# Контрольные и тестовые вопросы по ПР6

# «Изучение возможностей ML моделей в RapidMiner» по вариантам с ответами

**Вариант 1**

1. Какой алгоритм классификации наиболее чувствителен к масштабу признаков?
   A) Decision Tree
   B) k-NN
   C) Random Forest
   D) Naive Bayes

**Ответ**: **B**

2. В каком случае F1-score будет предпочтительнее Accuracy для оценки качества модели?
   A) При большом объёме данных
   B) При сбалансированных классах
   C) При несбалансированных классах
   D) При малом количестве признаков

**Ответ**: **C**

3. Какой оператор в RapidMiner используется для оценки качества моделей методом многократного разбиения данных?
   A) Apply Model
   B) Cross Validation
   C) Split Data
   D) Performance (Classification)

**Ответ**: **B**

4. Какой из перечисленных алгоритмов машинного обучения не требует обязательной нормализации данных?
   A) Logistic Regression
   B) k-Nearest Neighbors
   C) Support Vector Machine
   D) Decision Tree

**Ответ**: **D**

5. Какие параметры алгоритма k-Nearest Neighbors наиболее критично влияют на качество классификации?
   A) Количество слоев и эпох
   B) Количество соседей и метрика расстояния
   C) Глубина дерева и коэффициент регуляризации
   D) Тип ядра и C-параметр

**Ответ**: **B**

6. В каких случаях Random Forest обычно показывает лучшие результаты, чем одно дерево решений (Decision Tree)?
   A) Когда данные сбалансированы
   B) Когда данных мало
   C) Когда модель склонна к переобучению
   D) Когда данные высокоразмерные и шумные

**Ответ**: **D**

7. Что такое метрика Precision в задачах классификации?
   A) Доля верно классифицированных объектов
   B) Доля объектов, правильно отнесенных к положительному классу, среди всех отнесенных к нему
   C) Доля правильно классифицированных положительных объектов из всех действительно положительных объектов
   D) Усреднённая гармоническая мера полноты и точности

**Ответ**: **B**

8. Почему важно использовать кросс-валидацию при оценке качества моделей?
   **Ответ**: Кросс-валидация позволяет получить более надежную оценку качества модели за счет многократного разбиения данных на обучающую и тестовую выборки, предотвращая случайные результаты от единичного разбиения.

9. Какие гиперпараметры в SVM с ядром RBF наиболее критично влияют на качество модели и почему?
   **Ответ**: Критичны параметр регуляризации (C), контролирующий компромисс между правильной классификацией и гладкостью границы решения, и параметр ядра (gamma), определяющий радиус влияния одиночных точек данных на решение модели.

10. Как влияет изменение глубины дерева (max_depth) на качество модели Decision Tree?
    **Ответ**: Увеличение глубины дерева повышает способность модели к обучению сложных зависимостей, но увеличивает риск переобучения. Снижение глубины дерева упрощает модель, уменьшая риск переобучения, но может снизить точность на обучающих данных.

---

**Вариант 2**

1. Какая модель машинного обучения лучше всего справляется с задачей, если признаки имеют высокую степень нелинейности и шумность?
   A) Logistic Regression
   B) Decision Tree
   C) Support Vector Machine с линейным ядром
   D) Random Forest

**Ответ**: **D**

2. Какой параметр критично важен при использовании Logistic Regression в RapidMiner?
   A) Количество соседей
   B) Параметр регуляризации (C)
   C) Глубина дерева
   D) Тип активационной функции

**Ответ**: **B**

3. Какая из следующих метрик лучше всего подходит для оценки модели при несбалансированных данных?
   A) Accuracy
   B) Precision
   C) Recall
   D) F1-score

**Ответ**: **D**

4. Что такое Recall в задачах бинарной классификации?
   A) Доля верно классифицированных объектов
   B) Доля верно отнесенных к положительному классу объектов из реально положительных
   C) Усреднённая гармоническая мера точности и полноты

D) Доля объектов, правильно отнесенных к положительному классу, среди всех отнесенных к нему

**Ответ**: **B**

5. Какая из перечисленных моделей склонна к переобучению на малых наборах данных?
   A) Logistic Regression
   B) Random Forest
   C) k-Nearest Neighbors
   D) Support Vector Machine

**Ответ**: **C**

6. В каких случаях нормализация признаков оказывает наибольшее влияние на качество модели?
   A) При использовании алгоритмов, основанных на деревьях решений
   B) При использовании алгоритмов, основанных на расстоянии между точками (например, k-NN)
   C) Только при больших объемах данных
   D) Только при линейных задачах классификации

**Ответ**: **B**

7. Какой из перечисленных алгоритмов обычно требует наибольшего времени на обучение при увеличении размера набора данных?
   A) Decision Tree
   B) Logistic Regression
   C) k-Nearest Neighbors
   D) Support Vector Machine

**Ответ**: **D**

8. Как можно уменьшить вероятность переобучения модели Decision Tree?
   **Ответ**: Уменьшением глубины дерева (max_depth), увеличением минимального числа объектов в листьях (min_samples_leaf), использованием обрезки (pruning) или применением ансамблевых методов, таких как Random Forest.

9. Почему Precision и Recall важны для оценки качества модели в задачах медицинской диагностики?
   **Ответ**: Precision отражает вероятность того, что модель не выдаст ложноположительные результаты, а Recall показывает способность

модели обнаруживать положительные случаи, что критично в медицинских задачах.

