РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ

Факультет физико-математических и естественных наук

Кафедра прикладной информатики и теории вероятностей

Практикум по математической статистике

Лабораторная работа №5

Тема: «Дискриминантный анализ»

Вариант 10

Выполнил

Студент: Феоктистов Владислав

Группа: НПМбд-01-19б

№ с/б: 1032192939

Преподаватель: Матюшенко Сергей Иванович

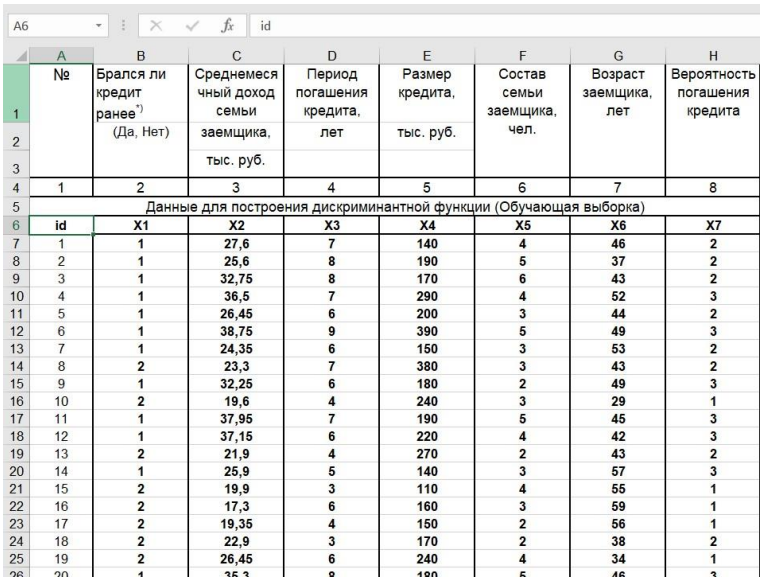
МОСКВА

2022 г.

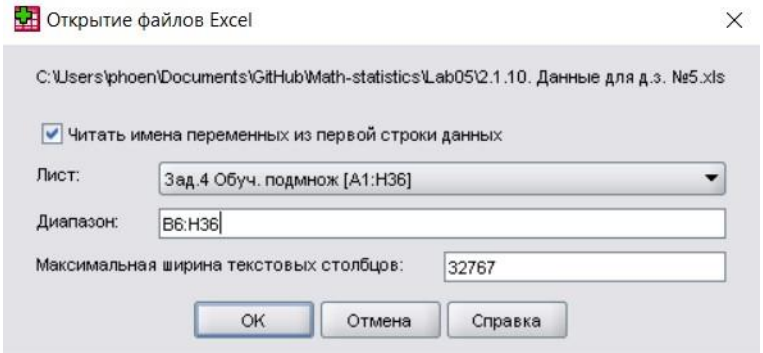
Цель работы: приобрести практические навыки применения дискриминантного анализа для решения конкретных задач с использованием статистического пакета SPSS.

Ход работы:

1. Изучил теоретические основы дискриминантного анализа, используя материалы учебного пособия.
2. Разобрал пример использования SPSS для реализации дискриминантного анализа.
3. Запустим программу SPSS и введем исходные данные по обучающей выборке.



