Белорусский государственный технологический университет

Кафедра «Информационных систем и технологий»

Лабораторная работа №3

**Деревья решений**

Выполнил студент

3 курса 3 группы

Процукович К.М.

Проверил

Колесников В. Л.

Минск 2018

# **1. Краткое описание объекта исследования и анализа**

Дерево – конечный связанный граф с множеством вершин, который не содержит циклов и имеет начальную вершину, называемую корнем дерева. Та вершина, которая не имеет выходящих ребер – лист. Все остальные вершины – внутренние.

Также существует отдельный вид деревьев – бинарное дерево. Это дерево, имеющие не более двух выходных вершин.

**Основные алгоритмы**

Можно выделить следующие основные алгоритмы:

* ID.3
* C4.5
* CART

**ID.3.** В основе этого алгоритма лежит понятие информационной энтропии - то есть, вероятности появление определенной информации. Работает только с категориальными атрибутами и дает только двух потомков.

**С4.5**. Этот алгоритм - усовершенствование предыдущего метода. Преимуществом его, по сравнению с предшественником, является то, что он может *усекать лишние вести деревьев*, что упрощает работу. А также он *может работать* и с *числовыми атрибутами*. И *нет ограничения на количество потомков*.

**CART**. Алгоритм разработан в целях построения так называемых бинарных деревьев решений. На каждом шаге идет разбиение на два потомка, в результате чего происходит перебор всех атрибутов и с выбором того, который максимизирует значения показателя.

Процесс построения дерева является нисходящим. И в ходе этого процесса алгоритм должен найти такой **критерий расщепления**, чтобы разбить множества на подмножества. Существуют различные критерии расщепления, я же остановлюсь на наиболее известном – **мера информационного выигрыша**. Основана она на энтропийном подходе и применяется в рассмотренных выше алгоритмах ID.3 и C4.5. Сам информационный выигрыш (уменьшение энтропии) считается по следующей формуле:



**2.Описиние проблемы**

В данной лабораторной работе нам будет необходимо с использованием математического пакета JMP SAS получить деревья решений для выходных параметров с максимальным расщеплением.

Направления расщепления при подавлении роста как на максимум, так и на минимум.

Результаты в виде упорядоченных последовательностей от корня к листьям и значения показателей в листьях поместить в предлагаемую таблицу и прокомментировать.

**3.Варианты решения проблемы**

Для решения поставленной проблемы мы будем использовать, созданной нами на первой лабораторной работы, базой данных, которую проанализировав, с помощью JMP SAS сможем сделать соответствующие выводы.

**4.Обработка результатов**

Мною были выбраны следующие деревья решения для анализа:

* Загрязнение воды
* Загрязнение атмосферы
* Пластичность

Выбор был сделан с упором на экологические факторы, так как нынче это стало глобальной проблемой, тем самым сильно меня заинтересовав. Так же и качественных характеристик была выбрана пластичность, где упор ставился на максимизацию данного параметра.

Сам же анализ и деревья будет строится в три этапа:

1. Выбор факторов, влияющие на корень дерева
2. Построение самого дерева решений
3. Анализ дерева

Первым моим шагом будет выбор атрибутов построения, ими будут все входные параметры, они будут выставлены как факторы (Factor), и один атрибут выходной параметр, выставленный корнем дерева (Response) (рис. 1).

