Data Preparation For Wine Quality Prediction

Ivan Kosarevych
kosarevych@ucu.edu.ua

Introduction 1

Human error takes place in every activity a person participates in. Wine quality classification is not an exception. There is interest whether a machine can distinguish wines better than a human. Within this paper, I am touching mostly the part of dataset preparation for such experiment and in the end present a simple model for wine quality prediction.

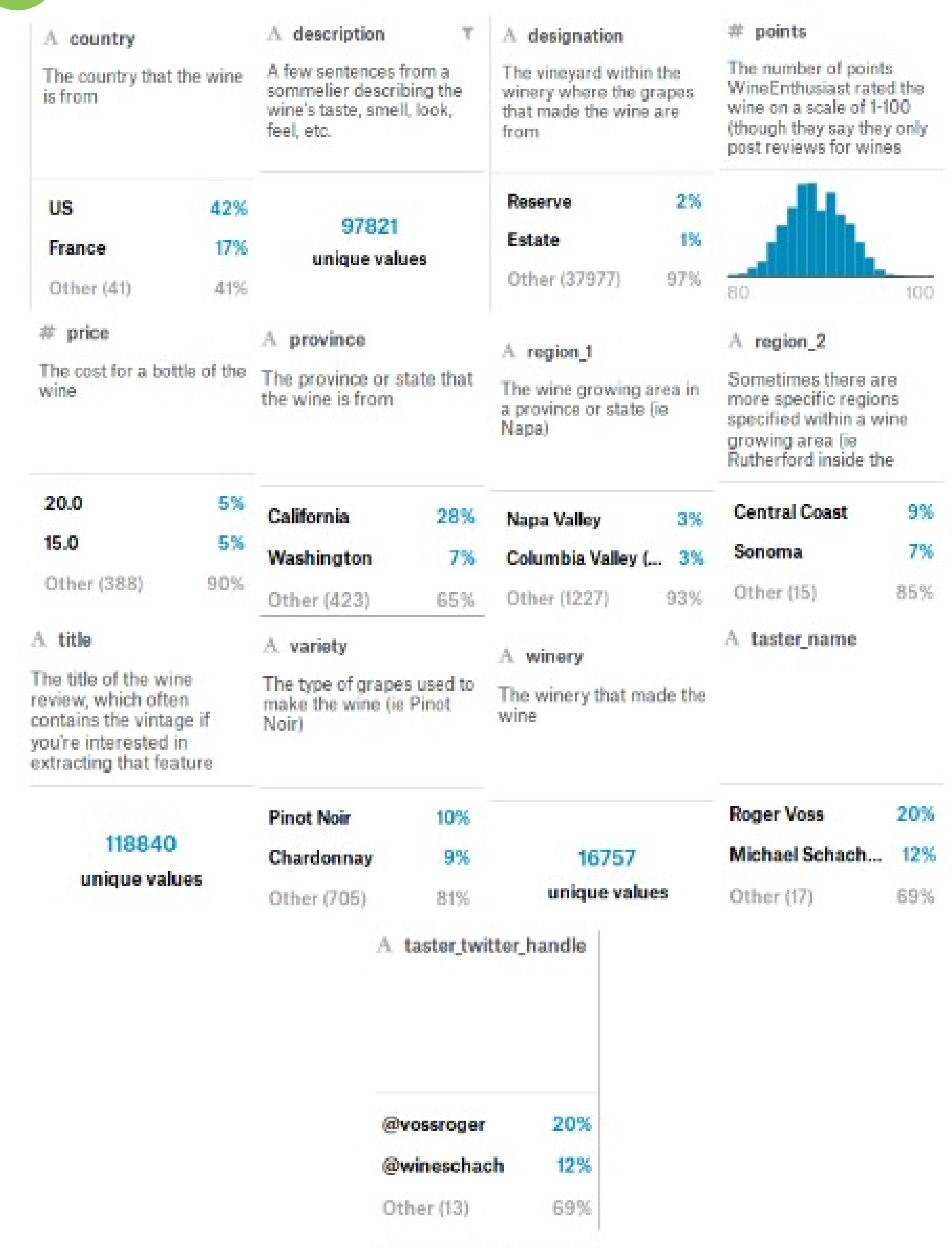
Methodology 2

The whole process consists of several stages:

- Data Cleaning
- Missing Values Imputation
- Data Transformation
- Outliers Investigation
- Data Normalization
- Dimension Reduction
- Model Selection

Experiment Setup 3

1 Data Parameters



A country
The country that the wine is from

US	42%
France	17%
Other (41)	41%

A description
A few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.