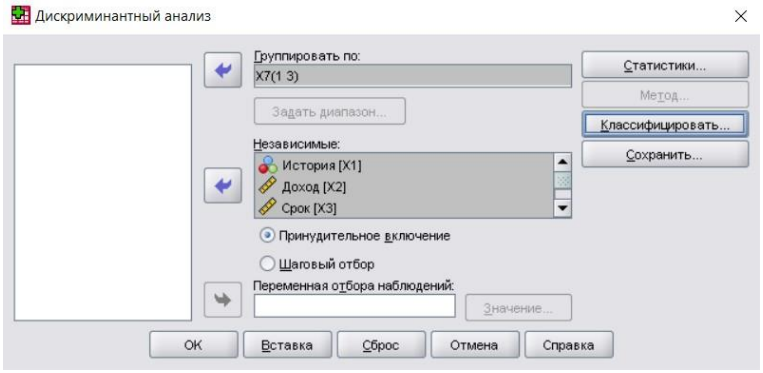
	A	B	C	D	E	F	G	H
1	№	Брался ли кредит ранее*)	Среднемесячный доход семьи	Период погашения кредита,	Размер кредита,	Состав семьи заемщика, чел.	Возраст заемщика, лет	Вероятность погашения кредита
2		(Да, Нет)	заемщика,	лет	тыс. руб.			
3			тыс. руб.					
4	1	2	3	4	5	6	7	8
5	Данные для построения дискриминантной функции (Обучающая выборка)							
6	id	X1	X2	X3	X4	X5	X6	X7
7	1	1	27,6	7	140	4	46	2
8	2	1	25,6	8	190	5	37	2
9	3	1	32,75	8	170	6	43	2
10	4	1	36,5	7	290	4	52	3
11	5	1	26,45	6	200	3	44	2
12	6	1	38,75	9	390	5	49	3
13	7	1	24,35	6	150	3	53	2
14	8	2	23,3	7	380	3	43	2
15	9	1	32,25	6	180	2	49	3
16	10	2	19,6	4	240	3	29	1
17	11	1	37,95	7	190	5	45	3
18	12	1	37,15	6	220	4	42	3
19	13	2	21,9	4	270	2	43	2
20	14	1	25,9	5	140	3	57	3
21	15	2	19,9	3	110	4	55	1
22	16	2	17,3	6	160	3	59	1
23	17	2	19,35	4	150	2	56	1
24	18	2	22,9	3	170	2	38	2
25	19	2	26,45	6	240	4	34	1

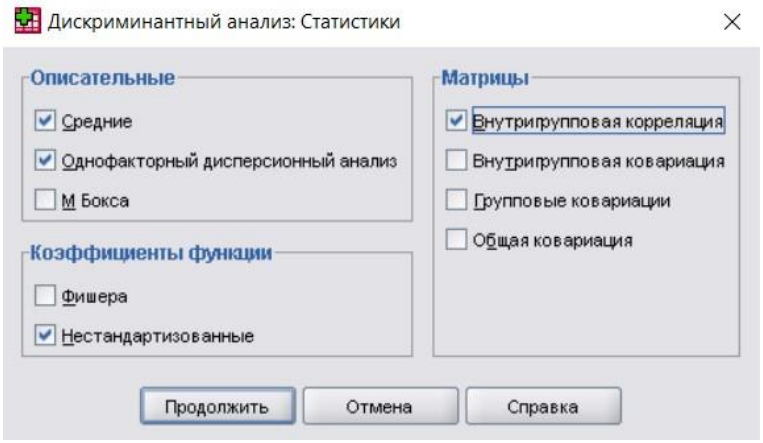


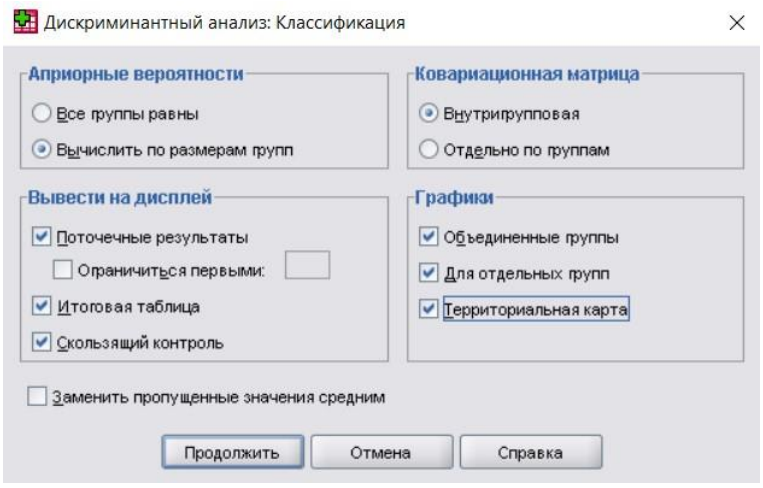
	X1	X2	X3	X4	X5	X6	X7
1	1	27,6	7	140	4	46	2
2	1	25,6	8	190	5	37	2
3	1	32,8	8	170	6	43	2
4	1	36,5	7	290	4	52	3
5	1	26,5	6	200	3	44	2
6	1	38,8	9	390	5	49	3
7	1	24,4	6	150	3	53	2
8	2	23,3	7	380	3	43	2
9	1	32,3	6	180	2	49	3
10	2	19,6	4	240	3	29	1
11	1	38,0	7	190	5	45	3
12	1	37,2	6	220	4	42	3
13	2	21,9	4	270	2	43	2
14	1	25,9	5	140	3	57	3
15	2	19,9	3	110	4	55	1
16	2	17,3	6	160	3	59	1
17	2	19,4	4	150	2	56	1
18	2	22,9	3	170	2	38	2
19	2	26,5	6	240	4	34	1
20	1	35,3	8	180	5	46	3
21	2	20,4	4	180	2	46	1
22	2	28,8	2	140	2	58	2
23	2	29,8	9	500	2	37	3
24	2	18,8	11	420	5	45	1

