УДК 004

К вопросу подготовки данных для создания системы анализа и прогнозирования основных показателей приемной кампании ФГБОУ ВО «Ярославский государственный технический университет»

Берсенев Александр Иванович
Ярославский государственный технический университет
Студент

Бойков Сергей Юрьевич
Ярославский государственный технический университет
К.т.н.

Аннотация
В статье рассматриваются пример подготовки категориальных данных, на которых впоследствии методами машинного обучения будет создана система для прогнозирования основанных показателей деятельности приемной кампании ЯГТУ.
Ключевые слова: дата-сет, подготовка данных, машинное обучение, база данных

On the issue of preparing data for the system of analysis and forecasting of the main indicators of the admission campaign of the Yaroslavl State Technical University

Bersenev Aleksandr Ivanovich
Yaroslavl State Technical University
Student

Bojkov Sergej Yurevich
Yaroslavl State Technical University
Candidate of technical sciences

Abstract
The article examines an example of the normalization of categorical data, on which a system will subsequently be created using machine learning methods to predict the performance indicators of the YSTU admission campaign.
Keywords: dataset, preparing data, machine learning, database

В эпоху информационных технологий на помощь работниками приемной комиссии вуза может прийти специализированные экспертные системы. Они не только позволяют провести анализ успешности приемной кампании после ее завершения, но и спрогнозировать её результат на основе данных, предоставленных от абитуриента. Стоит принять во внимание, что

данная система выступают не в качестве замены персонала, а в роли эксперта-консультанта в данной предметной области.

Основными источниками финансирования государственных образовательных учреждений являются средства федерального и местного бюджетов. Негосударственные образовательные учреждения вправе получать средства из этих источников после получения ими государственной аккредитации. Для успешного проведения приемной кампании, необходимо грамотно распределять силы на каждого абитуриента.

Целью работы является приведение данных к виду приемлемому для применения к нему алгоритмов машинного обучения. Эти данные впоследствии будут использованы для создания автоматизированного инструмента, который на основе предоставленных абитуриентом данных будет способен прогнозировать вероятность его последующего отчисления, а также определять попадает ли абитуриент в число тех, кому стоит уделять повышенное внимание в ходе проведения приемной кампании.

Впоследствии разработанная система позволит существенно снизить нагрузку на приемную комиссию вуза, а также поможет деятельности приемной комиссии на всех этапах проведения приемной кампании.

Для создания подобной системы вузом был предоставлен дата-сет в виде базы данных MSSQL Server с данными о проведении приемных кампаний до 2013г. Исходный дата-сет представлен в виде базы данных MS SQL Server. База данных содержит более 200 таблиц (Рисунок 1).

Дата-сет – это набор данных, которые компьютер обрабатывает как единое целое. Это означает, что набор данных содержит множество отдельных фрагментов данных при этом он может использоваться для обучения алгоритма с целью поиска предсказуемых закономерностей внутри всего набора данных.

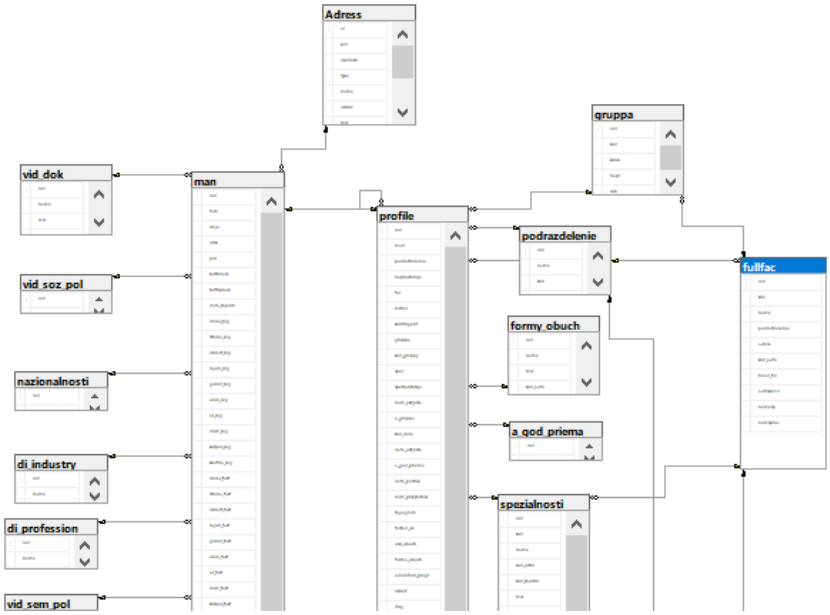

Рисунок 1 – исходный дата-сет в MS SQL server, таблицы, связанные с абитуриентами

Две основные таблицы из этой базы данных, на которых будет происходить прогнозирование успешности приемной кампании ЯГТУ – это таблицы man и profile, а также некоторые связанные с ними таблицы. Таблица man содержит информацию о человеке, а таблица profile личные дела студента ЯГТУ. В таблице 1 представлен список таблиц исходного дата-сета, связанных с абитуриентами.

