Exercise 3

1 Bayes Theorem (1 point)

99.5% of the population do not use drugs and 0.05% are drug users. A test allowing to detect the presence of drugs in the blood gives the following results

	user	**non-user**
positive	99%	1%
negative	1%	99%

What is the probability that a randomly selected individual that had a positive test is actually not a drug user?

2 A-B Testing (2 points)

A company designed a new version of the landing page of their web-site and would like to test if this will improve their sales. To this end, they split the traffic between two versions of the web-site and collected the corresponding data given in **ab_data.csv**. The column **timestamp** describes the time of the visit of the landing page, the column **converted** describes if a session terminated with a sale, and the column **landing_page** indicates the version of the landing page.

- In what proportion did the company split the traffic? What is the conversion rate for each version of the landing page?
- Under the hypothesis that the probability of conversion does not depend on the version of the landing page with the help of a binomial-test find how likely it is to observe the number of conversions as extreme as the one for the old landing page and the one for the new landing page.
- Under the same null hypothesis use the χ-squared test and the normal-test to measure the significance of the difference in the conversion-rates of the landing page versions.

3 Logistic Regression (5 points)

Use the dataset **winequality_binary.csv** to implement logistic regression using the python library numpy.

- Use cross-entropy as the cost function
- Implement the gradient decent algorithm that includes learning rate as hyperparameter
- The implementation should mimick the behavior of sklearn models: include a method "fit" that accepts as input a matrix X of predictors and a vector y of targets and uses the gradient decent algorithm to find weights

that minimize the cost-function. The class should also contain a "predict" method that accepts as input a matrix X of predictors and returns as output a vector of predicted classes.

- Split the data into train and test parts and find the classification accuracy of your logistic regression algorithm. How can you check if the gradient decent method worked well? How did you choose the value of the learning rate?

4 Confidence interval of the prediction accuracy (2 points)

Using the dataset **winequality_binary.csv** find a 95% bootstrap-confidence interval for the the test classification accuracy score of the logistic regression classifier. (See course lectures for the procedure).