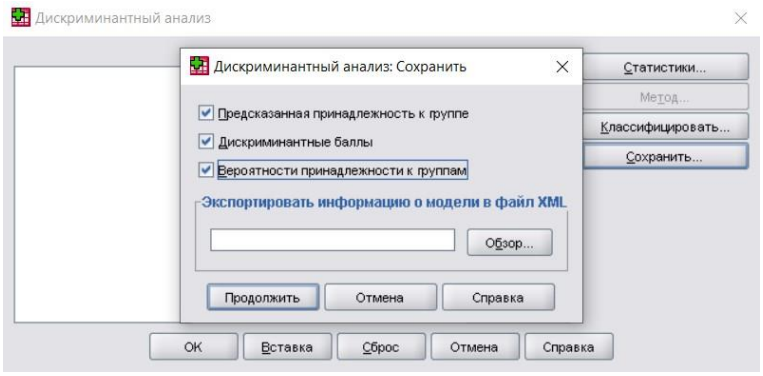
	Имя	Тип	Ширина	Десятич...	Метка	Значения	Пропуски	Столбцы	Выравнивание	Шкала
1	X1	Числовая	11	0	История	{1, Да}...	Нет	11	По право...	Номинальная
2	X2	Числовая	11	1	Доход	Нет	Нет	11	По право...	Количестве...
3	X3	Числовая	11	0	Срок	Нет	Нет	11	По право...	Количестве...
4	X4	Числовая	11	0	Кредит	Нет	Нет	11	По право...	Количестве...
5	X5	Числовая	11	0	Семья	Нет	Нет	11	По право...	Количестве...
6	X6	Числовая	11	0	Возраст	Нет	Нет	11	По право...	Количестве...
7	X7	Числовая	11	0	Вероятность	{1, Низкая}...	Нет	11	По право...	Номинальная

Далее проделаем дискриминантный анализ.









После нажатия на кнопку «ОК» получаем следующий вывод:

Сводка результатов обработки наблюдений

Невзвешенные наблюдения		N	Процент
Валидные		30	100,0
Исключенные	Пропущенные или лежащие вне диапазона коды группирующей переменной	0	,0
	По крайней мере одна пропущенная дискриминантная переменная	0	,0
	Оба групповых кода пропущены или лежат вне диапазона, и отсутствует по крайней мере одна дискриминантная переменная.	0	,0
	Итого искл.	0	,0
Всего набл.		30	100,0

Групповые статистики

Вероятность		Среднее	Стд. отклонение	Кол-во валидных (искл. целиком) Невзвешенные	Взвешенные
Низкая	История	2,000	,0000	9	9,000
	Доход	20,444	3,0025	9	9,000
	Срок	5,444	2,4552	9	9,000
	Кредит	218,889	94,7951	9	9,000
	Семья	3,111	1,0541	9	9,000
	Возраст	44,778	12,0807	9	9,000
Средняя	История	1,417	,5149	12	12,000
	Доход	26,242	3,4670	12	12,000
	Срок	5,333	2,0597	12	12,000
	Кредит	211,667	76,3763	12	12,000
	Семья	3,750	1,6026	12	12,000
	Возраст	45,750	7,1748	12	12,000
Высокая	История	1,111	,3333	9	9,000
	Доход	34,089	4,2316	9	9,000
	Срок	7,222	1,3944	9	9,000
	Кредит	262,222	116,8094	9	9,000
	Семья	3,778	1,2019	9	9,000
	Возраст	47,889	6,1734	9	9,000
Итого	История	1,500	,5085	30	30,000
	Доход	26,857	6,4137	30	30,000
	Срок	5,933	2,1324	30	30,000
	Кредит	229,000	94,6263	30	30,000
	Семья	3,567	1,3309	30	30,000
	Возраст	46,100	8,4786	30	30,000