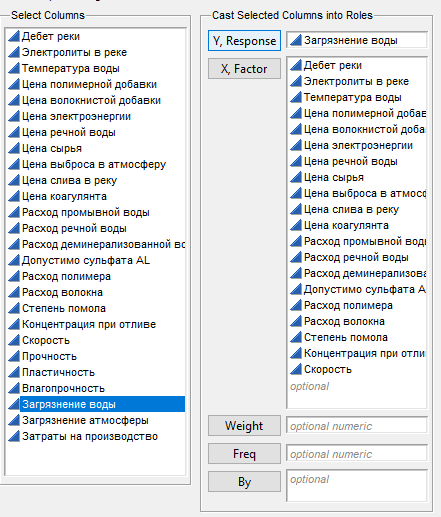


Рисунок 1 – Выбор атрибутов

После выбора всех параметров будет нужно лишь нажать на кнопку «ОК», и программа сама проведет построение и нам останется лишь раскрыть полученное дерево в сторону необходимой нам величины. Так как первым я проводил анализ загрязнения воды, то естественно я пошел в сторону меньшего загрязнения, в данной программе это левая ветвь. Для анализа были выбраны все 200 строк базы данных полеченной еще в первой лабораторной работе (рис. 2).

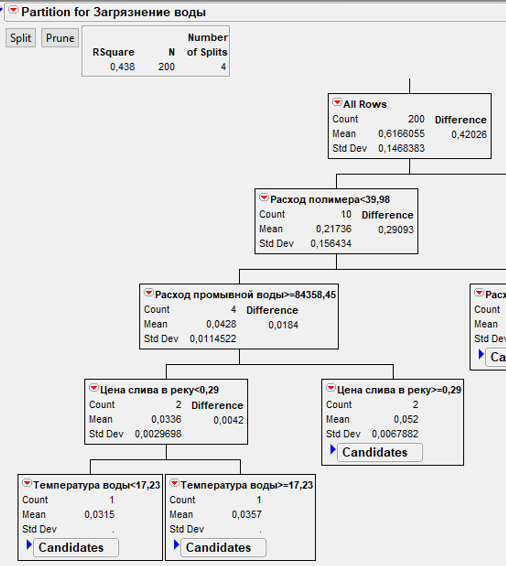


Рисунок 2 – Дерево решений для загрязнения воды

После построения я получил, что дерево разветвилось четыре раза в сторону минимального загрязнения, получив цифру в 0.0315. Следовательно, можно сказать, что на это влияет четыре параметра, а именно:

* Расход полимера (<39.98)
* Расход промывочной воды (>=8458.45)
* Цена слива в реку (<0.29)
* Температура воды (<17.23)

Где расход полимера и промывочной воды — это управляемые параметры, которые предприятие может регулировать, а остальные – цена слива в реку и температура воды, естественные (природные) факторы и предприятие не в силе на них повлиять, следовательно, можно сделать вывод, что человек не может в полной мере уменьшить загрязнение, однако всё же больше 190 из 200 строк оказались под управляющих факторов, что говорит о многом.

Затем было построено дерево для загрязнения атмосферы по тому же принципу и упор делался так же на минимизацию данного значения (рис. 3).

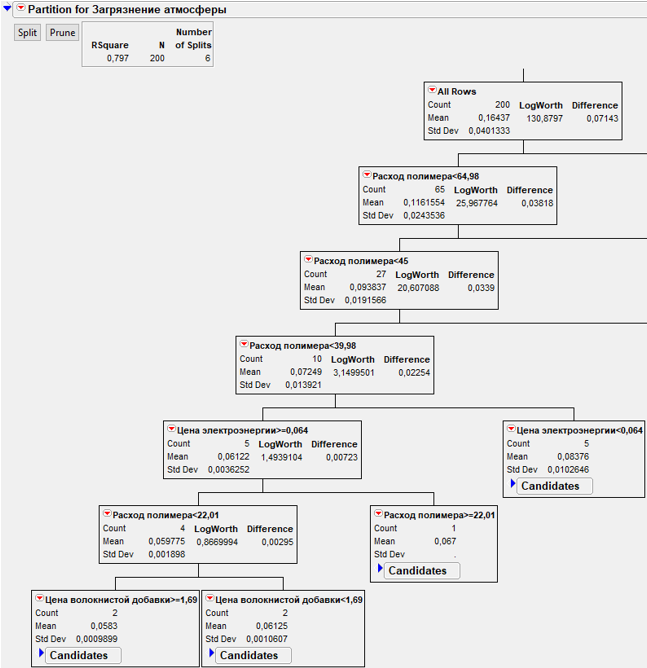


Рисунок 3 – Дерево решений для загрязнения атмосферы

В данном случае дерево разветвилось шесть раз четыре из которых указывает на расход полимера, из-за чего это может быть я не совсем уверен. Остальных же два фактора это цена электроэнергии и волокнистой добавки (природные факторы).

