Изготвили:

Сияна славова, 24963

Иван Капукаранов, 24958

Проект за курса „Откриване на знания в данни“

Задача: wine quality classification

# Декларация за липса на плагиатство

1. Тази курсова работа е моя работа, като всички изречения, илюстрации и програми от други хора са изрично цитирани.

2. Тази курсова работа или нейна версия не са представени в друг университет или друга учебна институция.

3. Разбирам, че ако се установи плагиатство в работата ми ще получа оценка “Слаб”.

Иван Капукаранов, ФН: 24958, 1 курс, ИИОЗ

Сияна Славова, ФН: 24963, 1 курс, ИИОЗ

Съдържание

[Декларация за липса на плагиатство 1](#_Toc454991495)

[Мотивация, Задача на курсовата работа 3](#_Toc454991496)

[Съществуващо решение 3](#_Toc454991497)

[Избрано решение 3](#_Toc454991498)

[Програмна реализация 3](#_Toc454991499)

[Предварителна обработка 3](#_Toc454991500)

[Данни 3](#_Toc454991501)

[Реализация 3](#_Toc454991502)

[Резултати от експерименти 3](#_Toc454991503)

[Заключение и бъдещо развитие 3](#_Toc454991504)

[Разпределение на задачите 3](#_Toc454991505)

[Литература и използвани източници 3](#_Toc454991506)

# Мотивация, Задача на курсовата работа

Данните представлява две множества с информация за химическия състав на различни видове вино. Едното множество е за бяло вино, а другото - за червено. Един екземпляр се определя със следните атрибути:

fixed acidity

volatile acidity

citric acid

residual sugar

chlorides

free sulfur dioxide

total sulfur dioxide

density

pH

sulphates

alcohol

Целта е да се класифицира виното според неговото качество. Категориите за качество са общо 10, но реално и в двете множества данните са с качество между 3 и 9.

# Съществуващо решение

В намереното [съществуващо решение](http://www.r-bloggers.com/predicting-wine-quality-using-random-forests/) е използван алгоритъмът Random forest. Той е приложен само върху White Wine dataset.

Като предварителна обработка качеството на виното се разпределя в три категории:

* Лошо (с качеството, по – малко от 6)
* Нормално (с качество, което е равно на 6)
* Добро (с качество, по – голямо от 6)

Тази предварителна обработка е направена поради липсата на най – много обучаващи данни с качество 6 и доста по – малко в другите категории.

В това решение ефективността на алгоритъма е измерена като accuracy, която е приблизително 71,5%.

# Избрано решение

Ние избрахме да сравним представянето на два алгоритъма върху така поставената задача, а именно random forest и neural network. За реализация използвахме [R Project](https://www.r-project.org/). За разлика от предварителното разделяне в три категории на качеството в съществуващото решение, ние решихме да работим с оригиналните степени на качество от 1 до 10.