Из данных таблицы «Критерий равенства групповых средних» следует, что переменные «Кредит» («Размер кредита»), «Семья» («Состав семьи заемщика»), «Возраст» («Возраст заемщика») незначимо различаются по группам, поскольку для них уровень значимости Знч. > 0.05, поэтому классификацию заемщиков целесообразно проводить по первым двум переменным: «История» («Брался ли кредит») и «Доход» («Среднемесячный доход семье заемщика»).

Критерий равенства групповых средних

	Лямбда Уилкса	F	ст.св1	ст.св2	Знч.
История	,507	13,106	2	27	,000
Доход	,291	32,831	2	27	,000
Срок	,838	2,619	2	27	,091
Кредит	,944	,796	2	27	,461
Семья	,948	,741	2	27	,486
Возраст	,978	,305	2	27	,740

Анализ матрицы коэффициентов в таблице «Объединенные внутригрупповые матрицы» свидетельствует об отсутствии мультиколлинеарности, поэтому коэффициенты корреляции малы.

Объединенные внутригрупповые матрицы

		История	Доход	Срок	Кредит	Семья	Возраст
Корреляция	История	1,000	-,455	-,141	,486	-,626	-,234
	Доход	-,455	1,000	,026	-,176	,621	,053
	Срок	-,141	,026	1,000	,583	,341	-,352
	Кредит	,486	-,176	,583	1,000	-,037	-,469
	Семья	-,626	,621	,341	-,037	1,000	,099
	Возраст	-,234	,053	-,352	-,469	,099	1,000

Данные таблицы «Собственные значения» показывают, что первая функция учитывает 95,2% дисперсии, а корреляция между исходными данными и данными, полученными по модели, высокая и составляет 0,929. Для второй функции эти значения намного меньше.

Собственные значения

Функция	Собственное значение	% объясненной дисперсии	Кумулятивный %	Каноническая корреляция
1	6,327[a]	95,2	95,2	,929
2	,318[a]	4,8	100,0	,491

a. В анализе использовались первые 2 канонические дискриминантные функции.

Оценка значимости дискриминантных функций проводится по коэффициенту Уилкса (λ). Из данных таблицы «Лямбда Уилкса» видно, что для первой функции значимость Знч. < 0,001, следовательно, она позволяет значимо и надежно дискриминировать наблюдения. В то же время значимость второй функции составляет лишь 0,239. Поэтому в дальнейшем для классификации целесообразно использовать только первую дискриминантную функцию.

Лямбда Уилкса

Проверка функции(й)	Лямбда Уилкса	Хи-квадрат	ст.св.	Знч.
от 1 до 2	,104	55,549	12	,000
2	,759	6,757	5	,239

Нормированные коэффициенты канонической дискриминантной функции

	Функция	
	1	2
История	-,963	1,235
Доход	1,039	,543
Срок	,019	1,276
Кредит	,807	-,837
Семья	-1,177	-,310
Возраст	,281	,383

Структурная матрица

	Функция	
	1	2
Доход	,620*	-,012
Возраст	,059*	,036
История	-,375	,507*
Срок	,152	,388*
Кредит	,080	,244*
Семья	,076	-,243*

Объединенные внутригрупповые корреляции между дискриминантными переменными и нормированными каноническими дискриминантными функциями.
Переменные упорядочены по абсолютной величине корреляций внутри функции.

*. Максимальная по абсолютной величине корреляция между переменными и дискриминантными функциями.

