

Logistic Classification

The Logistic classification aim is to determine whether or not an object x belongs to a class. To determine this, object x have multiple characteristics that help a model to determine its class.

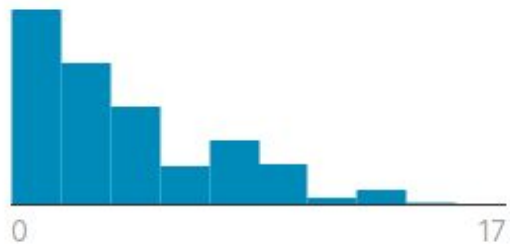

In order to evaluate the object x , it must be evaluated with a hypothesis function, very similar as linear or multivariate regression. However, the hypothesis function must be changed, which must be:

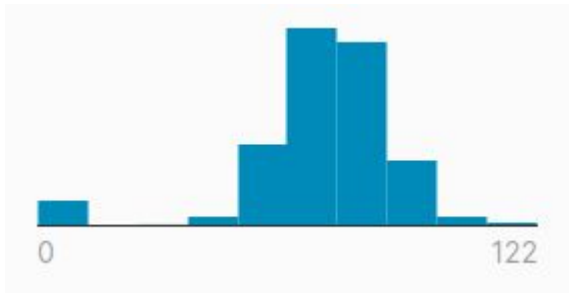
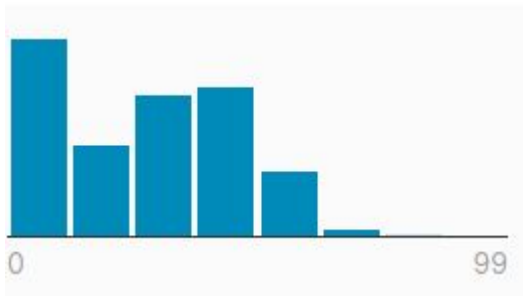
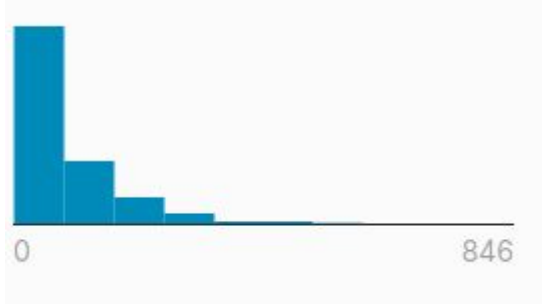
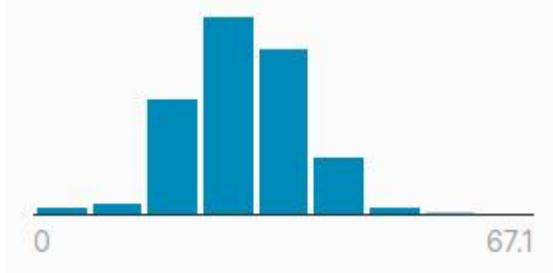
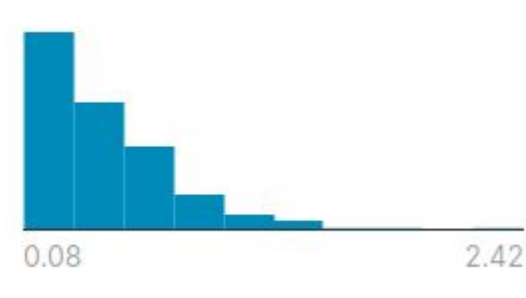
$$L_w(x) = \frac{1}{1 + e^{-w^T x}} = P(y|x; w)$$