# 

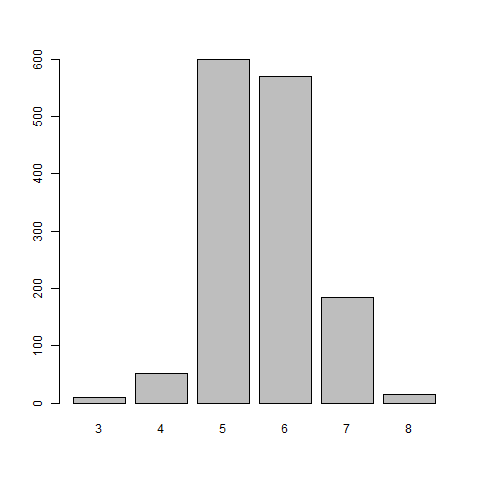
# Програмна реализация

## Предварителна обработка

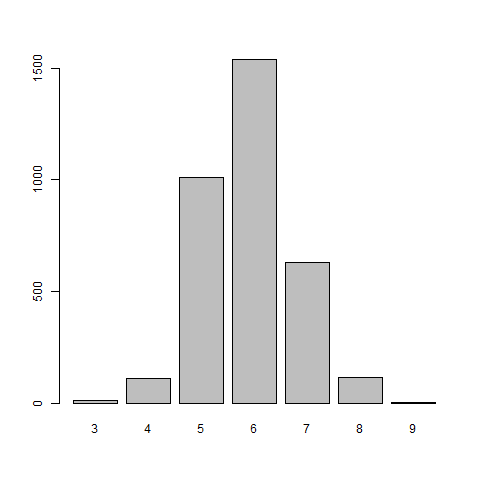
Тъй като множествата, които използвахме за задачата не бяха разделени на тестово и обучаващо, ние отделихме 10% от всяко едно от тях за тестване и 90% за обучение.

## Анализ на данните

Както се вижда от двете графики на разпределенията (Фигура 1 и Фигура 2), данните и за бяло и за червено вино са нормално разпределени със сравнително малко стандартно отклонение, което води на струпване на данни за средна степен на качество (степен 5 или степен 6). Вижда се също, че има много малко на брой обучаващи данни за екстремни степени на качество (например много ниско или много високо). От графиките се вижда и че реално степените на качество в множествата са от порядъка на 3 до 9.

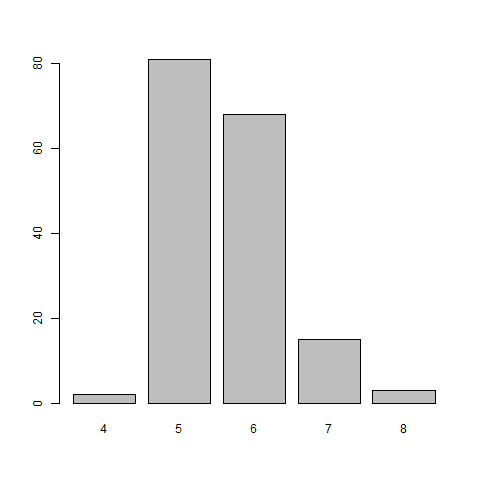


Фигура 1:Разпределение обучаващо множество – червено вино

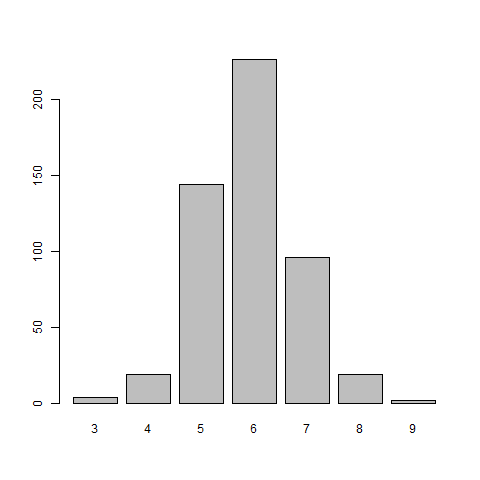


Фигура 2:Разпределение обучаващо множество– бяло вино

Ситуацията е подобна и при разпределението на тестовите множества (Фигура 3 и Фигура 4).