Формально по данным таблицы «Коэффициенты канонической дискриминантной функции» можно построить две дискриминантные функции:

$$D_1(X) = -4{,}286 - 2{,}566x_1 + 0{,}290x_2 + 0{,}009x_3 + 0{,}008x_4 - 0{,}876x_5 + 0{,}032x_6;$$

$$D_2(X) = -11{,}943 + 3{,}291x_1 + 0{,}151x_2 + 0{,}631x_3 - 0{,}009x_4 - 0{,}231x_5 + 0{,}044x_6$$

Коэффициенты канонической дискриминантрой функции

	Функция	
	1	2
История	-2,566	3,291
Доход	,290	,151
Срок	,009	,631
Кредит	,008	-,009
Семья	-,876	-,231
Возраст	,032	,044
(Константа)	-4,286	-11,943

Ненормированные коэффициенты

Однако поскольку значимость второй функции более 0,001, ее для дискриминации использовать нецелесообразно.

Координаты центроидов по группам приведены в таблице «Функции в центроидах групп». Они используются для нанесения центроидов на карту восприятия.

Функции в центроидах групп

	Функция	
Вероятность	1	2
Низкая	-2,873	,503
Средняя	-,289	-,652
Высокая	3,258	,366

Ненормированные канонические дискриминантные функции вычислены в центроидах групп.

Классификационные статистики

Сводка классификации

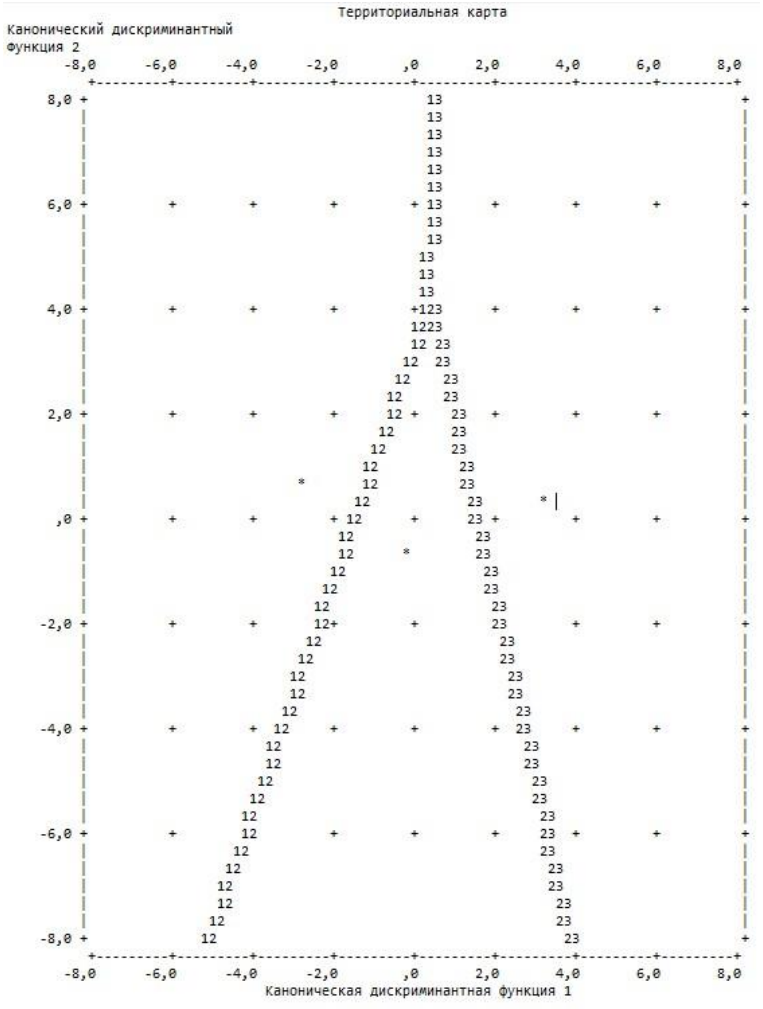
Обработано		30
Исключенные2	Пропущенные или лежащие вне диапазона коды групп2	0
	По крайней мере одна дискриминантная переменная пропущена.	0
Используется в выводе		30

Априорные вероятности для групп

		Наблюдения, использованные в анализе	
Вероятность	Априорные	Невзвешенные	Взвешенные
Низкая	,300	9	9,000
Средняя	,400	12	12,000
Высокая	,300	9	9,000
Итого	1,000	30	30,000

Карта восприятий визуализирует разделение наблюдений функциями. Так, первая функция $D_1(X)$ делит наблюдения на две группы: 1, 2 и 2, 3, вторая функция $D_2(X)$ отделяет наблюдения 2 от всех остальных.

