**Week-5 Individual Deliverable Report**

**By Shaurya Parshad**

**Executive Summary**

I developed three machine learning models to predict chronic kidney disease using 200 patient medical records. After addressing several data quality issues, the final models achieved 97.5% to 100% accuracy. This report documents the process, challenges encountered, and results.

**Project Overview**

**The Dataset**

* **Total patients:** 200 (128 with CKD, 72 healthy)
* **Medical measurements:** 26 different health indicators
* **Data split:** 64% CKD patients, 36% healthy (reasonably balanced)

**The Process & Challenges**

**Challenge 1: Messy Data Format**

The dataset came with values in ranges and special characters that computers couldn't understand:

* Ages like "20-30" or ">=74"
* Urine measurements like "1.019-1.021"
* Excel accidentally converted "3-4" to "3-Mar" (a date!)

**Solution:** I wrote a conversion function that:

* Turns ranges into averages (e.g., "20-30" becomes 25)
* Handles comparison symbols (e.g., ">=74" becomes 74)
* Fixes Excel's date conversions (e.g., "3-Mar" back to 3.5)

**Challenge 2: Missing Class Labels**

One patient record had no diagnosis label - we didn't know if they had CKD or not. This single missing value caused the entire system to crash.

**Solution:** Removed that one incomplete record before training. Final dataset: 200 clean records.

**Challenge 3: Data Leakage (The Big Problem)**

My first models showed 100% perfect accuracy, which sounds great but was actually wrong! I discovered two features that were "cheating":

1. **"Affected" feature** - This literally means "has CKD." Using it to predict CKD is like asking "Does this sick person have the disease?" to predict if they're sick. Circular logic!
2. **"Stage" feature** - CKD stage (1-5) is determined AFTER a doctor diagnoses you. You can't use it to make the diagnosis in the first place.

**Solution:** Removed both features. Accuracy dropped from 100% to 97.5%, which is more realistic.

**Challenge 4: Train-Test Split**

I split my data into two groups:

* **Training set (160 patients, 80%):** The models learn patterns from this data
* **Test set (40 patients, 20%):** Used to check if the models actually learned or just memorized

This prevents the model from "studying with the answer key" for the final exam.