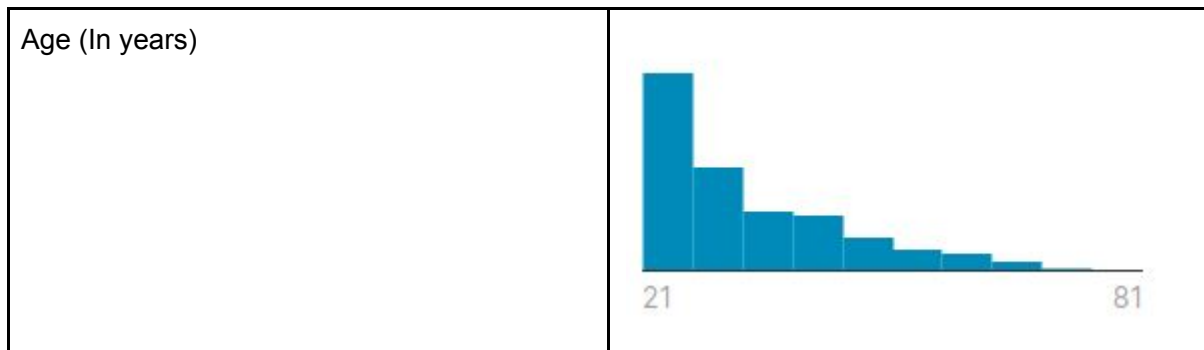
Diabetes Prediction:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Characteristics:

Characteristics	Histogram
Pregnancies (Number of times pregnant)	
Glucose (Plasma glucose concentration a 2 hours in an oral glucose tolerance test)	

<p>BloodPressure (Diastolic blood pressure (mm Hg))</p>	 <p>A histogram showing the distribution of diastolic blood pressure. The x-axis ranges from 0 to 122. The distribution is roughly bell-shaped, centered around 100-110 mm Hg, with a few outliers at lower values.</p>
<p>SkinThickness (Triceps skin fold thickness (mm))</p>	 <p>A histogram showing the distribution of triceps skin fold thickness. The x-axis ranges from 0 to 99. The distribution is skewed to the right, with a peak at the lowest values (around 10 mm) and a long tail extending towards 90 mm.</p>
<p>Insulin (2-Hour serum insulin (mu U/ml))</p>	 <p>A histogram showing the distribution of 2-hour serum insulin levels. The x-axis ranges from 0 to 846. The distribution is highly skewed to the right, with a very high frequency at the lowest values (below 100 mu U/ml) and a long tail extending towards 800 mu U/ml.</p>
<p>BMI (Body mass index (weight in kg/(height in m)^2))</p>	 <p>A histogram showing the distribution of Body Mass Index (BMI). The x-axis ranges from 0 to 67.1. The distribution is roughly bell-shaped, centered around 25-30 kg/m², with a few outliers at lower values.</p>
<p>DiabetesPedigreeFunction (Diabetes pedigree function)</p>	 <p>A histogram showing the distribution of the Diabetes Pedigree Function (DPF). The x-axis ranges from 0.08 to 2.42. The distribution is skewed to the right, with a peak at the lowest values (around 0.1) and a long tail extending towards 2.4.</p>



Discussion:

Data set:

There are some features that may be not applicable. For example, I don't know how much the Pregnancies feature can contribute to the model. Excessive features are likely cause that the generated model may not be as precise and accurate as it could be. Since the dataset has a lot of features, the data set should be scaled in order to improve time performance.

Results:

The data is shuffled each time the program is executed, so it was decided to make a table with the variables of confusion matrix, running the program 10 times. These data were made with a learning rate of 0.005 and stopping criteria of 0.01.

C.M.V	1	2	3	4	5	6	7	8	9	10	Avg.
TP	24	20	31	31	28	39	37	37	29	37	31,3
TN	82	102	85	93	83	81	86	86	84	83	86,5
FP	10	8	11	13	13	14	10	11	17	10	11,7
FN	37	23	26	16	29	19	20	19	23	23	23,5

Metric	Accuracy	Precision	Recall	Specificity	F1 score
Average	0,76993	0,72791	0,57117	0,88086	0,64008

In conclusion, the model is far from being precise and exact. The main problem is that there are too many false negatives. This may be caused by features that are not necessary or not as relevant in the model.