10. Каковы преимущества алгоритма Random Forest перед отдельным деревом решений?
    **Ответ**: Random Forest снижает вероятность переобучения за счет использования множества деревьев и голосования большинства, стабилизируя результаты и повышая обобщающую способность модели.

---

**Вариант 3**

1. Какой параметр в алгоритме Support Vector Machine отвечает за ширину разделяющей границы (margin) между классами?
   A) Gamma
   B) Max depth
   C) k-neighbors
   D) Параметр регуляризации (C)

**Ответ**: **D**

2. Какая модель машинного обучения лучше подходит для интерпретации результатов и понимания структуры данных?
   A) Support Vector Machine
   B) Logistic Regression
   C) Random Forest
   D) Decision Tree

**Ответ**: **D**

3. Какой из методов лучше всего использовать для надежной оценки модели при ограниченном объеме данных?
   A) Random split
   B) Single train-test split
   C) Cross-validation
   D) Bootstrap

**Ответ**: **C**

4. Какая из следующих метрик наиболее чувствительна к дисбалансу классов в данных?

A) Accuracy
B) Precision
C) Recall
D) F1-score

**Ответ**: **A**

5. Для какого алгоритма критически важно предварительно нормализовать числовые признаки?
   A) Decision Tree
   B) Random Forest
   C) k-Nearest Neighbors
   D) Naive Bayes

**Ответ**: **C**

6. Как влияет увеличение параметра k (количества соседей) на модель k-NN?
   A) Повышает риск переобучения
   B) Повышает устойчивость модели, снижает влияние шума
   C) Снижает точность на больших наборах данных
   D) Увеличивает вероятность использования нелинейных границ решений

**Ответ**: **B**

7. Что такое метрика F1-score в задачах классификации?
   A) Средняя точность всех классов
   B) Усредненная гармоническая мера точности (precision) и полноты (recall)
   C) Доля верно классифицированных объектов
   D) Средняя полнота всех классов

**Ответ**: **B**

8. Какие гиперпараметры Random Forest можно настроить, чтобы попытаться снизить риск переобучения?
   **Ответ**: Уменьшение максимальной глубины деревьев (max_depth), увеличение минимального количества объектов в узле (min_samples_leaf), ограничение количества признаков, используемых для разделения узла.

9. Почему при использовании алгоритма k-NN важно выбрать оптимальное значение k?

**Ответ**: Слишком малое k делает модель чувствительной к шуму и выбросам, увеличивая риск переобучения, а слишком большое k может привести к недостаточной точности и игнорированию важных локальных структур данных.

10. Чем отличается использование ядра RBF от линейного ядра в алгоритме SVM?
    **Ответ**: Ядро RBF позволяет моделировать нелинейные границы решений за счет преобразования пространства признаков, а линейное ядро подходит только для линейно разделимых задач.

---

**Вариант 4**

1. Какой алгоритм классификации менее всего подвержен влиянию выбросов в данных?
   A) k-Nearest Neighbors
   B) Support Vector Machine
   C) Logistic Regression
   D) Decision Tree

**Ответ**: **D**

2. Какой метод проверки качества модели дает наиболее стабильную оценку при повторных запусках на одних и тех же данных?
   A) Single train-test split
   B) Bootstrap
   C) Holdout
   D) Cross-validation

**Ответ**: **D**

3. Какой тип проблемы решает алгоритм Logistic Regression?
   A) Регрессия
   B) Кластеризация
   C) Классификация
   D) Ранжирование

**Ответ**: **C**

4. Что измеряет метрика Recall в задачах классификации?
   A) Долю верных предсказаний среди всех объектов
   B) Долю верных положительных предсказаний среди всех

действительно положительных
C) Долю верных отрицательных предсказаний среди всех отрицательных
D) Среднюю гармоническую точности и полноты

**Ответ**: **B**

5. Какой из параметров модели Decision Tree наиболее непосредственно влияет на переобучение?
   A) Максимальная глубина (max_depth)
   B) Количество соседей (k)
   C) Тип активации
   D) Параметр регуляризации (C)

**Ответ**: **A**

6. Какой из следующих алгоритмов использует понятие гиперплоскости для разделения классов?
   A) Decision Tree
   B) Random Forest
   C) Support Vector Machine
   D) k-Nearest Neighbors

**Ответ**: **C**

7. Какая проблема возникает при применении метрики Accuracy в задачах с сильным дисбалансом классов?
   A) Недооценка количества классов
   B) Неспособность правильно интерпретировать ошибки классификации
   C) Завышение точности за счет доминирующего класса
   D) Недооценка общего качества модели

**Ответ**: **C**

8. Почему важна регуляризация в модели Logistic Regression?
   **Ответ**: Регуляризация помогает избежать переобучения, уменьшая влияние отдельных признаков с высокими весами, улучшая обобщающую способность модели на новых данных.

9. Как выбор функции ядра влияет на возможности алгоритма SVM?
   **Ответ**: Выбор ядра определяет, может ли SVM разделять данные только линейно (линейное ядро) или моделировать нелинейные зависимости (RBF, полиномиальные ядра), существенно влияя на точность классификации.

10. Как влияет число деревьев (number of trees) в модели Random Forest на ее производительность и обобщающую способность?
    **Ответ**: Увеличение числа деревьев повышает устойчивость модели и улучшает ее обобщающую способность, но после определенного количества прирост точности становится незначительным, а затраты на вычисления возрастают.