**Understanding the Results**

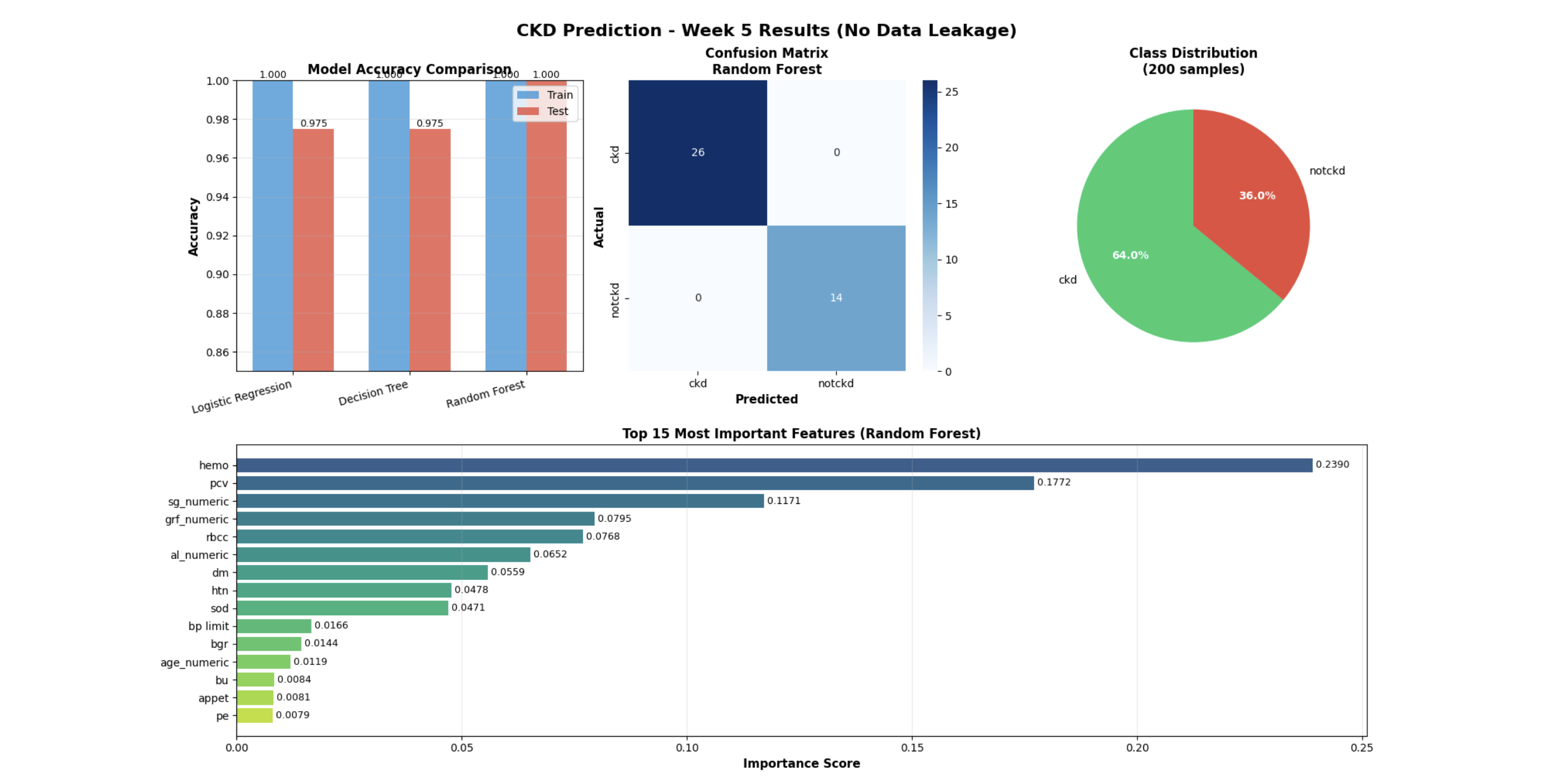
**Model Performance**

| **Model** | **Training Accuracy** | **Testing Accuracy** |
| --- | --- | --- |
| Logistic Regression | 100% | 97.5% |
| Decision Tree | 100% | 97.5% |
| Random Forest | 100% | 100% |

**What This Means:**

* **Logistic Regression (97.5%):** Correctly identified 39 out of 40 patients. Missed 1 patient.
* **Decision Tree (97.5%):** Also got 39 out of 40 correct. This consistency with Logistic Regression is a good sign.
* **Random Forest (100%):** Got all 40 test patients correct, but this might be "overfitting" - the model may have memorized patterns instead of learning general rules. With only 40 test cases, it's hard to tell. I'll test this more thoroughly in Week 6 using cross-validation.

**Understanding the Visualizations**

****

**Chart 1: Model Accuracy Comparison (Top Left)**

Shows blue bars (training) and red bars (testing) for each model. All models learned the training data perfectly (100%), but only Random Forest maintained perfect accuracy on new patients.

**Chart 2: Confusion Matrix (Top Middle)**

This grid shows Random Forest's predictions:

* **Top-left (26):** Correctly identified 26 CKD patients as having CKD
* **Bottom-right (14):** Correctly identified 14 healthy people as healthy
* **Top-right (0):** Never mistakenly said a healthy person had CKD
* **Bottom-left (0):** Never missed a CKD patient

Perfect score, but again, this needs more testing.

**Chart 3: Class Distribution (Top Right)**

The pie chart shows our dataset has almost twice as many CKD patients (64%) as healthy patients (36%). This is better than my original dataset which was 94% CKD - too imbalanced to train properly.

**Chart 4: Feature Importance (Bottom)**

This bar chart ranks which medical tests matter most for prediction. Longer bars = more important for detecting CKD.

**The Medical Indicators: What They Mean**

Here are the top 10 most important health measurements for detecting CKD, explained in plain English:

**1. Hemoglobin (23.9% importance)**

**What it is:** The protein in red blood cells that carries oxygen throughout your body.

**Normal range:** 13.5-17.5 g/dL (men), 12-15.5 g/dL (women)

**Why it matters for kidneys:** Healthy kidneys make a hormone called erythropoietin that tells your body to produce red blood cells. When kidneys fail, they make less of this hormone, so you end up with less hemoglobin. Low hemoglobin = possible kidney damage.

**Effect:** Patients with CKD often have anemia (low hemoglobin), causing fatigue and weakness.

**2. Packed Cell Volume / PCV (17.7% importance)**

**What it is:** The percentage of your blood that's made up of red blood cells.

**Normal range:** 40-50% (men), 36-44% (women)

**Why it matters for kidneys:** Directly related to hemoglobin. When kidneys don't signal for enough red blood cell production, PCV drops. It's like measuring how "thick" your blood is with cells.