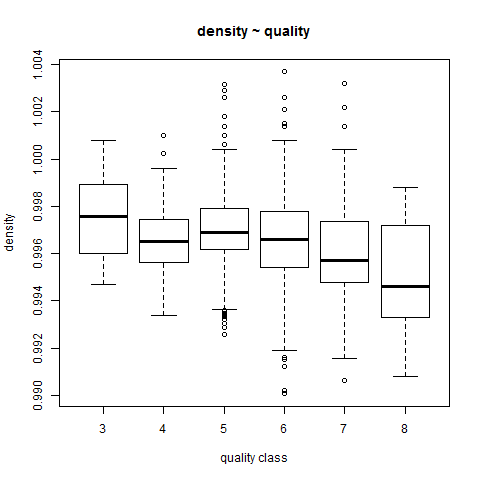
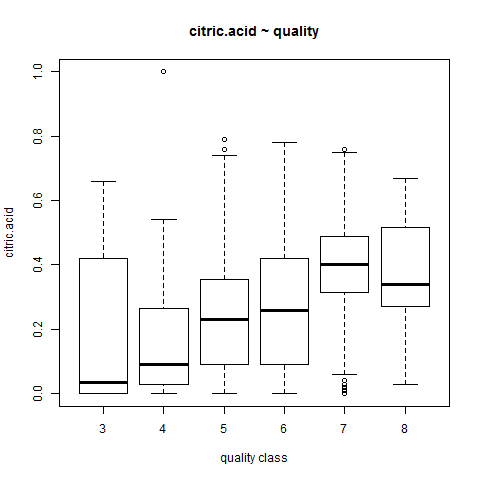
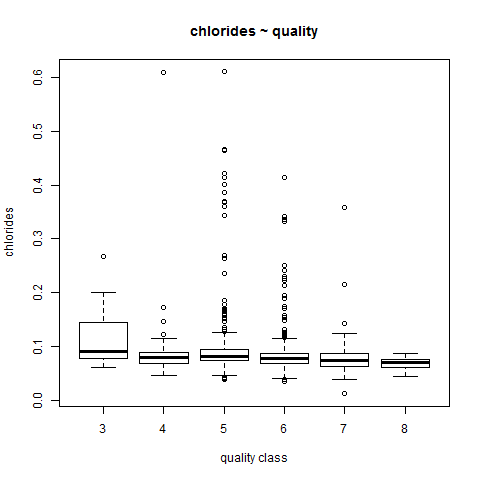
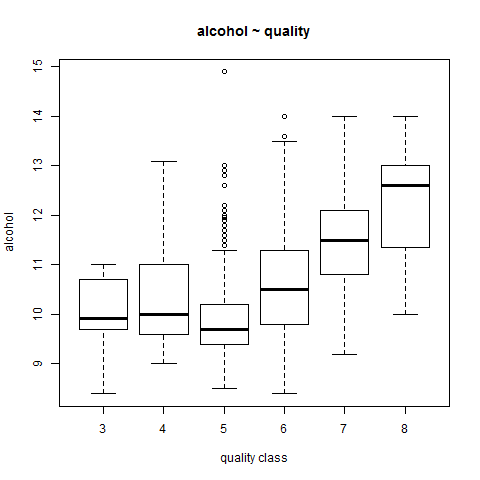
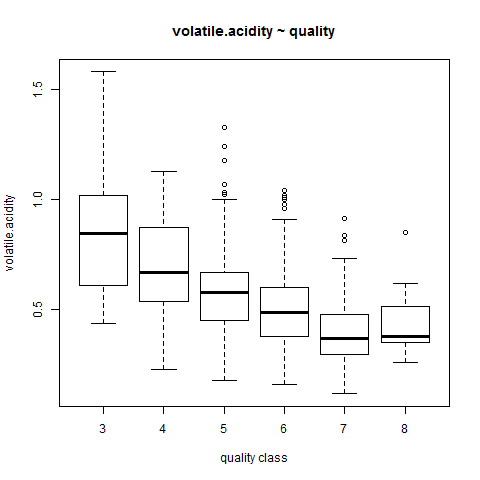
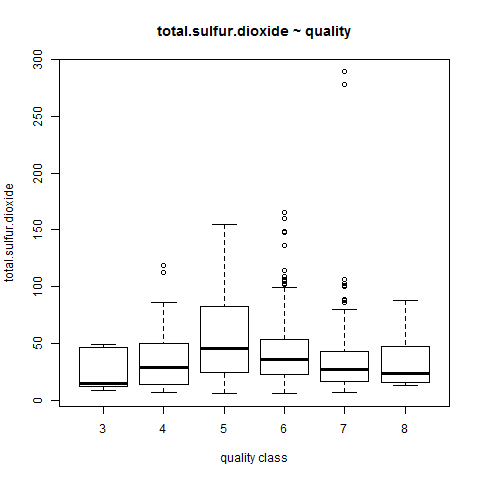
