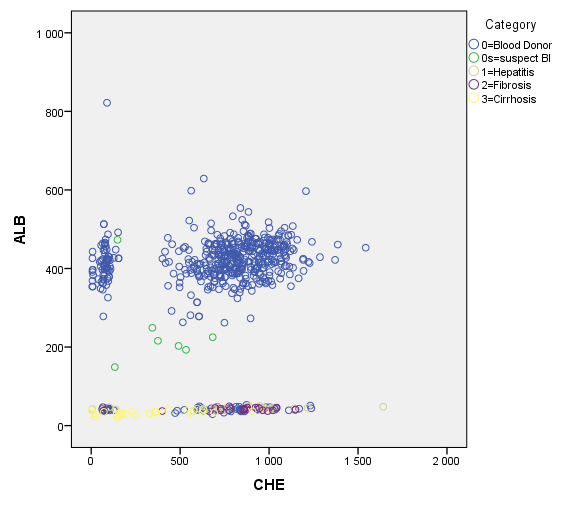
Анализ строится на основе диаграмм рассеяния. Результат анализа – выбор параметров для кластеризации выборки.

В первую очередь стоит рассмотреть комбинации параметров, выдавших бимодальное распределение: ALB, CHE, CHOL, PROT. Под вопросом ALP. Параметры, распределенные мономодально и приближены к нормальному распределению: ALT, AST, BIL, CREA, GGT.

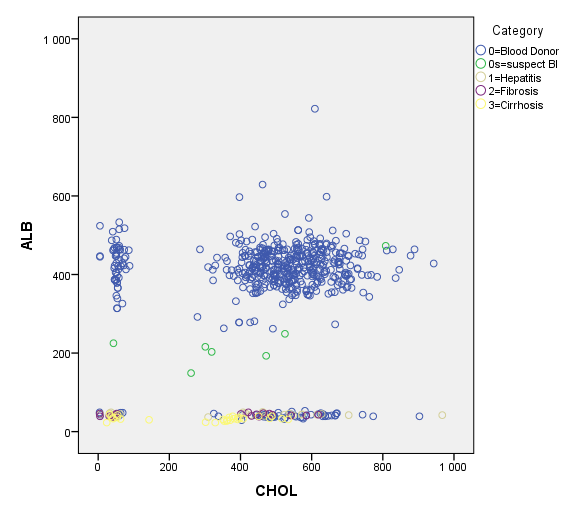
**Анализ в 2D**

ALB – CHE



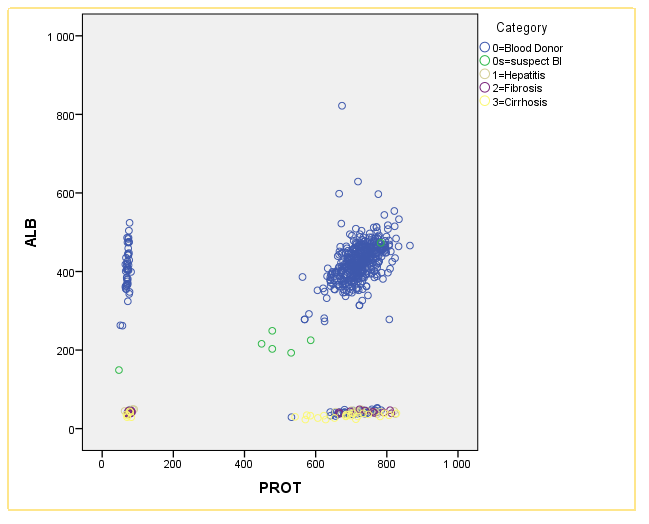
ALB довольно много здоровых (синий) и подозрительных (зеленый) отделил от больных (желтый, фиолетовый, серый), но недостаточно хорошо. CHE **дает не особо много информации**, так как и здоровые, и больные встречаются по всей оси значений.

ALB-CHOL



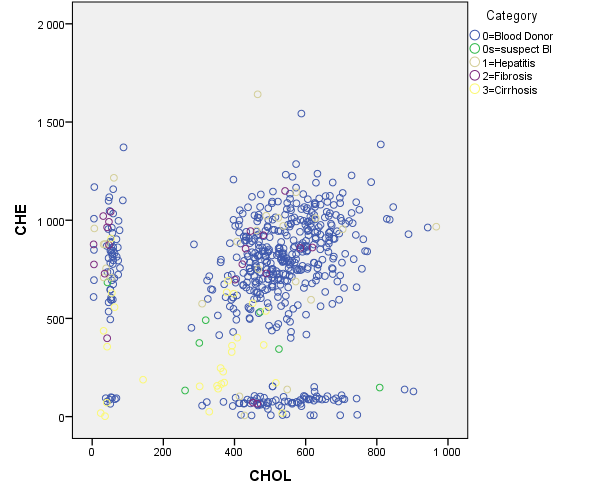
Ситуация аналогична предыдущей, однако конфигурация при малых значениях ALB заметно поменялась. Было бы хорошо посмотреть на комбинацию параметров CHE-CHOL.