**Effect:** Low PCV means thin, watery blood that can't carry enough oxygen, making you tired and dizzy.

**3. Specific Gravity / SG (11.7% importance)**

**What it is:** Measures how concentrated your urine is compared to pure water.

**Normal range:** 1.005-1.030

**Why it matters for kidneys:** Healthy kidneys concentrate urine by filtering out water and keeping important stuff. Damaged kidneys can't concentrate properly, so urine specific gravity stays around 1.010 (very dilute) no matter what.

**Effect:** If your urine specific gravity doesn't change throughout the day, your kidneys aren't adjusting fluid balance properly.

**4. GRF - Glomerular Filtration Rate (8.0% importance)**

**What it is:** Measures how much blood your kidneys filter per minute.

**Normal range:** 90-120 mL/min

**Why it matters for kidneys:** This is THE direct measure of kidney function. It's like measuring the speed of a water filter. Low GRF = kidneys are filtering slowly = kidney damage.

**Effect:** GRF below 60 for 3+ months = chronic kidney disease diagnosis. Below 15 = kidney failure, need dialysis.

**5. Red Blood Cell Count / RBCC (7.7% importance)**

**What it is:** The actual number of red blood cells in your blood.

**Normal range:** 4.5-5.5 million cells/µL (men), 4.0-5.0 million cells/µL (women)

**Why it matters for kidneys:** Same reason as hemoglobin and PCV - damaged kidneys don't signal for enough red blood cell production.

**Effect:** Low red blood cell count = not enough oxygen delivery = chronic tiredness.

**6. Albumin / AL (6.5% importance)**

**What it is:** A protein that should stay in your blood, not leak into urine.

**Normal range:** 0 in urine (none should be there)

**Why it matters for kidneys:** Healthy kidneys filter waste but keep important proteins like albumin in your blood. Damaged kidney filters become "leaky" and let albumin spill into urine.

**Effect:** Albumin in urine (albuminuria) is an early warning sign of kidney damage, often appearing years before other symptoms.

**7. Diabetes Mellitus / DM (5.6% importance)**

**What it is:** Whether the patient has diabetes (high blood sugar disease).

**Why it matters for kidneys:** High blood sugar damages tiny blood vessels in the kidneys over time. Diabetes is the #1 cause of kidney disease in developed countries.

**Effect:** About 1 in 3 people with diabetes develop kidney disease. Controlling blood sugar prevents this.

**8. Hypertension / HTN (4.8% importance)**

**What it is:** High blood pressure.

**Normal range:** Below 120/80 mmHg

**Why it matters for kidneys:** High pressure damages the delicate filtering units in kidneys (nephrons). It's like forcing water through a coffee filter at high pressure - eventually it tears. Also, damaged kidneys CAUSE high blood pressure, creating a vicious cycle.

**Effect:** Hypertension is the #2 cause of kidney disease after diabetes.

**9. Sodium / SOD (4.7% importance)**

**What it is:** Salt levels in your blood.

**Normal range:** 135-145 mEq/L

**Why it matters for kidneys:** Kidneys regulate sodium to control blood pressure and fluid balance. When kidneys fail, they can't regulate sodium properly, leading to swelling (edema) and high blood pressure.

**Effect:** Abnormal sodium can indicate the kidneys aren't balancing fluids correctly.

**10. Blood Pressure Limit (1.7% importance)**

**What it is:** Classification of blood pressure as normal or high.

**Why it matters for kidneys:** As mentioned with hypertension, high BP both causes and results from kidney damage.

**Effect:** Monitoring BP helps track kidney disease progression.

**Other Important Indicators (Not in Top 10 but Still Measured)**

* **Blood Urea (BU):** Waste product kidneys should filter out. High levels = kidneys not cleaning blood properly.
* **Serum Creatinine (SC):** Another waste product. High levels = reduced kidney function.
* **Potassium (POT):** Kidneys regulate this. Too high or low = kidney problems.
* **White Blood Cell Count (WBCC):** Can indicate infection or inflammation affecting kidneys.
* **Pus Cells, Bacteria:** Signs of kidney infection.
* **Red Blood Cells in Urine:** Shouldn't be there; indicates kidney damage or stones.
* **Coronary Artery Disease (CAD):** Heart and kidney health are connected.
* **Appetite:** CKD patients often lose appetite due to toxin buildup.
* **Pedal Edema (PE):** Swollen feet/ankles from fluid retention when kidneys can't remove excess water.
* **Anemia (ANE):** Low red blood cell count, common in CKD.