Поле графика разделено дискриминантными функциями на три области: в левой части находятся преимущественно наблюдения первой группы с низкой вероятностью своевременного погашения кредита; в правой части – третьей группы с высокой вероятностью и в нижней части – второй группы со средней вероятностью.



Символы, используемые в территориальной карте

Символ	Грп.	Метка
1	1	Низкая
2	2	Средняя
3	3	Высокая
*		Указывает центр группы

В таблице «Поточечные статистики» размещена информация о фактических (*Actual Group*) и предсказанных (*Predicted Group*) группах для каждого заемщика и соответствующие дискриминантные баллы (*Discriminant Scores*), полученные при подстановке значений переменных в уравнениях дискриминантных функций $D_1(X)$ и $D_2(X)$.

Поточечные статистики

			Наивероятнейшая группа					Вторая вероятнейшая группа			Дискриминантные баллы	
				P(D>d | G=g)								
	Номер наблюдения	Фактическая группа	Предсказанная группа	p	ст.св	P(G=g | D=d)	Квадрат расстояния Махаланобиса до центра	Группа	P(G=g | D=d)	Квадрат расстояния Махаланобиса до центра	Функция 1	Функция 2
Исходные	1	2	2	,721	2	,982	,655	3	,014	8,635	,371	-,183
	2	2	2	,779	2	,949	,500	1	,051	5,757	-,942	-,922
	3	2	2	,506	2	,973	1,361	3	,017	8,890	,276	,369
	4	3	3	,497	2	1,000	1,398	2	,000	22,681	4,412	,111
	5	2	2	,202	2	,883	3,200	3	,117	6,666	1,349	-1,372
	6	3	3	,235	2	1,000	2,896	2	,000	28,771	4,956	,472
	7	2	2	,655	2	,983	,845	3	,016	8,509	,608	-,854
	8	2	2	,515	2	,895	1,326	1	,104	5,046	-,627	,449
	9	3	3	,794	2	1,000	,462	2	,000	18,130	3,896	,133
	10	1	1	,147	2	,943	3,833	2	,057	10,012	-3,365	-1,393
	11	3	3	,890	2	,996	,232	2	,004	11,800	2,883	,669
	12	3	3	,759	2	,999	,552	2	,001	15,874	3,675	-,248
	13	2	2	,697	2	,874	,722	1	,126	4,012	-1,116	-,459
	14	3	2**	,365	2	,927	2,018	3	,073	6,518	1,092	-,986
	15	1	1	,275	2	,999	2,585	2	,001	17,638	-4,424	,079
	16	1	1	,405	2	,999	1,810	2	,001	16,613	-3,720	1,548
	17	1	1	,880	2	,950	,257	2	,050	6,730	-2,451	,782
	18	2	1**	,433	2	,539	1,673	2	,461	2,558	-1,844	-,282
	19	1	1	,675	2	,892	,787	2	,108	5,593	-2,078	,895
	20	3	3	,397	2	,952	1,847	2	,048	8,404	2,072	1,031
	21	1	1	,778	2	,847	,503	2	,153	4,502	-2,216	,236
	22	2	2	,182	2	,944	3,403	3	,030	9,720	,232	1,118
	23	3	3	,447	2	,999	1,612	2	,001	15,605	2,957	1,600
	24	1	1	,531	2	,997	1,265	2	,003	13,699	-3,257	1,560
	25	3	3	,982	2	,999	,036	2	,001	14,788	3,376	,514
	26	2	2	,385	2	,598	1,907	1	,402	2,122	-1,587	-,182
	27	1	1	,539	2	,832	1,238	2	,168	5,007	-1,854	,948
	28	1	1	,761	2	,882	,547	2	,118	5,137	-2,495	-,133
	29	2	2	,048	2	,999	6,064	1	,000	20,750	-,094	-3,106
	30	2	2	,214	2	,999	3,083	1	,001	16,154	-,089	-2,396
Кросс-проверенные[a]	1	2	2	,554	6	,963	4,919	3	,031	11,239		
	2	2	2	,244	6	,866	7,923	1	,134	11,073		
	3	2	2	,074	6	,866	11,521	3	,111	15,058		
	4	3	3	,853	6	1,000	2,637	2	,000	23,888		
	5	2	2	,225	6	,631	8,188	3	,369	8,687		
	6	3	3	,178	6	1,000	8,921	2	,000	37,535		
	7	2	2	,277	6	,944	7,505	3	,055	12,606		
	8	2	2	,437	6	,783	5,878	1	,216	7,881		
	9	3	3	,459	6	1,000	5,686	2	,000	22,461		
	10	1	1	,047	6	,729	12,743	2	,271	15,297		
	11	3	3	,532	6	,990	5,093	2	,010	14,869		
	12	3	3	,805	6	,999	3,033	2	,001	17,355		
	13	2	2	,670	6	,794	4,046	1	,206	6,168		
	14	3	2**	,466	6	,999	5,627	1	,000	21,429		
	15	1	1	,212	6	,999	8,381	2	,001	23,699		
	16	1	1	,188	6	,999	8,755	2	,001	23,731		
	17	1	1	,690	6	,912	3,901	2	,088	9,146		
	18	2	1**	,529	6	,852	5,114	2	,148	9,190		
	19	1	1	,252	6	,702	7,820	2	,298	10,112		
	20	3	3	,243	6	,807	7,939	2	,193	11,378		
	21	1	1	,907	6	,796	2,129	2	,204	5,425		
	22	2	3**	,003	6	,600	19,944	2	,309	21,845		
	23	3	3	,019	6	,993	15,112	2	,007	25,605		
	24	1	1	,000	6	,994	25,041	2	,006	35,769		
	25	3	3	,965	6	,999	1,419	2	,001	15,394		
	26	2	1**	,686	6	,592	3,929	2	,408	5,248		
	27	1	1	,675	6	,717	4,013	2	,283	6,446		
	28	1	1	,103	6	,574	10,554	2	,426	11,730		
	29	2	2	,004	6	1,000	19,379	3	,000	35,465		
	30	2	2	,081	6	,998	11,247	1	,001	24,317		