97821
unique values

A designation
The vineyard within the winery where the grapes that made the wine are from

Reserve	2%
Estate	1%
Other (37977)	97%

# points
The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines

# price
The cost for a bottle of the wine

20.0	5%
15.0	5%
Other (388)	90%

A province
The province or state that the wine is from

California	28%
Washington	7%
Other (423)	65%

A region_1
The wine growing area in a province or state (ie Napa)

Napa Valley	3%
Columbia Valley (...	3%
Other (1227)	93%

A region_2
Sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the

Central Coast	9%
Sonoma	7%
Other (15)	85%

A title
The title of the wine review, which often contains the vintage if you're interested in extracting that feature

118840
unique values

A variety
The type of grapes used to make the wine (ie Pinot Noir)

Pinot Noir	10%
Chardonnay	9%
Other (705)	81%

A winery
The winery that made the wine

16757
unique values

A taster_name

Roger Voss	20%
Michael Schach...	12%
Other (17)	69%

A taster_twitter_handle

@vossroger	20%
@wineschach	12%
Other (13)	69%

Figure 1. Data Set columns

In this data set '*points*' is a target column as a measure of wine quality

7 Model Selection

It is important to notice that the problem is a classification one with 100 classes. However, as our dataset contains only up to 20 classes (from 81 to 100) the more appropriate is to use one of the regression models. Meanwhile classification models are to be considered as well.

The following models were studied: Linear Regression, Logistic Regression, SVM Classifier and MLP Regressor.

2 Data Cleaning

Columns '*taster name*', '*taster twitter handle*', '*title*', '*description*' were dropped. They were considered as not connected with a target column.

3 Missing Values Imputation

First approach is **global most common substitution**. Missings in categorical columns were filled with most frequent value(mode). Missed values in numerical columns were imputed with average of this column.

The second method is **K-Nearest Neighbours** Classifier. The columns of the data set were one by one imputed using KNN. Other columns in order to train the classifier were imputed using global most common substitution.

Column	Missing values
country	63
designation	37465
price	8996
province	63
region 1	21247
region 2	79460
variety	1

Table 1. Number of missing values by columns

4 Data Transformation

For every categorical column all its unique values were obtained. For every such value a unique integer identifier was assigned. Then every value in the column was replaced with its identifier.

5 Outliers Detection

Data has no significant number of outliers. However values in most of columns are very similar.