ALB-PROT



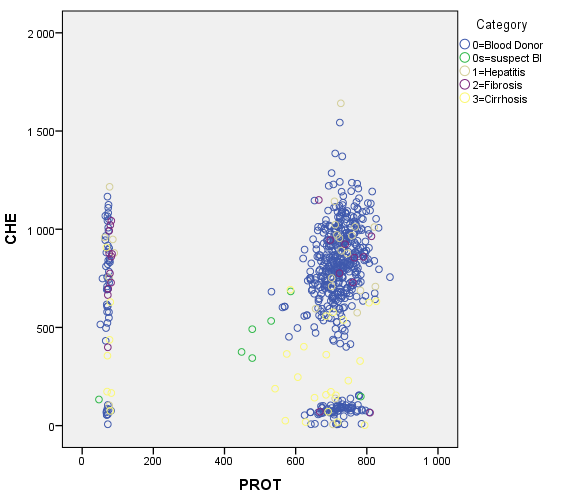
Здесь уже диаграмма получилась более интересной. Однозначно выделяется группа больных около нуля по обоим осям координат. Тем не менее, при высоких значениях PROT и низких значениях ALB наблюдается смешение здоровых и больных пациентов. Следует проверить разделение на больших значениях PROT с другими параметрами.

CHE-CHOL



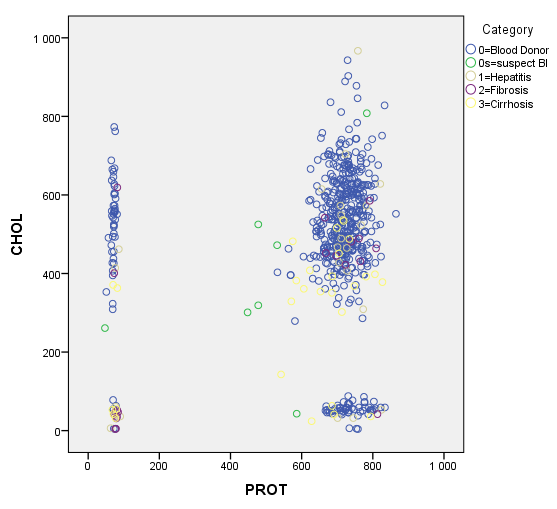
Плохой результат. Больные в больших количествах перемешались со здоровыми. Данная пара вряд ли может принести пользу для анализа. Вероятно, параметр CHE не дает нам никакой существенной информации на этой диаграмме.

CHE-PROT



Аналогичная предыдущей картина. Больные смешались со здоровыми и параметр CHE никак не повлиял на результат. Хотя и параметр PROT сильно перемешивает группы пациентов.

CHOL-PROT



Картина непонятная. Опять же из-за смешения больных и здоровых пациентов нельзя однозначно сказать, что здесь происходит.

**Вывод по парам параметров с бимодальным распределением**

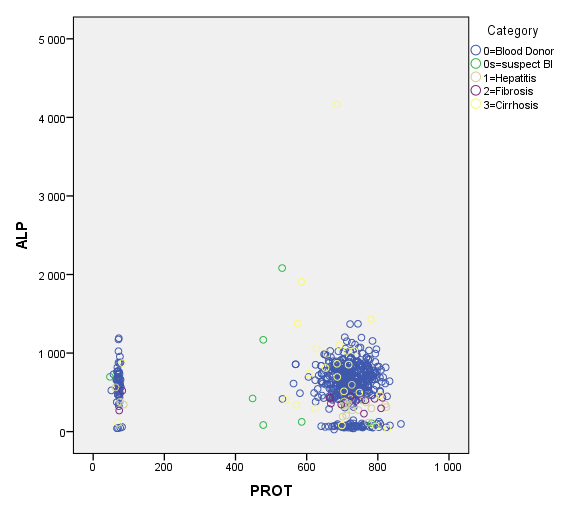
По полученным графикам однозначно можно сказать, что параметр ALB является критически важным при поиске людей, больных гепатитом, циррозом и фиброзом. В паре с параметром PROT он позволяет довольно точно выделить по крайней мере одну группу больных. Если удастся найти параметр, который на больших значениях PROT (>500) однозначно разделит группы пациентов на больных и здоровых – получится построить очень точную модель кластеризации.