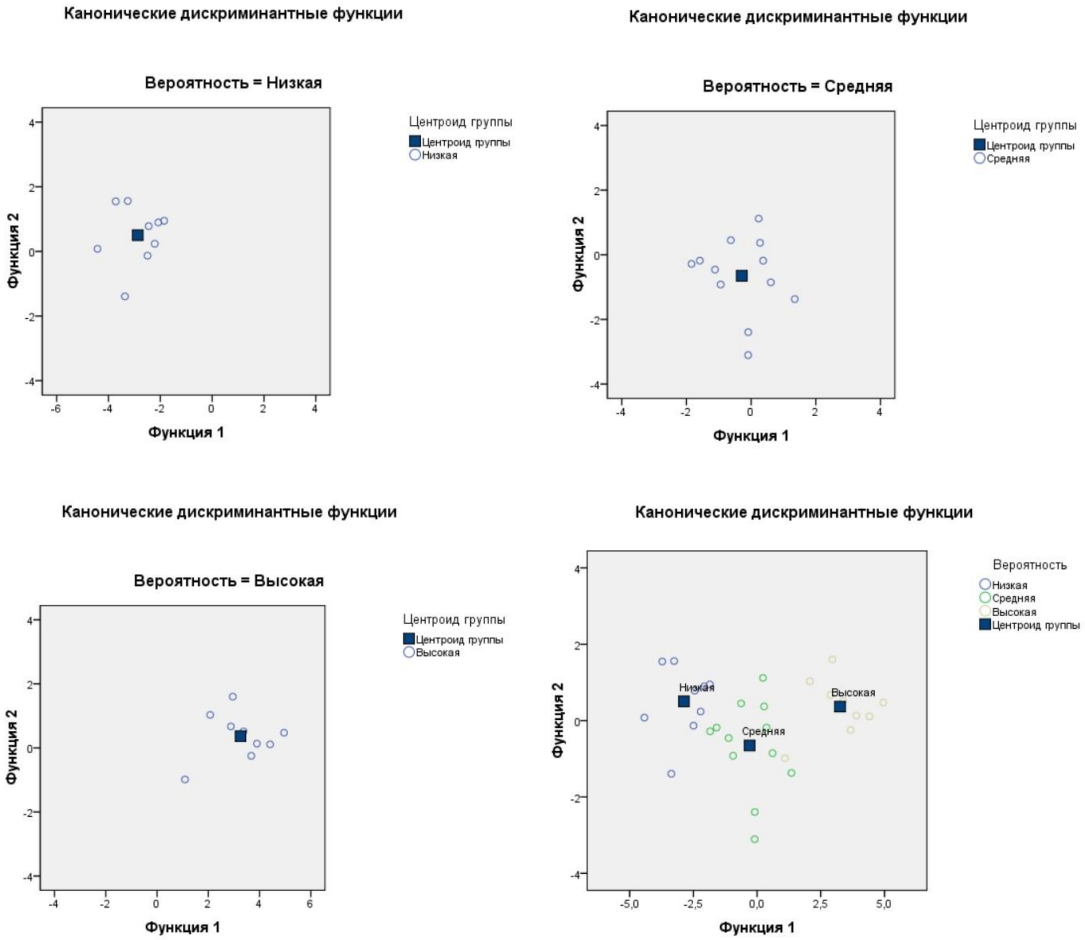
Для исходных данных квадрат расстояния Махаланобиса вычисляется по канонической функции
Для кросс-проверяемых данных квадрат расстояния Махаланобиса вычисляется по наблюдениям.