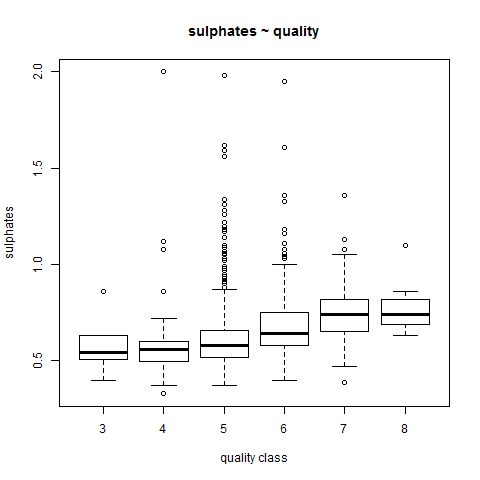
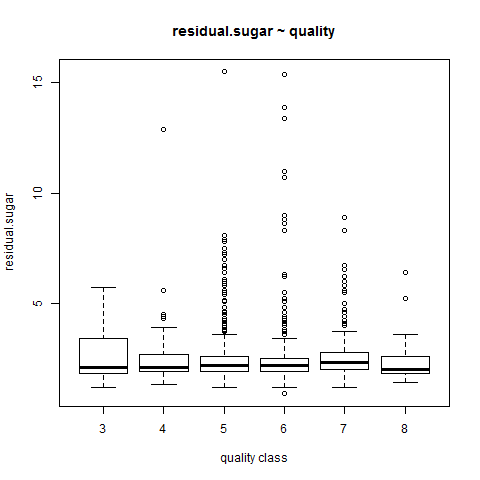
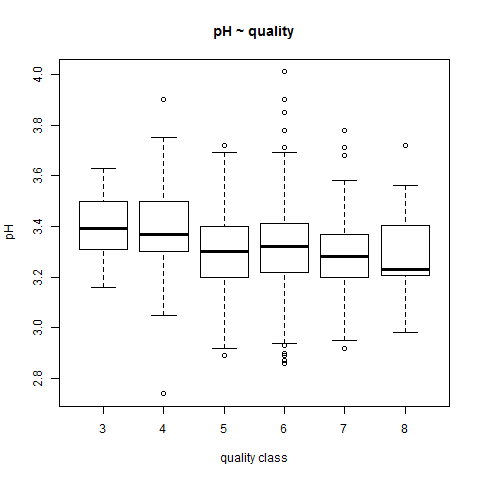
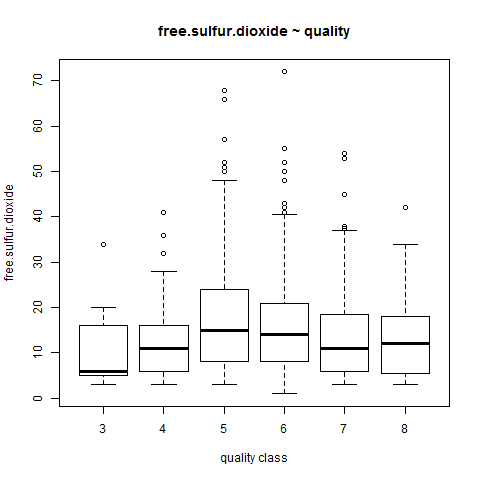
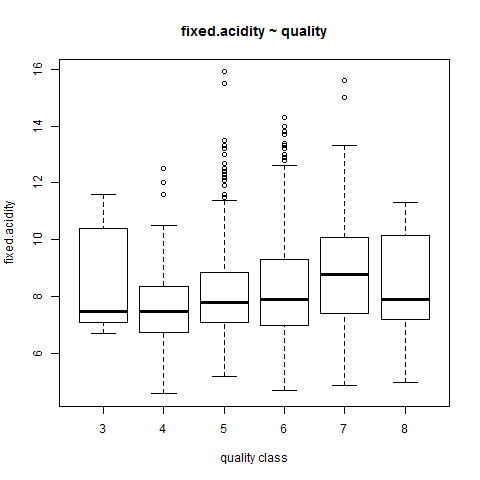
Фигура 3: Разпределение тестово множество - червено вино