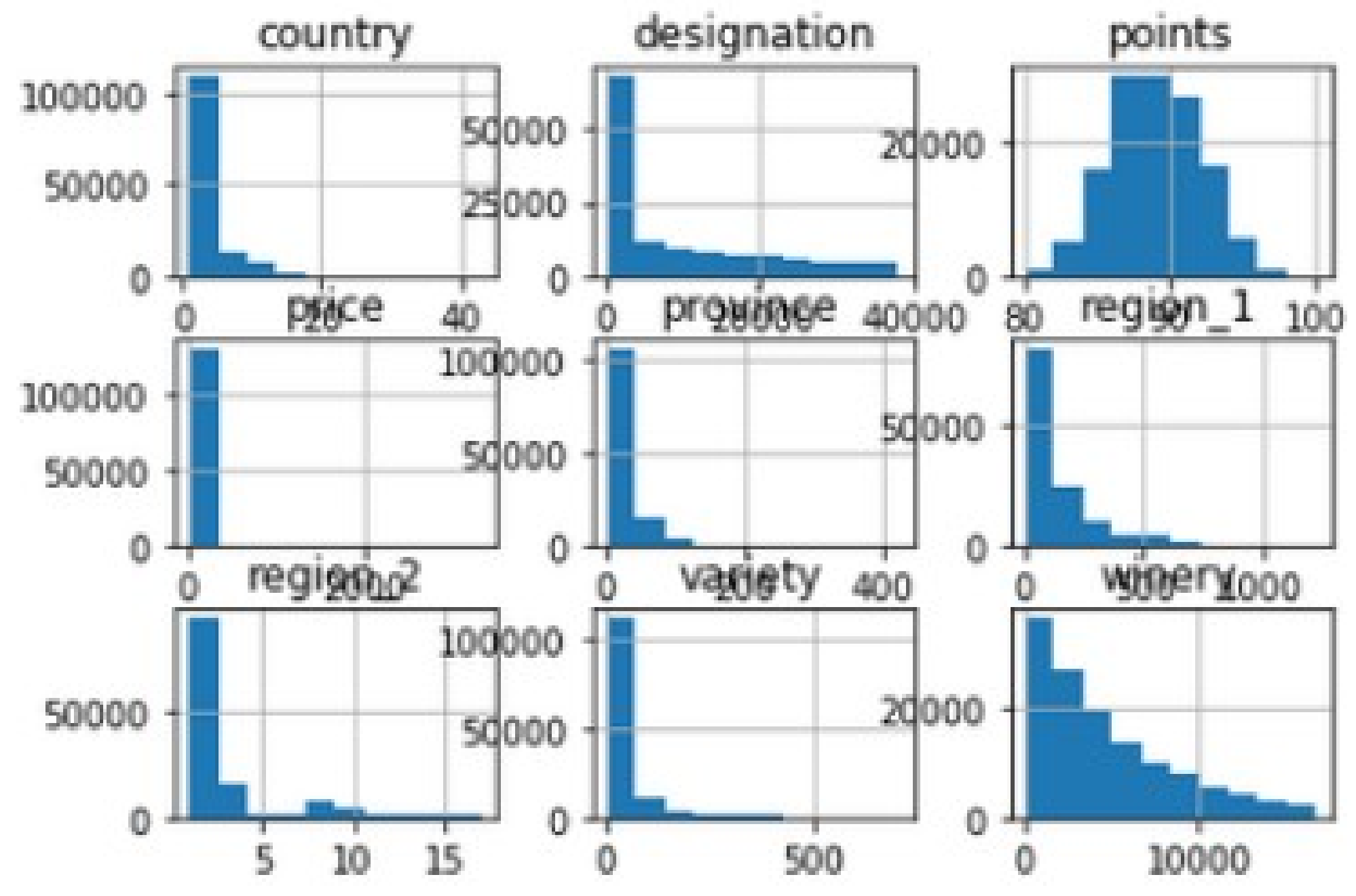


Figure 4. Column histograms

6 Dimension reduction

In order to accomplish dimension and *similarity* reduction I considered two ways. The quality of the reduction was measured by a simple Linear Regression.

Principal Component Analysis is a basic approach and was considered at first.

