# Scientific Writing Techniques project

# “Description of Experiment and Results”

VITALII CHIBRIKOV

2333275007

For this project, I chose a dataset from the UC Irvine Machine Learning Repository [1]. I have analyzed data, prepared data for neural network learning, created a classification model, and analyzed the results. The dataset I have selected is free and good for educational purposes. Experience I gained doing the work will help me with real projects.

## Experiment

This section of the paper contains a description of the data, the process of data preparation, and a description of the model.

**Dataset**

For this work, I chose Mushroom dataset [2]. The data contains 8124 descriptions of physical characteristics of mushrooms from the Agaricus and Lepiota Family. The total number of characteristics is 22. Each characteristic is one value from a fixed number of variants described by one letter. The example of original data shown in Table 1.

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описаниеTable 1. Illustration of mushrooms’ data presented by letters.

For each mushroom, the data also contained information about its edibility. Edibility is represented by the two letters ‘e’ for edible mushrooms and ‘p’ for poisonous. An example of classification characteristics is shown in Table 2.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описаниеTable 2. Mushroom can only be edible or poisonous.

**Data preparation**

The original dataset contains missing values. To prepare the data for classification, missing values should be removed or substituted. The results of the missing value analysis are shown in Table 3.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описаниеTable 3. Number of missing values by column.

As one can see, the column “stalk-root” contains, 2480 missing values. It is a quarter of all values. So, it is better to remove the column completely than to substitute missing values.

The next step in data analysis is to check the balance of characteristics and categories. To check the distribution, the number of values in each column was calculated. Presence of different values in the data shown in Table 4.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описаниеTable 4. Values of mushrooms’ characteristics for each column. Sorted from most presented to least.

The column “veil-type” contains only one value: "p." No other values were presented in the research dataset. So, this column does not add new information to the model and can be removed.

The next step in data preparation was the transformation from categorical to numerical. Sklearn preprocessing LabelEncoder [3] was used for this purpose. Both for matrix X with features and for vector y with targets.

After conversion to numerical form, the data can be analyzed for outliers. An outlier is a data point that is noticeably different from the rest. They represent errors in measurement or bad data collection. Outliers can’t contribute to the learning process and should be removed from the dataset. For this purpose, Sklearn neighbors LocalOutlierFactor was used. It detected 26 rows with outlier data. They were removed.

In terms of the balance of data for classification, the selected dataset is slightly unbalanced. It contains more edible mushrooms than poisonous ones. To fix it, the Synthetic Minority Over-sampling Technique (SMOTE) was used. Chart 1 shows the number of edible and poisonous mushrooms before SMOTE and after it.

Изображение выглядит как текст, График, число, линия

Автоматически созданное описаниеChart 1. Poisonous and edible mushrooms before and after SMOTE.

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описаниеNext step of data preparation was Principal Component Analysis (PCA). By applying of PCA to the data the following explained variance achieved: [0.34396141 0.51313973 0.63646173 0.704509 0.76411744 0.81415453 0.85786678 0.89747943 0.9340666 0.9597921 0.97945844 0.98608793 0.9906492 0.99413315 0.99618324 0.99770705 0.99909567 0.99958698 0.99994359 1. ] and it is shown on Chart 2.

Chart 2. PCA explained variance.

As one can see, 12 main dimensions of data are enough for less than 2% data loss after dimension reduction. But as far as we speak about the prediction of poisonous mushrooms, false negative results of the model, even in 2% of cases, are a serious concern. So, we can’t discard data if the life and health of a person may be in danger. As a result, I decided to reduce the dimensions from 20 to 19. Discarding less than 0.0001 of the original information. This precision, as will be shown in the Result section of the paper, is correlated with the precision of the model.

The prepared data have dimensions X: [8400,19] and y: [8400,1] and are shown in Table 4.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описаниеTable 4. The dataset after preparation.

**Neural Network Model**

Изображение выглядит как текст, снимок экрана, Шрифт, алгебра

Автоматически созданное описаниеAs a model, I chose Sequential Neural Network [4] with two hidden layers and one Dropout layer between them. The schema of the model shown in Figure 1. This simple model is enough to show 100% Recall on the prepared dataset after relatively fast training. The importance of the Recall will be discussed in the Results section of the paper.

Figure 1. Model schema.

This schema was selected after several experiments with the model performance. I started with a model without hidden layers with 4 nodes in the input layer and started to increase the number of nodes . The model shown recall 100% with 100 nodes and 2201 parameters. I decided to reduce the number of parameters, introducing hidden layers. Model with input layer with 8 nodes and two hidden layers with 16 and 8 nodes, with 449 parameters in total shown the same result as model with one layer and 100 nodes. To reduce underfitting I increased the number of nodes to 16, 32, and 16. And to reduce potential overfitting, I added Dropout layer.

As loss function, I used BinaryCrossentropy [5] it shown better result than MeanSquaredError and MeanSquaredLogarithmicError.

As optimizer Adaptive Moment Estimation with learning\_rate=0.001 was used. Learning rates 0.01 and 0.0001 show worse results.

## Results

In this section of the paper, the results of the ML process are described. It consists of two parts: a description of the gathered metrics and an interpretation of the results.

**Metrics**

During the fitting process, metrics such as loss, precision, false negative number, and recall were measured. Loss (binary crossentropy) was used in the fitting process and represents the performance of the model. Precision shows the value of the mode to the final user. Recall and false negatives show the safety of the model. Even one false negative scenario may be very dangerous for a final user because it is a recommendation to eat poisonous mushroom. So, train parameters were tuned to show as high Recall as possible.

The result of fitting is shown on Chars 1-3.

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описаниеChart 1. Model binary crossentropy loss.

Chart 2. Model Precision.

Chart 3. Model Recall.

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание**Interpretation**

Neural network for prediction of poisonous mushrooms on base of description of their physical characteristics was developed. To use the model, one should input 20 parameters of evaluated mushroom. The result of identification could be obtained nearly immediately.

On base of Charts 1-3 one can see, that the model is fitting data. Selected parameters of the model are good for the task. The developed model shows 100% Recall and 100% Precision on the provided data. It can be used to evaluate the edibility of mushrooms with a very high probability of a correct result. The model was built with special care to Recall, so False Negative cases should be minimal. At the same time, the “price” of a wrong decision is too high (one can eat the poisonous mushroom, and it can lead to a fatal case), it should be treated as good advice to eat only mushrooms you know and discard any not-clear cases.

**References**

[1] <https://archive.ics.uci.edu>

[2] <https://archive.ics.uci.edu/dataset/73/mushroom>

[3] <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>

[4] <https://keras.io/guides/sequential_model/>

[5] <https://keras.io/api/losses/probabilistic_losses/>