy of 71%. So, as the final model, I build a model with 25 features. The figure below shows the evaluation metrics for this model.

Classification Performance:



	Value
accuracy	0.715017
precision	0.701890
recall	0.715017

Figure 10: Model Evaluation Metrics

These metrics show that the model is able to accurately classify 75% of water pumps present in the test dataset. This performance of the models can be further improved with techniques of Cross-Validation, Feature extraction, dealing with class imbalance or choosing other classifiers such as Random Forest. This optimization task is out of scope for this project and is not explored in this report.

5 References

- 1. Lecture Notes Principles of Statistical Modeling 2019: Herbert Jaeger.
- 2. Lecture Notes Machine Learning 2019: Herbert Jaeger.
- 3. Lecture Notes Principles of Statistical Modeling 2020: Stefan Kettemann.