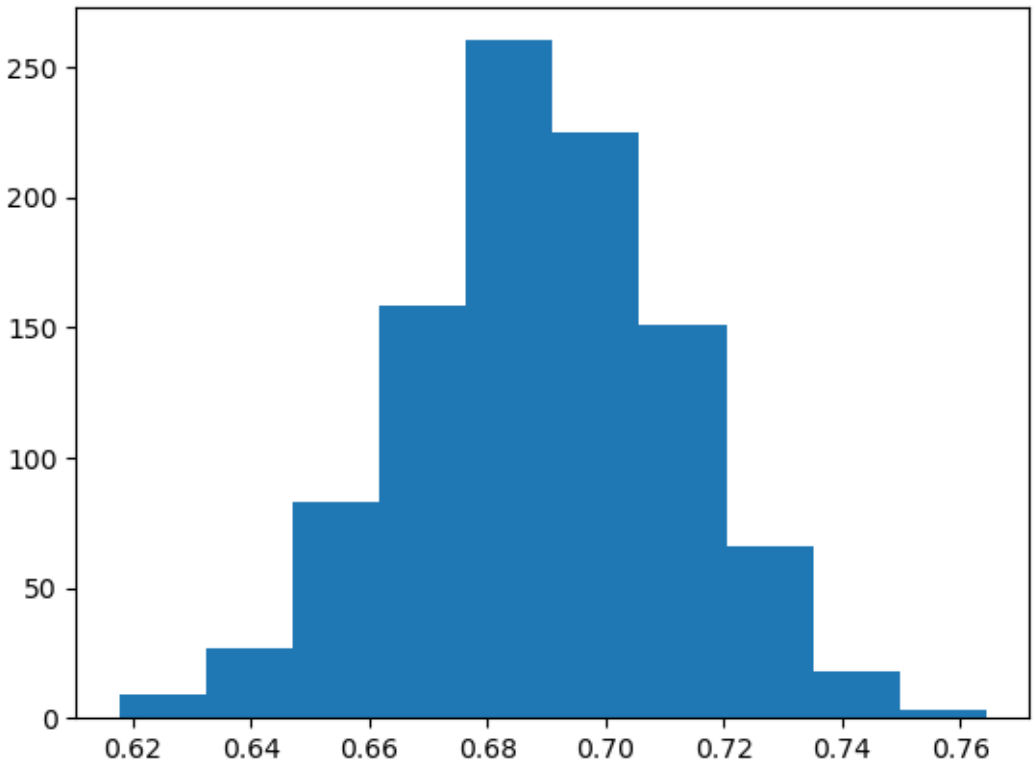


```
95.0 confidence interval 64.3% and 73.2%
```

Используя бутстрэп, мы можем рассчитать доверительные интервалы.
Для этого сначала упорядочиваем статистику, а затем выбираем значения в выбранном процентиле для доверительного интервала. Выбранный процентиль в этом случае называется альфа.
Доверительный интервал часто называют эмпирическим доверительным интервалом.
Бутстрэп можно использовать для оценки производительности алгоритмов машинного обучения.

Рассчитаем интервал точности классификации

Источник данных: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

In [5]:
```
data=pd.read_csv(r'C:\pima-indians-diabetes.data.csv.csv', delimiter=',')
data.info()
data
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Out[5]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

Я использовала набор данных с помощью Pandas.
Можно сделать 1000 итераций начальной загрузки и выбрать выборку, размер которой составляет 50% от размера набора данных.
Используется доверительный интервал 95 %, поэтому выбираются значения в процентилях 2,5 и 97,5.
При выполнении примера выводится точность классификации на каждой итерации начальной загрузки.
Создается гистограмма 1000 оценок точности, показывающая распределение, подобное Гауссу.
Наконец, приводятся доверительные интервалы, показывающие, что существует 95% вероятность того, что доверительные интервалы 64,3% и 73,2% охватывают истинный навык модели.
Этот же метод можно использовать для расчета доверительных интервалов любых других оценок ошибок, таких как среднеквадратическая ошибка для алгоритмов регрессии.

```
# confidence intervals

alpha = 0.95

p = ((1.0-alpha)/2.0) * 100

lower = max(0.0, numpy.percentile(stats, p))

p = (alpha+((1.0-alpha)/2.0)) * 100

upper = min(1.0, numpy.percentile(stats, p))

print('%.1f confidence interval %.1f%% and %.1f%%' % (alpha*100, lower*100, upper*100))

Accuracy=0.739
```