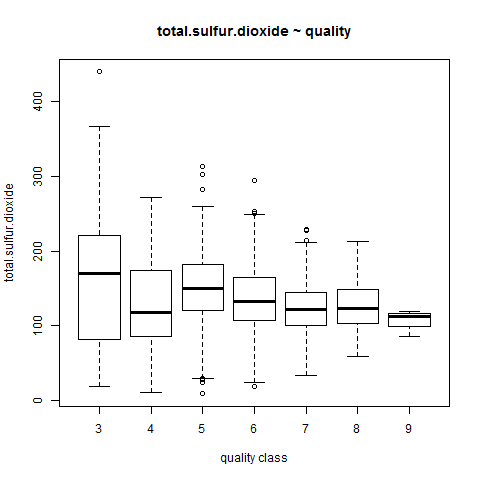
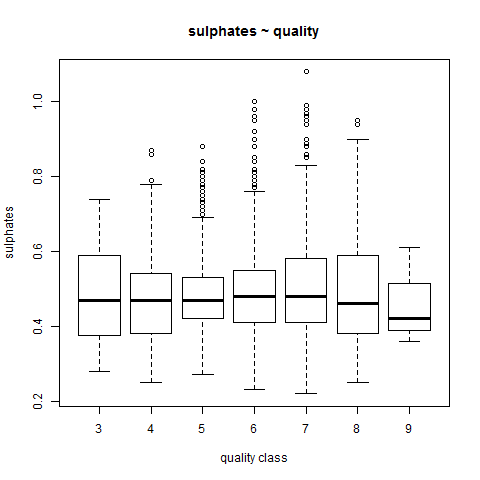
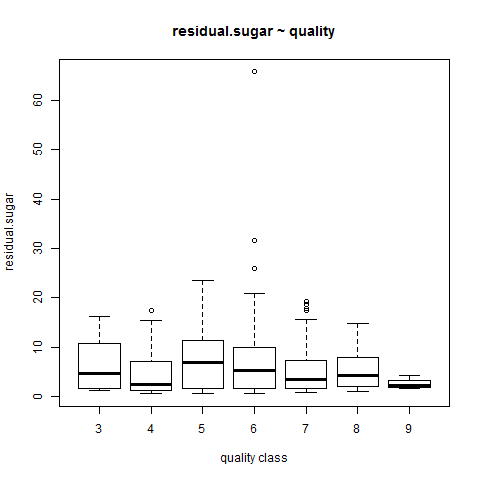
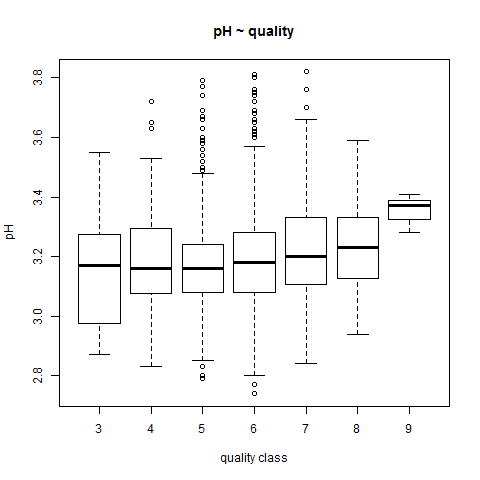
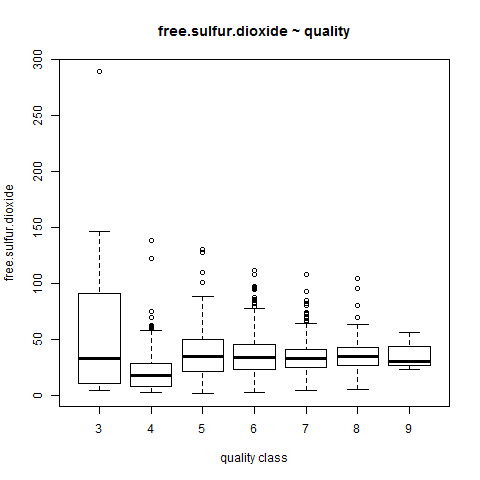
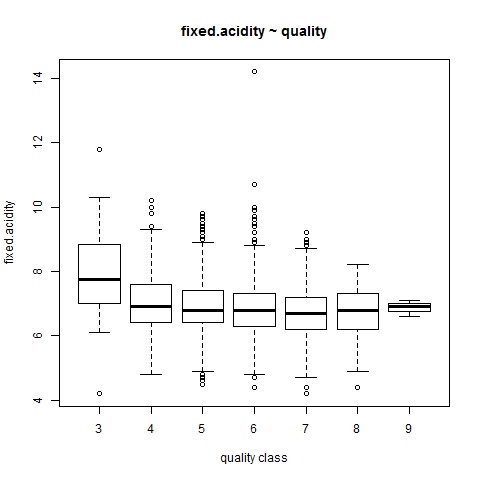