**Поиск парного параметра для PROT**

Из неисследованных параметров остались ALP, ALT, AST, BIL, CREA, GGT.

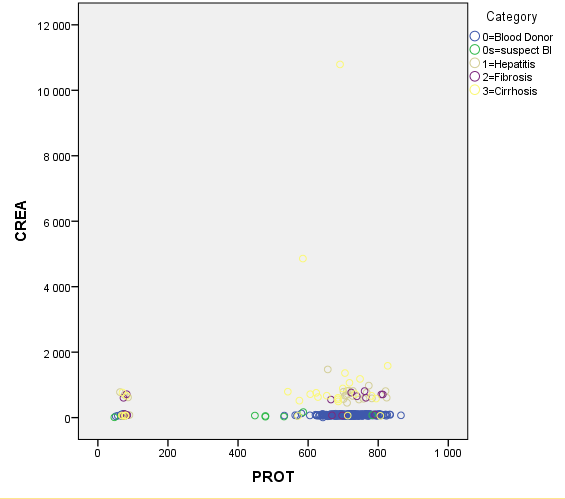
Проведем анализ с помощью диаграмм рассеяния на связь с параметром PROT.

ALP-PROT



Гипотетически может очень хорошо разбросить значения. Но не факт.

CREA-PROT



Выглядит очень многообещающе. Многие пациенты при уровне PROT > 500 показывают хорошее разделение на больных и здоровых. Потенциально может служить третьим параметром для выделения кластеров из выборки.

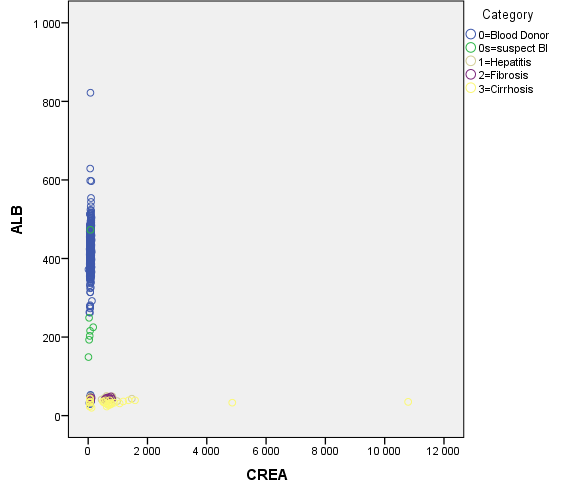
Остальные параметры в паре с PROT не показали никаких выдающихся результатов. Группы пациентов оказывались перемешаны между собой.

Сейчас можно выделить **тройки параметров**, которые можно считать на данном этапе **наиболее перспективными** в решении задачи кластеризации:

* ALB-PROT-CREA
* ALB-PROT-ALP

**Оставшиеся пары параметров**

ALB-CREA



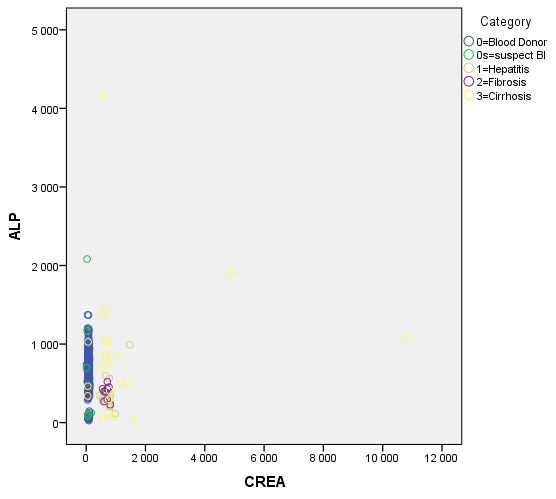
В подтверждение сказанного ранее приведена диаграмма рассеяния ALB-CREA. Здесь отчетливо видно, как у большой группы больных пациентов значение параметра CREA повышено в сравнении со здоровыми пациентами и малой частью больных. Этот параметр однозначно должен участвовать в построении модели.

Остальные параметры в паре с параметром ALB не принесли какой-либо полезной информации на диаграммах рассеяния. Следовательно, из перспективной тройки параметров удаляется ALB-PROT-ALP.