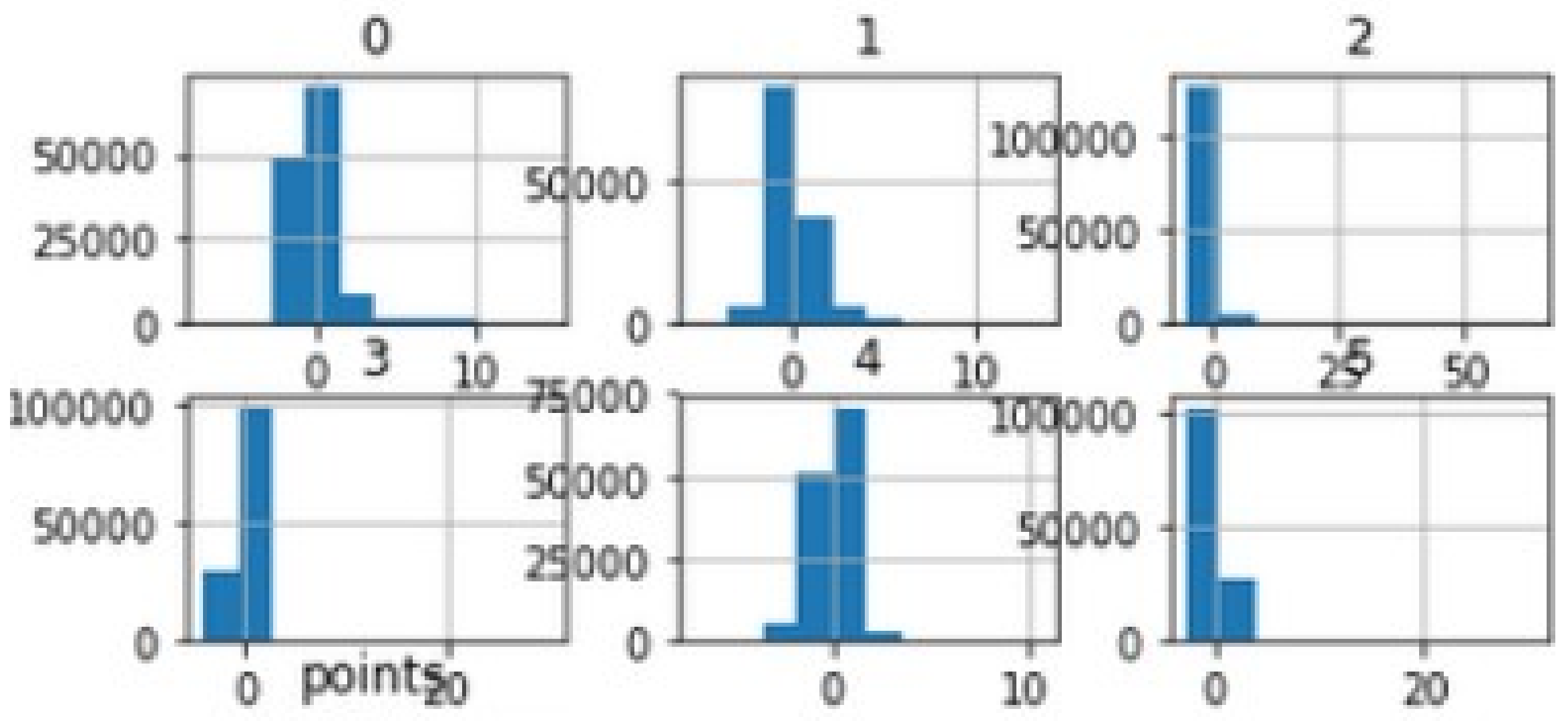


Figure 6. PCA reduced data

The second way is **Decision Trees** Classifier with 50 estimators. With this model I also found how each feature is important.

Results 4

Approach	Train Accuracy(%)	Test Accuracy(%)
Global most common substitution	18.01	17.18
K-Nearest Neighbours	17.57	16.70

Table 3. Missing value imputation prediction accuracy

Approach	Time performance
Global most common substitution	0.35 sec.
K-Nearest Neighbours	> 1hr.

Table 4. Missing value imputation time performance

Approach	Train Accuracy(%)	Test Accuracy(%)
PCA	17.83	17.04
DT	17.84	16.93

Table 5. Dimension reduction prediction accuracy

Approach	Time performance
PCA	0.7816 sec.
DT	> 1 min.

Table 6. Dimension reduction time performance

Approach	Train Accuracy(%)	Test Accuracy(%)
Linear Regression	17.83	17.04
Logistic Regression	15.66	15.54
SVM Classifier	17.71	16.93
MLP Regressor	35.68	35.81

Table 7. Selected model prediction accuracy

Approach	Time performance
Linear Regression	1.28 sec.
Logistic Regression	> 26.39 sec.
SVM	> 30 min.
MLP Regressor	171.18 sec.

Table 8. Dimension reduction time performance

Code on GitHub:

Conclusions 5

- For missing value imputation stage Global Most Common Substitution works better than K-Nearest Neighbours approach
- PCA successfully reduces similarity and dimension of data and does it better than Decision Trees.
- Neural network model overperformed others in prediction accuracy.
- More complex approaches can be applied on the every stage. Other data sets can be found, studied and possibly merged with this one. Own multilayer neural network can be composed and trained using more advanced technologies.

Applied Sciences Faculty, Ukrainian Catholic University, Lviv, Ukraine 2019