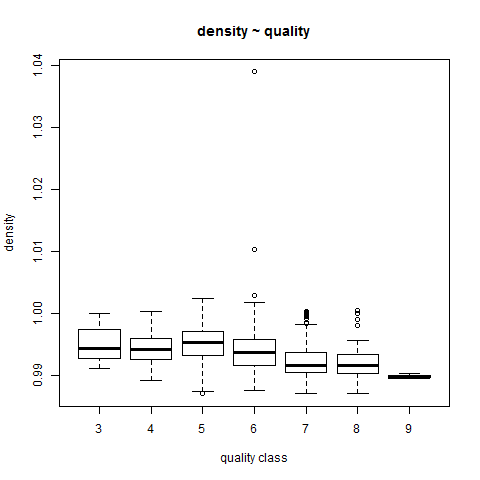
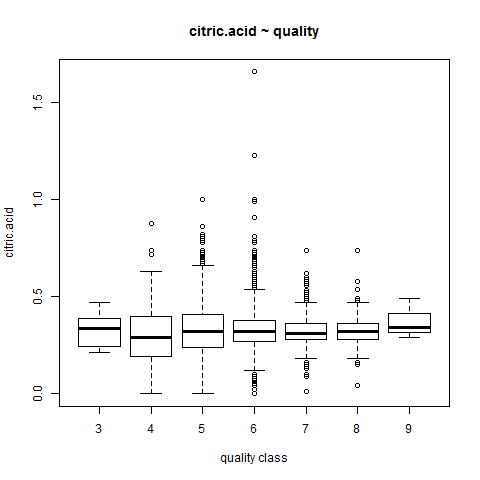
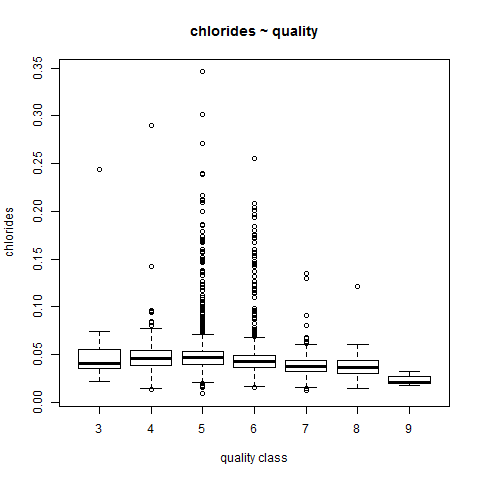
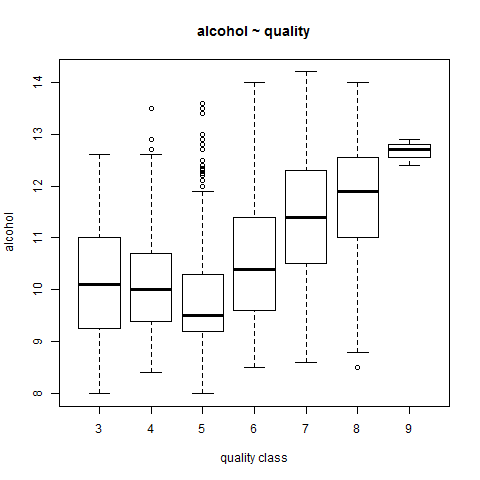
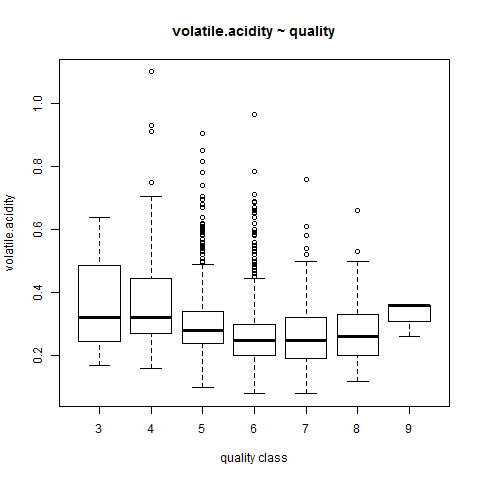
Фигура 4: Разпределение тестово множество - бяло вино