**Поиск парных параметров для CREA**

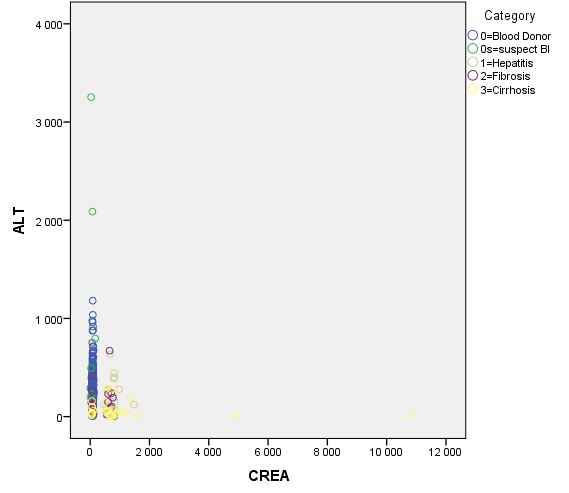
Так как параметр CREA хорошо показал себя в связке с параметрами ALB и PROT, то следует проверить остальные параметры и постараться выделить дополнительные критерии. В случае успеха модель будет построена более точно.

ALP-CREA



Здесь плохо видно, но похоже, что околонулевые значения параметров ALP и CREA позволяют утверждать о здоровье пациента. То есть параметр ALP может помочь отсеять больных от здоровых при околонулевых значениях CREA. Картина, полученная на диаграмме рассеяния при ALB-CREA должна показывать большее разделение между больными и здоровыми.

ALT-CREA



Потенциально может помочь разделить больных и здоровых, но шансы очень малы.

Остальные параметры оказались малоинформативными и пользы в построении модели не принесут.

**Вывод по 2D анализу компонентов для кластеризации**

В ходе анализа были получены следующие комбинации параметров:

* ALB-PROT-CREA-ALP
* ALB-PROT-CREA-ALT
* ALB-PROT-CREA

Данные комбинации выглядят наиболее перспективными для построения. Сначала стоит попробовать построить кластеры на основе этих параметров. Если создать достаточно точную модель не получится, то стоит провести для этих параметров 3D анализ.

**Кластеризация методом k-средних**

ALB-PROT-CREA

Беглый анализ полученных результатов при указании 4 и 5 кластеров дает неудовлетворительный результат. Получается так, что здоровые оказываются в одной группе с больными. Если же проводить анализ на выделение 2 кластеров, то окажется, что и больные, и здоровые пациенты находятся в одном кластере.

ALB-PROT-CREA-ALP

Беглый анализ также показывает неудовлетворительный результат, как и в прошлом случае. Для задачи на выделение 2, 4 и 5 кластеров больные и здоровые оказываются в одной и той же группе. Разделения не происходит. Вероятно, метод k-средних не подходит для выборки.

ALB-PROT-CREA-ALT

Визуальный анализ результата показал, что пара выбросов в данных перетягивает на себя кластеры целиком. Поэтому метод k-средних не подходит под выборку.

**Иерархический кластерный анализ**

Сразу стоит сделать пояснение: в данном методе для всех проведенных анализов использовалась стандартизация данных к диапазону от 0 до 1.

ALB-PROT-CREA

Анализ c использованием метода связи Уорда и Евклидовой метрики показал достаточно хорошее выделение двух кластеров здоровых пациентов. Два других кластера содержат всех больных пациентов и крайне малое количество здоровых- около 20 человек были ошибочно приняты за больных. Причем в кластеры со здоровыми пациентами попали только здоровые и находящиеся под подозрением. Данный результат при первичной оценке можно считать удовлетворительным.

При использовании метода центроидной кластеризации метод начинает показывать не такие хорошие результаты и, судя по всему, некоторые выбросы начинает воспринимать как отдельные кластеры. Несмотря на это, полученный результат не сильно отличается от предыдущего: кластеры здоровых пациентов не включают в себя ни одного больного, однако в кластеры с больными пациентами попадает небольшое количество здоровых пациентов, около 20 человек.

ALB-PROT-CREA-ALP

Добавление параметра ALP к исследованной ранее тройке не принесло существенной пользы – в кластерах с больными пациентами все еще появляются здоровые. Однако при использовании метода центроидной кластеризации удалось выделить в отдельные кластеры пациентов с подозрением на заболевания. В кластерах, где здоровых пациентов больше половины больные пациенты не попадают. Этот результат также можно считать удовлетворительным.

ALB-PROT-CREA-ALT

Добавление параметра ALT к анализу с использованием метода Уорда позволило выделить 4 крупных кластера пациентов. Причем 2 кластера целиком состояли из здоровых пациентов и пациентов с подозрением на болезни, один кластер состоял из больных пациентов за исключением одного здорового и один кластер, в котором ошибочно оказалось около 15 пациентов.

Используя метод центроидной кластеризации был получен результат, аналогичный полученному с помощью метода Уорда, однако теперь в отдельные кластеры выделились пациенты с подозрением на болезни.

Полученные результаты также можно считать удовлетворительными.