Последним, что я анализировал стала пластичность, выбрав всё те же факторы я преступил к построению дерева, однако в этот раз шёл в противоположную сторону (правую) для поиска экземпляра с максимальным значением (рис. 4)



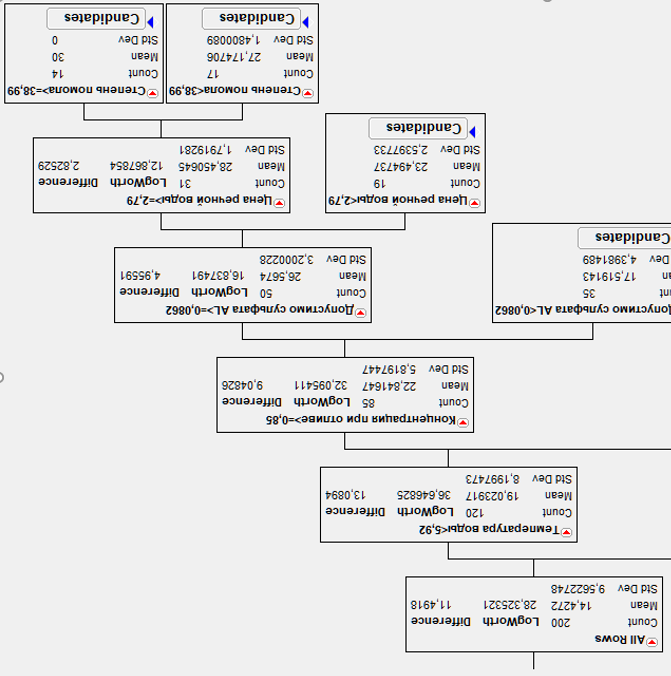


Рисунок 4 – Дерево решений для пластичности

Как можно увидеть данное дерево разветвилось пять раз, что дало нам пять факторов, влияющих на пластичность, а именно:

* Температура воды
* Концентрация при отливе
* Допустимо сульфата
* Цена речной воды
* Степень помола

Три из них (степень помола, допустимо сульфата, концентрация при отливе) являются управляющими, остальные же два (цена речной воды и её температура) возмущающими.

**5.Выводы**

В результате проведенного анализа можно сделать выводы о том, что на результирующие параметры достаточно сильно влияют как управляющие, так и возмущающие параметры. Так же можно сказать, что все выходные параметры не имеют линейных зависимостей, а их значение является результатом воздействия множества факторов, некоторых в большей, некоторых в меньшей мере.

Если рассмотреть сами факторы, то действительно влияющие на результат делятся примерно поровну, и лишь на одну из них человек на предприятии может напрямую воздействовать, будь то расход полимера и промывочной воды для загрязнения воды или концентрация при отливе, допустимо сульфата и степень помола для пластичности, все из вышеперечисленных параметров человек может регулировать, а в следствии напрямую влиять на результат. Но следует рассмотреть и обратную сторону медали, имеются также и возмущающие факторы, как цена на речную воду или её температура, на которую предприятие повлиять не в состоянии. Из этого можно сделать вывод, что предприятие и выходные параметры являются наполовину зависимыми от природного фактора, на который трудно как бы то ни было повлиять.

В заключении хочу сказать, что хоть мы и сумели разглядеть взаимосвязи, полной же картины у нас всё равно не имеется, так как мы анализировали лишь один показатель, в отрыве от остальных, и что для одного в радость, для другого в печаль. Следовательно, могут возникать конфликты приводящие к потере качества и/или увеличению загрязнения окружающей среды.