Таблица 1 – Таблицы исходного дата-сета

Наименование сущности	Первичный ключ	Пояснение
profile	oid	Личные дела студентов.
man	oid	Учащийся
vid_doc	oid	Справочник видов документов
vid_soz_pol	oid	Справочник видов социального положения
nazionalnosti	oid	Справочник национальностей
di_industry	oid	Справочник видов областей деятельности
di_profession	oid	Справочник видов должностей
vid_sem_pol	oid	Справочник видов семейного положения
v_godnosti	oid	Справочник категорий годности
Adress	id	Адреса
gruppa	oid	Справочник групп
podrazdelenie	oid	Справочник подразделений
formy_obuch	oid	Справочник форм обучений
a_god_priema	oid	Справочник годов приема
spezialnosti	oid	Справочник специальностей
fullfac	oid	Полное направление
kafedry	oid	Справочник кафедр

Первое, что было сделано это – собраны все справочники назад в таблицы, где они используются, иными словами, была проведена деморализация базы данных. Такой подход построения модели необходим и оправдан в реляционных базах данных, где нужно обеспечение целостности данных, однако для задачи прогнозирования — это лишнее и затруднит дальнейший анализ.

Для прогнозирования данных по экзаменам средствами MS SQL Server было создано представление, следующим запросом:

```
CREATE VIEW sum_ball AS
SELECT a_ext.profile, SUM(a_ext.ball)
FROM a_ext
GROUP BY a_ext.profile
```

Помимо это в ходе отчистки дата-сета были выполнены еще и следующие операции:

1. Переименованы все имена сущностей и столбцов
2. Произведена отчистка входного дата-сета от лишних и не нужных для анализа данных и связей
3. Исправлены некоторые неточности и ошибки в данных.

После применения к исходному дата-сету всех описанных выше преобразований он приобрел вид, изображенный на рисунке 2.

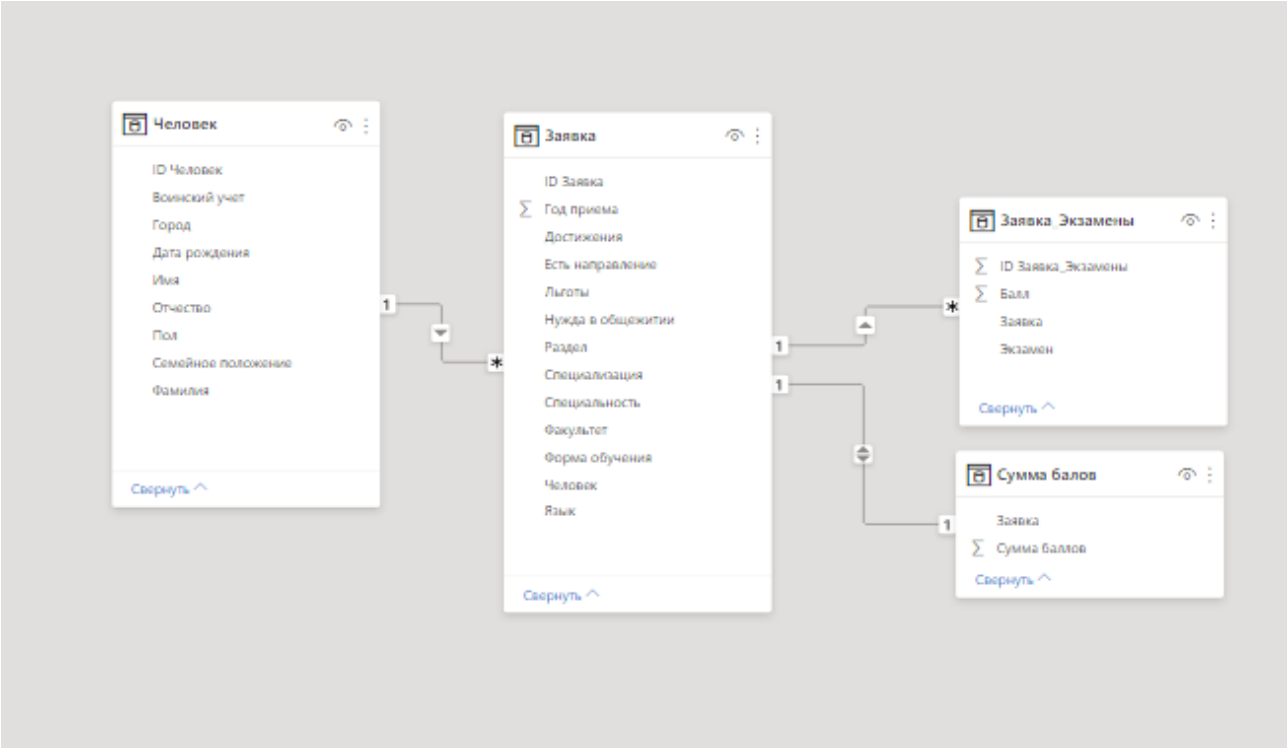

Рисунок 2 - Конечный вид дата-сета для дальнейшего применения его в моделях машинного обучения

В результате проделанной работы исходный датасет был приведен к виду приемлемому для дальнейшего применения к нему алгоритмов машинного обучения. В последствии данный дата-сет будет использован для создания автоматизированного инструмента, который будет прогнозировать основные показатели деятельности приемной кампании ЯГТУ на основе данных, предоставленных абитуриентами.

Библиографический список

1. Элбон К. Машинное обучение с использованием Python. Сборник рецептов. СПб.: БХВ-Петербург, 2019. 384 с.
2. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. М.: Альфа-книга, 2018. 418 с.