**Кластерный анализ с помощью нейросетей**

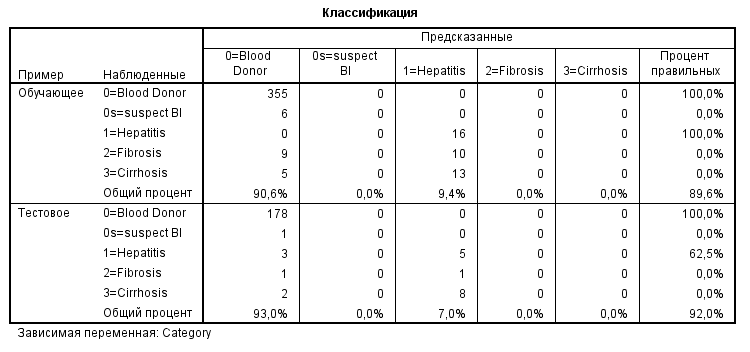
В SPSS реализована возможность выделения кластеров с помощью нейросетей. Работает это следующим образом: вся выборка делится на обучающую и тестовую. Обучающая необходима для настройки весовых коэффициентов внутри нейросети. Тестовая часть выборки требуется для выявления точности решаемой задачи на данных, которые ранее в нейросеть не подавались. В моем случае выборка случайным образом разбивалась на обучающую и тестовую, причем в обучающую попадало 70% от исходной выборки, а оставшиеся 30% в тестовую. Также стоит упомянуть про архитектуры нейросетей, которые были использованы при решении задачи с различными параметрами. Для того, чтобы все наборы параметров были в одинаковых условиях я зафиксировал следующие варианты структуры нейросети:

* Один скрытый слой с 1 нейроном, функция активации скрытого слоя – гиперболический тангенс, функция активации выходного слоя – тождественная
* Один скрытый слой с 6 нейронами, функция активации скрытого слоя – гиперболический тангенс, функция активации выходного слоя – тождественная
* Два скрытых слоя с 6 нейронами в каждом, функция активации скрытого слоя – гиперболический тангенс, функция активации выходного слоя – тождественная
* Два скрытых слоя с 6 нейронами в каждом, функция активации скрытого слоя – гиперболический тангенс, функция активации выходного слоя – softmax
* Два скрытых слоя с 6 нейронами в каждом, функция активации скрытого слоя – сигмоида, функция активации выходного слоя – softmax

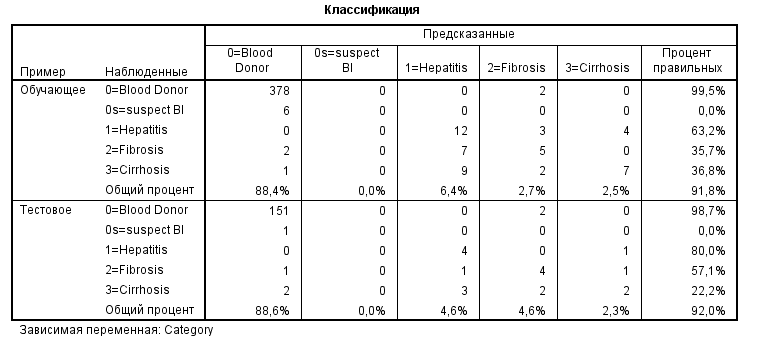
Далее приведены результаты кластеризации для выбранных наборов параметров выборки.

ALB-PROT-CREA

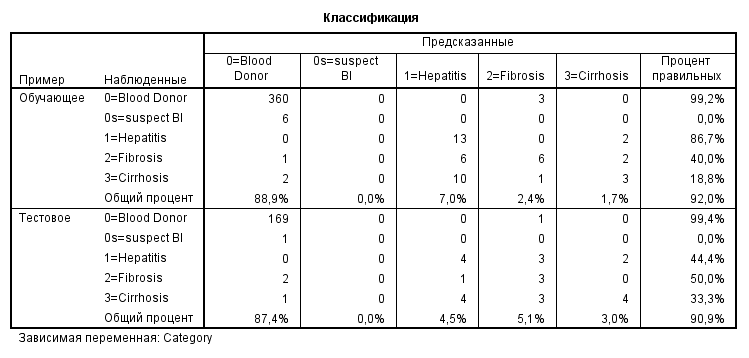
Первый эксперимент:



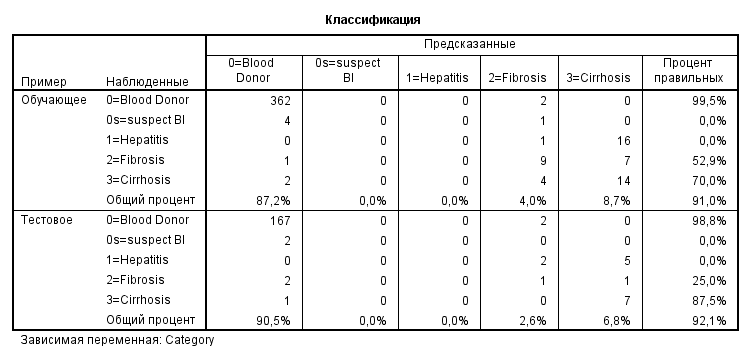
Второй эксперимент:



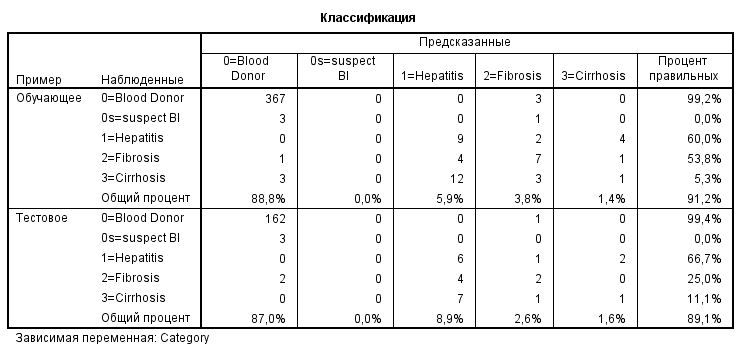
Третий эксперимент:



Четвертый эксперимент:



Пятый эксперимент:

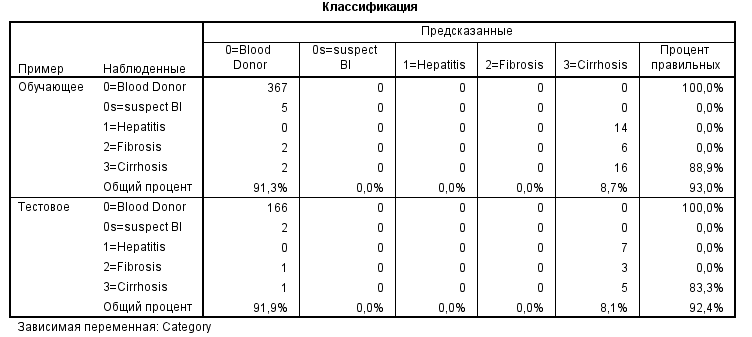


Результаты кластеризации показывают точность более 89% во всех пяти экспериментах. Такой высокий результат можно объяснить тем, что в выборке большинство пациентов является здоровыми. Если посмотреть на процентное соотношение здоровых пациентов к остальным, то можно заметить, что их более 85%. Это значит, что если бы нейросеть всех пациентов помечала как здоровых, то она бы выдавала более 85% правильных ответов. Нейросеть даже без обучения может выдавать очень хорошие в процентном соотношении результаты. Следовательно, в оценке экспериментов десятая доля процента точности может оказаться существенной.

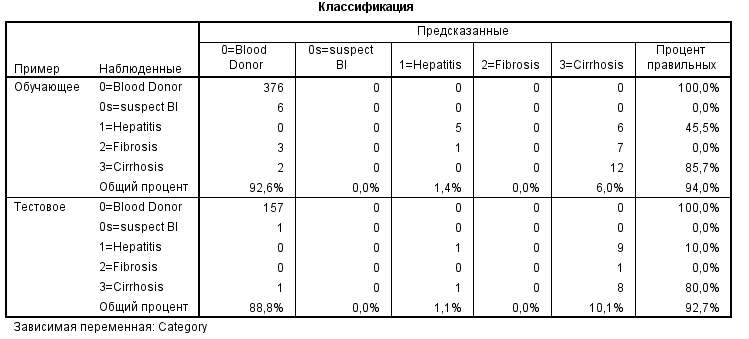
Наиболее точные результаты (более 92%) для набора ALB-PROT-CREA были получены с помощью нейросетей, использующих один скрытый слой, а также нейросетью с двумя скрытыми слоями, использующей функцию активации выходного слоя softmax и функцию активации скрытых слоев – гиперболический тангенс.