При анализа на данните установихме още, че не всички атрибути характеризират достатъчно добре всеки един клас вино. Имаше доста атрибути, които почти не се различаваха по стойности за различните степени на качество. Имаше и такива, при които се наблюдават екстремни стойности (outliers).

### Червено вино (Атрибут ~ Качество)



### Бяло вино (Атрибут ~ Качество)



## Реализация

За програмната реализация използвахме R Project. Използвахме следните два алгоритъма:

* Random forest ([R имплементация](https://cran.r-project.org/web/packages/randomForest/randomForest.pdf))
  + Направихме експерименти с различен брой дървета в интервала [10,100]
* Neural Networks ([R имплементация](https://cran.r-project.org/web/packages/RSNNS/RSNNS.pdf))
  + Използвахме персептрон с един скрит слой. Броят неврони в скрития слой е 100. Броят максимални итерации за обучение е 1000. За първоначални тегла са използвани случайно избрани. Обучаващата функция е backpropagation.
  + Направихме експерименти с различна стойност на параметъра на функцията за обучение, като използвахме стойности в интервала [0,1] с промяна в стотните.

# Резултати от експерименти

Пишеш видове екперименти – обучаващо бяло, тестово червено; об бяло, тестово бяло; об миксирано, тестово…

Иван

# Заключение и бъдещо развитие

Иван

# Разпределение на задачите

Иван

# Литература и използвани източници

* Wine dataset:
  + <https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/>
* Съществуващо решение:
  + <http://www.r-bloggers.com/predicting-wine-quality-using-random-forests/>
* Random forest algorithm:
  + <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>
* RSNNS – neural networks algorithm
  + <https://cran.r-project.org/web/packages/RSNNS/RSNNS.pdf>
  + <https://www.jstatsoft.org/article/view/v046i07>
* R Project
  + <https://www.r-project.org/>