**. Неправильно классифицированное наблюдение

a. Кросс-проверка проводится только для наблюдений в анализе. При кросс-проверке каждое наблюдение классифицируется функциями, выведенными по всем наблюдениям, за исключением его самого.

На рисунках с 1го по 3ий ниже отражено расположение заемщиков каждой из трех групп на плоскости двух дискриминантных функций $D_1(X)$ и $D_2(X)$. По этим графикам можно проводить детальный анализ вероятностей погашения кредита внутри каждой группы, судить о характере распределения заемщиков и оценивать степень их удаленности от соответствующего центроида.

Кроме того, на 4ом рисунке в той же системе координат приведен объединенный график распределения всех групп заемщиков вместе со своими центроидами; его можно использовать для проведения сравнительного визуального анализа характера взаимного расположения групп заемщиков банка с разными вероятностями погашения кредита. В левой части графика расположены заемщики с низкой вероятностью погашения кредита, в правой – с высокой, а в средней части – со средней вероятностью. Поскольку по результатам расчета вторая дискриминантная функция $D_2(X)$ оказалась незначима, то различия координат центроидов по этой оси незначительны. Этот факт подтверждается картой восприятия, которая была расположена выше.



Данные таблицы «Результаты классификации» свидетельствуют о том, что для 93,3% исходных и перекрестно-проверяемых 86,7% сгруппированных наблюдений классификация проведена корректно, высокая точность достигнута в каждой из групп, но в первой она максимальная – 100%, а в третьей несколько ниже – 88,9%.

Результаты классификации[b,c]

		Вероятность	Предсказанная принадлежность к группе			Итого
			Низкая	Средняя	Высокая	
Исходные	Частота	Низкая	9	0	0	9
		Средняя	1	11	0	12
		Высокая	0	1	8	9
	%	Низкая	100,0	,0	,0	100,0
		Средняя	8,3	91,7	,0	100,0
		Высокая	,0	11,1	88,9	100,0
Кросс-проверенные[a]	Частота	Низкая	9	0	0	9
		Средняя	2	9	1	12
		Высокая	0	1	8	9
	%	Низкая	100,0	,0	,0	100,0
		Средняя	16,7	75,0	8,3	100,0
		Высокая	,0	11,1	88,9	100,0

a. Кросс-проверка проводится только для наблюдений в анализе. При кросс-проверке каждое наблюдение классифицируется функциями, выведенными по всем наблюдениям, за исключением его самого.

b. 93,3% исходных сгруппированных наблюдений классифицировано правильно.

c. 86,7% перекрестно-проверяемых сгруппированных наблюдений классифицировано правильно.

Вывод: приобрёл практические навыки применения дискриминантного анализа для решения конкретных задач с использованием статистического пакета SPSS.