ALB-PROT-CREA-ALP

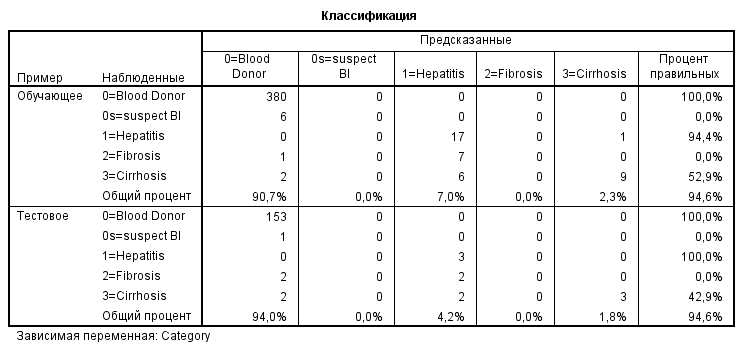
Первый эксперимент:



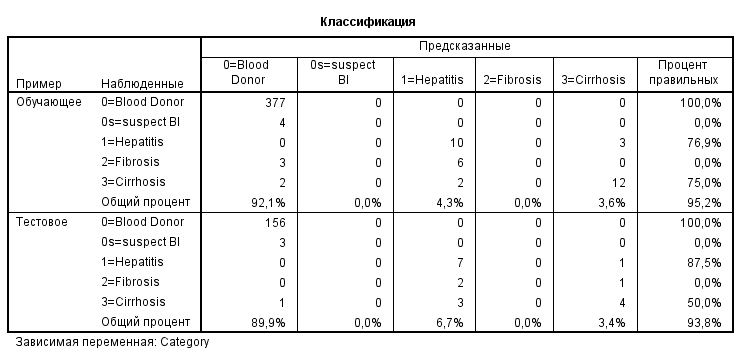
Второй эксперимент:



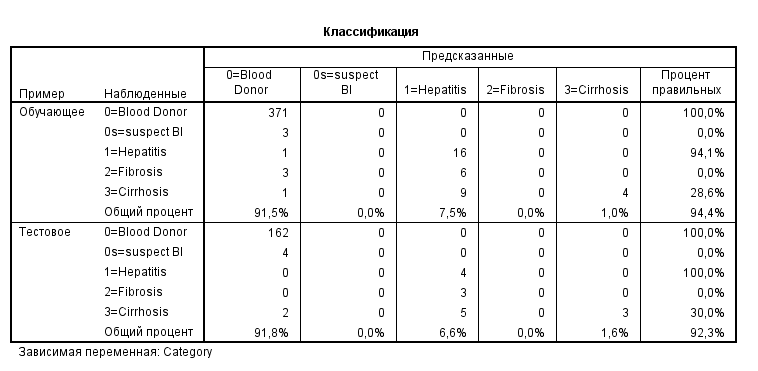
Третий эксперимент:



Четвертый эксперимент:



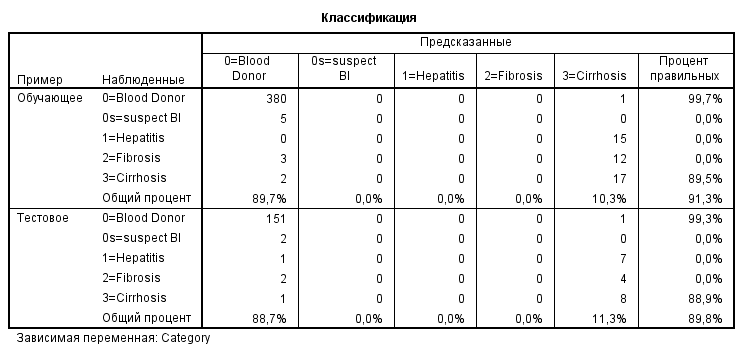
Пятый эксперимент:



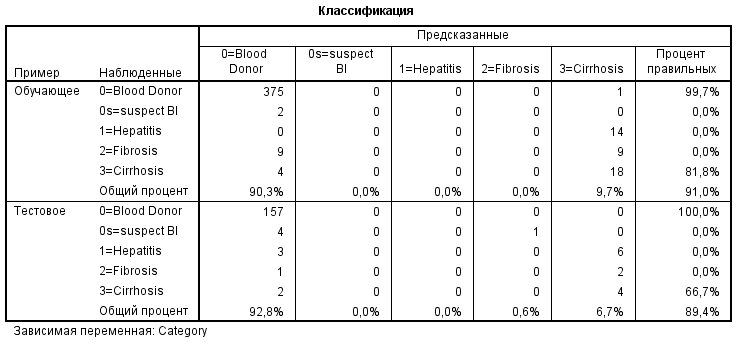
Здесь уже видно, что любой эксперимент в наборе ALB-PROT-CREA-ALP дает лучшие результаты, чем в наборе ALB-PROT-CREA. Результат в 94.6% был получен в третьем эксперименте, что на данном этапе является лучшей точностью кластеризации с помощью нейросети. Такую модель можно считать удовлетворительной.

ALB-PROT-CREA-ALT

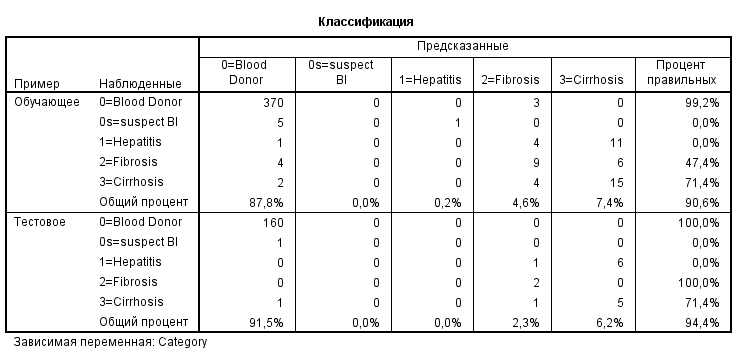
Первый эксперимент:



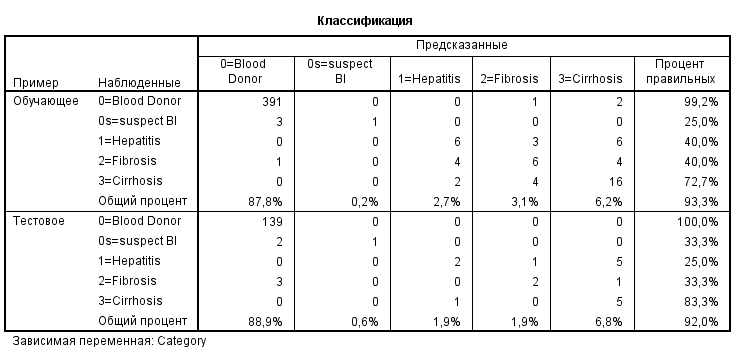
Второй эксперимент:



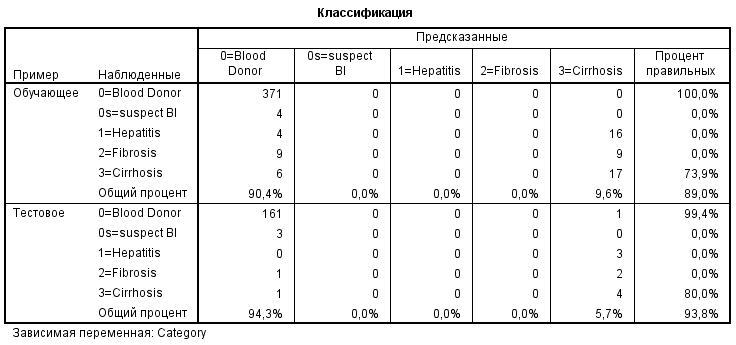
Третий эксперимент:



Четвертый эксперимент:



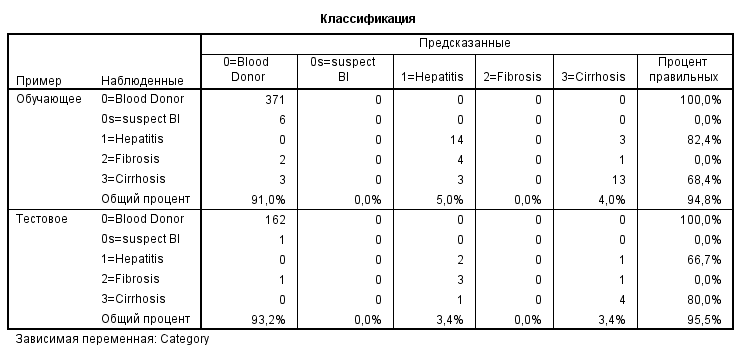
Пятый эксперимент:



Здесь самый лучший результат был получен в третьем эксперименте. Нейросеть на тестовой выборке дала 94.4% правильных ответов, что тоже является достаточно высоким уровнем. Такую модель также можем принять за удовлетворительную.

Из всех проведенных экспериментов можно сделать вывод, что на выборке достаточно легко построить нейросеть с точностью ответов 90%. Однако до 94% точных ответов удалось довести лишь одну конфигурацию нейронной сети. При этом нейросеть выдала такой результат при наборе параметров ALB-PROT-CREA-ALP и ALB-PROT-CREA-ALT.

Очевидно, из этого результата рождается вопрос – если одна и та же конфигурация нейросети выдала на разных наборах параметров выдающийся результат, то что произойдет, если эти параметры смешать? То есть взять набор параметров ALB-PROT-CREA-ALP-ALT и на их основе построить нейросеть, использованную для третьего эксперимента. Ниже приведен результат работы такой нейросети:



Точность работы нейросети составила 95.5% - это лучший результат среди всех экспериментов, проведенных с нейросетью. В итоге был получен результат, который можно считать достаточно хорошим.

**Двухэтапный кластерный анализ (???)**